TEXT FLY WITHIN THE BOOK ONLY

UNIVERSAL
LIBRARY
OU_162046
UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 500 / K91F Accession No. G 8867

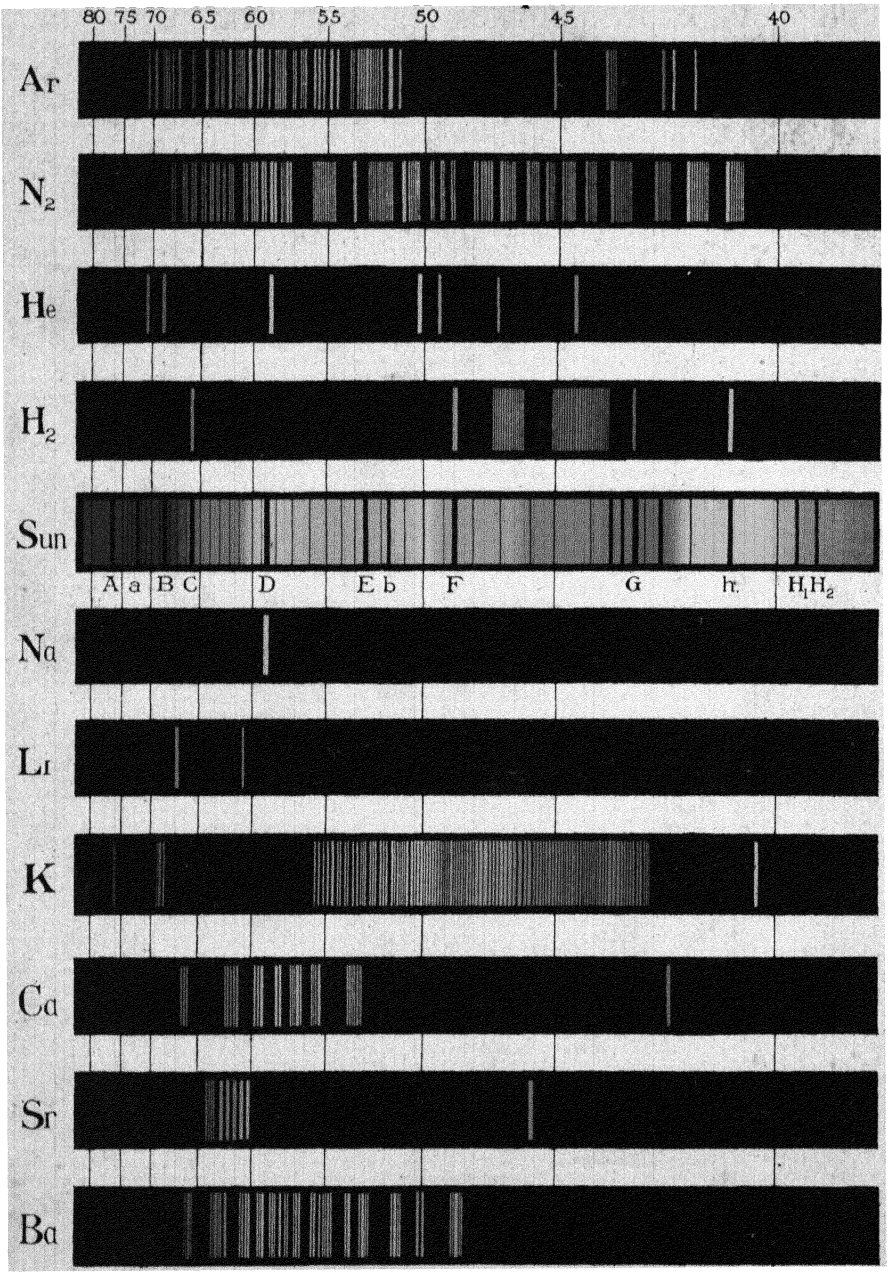
Author Krauskopf, Konard Bates.

Title Fundamentals of Physical Science. 1948

This book should be returned on or before the date last marked below.

FUNDAMENTALS OF PHYSICAL SCIENCE

Plate of Spectra



FUNDAMENTALS OF PHYSICAL SCIENCE

An Introduction to the Physical Sciences

BY KONRAD BATES KRAUSKOPF
Associate Professor of Geology, Stanford University

SECOND EDITION
FIFTH IMPRESSION

New York Toronto London
McGRAW-HILL BOOK COMPANY, INC.
1948

FUNDAMENTALS OF PHYSICAL SCIENCE

Copyright, 1941, 1948, by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

Preface to the Second Edition

THE preparation of a second edition has been prompted chiefly by a desire to bring the book up to date in several rapidly advancing fields of physical science. Notable changes are a complete rewriting of the chapter on the atomic nucleus, addition of a section on the uncertainty principle, introduction of Brönsted's theory in the discussion of acids and bases, and increased emphasis on air-mass analysis in weather forecasting. The new edition has also provided a welcome opportunity to correct a multitude of minor errors, ambiguities, and awkward phrasings, to add some necessary problems, and to improve several illustrations. A few sections of the first edition have been omitted or shortened because they did not contribute directly to the central theme of the book.

For pointing out errors in the first edition the author is indebted to many students and friends. Particular help in preparing the second edition has been most generously given by Profs. F. O. Koenig, Paul Kirkpatrick, and C. Alvarez-Tostado of Stanford University and Prof. F. C. Krauskopf of the University of Wisconsin.

KONRAD BATES KRAUSKOPF

TOKYO, JAPAN
May, 1948.

Preface to the First Edition

THIS book is addressed primarily to college students who wish a general knowledge of the physical sciences rather than detailed knowledge in any one science. It should likewise fill the need of the general reader, college trained or not, who seeks information about the methods of science and the place of science in our modern world. From either college student or general reader the book requires nothing in the way of preparation beyond a lively curiosity and a willingness to make some effort to train his mind in unaccustomed ways of thinking.

Many books in recent years have presented to the general reader the more spectacular findings and achievements of modern science in readable and entertaining form. If such books are designed to awaken the reader's interest in science, this one is designed to deepen and enlarge that interest. Science in these chapters is not presented as a parade of marvels, but as a method of thought that leads to understanding and control of natural processes. Emphasis is placed less on the specific accomplishments of science than on how these accomplishments were made possible. By stressing the methods of scientific reasoning rather than their results the book attempts to give its readers a truer picture of the relationship of science to modern life and thought, a better appreciation of the limitations as well as the extraordinary power of the scientific method.

Materials for the book are taken from the four sciences of astronomy, physics, chemistry, and geology. Many fascinating branches of each one must be omitted from a volume of this size, but all the more important parts of each have been included in what the author hopes is a proper perspective. The different sciences are not treated separately, since it seems desirable in a book of this kind to emphasize the unity of physical science as a field of knowledge rather than to stress its arbitrary divisions.

A few words are necessary regarding use of the book as a college text. It is intended for a course *in* science, not for a superficial course *about* science. In covering so wide a field it must of necessity touch many topics

lightly; but the more important ones are discussed as rigorously, and should require as much of the student's time and effort, as those in any elementary science course.

The book is designed to be adaptable to courses of different lengths. The fourth section is especially planned to make the book flexible: its chapters (except for the first two) depend so little on each other that one or several may be omitted without destroying continuity. In Part III the long chapter on electric currents (Chap. XVIII) may be safely omitted or left to the better students; in Part V, Chap. XXXIII (Weather and Climate) has little direct connection with the others, and a coherent picture of the physical side of geology is possible without Chaps. XXXIV, XXXVII, and XXXIX; in Part VI the final chapter may well be left as optional reading.

The material of the book has served satisfactorily for a three-unit course (two lectures and one three-hour laboratory per week) at Stanford University for the last twelve years. The content of the course varies somewhat from year to year, but the arrangement is approximately as follows: the first two parts of the book are covered during the autumn quarter, a few sections in Chaps. V, VI, and X being used as optional reading; Parts III and IV occupy the winter quarter, with about three chapters in Part IV omitted; Parts V and VI come in the spring, the final chapter being assigned only to better students. This is suggested, of course, as only one possible adaptation of the book to a three-unit program.

How much mathematics should be included in a course of this sort is a troublesome question. Most teachers would probably agree that some use of mathematical symbolism is desirable in an elementary course in physical science. Unfortunately the mathematical background of many college students, even in the elementary operations of arithmetic, is woefully inadequate. To give such students any idea of the place of mathematics in physical science requires an unconscionable amount of time and effort; it may well be argued that the time could be better spent on other topics.

The author has not tried to answer this question, but rather has attempted in this respect also to make the book sufficiently flexible to fit differing points of view. The two important mathematical ideas in the book, proportionality and graphic representation, are both introduced in Chap. IV. These are important, not only because they show the student how experimental data are expressed in algebraic form, but also because they have wide application outside the domain of physical science. These ideas are amplified and applied in Chaps. V and VI, but thereafter throughout the book mathematics is used sparingly. A number of problems in later chapters involve simple mathematics, but these are intended

for better-prepared students who enjoy exercising their mathematical ingenuity.

Thus discussions from the book may be made to include as much or as little mathematics as the instructor desires. If it seems best to omit mathematics entirely, Chap. IV should be skipped and some sections in Chaps. V and VI should be omitted or modified. In the rest of the book very little alteration will be necessary.

It is a pleasure for the author to acknowledge his indebtedness to the many persons who have assisted in the preparation of this book. Professor F. O. Koenig, of Stanford University, read critically the entire manuscript. Parts of the manuscript were read and criticized by Professors Paul Kirkpatrick, A. C. Waters, and Claudio Alvarez-Tostado of Stanford University, and by Professors Farrington Daniels and F. C. Krauskopf, of the University of Wisconsin. For the many valuable suggestions of these critics the author is sincerely grateful, but responsibility is entirely his for any errors which remain in the text.

The author would like to express his thanks also to the many organizations and individuals who furnished illustrations.

To his wife must go a special word of thanks for her valuable criticisms and constant assistance in all stages of the writing of this book.

KONRAD BATES KRAUSKOPF

PALO ALTO, CALIF.
May, 1941

Contents

PREFACE TO THE SECOND EDITION v

PREFACE TO THE FIRST EDITION vii

PART I. THE SOLAR SYSTEM

I. Ptolemy and Copernicus 3
II. The Sun and Its Family 16
III. Force and Motion 37
IV. The Language of Mathematics 47
V. Forces in Combination 62
VI. The Law of Gravitation 74
VII. Origin of the Solar System 90

PART II. MATTER AND ENERGY

VIII. Energy 101
IX. Solids, Liquids, and Gases 113
X. The Kinetic Theory 127
XI. Chemical Change 145
XII. Weight Relations in Chemical Reactions 156
XIII. The Atomic Theory 166
XIV. The Language of Chemistry 180
XV. The Periodic Law 194

PART III. THE STRUCTURE OF MATTER

XVI. Electricity and Magnetism 213
XVII. The Electron 227
XVIII. Electric Currents 237
XIX. Light Waves 255
XX. X Rays and Radioactivity 272
XXI. The Atomic Nucleus 281
XXII. Radiation 305
XXIII. Subatomic Chemistry 321

PART IV. FUNDAMENTAL PROCESSES

XXIV. Ionic Reactions 337
XXV. Acids, Bases, and Salts 349
XXVI. Chemical Energy 366
XXVII. Reaction Rates and Equilibrium 377
XXVIII. Oxidation and Reduction 389
XXIX. Carbon Compounds 401
XXX. Silicon Compounds 419
XXXI. The Colloidal State 430

PART V. THE BIOGRAPHY OF THE EARTH

XXXII. Earth Materials 441
XXXIII. Weather and Climate 450
XXXIV. Rocks and Minerals 471
XXXV. Erosion and Sedimentation 488
XXXVI. Vulcanism and Diastrophism 518
XXXVII. The Law of Uniform Change 545
XXXVIII. Interpreting the Rock Record 553
XXXIX. Earth History 569

PART VI. STARS AND GALAXIES

XL. The Sun 603
XLI. The Stars 618
XLII. The Nebulae 634
XLIII. Frontiers of Physical Science 648

APPENDIX 661

INDEX 663

PART I

THE SOLAR SYSTEM

IN ONE sense this book will be a summary of man's knowledge about the world in which he lives. It will not be a complete summary, for the important part of the world made up of living things will be mentioned only casually. Even in those branches of science which deal primarily with inanimate material—astronomy, physics, chemistry, geology—the summary must leave many phenomena untouched. Physical science in the twentieth century encompasses too wide a field to be brought within the covers of a single volume. So this must be a special kind of summary, a selective summary which will handle stars and atoms, rocks, light waves, and electric motors with an eye to their importance and their relationships rather than their intricate details.

In a second sense this book will be history on a grand scale—not a human history of wars, laws, and governments, but the longer history of the earth itself. Science is deeply concerned with origins, and all branches of science combine to give us a picture not only of the earth today but of the earth and the universe as they have appeared in past ages.

And finally, the chapters to come will give a record of a great human adventure. Science deals so persistently and so coldly with facts and objects that one easily forgets the simple truth that all science is a product of the human mind. The search for knowledge is an adventure of the intellect whose beginnings go back to the days before written history and which today is still leading us on to explore dark corners of the universe. Men of many nations and many races have taken part in the great adventure, and the modern science of which we boast is largely a heritage from their mental labor.

To understand the world of which we are a part, to read the pages of the earth's long history, to learn how man's ideas about the universe have developed—these, then, are the objects we shall set ourselves in the pages to follow.

Now where in the wide range of natural phenomena shall we look for a starting point in attaining these objects? The scientist is essentially a simplifier, one whose constant effort is to see behind the complexities of natural events the basic principles and relationships which they exhibit; to understand his methods, we need a set of events familiar enough and simple enough so that we can detect their fundamental relationships without too intricate an analysis. We shall not look first, for instance, at a growing plant, a volcano in eruption, or the motions of the sea. These, like most events of our experience, are too erratic, too infinitely complex for analysis in simple terms. We shall do best to follow the example set by men of early civilizations. They too sought simplicity and order in the universe, and found them most clearly shown in the motions of sun and stars across the sky. Nowhere else in our immediate experience do events follow each other according to such precise and unchanging rules. So we shall begin as our ancestors began, with the sun and moon, the planets and stars.

CHAPTER I

Ptolemy and Copernicus

THE most obvious motion of the heavenly bodies is their daily east-west crossing of the sky. The rising of the sun and moon, their steady progress across the sky, their setting at the western horizon, are familiar observations. That the stars move in similar fashion becomes evident from a few minutes' watching of the night sky.

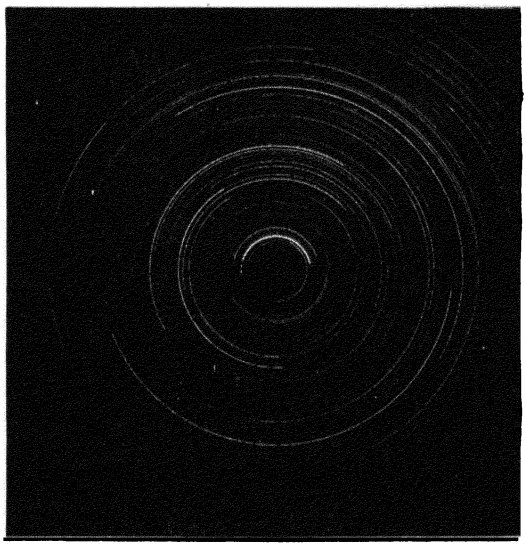


FIG. 1. *Paths of stars in the northern sky, photographed by leaving a camera pointed toward Polaris for 11½ hours. The trail of Polaris is the bright one slightly above the true pole. (Photographed by Wilson at the Goodsell Observatory.)*

One star in the northern sky appears to move scarcely at all. This is the *North Star*, or *Polaris*, for centuries used as a guide by travelers simply because of its unchanging position. Stars in its vicinity do not rise or set, but move around it in circles, circles which carry them under Polaris from west to east and over it from east to west. Farther from Polaris the circles

get ever larger, until presently they dip below the northern horizon (Fig. 1). Paths of the stars in the southern sky, together with the paths of the sun and moon, are simply very large circles, all with their centers near Polaris. Sun, moon, and stars "rise" and "set" because their circles lie in part beneath the horizon. Thus the whole sky appears to revolve once a day about this not very conspicuous star.*

That Polaris occupies a central position in this huge merry-go-round is no fault of its own. Very early in our scholastic careers most of us learn that the earth rotates once a day on its axis and that Polaris simply happens to lie almost directly over the north pole. Thus, as the earth turns, everything about it appears to be moving past it except this one star which is on the line of its axis. To us this explanation seems obvious enough, since we were brought up to believe it. But let us, for this chapter, forget our modern beliefs and imagine ourselves in the position of men who lived several centuries ago, when belief in a rotating earth was considered dangerously radical. Let us examine the motions of heavenly objects in some detail and see why scientific ideas regarding them have changed so greatly.

Motions of Sun, Moon, and Planets

Except for their circular motion around Polaris, the stars appear to remain fixed in their positions with respect to each other. Stars of the Big Dipper move about halfway around Polaris between every sunset and sunrise, but the shape of the Dipper itself has not changed since our grandfathers were young. To emphasize the fact that stars do not change their relative positions, they are often referred to as "fixed" stars.

Easily recognized groups of fixed stars, like those which form the Big Dipper, are called ***constellations***. Near the Big Dipper is the much less conspicuous Little Dipper, with Polaris at the end of its handle. On the other side of Polaris from the Big Dipper is the W-shaped constellation Cassiopeia, named for an ancient Ethiopian queen. The legendary hunter Orion, a bright constellation conspicuous on winter nights, consists of four stars in a slightly warped rectangle with a row of three stars across its middle. Scorpio, visible above the southern horizon in summer, is a fan-shaped group of stars with a curving tail. With a little patient observation many other animals and heroes and beautiful women can be found hidden among the fainter stars. Most of the constellations bear little resemblance to the figures they are supposed to represent, but they serve as convenient names for definite regions of the sky.

* These sentences apply specifically only to stars visible in the northern hemisphere. To an observer south of the equator celestial objects appear to describe circles about a point in the southern sky directly over the earth's south pole.

In their daily east-west crossing of the sky, the sun and moon move more slowly than do the fixed stars, and so appear to drift *eastward* among the constellations. The slow eastward motion is most easily observed for the moon: if the moon is seen near a bright star on one evening, by the next evening it will be some distance east of the star, and on succeeding nights it will be farther and farther to the east. In about a month the moon drifts eastward completely around the sky and returns to its starting point.

The sun's motion is less easily followed, because we cannot observe directly what stars it is near. But if we note what constellations appear just after sundown, we can estimate the sun's approximate location among the stars and follow it from day to day. We find that the sun moves eastward more slowly than the moon, so slowly that the day-to-day change is scarcely noticeable. Because of the sun's motion each constellation appears to rise a few minutes earlier each night, so that after a few weeks or months the appearance of the night sky becomes markedly different. By the time the sun has migrated eastward completely around the sky, we find that a year has elapsed. The length of the year, in fact, is determined from this apparent motion of the sun among the stars.

The precise length of the year, according to modern measurements of the sun's motion, is 365 days 5 hr 48 min 46 sec. Calendar makers have known for three millenniums that the year should be about 365 days long, but they have disagreed widely in handling the awkward residue of nearly 6 hours and in apportioning 365 days among ten to twelve months. Our present calendar is a heritage of Roman days, when Julius Caesar revised the old ten-month Roman calendar, which permitted January to occur sometimes in summer, sometimes in winter. Caesar's improved calendar was still so far from perfect that in 1,600 years its error amounted to eleven days. Further changes designed to eliminate this discrepancy were introduced by Pope Gregory XIII in 1582. Pope Gregory's calendar, in common use throughout the Western world today, is in error by only one day in 3,300 years.

Five other objects in the sky, the ancients found, shift their positions with respect to the stars. These objects, to all appearances like five bright stars, were called ***planets*** (Greek for "wanderer") and were named for the Roman deities Mercury, Venus, Mars, Jupiter, and Saturn. The strange motions of the planets have excited the interest and wonder of all civilized peoples; we find records of planetary movements even from the third millennium B.C. in Babylonia. From that day to this the planets have served both as a fruitful source of income to fortunetellers and as objects of scrupulous study by astronomers. Like the sun, the planets shift their positions so slowly that their day-to-day motion is difficult to detect, but unlike the sun they move in complex paths. In general each

planet moves eastward among the stars, but its rate of motion is variable and at intervals it stops and moves for a brief time westward. Thus the path of a planet is characterized by "loops" (Fig. 2) which recur after definite periods of time.

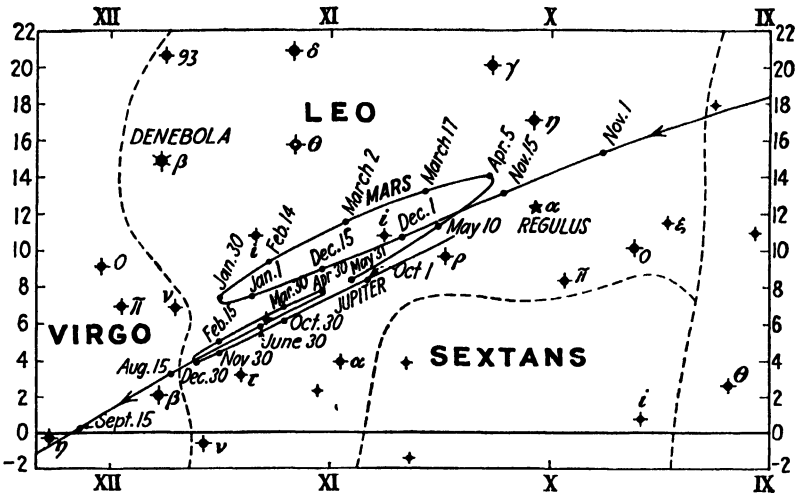


FIG. 2. *Loops in the apparent paths of Mars and Jupiter among the fixed stars. Each planet moves in general eastward (from right to left on the diagram), but for a few weeks reverses its motion. (From An Orientation in Science by Watkeys and associates.)*

To summarize the motions of celestial objects: The sun and moon, planets, and stars all appear to move in circles around a line connecting the earth and Polaris. Because of this motion all objects except stars in the northern sky rise in the east and set in the west, approximately once a day. The east-west motions of the sun, moon, and planets are in general a bit slower than the motions of the stars, so that these bodies slowly shift their positions eastward with respect to the stars. The eastward movements of the sun and moon are approximately uniform, but the planets move with variable speeds and occasionally even change their direction of motion for a few weeks or months.

The Ptolemaic Hypothesis

So far we have been simply listing *observational facts*, facts which anyone can discover for himself by keeping track of the changing positions of objects in the sky. We have described the motions of celestial objects just as we see them today and just as other observers saw them many centuries ago. These observed motions are complex, but far from erratic: each object which we have considered follows a definite path across the sky which it repeats faithfully time after time. Complicated motions with an underlying regularity—to a scientific mind, such a series of motions

demands an explanation. We need a *hypothesis*, a guess concerning the actual space relations of objects in the sky, which will account for the observed motions as combinations of simpler real motions.

If we forget our modern prejudices, we find two possible bases on which a satisfactory hypothesis may be framed: either the earth is stationary, as it appears to be, with the celestial objects revolving about it; or the earth is moving, its motion being then responsible for a part of the apparent motion of other objects. Thus the apparent daily rotation of the sky may represent an actual motion of sun, moon, planets, and stars, or it may be explained by a rotation of the earth on its axis. The apparent eastward shift of the sun's position among the constellations may be a real motion of the sun, or it may be due to another motion of the earth. These alternatives were clear to the philosophers of ancient Greece; some believed in a stationary earth, a few argued for a moving earth. In their day scientific knowledge was not sufficiently advanced to settle the matter. Except for crude estimates regarding the moon and sun, they had no idea of the sizes or distances of celestial objects, no certain knowledge even of the shape of the earth. It is hardly surprising, therefore, that most of the Greeks favored the common-sense view that the earth is stationary.

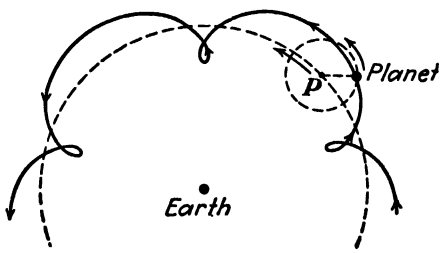


FIG. 3. *Motion of a planet according to the Ptolemaic system. The planet moves in a small circular path about a point P, which moves in a larger circle around the earth. Combination of these circular motions gives the looped path shown by the solid line.*

The hypothesis most widely accepted by the later Greek and Roman scholars was devised originally by Hipparchus and elaborated by Ptolemy of Alexandria. After this latter personage, about whom little is known except that he lived during the reigns of the Roman emperors Hadrian and Antoninus Pius, in the last declining years of Greek culture, Hipparchus's picture of the universe is called the Ptolemaic system.

An intricate and ingenious system it was. Our earth stands at the center, motionless, with all other objects in the universe revolving about it in paths which are either circles or combinations of circular motions—since to the Greeks the circle was the only "perfect" curve, hence the only conceivable path for a celestial object. Enclosing all is a gigantic crystal sphere studded with the fixed stars, making approximately one revolution each day. Somewhere inside is the sun, moving around the earth exactly once a day. Between the motions of sun and stars is just enough difference so that the sun appears to move among the constellations, completing its circuit once a year. Near the earth in a small orbit is the moon, revolving

more slowly than the sun. The planets Venus and Mercury come between moon and sun; the planets Mars, Jupiter, and Saturn between sun and stars. To account for the observed peculiarities of planetary motion, Ptolemy imagined each planet to move in a small circle about a point which in turn described a large circle about the earth (Fig. 3). By a combination of these circular motions each planet travels in a series of loops; since we observe these loops edgewise, it appears to us as if the planets move with variable speeds and sometimes reverse their motions.

From observations made by himself and his predecessors Ptolemy calculated the relative speed with which each celestial object moved in its orbit. Using these speeds he could then compute the location of an object in the sky for any date in the past or future. These *computed* positions checked well with *observed* positions which had been recorded for several centuries before his time, and his *predictions* agreed with *observations* made in succeeding years. So Ptolemy's hypothesis fulfilled all the requirements of a scientific hypothesis: it was based solidly on observational facts, it explained adequately all the facts about celestial motions which were known in his time, and it made possible the prediction of facts which could be verified in the future.

For 1,400 years no one seriously questioned Ptolemy's hypothesis. Generations of Arabian and Persian scholars found it adequate for their limited observations. Later on it was accepted without question in the schools and monasteries of Europe. Men sought neither to inquire into its theoretical foundation nor to verify its predictions by observation. So long a time of blind acceptance gave to Ptolemy's universe, as to many other ideas of Greek science, a completely undeserved authority. That such an age-old tradition should be slow to yield to new ideas is quite understandable, although the zeal with which Christian scholars defended a product of pagan logic seems at times a bit extreme.

The Copernican Hypothesis

By the sixteenth century it had become obvious that something was not quite right in the Ptolemaic system. Observed positions of the planets simply did not agree with the positions calculated from Ptolemy's complicated orbits. Discrepancies were not large, but could be detected even by inexpert observers. There were two possible ways for removing the discrepancies: either slight changes could be introduced into the Ptolemaic orbits, making the system still more complicated; or the Ptolemaic hypothesis could be discarded in favor of a completely new hypothesis based on different assumptions.

The first to defy tradition by setting up a new explanation for the universe was Nikolaus Copernicus, a versatile and energetic Pole of the early sixteenth century. Copernicus lived in the years following Colum-

bus's great discovery, years when mental as well as geographical horizons were receding before eager explorers. In Italy it was the time of Leonardo da Vinci and Michelangelo; a time of commercial expansion and incessant wars between rival cities; a time of great fortunes and fantastic corruption in government: a time of brilliant thinkers and inspired artists. To this

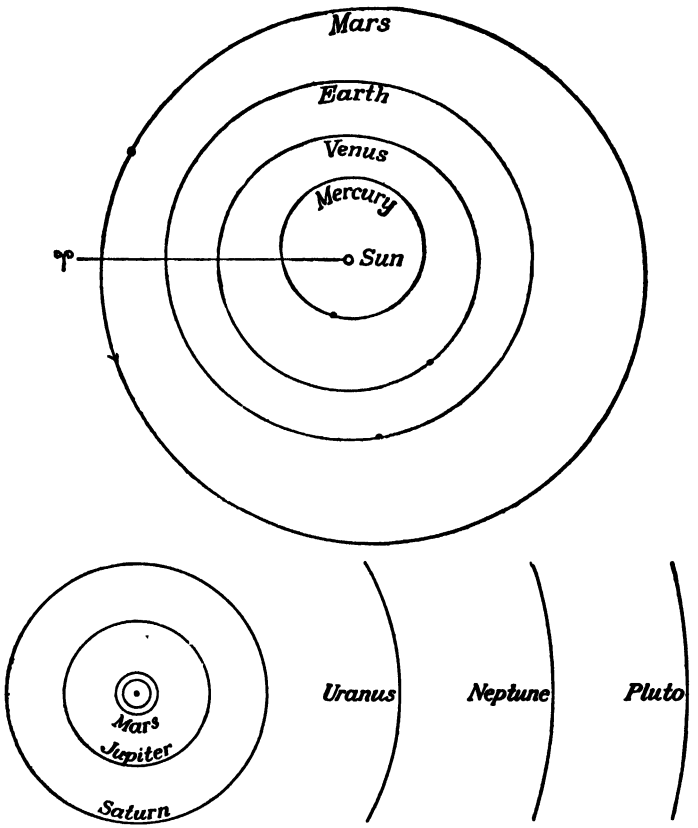


FIG. 4. *The Copernican system (modified according to present knowledge). Orbits of the four inner planets are shown in the upper diagram, and orbits of the planets from Mars to Pluto are shown on a smaller scale in the lower diagram. Orbits of the inner planets are drawn as ellipses rather than the true circles which Copernicus assumed. The three outermost planets were unknown in Copernicus's time. (From Elements of Astronomy by Fath.)*

Italy of the Renaissance went Copernicus as a student, learning in its universities medicine, theology, and mathematics and acquiring a deep distrust of Ptolemaic astronomy. Back in his native Poland he practiced medicine and interested himself in currency reform, but he devoted most of his time to developing an idea which had germinated in Italy: the idea that the universe could be vastly simplified if the sun rather than the earth were taken as its center.

Let us consider the earth, said Copernicus, as one of the planets, a sphere rotating once a day on its axis. Let us further imagine that the planets, including the earth, revolve in circular orbits about the sun (Fig. 4), that the moon is relatively close to the earth and revolves about it, and that the stars are situated at great distances beyond. In this picture, rotation of the earth on its axis explains the daily rising and setting of celestial objects. The apparent motion of the sun among the stars is due to the earth's motion in its orbit: as we swing around the sun, it appears to us as if the sun were constantly shifting its position against the background of fixed stars. The moon's eastward drift is in large part due to its actual orbital motion. Apparent movements of the planets are explained as a combination of their actual motions around the sun and our shift of

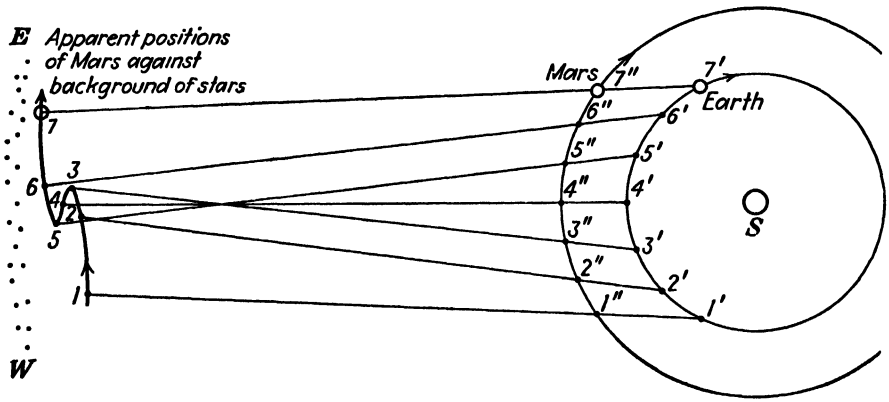


FIG. 5. *Real and apparent motions of Mars, according to the Copernican system.*

position as the earth moves. Figure 5 shows how such a combination of motions can account for changes of direction in the apparent motion of Mars. The numbers 1′, 2′, 3′ . . . represent positions of the earth separated by equal time intervals, and 1″, 2″, 3″ . . . represent corresponding positions of Mars. We see Mars against a background of stars, as if it occupied the positions 1, 2, 3. . . . While the earth is moving from 1′ to 3′ and Mars from 1″ to 3″, the apparent position of the planet shifts eastward among the stars from 1 to 3. But as we overtake Mars and move past it (positions 3′ to 5′), the apparent position shifts in the opposite direction from 3 to 5. Thereafter the motion is again eastward (positions 5 to 7). Thus Copernicus's hypothesis accounts satisfactorily for the general eastward movements and occasional brief westward movements of the planets.

The idea behind Copernicus's hypothesis was not wholly new, for some of the Greeks had realized that apparent celestial motions might possibly be the result of motions of the earth. But Copernicus went far beyond these earlier speculations by working out the planetary motions

mathematically. From observations of the positions of the planets he calculated how big each orbit must be in comparison with that of the earth and how fast each planet must be moving. With these figures he could compute the apparent positions for any time in the past or future, just as Ptolemy had done with figures based on a different hypothesis. Copernicus found that the calculated positions agreed with observations fairly well, but not much better than those calculated from the Ptolemaic system.

Despite the lack of a complete check with observation, Copernicus believed that his simple circular orbits gave a more truthful picture of the universe than the complex orbits of the Ptolemaic system. Such a belief in those days was dangerous, for the all-powerful church was reluctant to see the earth removed from its important place at the hub of the universe. Copernicus, having no wish to become a martyr, wisely did not advertise his views during his lifetime. Only in the year of his death, 1543, did he allow his work to be published.

Establishment of the Copernican System

With the publication of Copernicus's manuscript there began a long and bitter argument. To us, growing up in the calm assurance that the earth moves, it seems odd that this simple idea was so long and so violently opposed. But in the sixteenth century, before the invention of the telescope, a decision between the Ptolemaic and Copernican systems was by no means easy. Predictions from both hypotheses agreed only moderately well with observations. Good scientific arguments could be brought forward on both sides. To settle the debate, precise observations were necessary, both of celestial objects and of the earth itself; and instruments to make these observations simply were not available.

Consider, said proponents of Ptolemy's hypothesis, how rapidly the surface of the earth must travel to complete a rotation every 24 hours. Would not all loose objects be flung into space by this whirling sphere, much as mud is thrown from the rim of a carriage wheel? And would not so dizzy a speed produce great winds to raze all rooted things, like buildings, trees, flowers? Admittedly the earth spins rapidly, answered the disciples of Copernicus, but probably the centrifugal force of its rotation is counterbalanced by the force of gravity which holds our feet to the ground. Besides, the speed of such a rotation is not so unimaginably great as that which the sun and stars and planets would need in order to revolve, as Ptolemy pictured them, once a day around a stationary earth. From Aristotle, supporters of the older theory derived another cogent argument: if the earth moves through space around the sun, why do the stars not change position relative to one another, as trees on a distant hillside appear to change position when we drive past? Perhaps, countered the

newer school, the stars are very remote. So went the dispute. The Ptolemaic hypothesis, in addition to theological support, had strong logical arguments behind it; the Copernican view rested on equally good logic, and had the further merit of greater consistency and simplicity.

Nearly a century elapsed before the labor of three men—Tycho Brahe, Johannes Kepler, and Galileo Galilei—produced unanswerable evidence for the superiority of Copernicus's idea. The argument was completely silenced only toward the end of another century, when Sir Isaac Newton showed that motion of the planets around the sun was a necessary consequence of his law of gravitation.

The precise observations of planetary positions, which were the first requisite for a decision between the rival hypotheses, were made by Tycho Brahe. A Dane of noble descent, Tycho scandalized his family and friends by getting his nose cut off in a duel, by marrying a peasant girl, and later in life by devoting himself to astronomy. Although a poor theorist, Tycho was a remarkably able observer. When he died in 1601, the records of his meticulous observations were bequeathed to his young German assistant Johannes Kepler—a singularly fortunate circumstance for the development of astronomy.

Kepler possessed what Tycho lacked, the imagination to conceive various hypotheses of planetary relationships and the perseverance to check each hypothesis against observational data. He found at once that neither the Ptolemaic nor the Copernican system in its simple form would fit Tycho's figures, and he set himself the task of finding some modification of one or the other with which the figures would agree. It was an enormous undertaking; Kepler wrestled for years with Tycho's data, trying and discarding one hypothesis after another, before he was rewarded with a satisfactory solution. The accomplishment seems the more remarkable in view of Kepler's unhappy life: through the years of mental labor he was beset again and again by ill-health, by poverty, and by domestic trouble.

Kepler's great discovery was that the Copernican hypothesis gives excellent agreement with Tycho's data, *provided that the planets are assumed to move in ellipses* rather than in circles.* Kepler abandoned circu-

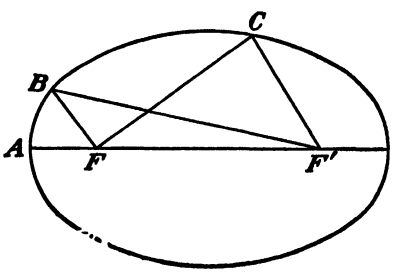


FIG. 6.

* An ellipse is a curve drawn so that the sum of the distances from every point on it to two fixed points is constant. Thus in the accompanying figure

$$BF + BF' = CF + CF' = AF + AF'.$$

The two fixed points (F and F') are called the *focuses* of the ellipse. An ellipse of given area is nearly circular if the focuses are close together, long and flat if the focuses are far apart.

ar orbits reluctantly; for he was something of a mystic and believed, like his predecessors, that circles were the most fitting type of path for celestial objects. But no circles or reasonable combinations of circles could fit Tycho's data as precisely as did simple ellipses. Kepler had too much faith in the correctness of his master's observations to be influenced by his own prejudices.

Kepler not only improved the Copernican system by the introduction of ellipses, but succeeded in winning from Tycho's figures some precise information about the speeds of the planets in their orbits. His findings are summarized in three simple statements which have come to be known as Kepler's laws:

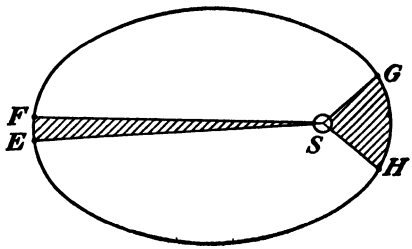


FIG. 7. *Kepler's second law. Area EFS = area GHS. E, F, G, H represent successive positions of a planet in its orbit; the planet covers the distance EF in the same time as it covers the distance GH. (Flatness of orbit greatly exaggerated.)*

1. *Every planet moves in an elliptical orbit around the sun, the sun occupying one focus of the ellipse.* The ellipses are not far from true circles (Fig. 4).
2. *The line connecting each planet with the sun sweeps over equal areas in equal times.* This means that a planet moves fastest in the part of its orbit nearest the sun, most slowly in the part of its orbit far from the sun (Fig. 7).
3. *The squares of the times required for the different planets to move completely around the sun are proportional to the cubes of their average distances from the sun.* Roughly this means that planets need more time to get around their orbits the farther they are from the sun.

Thus finally the solar system was explained in terms of simple motions. Planetary positions computed from Kepler's ellipses agreed not only with Tycho's data but with observations made thousands of years earlier. Predictions could be made regarding positions of the planets in the future —accurate predictions this time, no longer mere approximations. Furthermore, Kepler's laws showed that the speed of a planet in different parts of its orbit was governed by a simple rule and that the speed was definitely related to the size of the orbit. Beside this simple, consistent picture Ptolemy's ungainly structure looked ludicrous indeed.

To convince the most skeptical that Copernicus's hypothesis had triumphed, one more type of evidence was desirable—direct telescopic observations of the planets. Such evidence was supplied by Kepler's contemporary, Galileo Galilei. Of this brilliant Italian and his pioneer work with the telescope we shall hear more in the next chapter.

A Basic Assumption of Science

The immediate successors of Copernicus had no adequate basis for choosing between his hypothesis and Ptolemy's. Each hypothesis was capable of explaining fairly well the observed motions of sun, moon, and planets; each met some grave objections, and each had strong arguments in its favor. The work of Tycho Brahe and Kepler changed the situation abruptly, giving one hypothesis a great advantage over the other. Ptolemy's system might still have been patched up by introducing more complicated circular motions, but scientists since Kepler have chosen his picture unreservedly rather than attempt such patchwork. Let us examine more critically the reasons behind this choice.

Kepler, we say glibly, proved that the Ptolemaic system was "wrong," the Copernican system "right." In one sense this is true, but we are giving to "right" and "wrong" a very special meaning. Kepler did not prove that Ptolemy's picture of the universe could not be made to work, but only that the structure proposed by Copernicus made the universe immensely simpler. He showed that the Ptolemaic system was "wrong" by pointing out its absurd complications.

Thus our choice between the two explanations depends on their relative simplicity. Always in science similar choices are made on this same basis. Given several hypotheses to account for the same set of observed facts, we select the one which is simplest, which involves the fewest assumptions, and which can best relate this set of facts to others.

We encounter here one of the fundamental assumptions on which science is built: *the simplest explanation is always the best explanation.* The word "simplest," however, requires qualification. Ptolemy's description of the universe is far simpler than the complete description which modern astronomers believe correct, but scientists do not therefore accept the Ptolemaic system. Ptolemy's universe is simpler only because it does not attempt to explain so many facts. Details of planetary motion, the existence of the three outer planets and of moons belonging to other planets than the earth, direct evidences of the earth's motion from stellar parallaxes (page 21) and the Foucault pendulum (page 20) were all completely beyond Ptolemy's range of observation. An attempt to fit these and other data of modern astronomy into the Ptolemaic system, while not impossible, would lead to fantastic complications. Ptolemy's hypothesis is simple only for the few facts that he knew in the second century A.D.; additional facts do not fit the hypothesis but require new assumptions and qualifications. Fairly simple at first, but growing more complex with each new observation, the Ptolemaic system has long since proved inadequate for modern astronomy. Thus when we say that science seeks always the

simplest explanation, we mean *the simplest explanation that will account for all known facts.*

The assumption that the simplest explanation is best merely expresses the faith of scientists in the simplicity and orderliness of the universe. To men untroubled by a conviction that the universe is simple and orderly, some other explanation than the simplest may be more satisfying. An eclipse of the moon to a savage represents a hungry beast trying to use the moon for his dinner and being dissuaded by frantic incantations; to a medieval churchman it was an act of God designed as a warning to erring humanity; to a scientist it represents the passage of the moon through the earth's shadow (page 25). The scientist chooses this last explanation because it involves no assumptions beyond those used in his picture of the solar system and because it is entirely consistent with that picture, whereas the other two explanations involve a multitude of assumptions regarding the nature and attributes of a supernatural being. This is no reflection on the correctness or the usefulness of the first two explanations or the men who devise them. But in science they can have no place.

Questions

1. If you were at the earth's equator, you would see Polaris on the northern horizon. Along what paths would the fixed stars appear to move?
2. Along what paths would the stars move if you were standing at the north pole, with Polaris directly overhead?
3. How is the sun's apparent eastward motion among the stars explained by the Ptolemaic hypothesis? By the Copernican hypothesis?
4. Why doesn't the moon rise at the same hour every night?
5. Describe the apparent motion of Saturn among the stars. Explain with the aid of simple diagrams how this motion is accounted for by (*a*) the Copernican system and (*b*) the Ptolemaic system. In what way is the Copernican explanation superior?
6. What change did Kepler introduce into the Copernican system? Why was this change necessary?
7. The sun, moon, and planets all follow approximately the same path across the sky, from east to west; none of these objects ever appear in the far northern sky or the far southern sky. What does this indicate regarding the arrangement in space of the members of the solar system?
8. From observations of the moon's motion, how could you prove that it is not another planet revolving around the sun, instead of a small body revolving about the earth?

CHAPTER II

The Sun and Its Family

WHILE Kepler was slowly and painfully working out the orbits of the planets, his Italian contemporary Galileo (Fig. 8) found support for the Copernican theory from a different angle. Hearing that a Dutch lens maker had invented a device for producing enlarged images of distant objects, Galileo undertook to construct a similar instrument. No plans or exact descriptions were available, but Galileo was familiar enough with lenses to guess the secret. Within a few weeks he succeeded in building a crude but practical telescope.

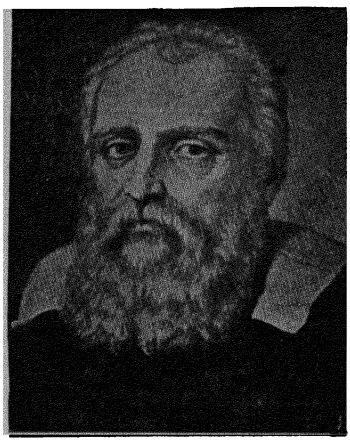


FIG. 8. *Galileo Galilei (1564–1642).* (*Courtesy of Gramstorff Bros., Inc.*)

To his friends the telescope was a marvelous and fascinating toy, but to Galileo it was a splendid new tool for astronomical observation. Quickly he scanned the moon, the sun, the planets, finding one fact after another which agreed better with Copernicus's view than with Ptolemy's. The moon, far from being a "perfect" heavenly sphere, showed a surface scarred with craters and jagged mountains. The sun's brilliant face was blemished with small black spots, spots which moved from day to day as if the sun were rotating—most unseemly behavior for a celestial body. The planets appeared in Galileo's telescope as small round disks, in contrast to the fixed stars which remained mere points of light. Venus, watched for several months, showed changes of phase like those of the moon [crescent, half-moon, full moon, etc. (Fig. 9)] and a great variation in size—all easily explained on the Copernican assumption that Venus follows an orbit within that of the earth. Near Jupiter, Galileo saw four small bodies which moved slowly from one side

to the other of the planet's disk. and he rightly concluded that they were moons following orbits around Jupiter; here indisputably were celestial objects which did not revolve about the earth.

All this was bitter medicine to adherents of the stationary-earth, perfect-sphere hypothesis, more bitter than the elliptic orbits of Kepler, since these observations required no mathematical skill to understand. Galileo was quick to follow up his advantage, deriding with eloquent tongue and witty pen the willful blindness of the followers of Aristotle and Ptolemy. Of this long quarrel with the orthodox churchmen and philosophers of his day we shall learn more in the next chapter, after examining yet another field in which Galileo cast doubt on the authority of Aristotle.

FIG. 9. *Photographs of Venus, showing how the planet's appearance changes as it and the earth move in their orbits. (Photographed by E. C. Slipher at the Lowell Observatory.)*

The present chapter we shall devote to a closer examination of those objects which were the chief concern of Galileo's astronomical studies—the sun, the moon, and the planets. Our knowledge of these bodies has come in part through improvements in Galileo's crude telescope, in part through the invention of new instruments undreamed of in Galileo's day. But we shall be less interested here in the methods of astronomy than in its results. We need first of all an accurate description of astronomical objects, particularly of the earth's nearest neighbors in space.

The Solar System

The greatest modern telescope gives no more direct information about a star than did Galileo's instrument. Through a telescope, just as to the naked eye, a star is simply a tiny point of light. Most of the planets, on the

other hand, are magnified to clear disks by telescopes of smaller size than Galileo's. This does not mean that the stars are smaller than planets, but only that they are very much farther away.

It is roughly 4,000 million miles from the sun to the outermost planet, 27 million million miles from the sun to the nearest star. To give these distances meaning, let us try a drastic reduction of scale. Take a golf ball to represent the sun, and a dozen feet away put a small sand grain to represent the earth. The farthest planet, Pluto, will be another sand grain following an orbit 1,000 feet (ft) in diameter about the golf ball. Within this 1,000-ft orbit are all the other planets. But to place the nearest star in our model, we must take another golf ball 600 miles (mi) away!

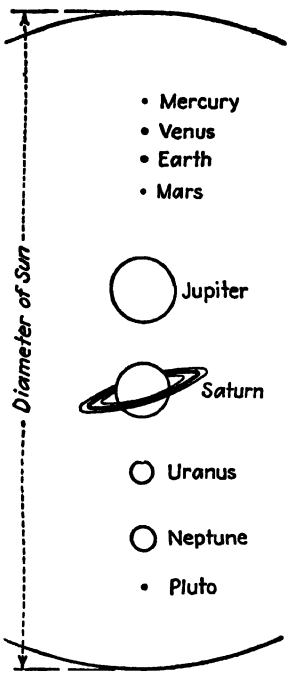


FIG. 10. *Relative sizes of planets and sun.* (*Watkeys.*)

We are isolated in empty space, we and the sun and the other eight planets. A little system of nine spheres, circling again and again about the bright sun, poised in emptiness, separated by unimaginable distances from everything else in the universe.

Because the sun is its central figure, this isolated whirligig is called the ***solar system.*** Before 1600 the system was known to contain eight parts: the sun, the moon, the earth, and five other planets. Galileo added four new bodies to the system—the moons, or ***satellites,*** which he found revolving about Jupiter. Since his time telescopic improvements have made possible the discovery of many more members of the sun's family. The list of planets now includes nine: in order from the sun they are Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, Pluto (Fig. 10). All except Mercury, Venus, and Pluto have satellites, smaller bodies revolving around them as the moon revolves about the earth. More than 1,500 small objects called *planetoids* (or *asteroids*), all less than 500 mi in diameter, follow separate orbits about the sun. Comets and at least some meteors, in Galileo's time thought to be atmospheric phenomena, are now recognized as still smaller members of the solar system.

Not only is the entire solar system isolated in space, but each of its principal members is separated from the others by distances which seem very large by everyday standards. From the earth to our nearest neighbor, the moon, is 240,000 mi (mean distance); from the earth to the sun is 93 million miles (mean distance). A rocket traveling away from the earth

at a steady speed of 100 miles per hour (mi/hr) would take more than three months to reach the moon, more than a century to reach the sun. Let us return for a moment to the model of a preceding paragraph, in which a golf ball represented the sun and a grain of sand 12 ft away the earth. On this scale the moon would be scarcely more than a dust speck, about ½-inch from the sand grain. The largest planet, Jupiter, would be the size of a small pebble, 60 ft from the golf ball. With three smaller pebbles, three more sand grains, and a few more dust specks, all within the 1,000-ft wide orbit of Pluto, the model is complete. An extremely empty structure, this solar system, its members separated by distances enormous compared with their size.

Planets *revolve* about the sun, and *rotate* on their axes. Their motion around the sun follows Kepler's three laws: Each planet describes an ellipse having the sun at one focus; the motion is fastest when the planet is nearest the sun, slowest when the planet is farthest away; planets with

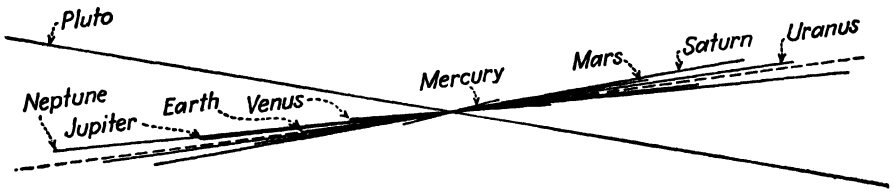


FIG. 11. *Orbits of the planets seen edgewise, to show that they all lie nearly in the same plane. (Sizes of orbits not to scale.)*

small orbits move rapidly, planets with large orbits slowly. Planetoids follow similar elliptical orbits, nearly all of them lying between the orbits of Mars and Jupiter. Satellites move in small ellipses around their planets. Comets describe ellipses around the sun, generally very flat, elongated ellipses in contrast to the nearly circular planetary orbits. Two facts about the motions of the solar system are especially important: (1) Nearly all the motions—revolutions of planets, planetoids, and satellites, axial rotations of sun and planets—are in the same direction; only the rotation of Uranus, and the revolutions of a few minor satellites, run contrary to the general motion. (2) All the orbits except those of comets lie nearly in the same plane (Fig. 11).

Planets, planetoids, and satellites shine only by reflected sunlight, so that observation of one of these objects is limited to that half which is directly exposed to the sun. Planets with orbits larger than the earth's never come between us and the sun, so that we can always see nearly the whole of their illuminated sides. Mercury and Venus, however, with orbits smaller than the earth's, are nearly between us and the sun for a good part of each revolution. In this position their dark sides are turned toward us, and we see them either not at all or as thin crescents (Fig. 9).

The Earth as a Planet

According to the figures in Table I (page 30), the earth is a very ordinary sort of planet. Neither size, nor position, nor rate of motion distinguishes it from its brethren. Probably our planet is unique in the possession of intelligent life on its surface, but even of this we cannot be quite sure.

The earth is not quite a perfect sphere, its diameter measured between the poles being about 26 mi shorter than the diameter at the equator. Such irregularities of the surface as mountains and ocean basins likewise mean a deviation from sphericity; but even the greatest of these irregularities, compared with the earth's diameter, amount to less than the roughness on the skin of an orange.

The earth's daily rotation on its axis and its yearly revolution about the sun we have discussed at some length in Chap. I. Most obvious evidence of the earth's rotation is the rising and setting of the sun and stars; most obvious evidence of the revolution is the movement of the sun and planets among the stars. Both of these observations can, of course, be explained, as Ptolemy tried to explain them, by movements of the heavens rather than the earth. But this explanation, worked out in detail to fit all the observed motions, becomes unconscionably complicated. Kepler's demonstration of the simplicity with which all detailed observations can be explained by the Copernican system is good evidence that the earth moves rather than the heavens.

Since Kepler's time other evidences of the earth's motions have come to light, which a supporter of Ptolemy would find even more difficult to explain than planetary movements. Of these lines of evidence we shall discuss here the two simplest.

The Foucault Pendulum—Evidence That the Earth Rotates. Hang a heavy weight from a long cord or wire, and set it to swinging as a pendulum. On the floor below the weight, mark the path along which it swings. Let the pendulum move undisturbed for an hour or so, then examine again the path of its swing. You will find that the pendulum's path has apparently turned through a small angle from its original direction. After each successive hour you will find that the path has shifted farther and farther around. Now there is nothing about the pendulum itself to cause a turning of this sort: once started, a pendulum should swing along the same path until friction brings it to a stop. If the pendulum does not turn, then the shift in its path must be due to a turning of the earth beneath it. If you could perform the experiment at the north or south pole, the pendulum's path would turn through 360° each 24 hours; at other points on the earth's surface the turning is slower. This tendency to turn because of the earth's rotation is not, of course, limited to the Foucault pendulum, but applies

to all objects not rigidly connected with the earth's surface. In a later chapter, for instance, we shall find that wind directions are greatly modified by the earth's axial motion.

Stellar Parallax—Evidence That the Earth Revolves around the Sun. One objection urged by Aristotle against any theory of the earth's motion may be rephrased in modern terms as follows: if the earth moves about the

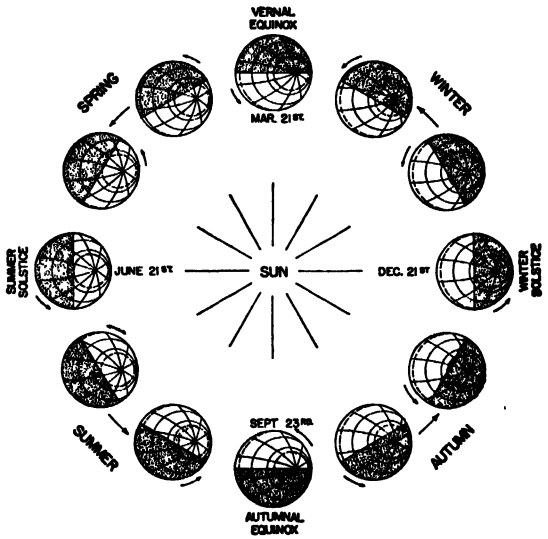


FIG. 12. *How the earth's tilted axis produces the seasons. In this diagram the north end of the axis is shown tilted toward a distant point above and to the right of the page. (From Elements of Geography by Finch and Trewartha.)*

sun, why should not the nearer stars appear to move with respect to the more distant ones, just as nearer objects in a landscape appear to move with respect to more distant ones when we ride past them? The same objection reappeared when Copernicus published his hypothesis. It is a good argument, and can be answered only by the assumption that the stars are so very far away that their shifts in position are not detectable with ordinary instruments. We know today that this assumption is correct: powerful modern telescopes show very slight shifts in the positions of the nearer stars as the earth moves in its orbit, but the movement is far too small for Galileo or his immediate successors to have discovered it. The change in position of objects against a distant background due to motion of the observer is called *parallax.* The measurement of stellar parallaxes is important in astronomy not only as evidence for the earth's revolution but as a means of determining the distances of the nearer stars (Fig. 317, page 621).

The earth's axis of rotation is inclined to the plane of its orbit at an angle of 66.5°. It maintains this angle throughout its journey around the

sun, so that the north pole for part of the year is tilted toward the sun, for the remainder of the year away from the sun (Fig. 12). This means that the vertical rays of the sun fall north of the equator for half the year, south of the equator for the other half. This shift of the sun's maximum illumination from one part of the earth to another gives the earth its seasons. When the north pole is inclined toward the sun, so that the maximum illumination is north of the equator, we in the northern hemisphere have our summer; when the north pole is inclined away from the sun, we have winter. Observing all this from the earth, we say commonly that the sun is high in the sky during summer, low in the sky during winter.

Our planet, then, is a spinning sphere, moving around the sun, its axis inclined to the plane of its motion. From the surface of this whirling ball we look out at the universe. We see our surroundings much as a man might see an earthly landscape from a Ferris wheel mounted at an angle on a moving train. It is inevitably difficult for the novice to adjust himself to observation from this moving and tilted platform, so different from the fixed, flat earth of our daily lives. But one cannot be "at home in the starry heavens" until this adjustment has been accomplished.

Perhaps a model will help. Use an apple or a rubber ball for the earth and an electric light for the sun. Hold the apple between thumb and forefinger, and rotate it about these points. Note how one portion after another of the apple's surface comes from darkness into light, then back into the shadow again. Now imagine yourself reduced to the size of a bacterium, looking out at the room from a point on your apple's surface halfway between its equator and north pole. Your tiny portion of the apple looks flat, and the flat surface determines your horizon in every direction. Above this horizon you see about half of the objects in the room. Now as the apple rotates, you see some objects disappearing below your horizon on one side, new objects "rising" above your horizon on the other. Presently a glow appears on the horizon, and slowly the electric light comes into view: it is the dawn of your bacterial day. The light climbs higher above your horizon, passes overhead, and "sets" below the horizon once more. Now, in your human incarnation, carry the apple around the light, keeping it rotating meanwhile. As a bacterium again, you will continue to experience day and night; but if you are an observant bacterium, you will find that each time the light "dawns," it appears against a different background. When the circuit is completed, you will observe that the light has apparently followed a circular path past all the objects in the room and has returned to its starting point. Finally, as a human once more, tilt the apple's axis toward the light, pointing it toward some distant object on the ceiling. Now carry the apple around its orbit again, rotating it as before but this time keeping its axis pointed toward the

same point on the ceiling. Your bacterial manifestation will find at first that the "days" are long, the "nights" short, and that the light stands high above his horizon at noon. When the apple is a quarter of the way around, "days" and "nights" will be equally long; when it is halfway around, the north pole will be inclined away from the sun, and you will have long nights and short days. Thus you will experience a change of seasons from "summer" to "winter."

Now if you will fit yourself again into the usual scale of things, remembering that the earth's surface looks to you like an apple to a bacterium; if you will keep in mind that your narrow horizon seems flat only because you are not tall enough to peer over its edge; if you will watch the sun set and the stars rise a few times, deliberately insisting to yourself that you are moving and not they; if you will keep some track of the stars that appear just after sunset for a month or so and note how they slowly change; if you will remember that Polaris corresponds to the spot on the ceiling toward which the apple's axis pointed, and note that the noonday sun is apparently closer to Polaris's position in summer than in winter; then you will truly comprehend the simple statement that "the earth is a planet."

The Moon

In an orbit half a million miles in diameter, the moon circles the earth approximately once a month. A giant among satellites, our moon (2,160 mi in diameter) is exceeded in size only by two of Jupiter's retinue. Like the earth, the moon turns on its axis as it revolves; but the rotation keeps pace with the revolution, so that the moon turns completely around only once during each circuit of the earth. This means that the same face of our satellite is always turned toward us, while the other side remains forever hidden.

Each month the moon completes its familiar cycle of *phases:* first a thin crescent in the western sky at sunset; growing and moving eastward (relative to the stars) with each successive night, through the stage of half-moon, until after two weeks the full moon rises in the east at sunset; then waning, becoming a thin crescent which rises just before the sun; finally vanishing altogether for a few days before its next appearance as a crescent. These different aspects represent the amounts of the moon's illuminated surface visible to us in different parts of its orbit. When the moon is full, it is on the opposite side of the earth from the sun, so that the side facing us is fully illuminated. In the "dark of the moon," it is moving approximately between us and the sun, so that the side toward the earth is in shadow. Figure 13 will help to make the phases clear, but the best way to understand them is to try the apple-and-electric-light setup once more, using a smaller apple for the moon.

When the moon is behind the earth (full moon), how can the sun illuminate it at all? Why doesn't the earth's shadow hide it completely? Again, when the moon passes between sun and earth, why isn't the sun hidden from view? The answer to these questions needs a three-dimensional figure, for which Fig. 14 is a poor substitute: the moon's orbit is tilted at a small angle to the earth's orbit, so that ordinarily the moon passes either slightly above or slightly below the direct line between sun and earth. On the rare occasions when the moon does pass directly before or behind the earth, an ***eclipse*** occurs—an eclipse of the moon when the earth's shadow obscures the moon, an eclipse of the sun when the moon's shadow touches the earth (Fig. 15).

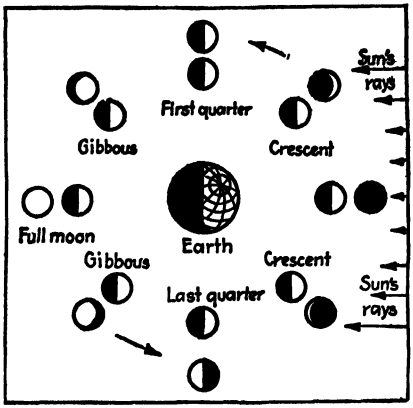


FIG. 13. *Phases of the moon. The inner ring of circles shows the moon as an observer far out in space would see it, with one half continually bathed in sunlight. The outer ring of circles shows the moon's appearance as seen from the earth. (From Handbook of the Heavens.)*

Observation of the moon with a small telescope brings out at once the chief features of its landscape (Figs. 16, 17): wide plains, jagged mountain ranges, and the innumerable mountain rings called "craters." Each mountain stands out in startling clearness. No cloud or haze hides the smallest detail. Mountain shadows are black and sharp-edged. When the moon passes before a star, the star remains bright and clear up to the moon's very edge. From these simple observations we conclude that the

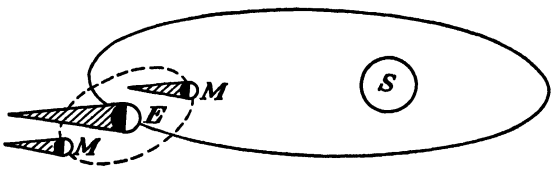


FIG. 14. *Eclipses do not occur every month, because the moon usually passes above or below the direct line from earth to sun. (Distances and sizes not to scale.)*

moon has no atmosphere. Without air, water could scarcely be present, since it would evaporate immediately; this inference is confirmed by the complete absence of lakes, oceans, or signs of erosion by rivers.

The first hardy pioneer who steps from his rocket onto the moon's surface will find himself in an eerie world. Without air to dim and scatter its light, the sun will blaze down from a black sky, with stars visible in broad daylight. There will be no wind, no sound. Temperatures, unmodi-

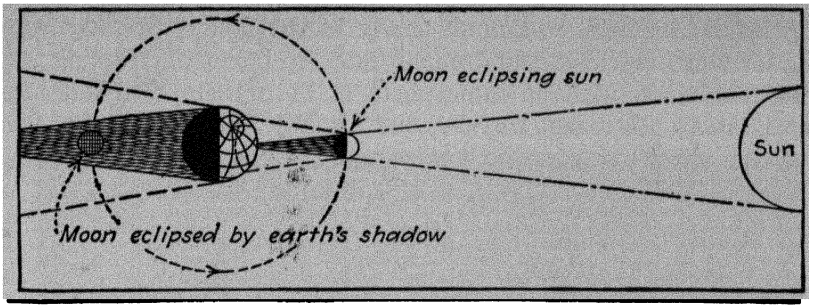


FIG. 15. *Eclipses of the sun and moon. (From Handbook of the Heavens.)*

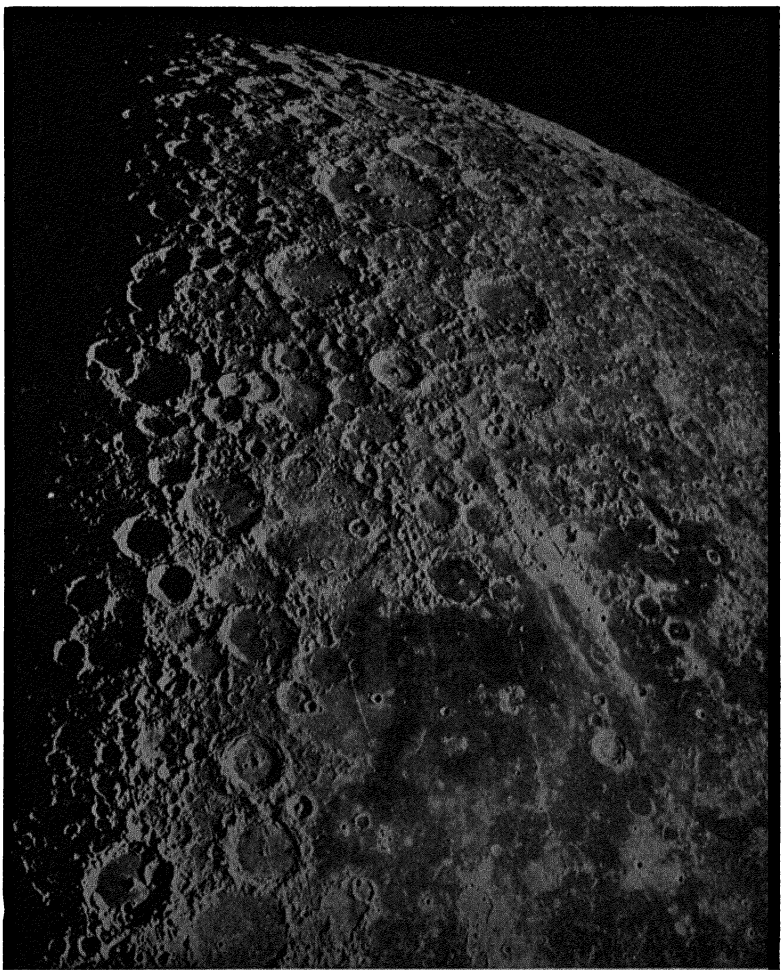


FIG. 16. *Part of the moon's surface, showing numerous craters. (Photographed at the Mt. Wilson Observatory.)*

fied by an atmosphere, will climb nearly to the boiling point of water during the long lunar day, and drop abruptly far below the freezing point at night. Meteors, with no atmosphere to burn them, will fall intermittently about the rocket traveler, and his stay will be abruptly termi-

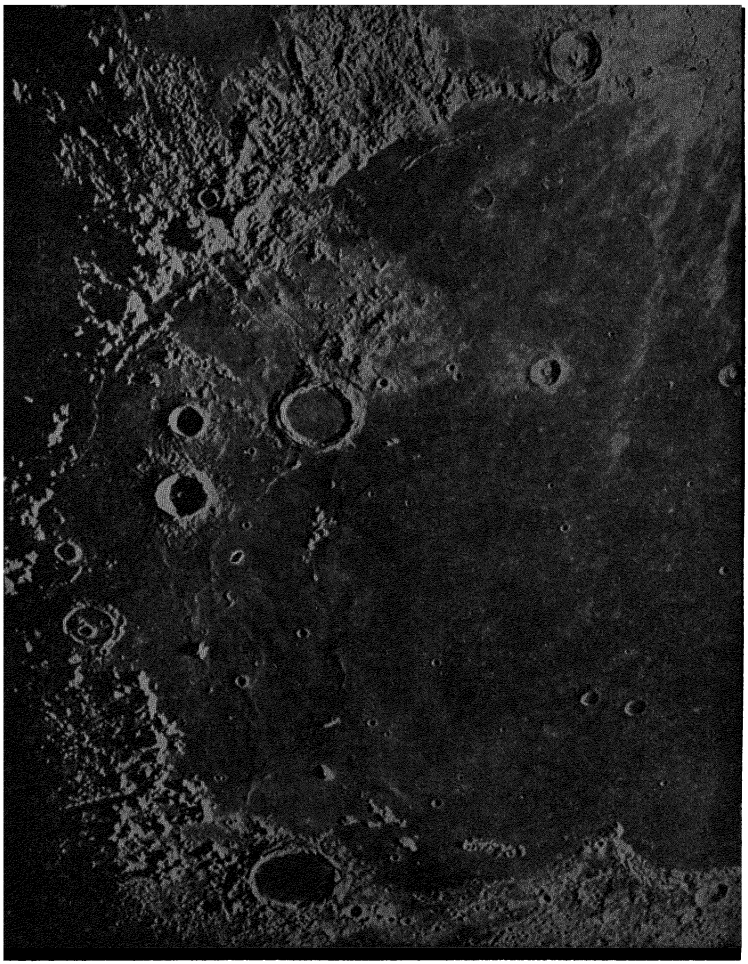


FIG. 17. *A great lunar plain bordered by high mountain ranges. (Photographed at the Mt. Wilson Observatory.)*

nated if one should chance to strike him. Probably these meteors will be the only moving objects he can find; all else will be motionless, desolate, a desert world beside which the earth's bleakest deserts must pale. Black shadows and glaring sunlit surfaces; great plains floored with fragments of volcanic rock, steep and rugged mountains rising out of them; nowhere any rounding of peaks or softening of slopes by vegetation and running water.

No major engineering difficulty stands in the way of a rocket expedition to the moon. But the enormous financial outlay required, and the slim chances for survival on the moon's inhospitable surface, will probably delay the attempt for many years.

The Sun

The sun is a star: a rather ordinary star, somewhat smaller than most. But to members of the solar system, the sun is a very large and important object. Its diameter of close to a million miles dwarfs all the planets, alone

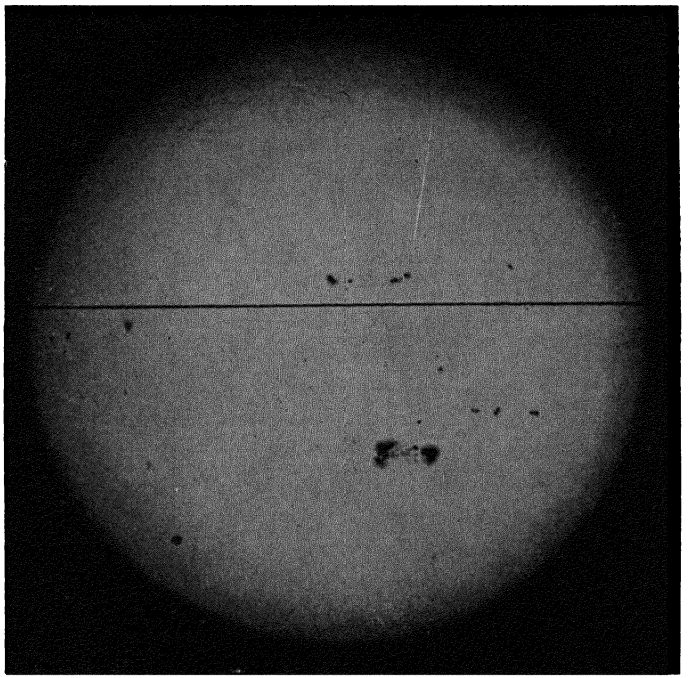


FIG. 18. *Photograph of the sun, showing sunspots. (Photographed by Wilson at the Goodsell Observatory.)*

or in combination (Fig. 10). It is the source of light without which planets and satellites could not shine, and the source of light and heat which make life possible on the earth.

Seen through a telescope, the sun appears as a yellowish disk, noticeably darker near the edge than at the center (Fig. 18). The darkening toward the edge is a hint that the material of the sun is gaseous for a considerable depth, since a solid or liquid surface should appear uniformly illuminated.

The only conspicuous markings on the sun are the small black spots called *sunspots* (Fig. 18). These change gradually in form, each one grow-

ing and then diminishing in size, a single spot lasting from a day or so to several months. The largest ones attain diameters of many thousand miles, large enough to engulf the earth with plenty of room to spare. Galileo, one of the first to observe sunspots, noted that they moved across the sun's disk, and he interpreted this observation as evidence that the sun rotates on an axis. Modern measurements show that points on the sun's equator complete a rotation about once in twenty-five days; curiously, other parts of the surface do not rotate as fast, points near the

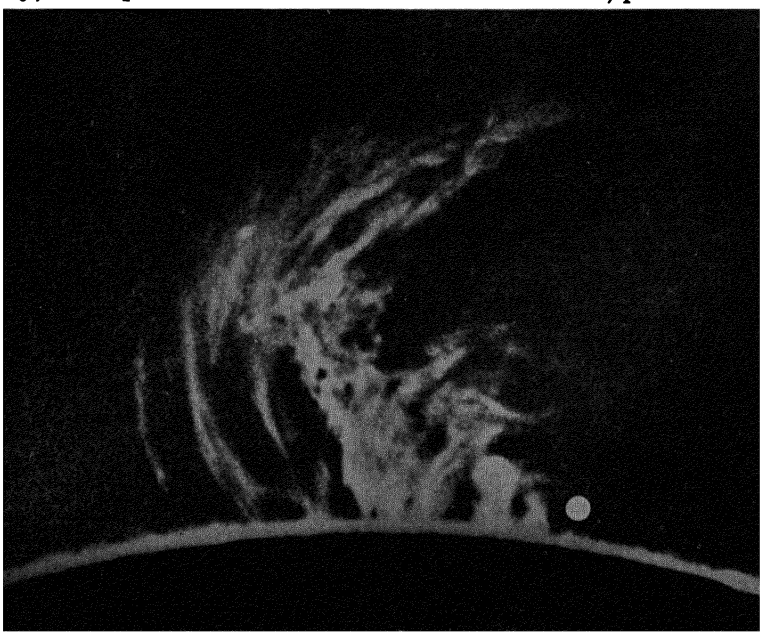


FIG. 19. *Photograph of prominences at the edge of the sun. The white spot shows the size of the earth to the same scale. (Photographed at the Mt. Wilson Observatory.)*

poles requiring about thirty-five days for each rotation. This fact is another indication that the outer part of the sun is fluid.

The temperature of the visible part of the sun is about 6000°C (10,000°F). At such temperatures all known substances are gaseous. Toward the interior of the sun, temperatures rapidly increase, up to several million degrees, so that the sun is presumably gaseous throughout —although whether any ordinary term like "gas" can apply to matter at the prodigious temperatures and pressures of the sun's deep interior is open to question.

The most favorable time for observing the cooler gases of the sun's atmosphere is during a total eclipse, when the moon shuts off the glare of the central disk. Total eclipses are rare, and the period of totality never

lasts more than a few minutes, but much of our knowledge of the sun has come from these brief periods. Near the edge of the sun during totality often appear reddish streamers and clouds of glowing gas, called *prominences*, which rise thousands, sometimes hundreds of thousands, of miles upward from the sun (Fig. 19). Recently methods have been devised for photographing prominences in broad daylight, but they are visible with ordinary instruments only during eclipses. Also limited to the time of a total eclipse are observations of the sun's *corona* (Fig. 20), the extreme upper part of the sun's atmosphere, extending outward a million miles and more.

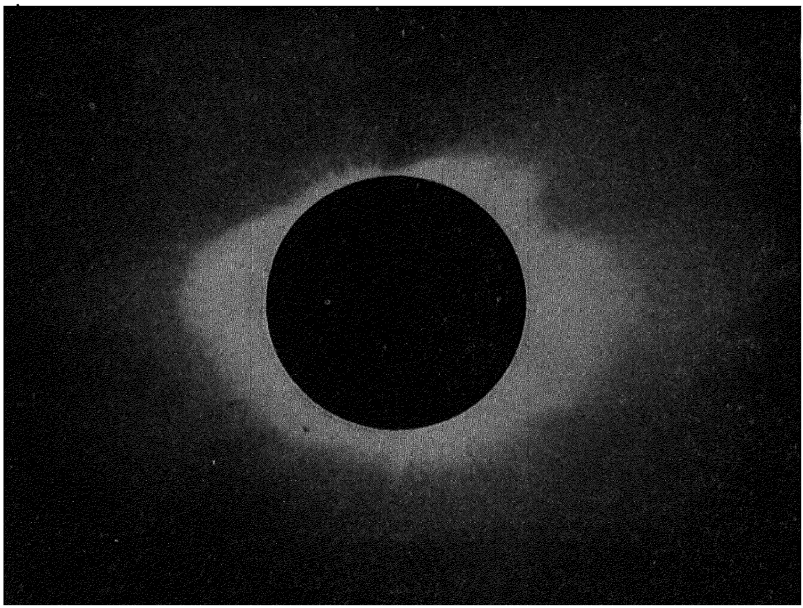


FIG. 20. *Photograph of the sun during the total eclipse of Aug. 31, 1932, showing the corona. (Photographed at Fryeburg, Maine, by Wright of the Lick Observatory.)*

By far the most impressive fact about the sun is its ability to radiate light and heat year after year, century after century. Of this prodigious output of radiation the earth receives but a tiny fraction, yet it is ample to supply warmth and light for the entire planet. How the sun can maintain this lavish expenditure of energy without exhausting its resources has only recently been explained, as a process involving slow conversion of the sun's mass into radiation.

Into the problems of the sun and its radiation we shall inquire again in a later chapter, after fortifying ourselves with much necessary information about matter and radiation in general.

The Planets

Table I summarizes a few important facts regarding the planets. The figures are somewhat rounded off, to make comparisons more obvious.

TABLE I. THE PLANETS

	Millions of miles from sun (mean)	*Mean diameter, miles*	*Mass, taking earth's mass equal to 1*	*Time of rotation*	*Time of revolution*	*Number of satellites*
Mercury	36	3,000	0.06	88 days	88 days	0
Venus	67	7,700	0.8	?	225 days	0
Earth	93	7,900	1.00	24 hr	365 days	1
Mars	142	4,200	0.11	24.5 hr	687 days	2
Jupiter	483	88,000	318.0	10 hr	12 yr	11
Saturn	886	75,000	95.0	10.2 hr	29.5 yr	9
Uranus	1,780	31,000	15.0	10.8 hr	84 yr	4
Neptune	2,790	33,000	17 0	15.8 hr	165 yr	1
Pluto	3,670	Small	0.8	?	248 yr	?

Mercury, smallest and swiftest of the planets, is too close to the sun for detailed observation of its surface. The planet rotates only once for each revolution, so that the same side perennially faces the sun. Mercury has either no atmosphere at all, or an exceedingly thin one. Surface temperatures on the sun-baked side are higher than the melting point of lead, while the other face remains very cold.

In size and mass the bright planet *Venus* resembles the earth more nearly than any other member of the sun's family. About the surface of Venus we can get little information, for the planet's atmosphere is filled perpetually with thick layers of cloud. In the part of the atmosphere above the clouds the spectroscope reveals an abundance of carbon dioxide, a heavy gas of which the earth's atmosphere contains only a small percentage. On the earth, carbon dioxide is important not only for the growth of plants, but as a "blanket" in preventing rapid loss of heat from the ground after sunset. Venus, blanketed more effectively than the earth, must retain more heat; recent estimates suggest a surface temperature near the boiling point of water. Since the temperature is so high, and since no oxygen can be detected in the atmosphere, the existence of life on Venus is extremely doubtful.

Mars has long had a peculiar fascination for astronomers and laymen alike, for it is the only other known body in all the universe on which surface conditions seem suitable for life as we know it. Yet Martian climates are exceedingly severe by our standards, and, if life exists, it has adapted itself to conditions that would soon destroy most earthly

organisms. Mars rotates on its axis in a bit over 24 hr; its revolution about the sun requires nearly two years; and its axis is inclined to its orbit at nearly the same angle as the earth's. These facts mean that the Martian day and night have about the same length as ours, that Martian seasons are six months long and at least as pronounced as ours. Farther from the sun than the earth, Mars receives considerably less light and heat; its atmosphere is very thin, so that a much smaller fraction of the sun's heat is retained after nightfall. Temperatures at midday in summer may rise to 80°F but at night must fall far below zero. Winter temperatures, and summer temperatures near the poles, remain much lower. Two other major difficulties which life must face on Mars are a scarcity of water and a very limited supply of oxygen in the atmosphere. That water exists is indicated by the white polar caps (Fig. 21) and by occasional clouds; oxygen has recently been detected spectroscopically. But the amount of each is very small.

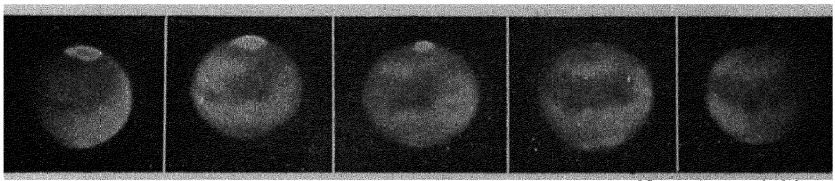


FIG. 21. *Photographs of Mars taken at intervals of a few weeks. Note the shrinking of the white polar cap and the change in appearance of the dark areas. (Photographed by E. C. Slipher at the Lowell Observatory.)*

Despite these frigid, desert conditions, one line of evidence suggests strongly that life does exist on Mars: certain areas show conspicuous seasonal color changes, from red-brown in winter to dark green in spring and summer. Other explanations are possible, but the most reasonable one ascribes these changes to the growth of vegetation during the warm season. As to what this vegetation is we can only guess; it may be some very primitive form, like mosses or lichens.

A few decades ago the Italian astronomer Schiaparelli and the American Lowell reported that the surface of Mars is covered with networks of fine dark lines, called "canals" (a poor English translation of the Italian "canali," meaning "channels"). The straightness and geometric patterns of these "canals" were considered evidence of the handiwork of intelligent beings. But the "canals" cannot be clearly photographed, and other observers have failed to find them. Such fine details, at the limit of visibility, are extremely difficult to observe, because of imperfections in telescope lenses and blurring due to currents in the earth's atmosphere. Probably the "canals" are illusions; certainly the existence of intelligent animal life on a planet so poorly supplied with air and warmth is not likely.

The giant planet *Jupiter*, like Venus, is eternally shrouded in clouds, so that we never see its solid surface. But Jupiter's clouds are conspicuously banded, and show other semipermanent markings which make possible a determination of the planet's period of rotation. This turns out to be less than 10 hr, which means that points on Jupiter's equator travel at the enormous speed of 28,000 mi/hr (compare the earth's equatorial speed, 1,040 mi/hr). Because of the rapid rotation (see page 81) Jupiter is much more conspicuously flattened at the poles than is the earth (Fig. 22). Jupiter's total volume is about 1,300 times that of the earth, but its mass is only 300 times as great. The material of the planet must therefore be much lighter on the average than the earth's material. From this fact and others, it is concluded that Jupiter's atmosphere is enormously thick.

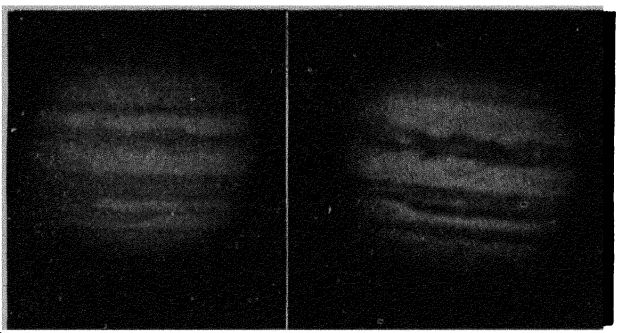


FIG. 22. *Photographs of Jupiter taken two years apart, 1915 and 1917. (Photographed by E. C. Slipher at the Lowell Observatory.)*

Its temperature is very low, about −200°F in the parts that we can observe. Evidently Jupiter's clouds cannot be made of water droplets. Spectroscopic observation shows that the clouds consist of liquid ammonia and methane, two substances which we know as gases on the earth. Probably the atmosphere which supports the clouds consists of the light gases hydrogen and helium.

The four satellites of Jupiter which Galileo discovered are conspicuous objects in a small telescope. The remaining seven are very small, two having escaped detection until recent years.

Saturn, in its setting of brilliant rings, is the most beautiful of the earth's kindred. The planet itself is much like Jupiter: similarly flattened at the poles by rapid rotation, similarly possessing an atmosphere thousands of miles thick, its solid surface similarly hidden by banded clouds. Farther from the sun than Jupiter, Saturn is considerably colder; ammonia is largely frozen out of its atmosphere, and its clouds consist mostly of methane.

The famous rings (Fig. 23), two bright ones and a fainter inner one, surround the planet in the plane of its equator. This plane is somewhat

year journey around the sun, we see the rings from different angles. Twice in the thirty-year period the rings are edgewise to the earth; in this position they are practically invisible, which suggests that their thickness is small—estimated at about 50 mi. The rings are not solid sheets as they appear, but consist of myriads of small bodies, each revolving about the planet like a tiny satellite. Saturn's nine ordinary satellites have orbits lying outside the outermost ring.

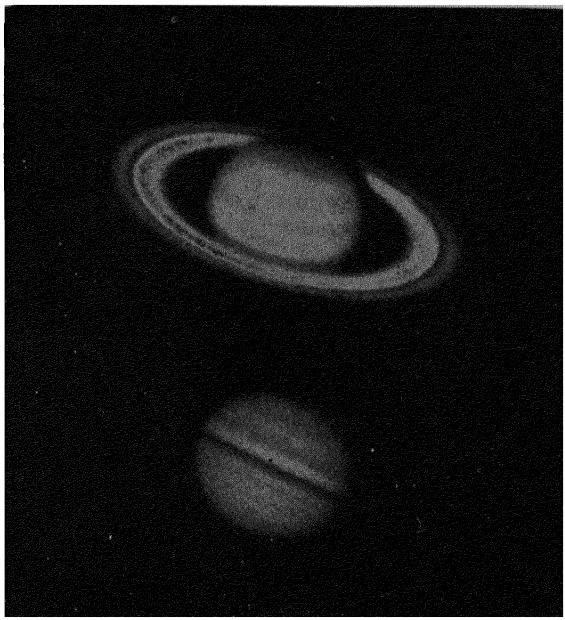


FIG. 23. *Photographs of Saturn with rings wide open in 1916 (above) and edgewise in 1921 (below) when they are too faint to be seen except as a dark band across the planet. (Photographed by E. C. Slipher at the Lowell Observatory.*

The three outermost planets, *Uranus*, *Neptune*, and *Pluto*, owe their discovery to the telescope. Uranus was found quite by accident in 1781, by the great English astronomer Herschel; Neptune (1846) and Pluto (1930) were found as a result of predictions based on their gravitational effects on other planets (see page 84). Uranus and Neptune are large bodies, each with a diameter about four times the earth's; Pluto is probably about the size of Mars. In their large size, their low density, their reflecting power for sunlight, Uranus and Neptune resemble Jupiter and Saturn, hence probably have similar thick, cloudy atmospheres. Whether Pluto has an atmosphere, how quickly it rotates, how many satellites it has even an exact figure for its size, must await further observation. The planet is so small, so far away, so feebly illuminated that exact informa-

Meteors and Comets

Meteors are small fragments of matter which the earth meets as it travels around the sun. Moving swiftly through the atmosphere, meteors are quickly heated to incandescence by friction, and we see their paths as glowing streaks ("shooting stars") against the sky. A few meteors fall to earth, but the great majority are disintegrated by heat and become atmospheric dust. If it were not for this heating and disintegration of meteors by the atmosphere, life on earth would be hazardous, for our planet is subject to a continual bombardment of several million meteors each day.

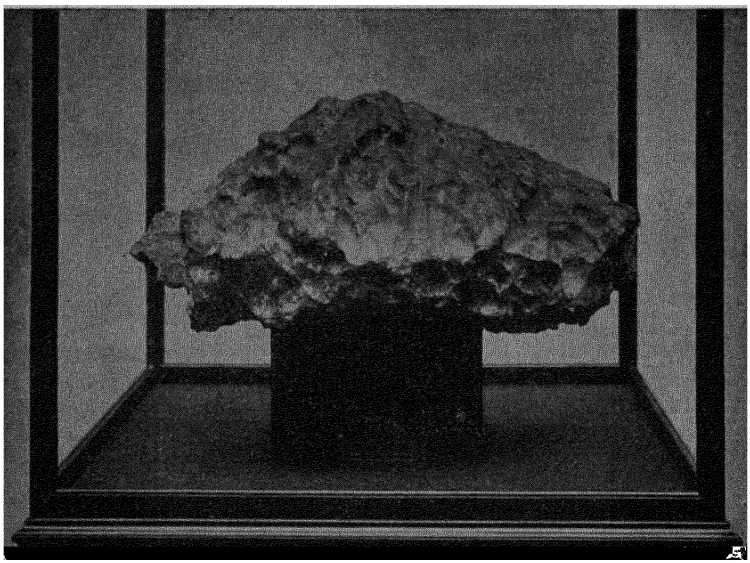


FIG. 24. *A large iron meteorite found near Tonopah, Nevada. Weight 1,486 kg (3,276 lb.). (Courtesy of the Field Museum.)*

The largest known fallen meteors, or *meteorites* (Fig. 24), weigh several tons. They range down to the size of small pebbles and sand grains, possibly smaller. Analyses of meteorites reveal no elements not found on the earth, although sometimes their elements are put together in unusual compounds. Nearly all meteorites which have been analyzed fall into one of two classes: (1) stony meteorites, with compositions like ordinary silicate rocks; (2) iron meteorites, consisting largely of iron with some nickel.

Meteors cannot be seen until they strike the earth's atmosphere, so that the problem of their ultimate origin is a difficult one. Probably the slower moving ones are members of the solar system, while the faster ones are intruders from outer space.

Comets appear as small, hazy patches of light, often accompanied by long, filmy tails (Fig. 25). Most comets are visible only telescopically, but occasionally one becomes conspicuous to the unaided eye. Watched for a few weeks or months, a comet at first grows larger, its tail longer and more brilliant; then it slowly fades, losing its tail and at length disappearing altogether. Paths followed by comets are quite different from the nearly circular planetary orbits. Usually a comet moves toward the sun

FIG. 25. *Halley's comet, photographed May 18, 1910. The bright object at the right is the planet Venus. (Photographed at the Lowell Observatory.)*

from far out in space beyond the orbit of Pluto, approaches the sun closely, swings around it, then disappears toward the outer parts of the solar system. Sometimes the orbit is a long, narrow ellipse, so that the comet returns at long intervals to the sun and is definitely a part of the solar system. Sometimes the orbit appears to be an open curve (parabola or hyperbola), along which the comet would move indefinitely outward into space and never return to the sun. Probably all comets belong originally to the solar system, and one is deflected into an open-curve orbit only if it chances to pass too close to a planet on its journey around the sun.

Comets are visible, and have tails, only when close to the sun The tails consist of gas and fine particles expelled by pressure of the sun's radiation. The light given out by a comet is produced by the sun, in part as reflected sunlight, in part by an excitation of the comet's substance somewhat similar to the excitation which makes neon glow in illuminated signs. Even when a comet is brightest, stars behind it can often be seen shining through both its tail and its head, showing that the comet is not a continuous solid body. Probably the heads are loose aggregates of meteor-like bodies—a suggestion supported by the fact that some comets with elliptic orbits have apparently disintegrated into swarms of meteors, which the earth meets whenever it crosses the original paths of the comets.

Among all the objects in the solar system, with their endless diversity of size, temperature, atmospheres, and composition, our own small planet seems to be the only one with that happy combination of moderate and uniform warmth, abundant liquid water, and an atmosphere with plenty of oxygen and carbon dioxide, which makes possible a prolific development of living things. But perhaps we should not be quite so serenely certain that these conditions are indispensable, for life has an amazing capacity to adapt itself to inhospitable environments. Quite probably life of some sort has found a foothold on the deserts of Mars. Perhaps even the hot, cloudy surface of Venus, or the cold ammonia-methane worlds of Jupiter and Saturn, have developed their own strange forms of life.

Questions

1. Draw a diagram to show the relative positions of earth, sun, and Mars when Mars is directly overhead at midnight. About how far is Mars from the earth in this position? How far is Mars from the earth when it is directly opposite the sun from us?
2. Draw a diagram to show the relative positions of Venus, sun, and earth when Venus appears as a thin crescent in our telescopes.
3. About how long a time elapses between sunrise and sunset on the moon? On Mars? On Jupiter? On Mercury?
4. Suppose you were on Mars, observing the earth through a telescope. Describe the changes in the earth's appearance as it moves around its orbit. Describe the apparent motion of the earth as viewed from Mars.
5. What sort of seasons would the earth have if its axis were perpendicular to the plane of its orbit? If its axis were parallel to the plane of its orbit, but maintained a fixed direction in space?
6. In what phase must the moon be at the time of a solar eclipse? At the time of a lunar eclipse?
7. Why would the danger of being struck by a falling meteorite be greater on the moon than on the earth?
8. Which planets can never be seen rising in the east at sunset? Why?
9. Why is the average temperature on Mars lower than that on the earth? Why is the Martian temperature more changeable?
10. Why isn't Mars eclipsed by the earth when the earth passes directly between Mars and the sun?

CHAPTER III

Force and Motion

Two outstanding questions regarding the solar system we have not yet touched upon. The first concerns its method of operation: What makes it work? What force or set of forces keeps the planets running smoothly in their orbits? The second involves its history: Where did the planets come from? What started them traveling around the sun in the first place?

A scientific answer to these questions is hardly possible without some knowledge of forces and moving bodies in general. So for the next few chapters we shall be concerned with the problem of motion—how it is produced, what factors influence it, and especially how it is related to the action of forces. We shall find in this study much to aid us not only in understanding the solar system, but in analyzing complex phenomena here on the earth.

Inertia, Mass, Weight

Imagine a ball at rest on a level table. Given a gentle push, the ball rolls a short distance and gradually comes to a stop. That is common experience. The harder and smoother ball and table top are made, the farther the ball rolls before stopping. Suppose that the ball could be made perfectly round, the table flawlessly smooth and perfectly level. In other words, suppose there were no friction between them; suppose further that all air could be removed from the path of the ball and that the table is infinitely long. If the ball were now set in motion, would it ever stop rolling?

The conditions of this experiment cannot, of course, be realized in our laboratories. But they can be closely approximated; and as resistance to its motion becomes less and less, the ball shows less and less inclination to stop. It is reasonable to conclude that under ideal conditions it would keep on rolling forever.

This conclusion was first expressed in the writings of Galileo. Later it was stated by Sir Isaac Newton in a general form which has come to be

known as Newton's first law of motion:

> *Every body continues in its state of rest or of uniform motion in a straight line unless acted upon by a force.*

In other words, objects about us do not start moving of their own accord; once set in motion, they continue with constant speed in a straight line until some resistance (*e.g.*, friction) makes them stop. In our daily life such influences as friction and air resistance cannot be eliminated, and consequently all moving bodies of our immediate acquaintance tend to stop. To keep them moving at constant speed it is necessary that something should push them continually, the push being used to overcome friction and air resistance. But here and throughout this chapter we are considering ideal conditions, under which friction and air resistance are absent. When we speak of balls we shall mean perfectly smooth, hard spheres; our tables and floors and roads will be smooth, level surfaces of indefinite extent. Granted these conditions, a moving ball has no reason to stop. A push is required to set it moving, but once started it continues of its own accord. Ideally, motion at constant speed in a straight line is a condition quite as natural as a state of rest.

Not only does a motionless body tend to remain at rest; it offers definite resistance to any attempt to make it move, as anyone discovers who tries to roll a heavy barrel. Once started, a body actively resists any effort to stop it or to change its motion, as the same barrel demonstrates as soon as it starts to roll. *This resistance which a material object offers to any change in its motion* is an important property of matter called ***inertia.*** Inertia is important because it gives us a means of measuring the quantity of matter present in an object.

The concept of *quantity of matter* deserves a moment's digression from our consideration of moving bodies. In everyday life we measure the amount of matter present in an object very simply, by weighing it. Thus an object weighing 2 pounds (lb) contains twice as much matter as one which weighs 1 lb. This method of measurement is unsatisfactory for a precise definition, because weight depends on the gravitational attraction of the earth for an object and this attraction is not the same everywhere. Even on the earth's surface it varies somewhat: objects weigh slightly less on mountaintops than at sea level, slightly less at the equator than near the poles.* The variation is extremely small, of course, so that for ordinary

* This statement is true only when the object is weighed by a device like a spring balance, in which the earth's pull is measured directly by its capacity for stretching a coiled wire. An ordinary balance does not measure gravitational attraction directly, but compares the earth's pull on an object with its pull on other objects ("weights") of known mass. Since object and weights are equally affected by changes in the earth's gravitational pull, an ordinary balance will give the same value for the "weight" of an object anywhere on the earth's surface.

purposes and even for most scientific purposes weight serves well enough as a measure of quantity of matter. But for distant bodies beyond the influence of the earth's gravitation the idea of "weight" becomes confusing. Thus a 150-lb man would "weigh" 25 lb on the moon, 2 tons on the sun, and nothing at all in empty space far from the solar system.

Imagine two balls of equal size, one made of lead, the other of wood. Here on earth we would say immediately that the lead ball contains more matter because it weighs more. If I were blindfolded and forbidden to weigh them I could still tell the balls apart by kicking them along a level floor, because the inertia of the lead ball would resist my kick more than that of the wooden ball. Now suppose that I, together with the two balls, were transferred to some point in the depths of space. Provided I were still interested in such matters, could I find any method of establishing the greater quantity of matter in the lead ball under these new conditions? Weighing them would be quite useless, for neither ball would exert any downward push on my hand or on the pan of a balance. But inertia would still give me an answer. The lead ball would again appear to contain more matter because my toe would again hurt more after kicking it. The resistance which the two balls offer to any attempt to make them move is a property quite independent of their weights—and much more fundamental than their weights, since it does not depend on their distance from the earth.

Because of the greater inertia of the lead ball, or its greater resistance to being set in motion, the physicist says that it has a greater ***mass*** than the wooden one, by which he means that it contains a greater *quantity of matter*. On the earth "mass" and "weight" often appear to mean about the same thing. The technical difference is that weight represents a pull exerted on an object by the earth, while mass is a property of the object itself. The weight of an object is different in different parts of the universe; its mass is the same everywhere. Unfortunately physicists express the two properties in the same units, a custom which naturally leads to confusion. The commonest unit of mass in scientific work is the gram;* since the easiest way of measuring mass on the earth's surface is by weighing, we say commonly—and correctly—that a one-gram mass has a weight of one gram. Similarly, a one-pound mass at the earth's surface weighs one pound, and so on. With a little practice it is easy to decide in a given discussion whether "gram" or "pound" or "kilogram" refers to one property or the other.

Force, Velocity, Acceleration

Let us now return to the last clause of Newton's first law of motion, which states that the motion of a body can be altered by the action of a

* The metric system is universally used in scientific work. Its units are described and compared with units in the ordinary British system in Table XXVIII, p. 661.

force. Most of us have a hazy notion of what force means: we think of a horse pulling a wagon, a man pushing a wheelbarrow or lifting a flour sack. Other familiar examples we have met in the preceding discussion: the force of gravity, which pulls us and objects about us to the earth's surface; the force of friction, which retards the motion of any object moving in contact with another. Centrifugal force, the pull of a magnet, the force of water pushing against the vanes of a turbine are further illustrations. In all these examples the central idea is one of pushing or pulling, lifting or throwing—a process either involving muscular effort, or producing the same results as the exertion of muscular effort. We shall speak of forces immensely greater than any muscle could produce, and forces immeasurably smaller than the most delicate touch could detect; but we call them "forces" only because they produce results, on a larger or smaller scale, similar to those accomplished by muscular effort. Force is thus a concept based on the direct evidence of our senses, and is difficult to define satisfactorily. The most workable definition is a restatement of the first law of motion and merely specifies the result of a force's action: *a force is any influence capable of producing a change in the motion of a body of matter.* Actual change of motion need not result from the application of a force. I may push with all my strength against a stone wall without affecting its motion in the slightest degree. Yet I still call my muscular exertion the application of a force, since the same exertion would be capable of producing motion if I should choose the object of its application more wisely.

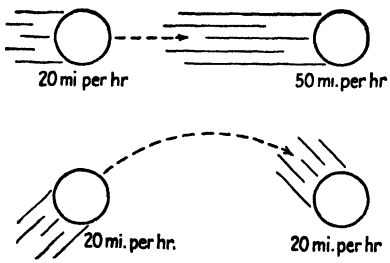


FIG. 26. *Velocity is changed by a change in either speed or direction.*

One idea in the above definition and in the first law remains to be clarified: the idea we have hitherto expressed by the vague term "change of motion." A moving body acted upon by no forces progresses, as stated above, by "uniform motion," or more precisely by "motion with constant velocity." Action of a force causes its motion to change so that its velocity is no longer constant; and, as long as the force acts, the velocity will continue to change. To give these statements a more precise meaning, we must devote a few sentences to the concept of velocity.

Like so many in physics, ***velocity*** and the related term *speed* are words with which we are all somewhat familiar but which in scientific work have a more special meaning than in everyday conversation. We say that a car travels with a speed or a velocity of 40 mi/hr, or that a man covers 100 yards in 10 seconds (sec) and so runs at a speed or a velocity of 10 yards a second. *Distance covered in a unit of time:* this is the usual

designation of a speed or a velocity. Physicists distinguish speed from velocity by including in velocity *direction* as well as distance covered per unit time; thus a train traveling east and covering 60 mi in 2 hr goes at a *speed* of 30 mi/hr, and at a *velocity* of 30 mi/hr eastward.

Constant velocity, therefore, implies constant speed along a straight line. Change in velocity produced by a force may affect either the speed or the direction (Fig. 26). Thus a man who steps on the accelerator of his car thereby applies more force to the drive shaft and increases the speed of the car. A small boy whirling a stone at the end of a string applies force to make the stone revolve in a circle, thereby changing its velocity continuously although its speed in the circle may be nearly constant. In whatever manner the motion is changed—whether speed is increased or decreased, or only the direction is altered—as long as the force acts the motion is said to be *accelerated.*

Suppose that a man traveling 30 mi/hr pushes down his accelerator and holds it down until his speedometer registers 50 mi/hr. He has changed his velocity by 20 mi/hr. Suppose that 2 minutes (min) are required for his car to reach the higher speed. Then every minute his velocity increases 10 mi/hr and we say that he travels with an *acceleration* of 10 miles per hour per minute, or 10 mi/hr/min. Or suppose that a ball starts to roll across the floor with a velocity of 10 centimeters per second (10 cm/sec), and suppose that friction retards its motion so that 3 sec later it is going only 4 cm/sec. Then in 3 sec its velocity has decreased by 6 cm/sec, and its change of velocity in each second, or its acceleration, is 2 centimeters per second per second (2 cm/sec/sec or 2 cm/sec^2). Thus ***acceleration*** is defined as *rate of change of velocity.* The change in velocity may be an increase, as in the first example above, or a decrease as in the second example. Or the change may take place in a third way, not so easily expressed in figures: by a change in direction. Technically a man may "accelerate" his car either with the throttle or with the brake, or by turning a corner, since any one of the three produces a change in his velocity. [The usual expression for acceleration, cm/sec/sec, is somewhat confusing at first. It is helpful to enclose the first part of the expression in parentheses, and to remember that this part represents the difference between the velocity at one time and the velocity at a later time. Thus an acceleration of 16.2 cm/sec^2, or (16.2 cm/sec) per second, means that in the course of 1 sec the velocity has changed by 16.2 cm/sec.]

It is highly important to remember that a body continues to be accelerated as long as a force acts upon it, and no longer. A force does not merely make a body move faster than it moved before; it causes the speed to increase steadily until the force is removed. Here again results which would be obtained under ideal conditions seem to contradict

everyday experience. Theoretically, as just stated, a small force acting on a ball rolling without friction would cause its speed to increase steadily, and the ball could be made to move with any conceivable speed if only the force acts for a long enough time. Theoretically again, the accelerator in an ideal automobile would not be needed at all to maintain a constant speed of 30 mi/hr or 60 mi/hr; once set in motion at either of these speeds, an ideal car would keep on moving without any further force being applied to its wheels. To change from 30 to 60 mi/hr would require use of the accelerator, that is, the application of a force. We might say then that the force had effected a change from one constant speed to another; but the direct effect of the force, the effect while the force was acting, was to make the speed increase continually, and unless the force was removed when 60 mi/hr was attained the speed would go right on increasing. An ideal car could be made to go at any desired speed, however fast, merely by a slight constant pressure on the accelerator. Ordinary cars cannot, because at high speeds friction and air resistance increase so that most of the force which the engine applies to the rear wheels is used up merely in counteracting these opposing forces.

From similar reasoning we conclude that no force at all, or an infinitely small force, is required to move an ideal ball from one position of rest on a level table to another. Application of any finite force would set the ball moving forever—with constant velocity if the force were removed, with steadily increasing velocity if the force were continuous. Practically we have to supply a force: we strike the ball, and let it roll to a stop in its new position. During the fraction of a second while we are striking it, the force we exert changes the ball's velocity from zero to some small value; when the striking force is removed friction begins to alter this velocity, reducing it to zero when the ball reaches its new position of rest. We apply a force to make the ball change position; but the force is used entirely in overcoming the opposing force of friction. Ideally the initial force needed would be infinitely small.

Newton's Second Law of Motion

Newton's second law of motion is a quantitative expression of the ideas we have discussed in the above paragraphs. It gives us a relation between force, mass, and acceleration which can be treated mathematically. We shall defer its mathematical details and its experimental justification until later, but the law itself will help in tying together ideas of this chapter:

> *The acceleration which a force can give an object is directly proportional to the magnitude of the force and inversely proportional to the mass of the object; and the acceleration is in the direction of the applied force.*

This means that if we measure the accelerations produced by different forces on the same mass, doubling the force will double the acceleration; and that if we let the same force act on different masses, doubling the mass will cut the resulting acceleration in half (Fig. 27). The law may be stated mathematically, according to rules we shall learn in the next chapter,

$$\frac{F}{m} = a \qquad \text{or} \qquad F = ma$$

F represents force, m mass, and a acceleration. In words, the second expression says that ***force*** is equal to *the product of mass and acceleration.*

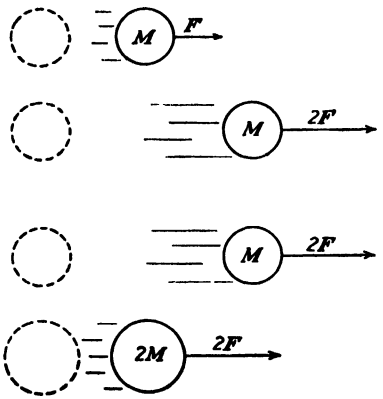


FIG. 27. *Greater force produces greater acceleration, if two masses are equal; greater mass receives smaller acceleration, if forces are equal and masses unequal.*

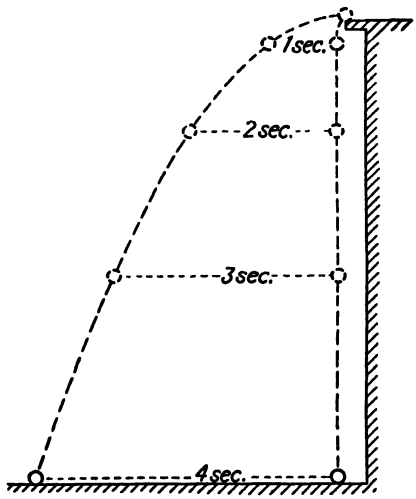


FIG. 28. *If one ball is thrown horizontally from the top of a building at the same time that a second ball is dropped vertically, the two will reach the ground together, since they are accelerated by the same constant force.*

This is a precise statement of the definition of force given above (page 40).

Perhaps the most familiar of simple forces is the attraction which holds us and other objects to the earth's surface. The attraction in general is called *gravity;* the force on any particular object is called its *weight.* Thus I may say, "Gravity pulls me to the earth with a force of 150 lb," or, "My weight is 150 lb." Let us use this force to illustrate and amplify the preceding discussion.

The weight of an object pulls it vertically downward, regardless of its position or its motion. That this force produces an acceleration we all

discover unconsciously at a very tender age, when we find out how much harder we hit the ground in jumping from a high elevation than from a low one. As long as a body is free to move toward the ground, its velocity increases very rapidly, so rapidly that direct measurement is difficult. The velocity increases by the same amount each second; in other words, the acceleration of gravity is constant. It matters not at all whether the body starts from a position of rest or is moving. If a ball is simply held in the air and dropped, its velocity increases steadily until it strikes the ground. If it is thrown in a horizontal direction, its motion is determined by its tendency to keep on moving in the direction it was thrown and by the pull of gravity. The latter accelerates its motion downward, so that the ball moves in a curved path which grows steeper as it nears the ground (Fig. 28). If the ball is tossed vertically upward, the acceleration, acting downward, at first is in a direction opposite to the ball's motion. Hence the ball's velocity steadily diminishes, becoming zero at the top of its climb, and then increases steadily downward, the acceleration remaining constant throughout.

For simplicity in the preceding paragraph we have neglected air resistance. In the motion of small light objects like raindrops it exerts a powerful retarding influence to the accelerating tendency of gravity; otherwise raindrops would attain bulletlike velocities, and we could not safely venture out in the lightest shower. But for moderately heavy objects dropping at moderate speeds, air resistance is only a minor correction.

The retarding action of the air often gives an erroneous idea of the relative rates with which different objects fall. A feather and a lead bullet indisputably fall with very different velocities. But the difference is wholly due to the lightness and the large surface of the feather; if the two objects, or any other two objects, are enclosed in an evacuated cylinder they fall at precisely the same rate. We might have predicted this result from the second law of motion: gravity pulls down on a 2-kilogram (2-kg) mass with twice as much force as on a 1-kg mass; but just twice as much force is required to accelerate a 2-kg mass to a given velocity; hence the acceleration of a 2-kg mass is the same as that of a 1-kg mass, and likewise the same as that of any other mass regardless of size. Since the acceleration of gravity is the same for any mass, the velocities of any two objects dropped simultaneously will increase at the same rate, and the objects will reach the ground together.

This conclusion, that all objects near the earth's surface fall with the same acceleration regardless of their weights, directly contradicts the teaching of Aristotle. With no conception of gravity as a force, Aristotle sought to explain falling in terms of "goodness" and "badness," after the fashion so common in Greek reasoning. Some materials, like fire, smoke, air, moved upward toward the heavens because they were intrinsically

"good"; earthy materials, like stones, metals, wood, moved downward toward the earth because they were "bad," or "imperfect." It followed that a heavy stone, containing more "badness," should move faster toward the earth than a light stone. We have no record that anyone seriously challenged this notion for nineteen centuries after Aristotle's death. The first to demonstrate its fallacy was that same Italian, Galileo Galilei, whose telescopic discoveries we followed in the last chapter.

We ascribe the laws of motion to Newton, since it was he who first stated them explicitly; but Galileo, nearly a century earlier, had reached a pretty clear understanding of the relation between motion and force. While still a young man, a professor at the University of Pisa, he realized that Aristotle's conclusion about falling bodies was wrong. He set out to demonstrate this in a radical and unheard-of manner: by actually performing the experiment of dropping a heavy and a light object simultaneously. From Pisa's famous leaning tower, before an assembly of his colleagues, students, and townspeople, Galileo let fall a cannon ball and a small bullet. They reached the ground together. To us, the experiment seems inescapable proof that Aristotle was wrong. But to Galileo's academic contemporaries, upholders of a thousand-year-old tradition, the experiment proved only that this young upstart had somehow bewitched either the cannon ball or their eyes.

Galileo lost his professorship at Pisa, but neither his faith in experiment nor his opposition to Aristotle was quieted. In other Italian cities he continued work on the problems of motion, his learning bringing him fame as a scholar, his personal charm winning him friends among the ruling families of Italy. Despite fame and friends, Galileo's continued and outspoken criticism of accepted opinions brought recurrent trouble with church authorities. Especially serious was his belief in the Copernican system, for which his telescope presently brought such convincing proof. For many years he managed to escape serious reprisals, but finally, as an old man, he was haled before the Inquisition at Rome and given the choice of torture or recanting his heretical beliefs. Sensibly he chose to recant, and so lived to complete his scientific work.

Galileo has a hallowed name in science for his studies of motion, for his telescopic observations, and particularly for his emphasis on experiment as a means of checking hypotheses and discovering new facts.

Questions

1. When a meteor approaches the earth from outer space, does its inertia increase, decrease, or remain constant before it reaches the earth's atmosphere? After it reaches the earth's atmosphere?
2. If a glass of water is placed on a piece of paper on a smooth table top, the paper can be removed by a sudden sideward jerk without spilling the water, while a less sudden pull will tip the glass over. Explain.

3. Is the statement that "a 200-lb mass weighs 200 lb" correct on the surface of the earth? Would it be correct on the surface of Mars? Explain.
4. Can a body be in motion if no forces are acting on it?
5. State clearly the relationship between force and change in velocity.
6. In which of the following instances is the motion of the car accelerated?
   a. A car climbs a steep hill, its speed decreasing from 50 mi/hr at the bottom to 20 mi/hr at the top.
   b. A car turns a corner at a constant speed of 30 mi/hr.
   c. A car climbs a long, straight hill at a constant speed of 40 mi/hr.
7. Is the moon's motion around the earth accelerated? Is a force acting on the moon?
8. A bullet travels half a mile in 3 sec. What is its average speed?
9. A car is moving at a speed of 10 ft/sec. Four seconds later its speed has increased to 32 ft/sec. What is its acceleration?
10. A bullet attains a speed of 300 m./sec during the 0.01 sec it is in the rifle barrel. What is its acceleration?
11. A stone dropped from a cliff falls with an acceleration of 980 cm/sec^2. How fast will the stone be moving 1 sec after it is dropped? Three seconds after it is dropped? What is its *average* speed during the first second? During the first 3 sec? How far will it fall during the first second? During the first 3 sec? (Assume that the original speed is zero.)

CHAPTER IV

The Language of Mathematics

VELOCITY, acceleration, force, and mass: with these basic concepts of physical science we have now a speaking acquaintance. So frequently will they appear in future chapters that a more intimate knowledge is necessary, a knowledge which can scarcely be gained without the use of elementary mathematical ideas and formulas.

Mathematics is a science in its own right, but one which touches the outside world only insofar as it is used as a tool in the other sciences. Terrifying as the subject often seems to the uninitiated, it is indispensable to progress in physics and chemistry and astronomy. With symbols and equations a scientist can express concisely ideas which would require many pages of careful writing to explain in ordinary language. When his ideas are thus expressed in simple mathematical form, new relationships and new channels of investigation often suggest themselves which would be lost in verbiage if mathematics were not used. To simplify, to clarify, to coordinate, to predict—in all sciences mathematics serves these several purposes.

For the present we shall need only two simple mathematical ideas. They happen to be ideas which find extensive use in ordinary life as well as in physical science.

First is the notion of *proportionality*. We read, for instance, that "pressure beneath a water surface increases with depth," that "at high speeds gasoline consumption in an automobile increases approximately with the cube of the speed," that "prices vary inversely with the value of gold." In the last chapter we learned that "acceleration is directly proportional to the force producing it and inversely proportional to the mass of the object being accelerated." These statements, all suggesting how one quantity changes in response to change in another, express various kinds of proportionality. Simple direct proportion implies that doubling one quantity doubles the second, tripling the first triples the second, etc. In more complex proportions, doubling one quantity may multiply the second by 4, by 8, by $\frac{1}{2}$, or by some less simple number.

The second mathematical idea which we shall take up is *graphic representation.* A proportion is an algebraic method of showing a certain relation between two quantities; a graph is a pictorial method. We have all seen graphs showing industrial progress from month to month, weather charts showing temperature changes and variations in rainfall over a period of time, medical charts showing increase or decrease in the incidence of various diseases. Frequently, as in the three examples just given, graphs show relationships too complex for simple representation in a mathematical formula. Even where such representation is possible, a graph often gives a more readily understandable idea of the relationship than the formula.

We shall devote this chapter, then, to an explanation of some simple mathematical devices. As an illustration of their usefulness we shall apply them to the problem of falling bodies, expanding the ideas in the last paragraphs of the preceding chapter. Later on these same mathematical devices will prove helpful in many other connections.

Velocity

Speed we have defined as distance covered in unit time, velocity as speed in a definite direction. Common time units are the hour, minute, and second, so speeds and velocities are given by expressions like 10 mi/hr, 2.5 mi/sec, 0.38 kilometers per minute (km/min), 5 cm/sec. Directly from the definition, using no mathematics except arithmetic, we can work out simple problems; for instance, a man walking 24 mi in 8 hr has a speed of 3 mi/hr, a bullet covering 700 meters (m.) in 2 sec a speed of 350 meters per second (m./sec). Thus speed or velocity is obtained by dividing total distance covered by time consumed. This statement will enable us to handle elementary problems, but more complicated ones are more easily solved if the relation is expressed in an equation. Let v represent speed or velocity, d the total distance, t the time.

$$\text{Speed} = \frac{\text{total distance covered}}{\text{time consumed}}, \qquad \text{or} \qquad v = \frac{d}{t} \tag{1}$$

Likewise from the definition or from common experience we know that a car going 40 mi/hr for 3 hr will traverse 120 mi, a horse running 6 m./sec for 10 sec will cover 60 m., etc. That is, total distance is equal to the product of velocity and time, or

$$d = vt \tag{2}$$

We could have obtained Eq. (2) from Eq. (1), of course, by simple algebra [multiplying both sides of Eq. (1) by t]. The two equations are merely different ways of expressing the same relationship.

In ordinary life a moving object seldom maintains the same velocity over long distances. When speed is variable, the quotient of total distance divided by time gives an average value of the speed over the interval. Thus a car may travel at a uniform rate of 30 mi/hr; more probably it will go 50 mi/hr on the open highway, 20 mi/hr around sharp turns, 0 mi/hr while stopping for gas; but, if it covers a total of 90 mi in 3 hr, its average speed is still 30 mi/hr. When this idea of average speed needs to be emphasized, it is customary to draw a line over v:

$$\bar{v} \text{ (average speed)} = \frac{d}{t} \tag{3}$$

This is the obvious, conventional method of handling the relation between distance, velocity, and time. So simple a relation needs no further treatment. But for purposes of illustration let us examine these quantities from another angle.

Direct Proportion

Imagine a number of balls which start moving along a level floor together and proceed with identical speeds in the same direction. Their velocities, then, are constant. Suppose we stop the balls at different times, and measure the distance each has covered. We obtain results like the following:

TABLE II

	Stopped after	*Distance covered**
First ball	2 sec	5.0 cm
Second ball	4 sec	10.1 cm
Third ball	6 sec	15.1 cm
Fourth ball	10 sec	24.9 cm
Fifth ball	20 sec	50.2 cm

* If the experiment could be performed under ideal conditions, and the measurements made by an ideal observer, the second distance should be *exactly* twice the first, the third *exactly* three times the first, etc. Actual physical measurements are always accompanied by more or less experimental error, as these figures indicate.

A ball rolling for 4 sec goes approximately twice as far as one rolling for 2 sec; one rolling for 6 sec goes three times as far, one rolling for 10 sec five times as far, etc. Two quantities so related are said to be ***directly proportional*** to one another. We may write this proportion in the manner most of us learn in high school,

$$\frac{d_1}{d_2} = \frac{t_1}{t_2}$$

That this equation fits the above data may be checked by substituting any pair of distances for d_1 and d_2 and the corresponding times for t_1 and t_2.

This formulation of a proportion is useful occasionally, but more often in scientific work we find it convenient to use the less cumbersome expression

$$d = Kt \tag{4}$$

in which K is a constant number called a ***proportionality constant.*** Introduction of this number K seems at first glance a bit mysterious. Let us for the moment disregard it, remembering simply that the equation $x = Ky$ is mathematical shorthand for the statement that "x is proportional to y." Why proportionality constants are convenient we shall discuss later.

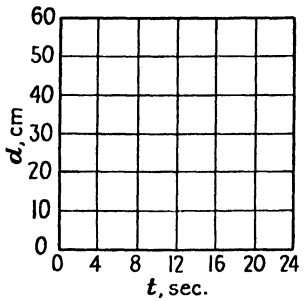


FIG. 29.

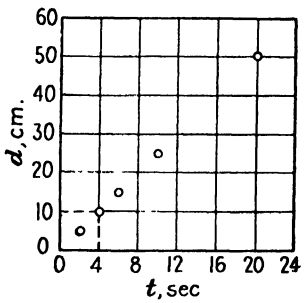


FIG. 30.

The graph of two quantities related by a direct proportion has a simple and characteristic form. As an example of such a graph, let us plot the figures in the above table. We begin by drawing two lines perpendicular to each other. These are called the *horizontal axis* and the *vertical axis*, and their intersection is the *origin*. On one axis, say the horizontal one, we lay off regular intervals away from the origin to represent different values of t, and on the vertical axis we lay off values of d (Fig. 29). Then for each pair of numbers like (4 sec, 10.1 cm) we plot a point, by measuring vertically from an appropriate point on the horizontal axis a distance corresponding to the value of d (Fig. 30). With a line through these points the graph is complete (Fig. 31). From such a graph the time required for a ball to travel any desired distance may be read at once; thus point A indicates that a ball would roll 18 cm in a little over 7 sec.

Note that the graph is a *straight line* passing *through the origin.* The graph of any direct proportion satisfies these two conditions, and conversely any graph which is a straight line through the origin must represent a direct proportion. Ordinarily the easiest way to discover whether or not one quantity is directly proportional to another is to find a series

of values of one quantity corresponding to values of the other, to plot them on a graph, and to see if the resulting line satisfies the two necessary conditions.

Direct proportionality is a very special relationship, not shown by many quantities in everyday life. For instance, a man's height increases with his age, but height and age are not directly proportional, as Fig. 32

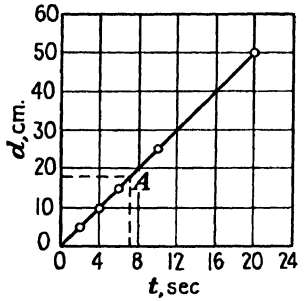


FIG. 31.

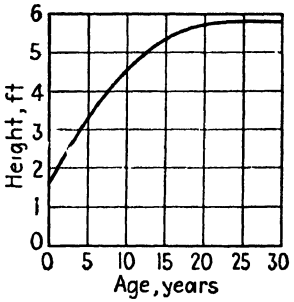


FIG. 32.

shows. Distance and time are related by a direct proportion only as long as velocity is constant. When a ball rolls downhill or falls freely, the distance it travels increases with time, but the two quantities are not proportional since the ball's velocity is increasing. The graph obtained by plotting d against t for such a motion is shown on page 54. Whenever one quantity becomes larger as another becomes larger, we may speak of a *direct relationship* between them, but it is a *direct proportion* only if one quantity is doubled when the other is doubled, tripled when the other is tripled, and so on.

Inverse Proportion

Often two quantities are so related that one increases as the other decreases. One's desire for food, for instance, diminishes as the quantity of food one eats grows larger. The value of an automobile gets steadily smaller with increasing age. The loudness of an explosion is greater the smaller its distance from you. These are examples of *inverse relationships*. An ***inverse proportion*** is a special case in which doubling one quantity makes the second shrink to half its former size, increasing the first tenfold decreases the second to one-tenth its value, etc. Perhaps the simplest example is the relation between velocity and time for moving objects which travel the same distance.

Suppose five balls are set moving with different speeds, and the time taken to cover a fixed distance is recorded for each. Obviously the greater a ball's speed, the shorter will be the time required. Suppose the following data are obtained:

TABLE III

	Speed	*Time required*
First ball	3 cm/sec	72.1 sec
Second ball	6 cm/sec	36.0 sec
Third ball	9 cm/sec	24.2 sec
Fourth ball	15 cm/sec	14.3 sec
Fifth ball	30 cm/sec	7.2 sec

Inspection of these figures shows that they represent an inverse proportion between v and t. As an equation this may be written

$$v = \frac{K}{t} \quad \text{or} \quad v = K\left(\frac{1}{t}\right) \tag{5}$$

We may read from Eq. (5) either that "v is inversely proportional to t" or that "v is directly proportional to $1/t$." That the two statements are

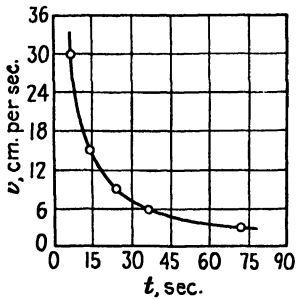


FIG. 33.

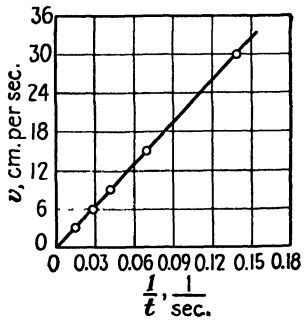


FIG. 34.

equivalent may be shown graphically by plotting from the above table v against t and v against $1/t$ (Figs. 33, 34). The curve shown in the first graph is characteristic of an inverse proportion; but such a proportion is best established by a straight line through the origin when one quantity is plotted against the reciprocal of the other, as in the second graph.

Freely Falling Bodies

A more complicated sort of proportion is illustrated by the relation between distance and time for a freely falling body. In Chap. III we learned that the force of gravity gives to any freely falling body an acceleration toward the earth, an acceleration so large that direct measurement of the body's position or velocity at any instant requires elaborate experimental technique. Perhaps the best instrument for such measurements is an electrical device which causes a spark to jump thirty times a second between a falling steel ball and an adjacent wire, each spark leaving a

mark on a paper strip hung along the path of the ball (Fig. 35). Measurement of the distance between each spark mark on the paper and the ball's original position gives the distances a freely falling body covers in known intervals of time. Data from such an experiment are given in Table IV and the graph of d against t in Fig. 36.

TABLE IV

Time	*Distance from Starting Point to Spark Mark*
0	0
$\frac{1}{30}$ sec	0.5 cm
$\frac{2}{30}$ sec	2.2 cm
$\frac{3}{30}$ sec	4.7 cm
$\frac{4}{30}$ sec	8.6 cm
$\frac{5}{30}$ sec	34.5 cm

That d and t are not directly proportional is at once obvious. From the table, when t is doubled from $\frac{1}{30}$ to $\frac{2}{30}$ sec, d increases from 0.5 to 2.2 cm, approximately a four-fold increase. When t increases to $\frac{3}{30}$ sec, d increases to 4.7 cm, about nine times its value at $\frac{1}{30}$ sec. Multiplying t by 2 thus multiplies d by 4 or 2^2, multiplying t by 3 multiplies d by 9 or 3^2, etc. Evidently d is proportional not to t but to the square of t, a surmise borne out by the graph of d against t^2 in Fig. 37. This proportionality is indicated by the equation

$$d = Kt^2 \qquad (6)$$

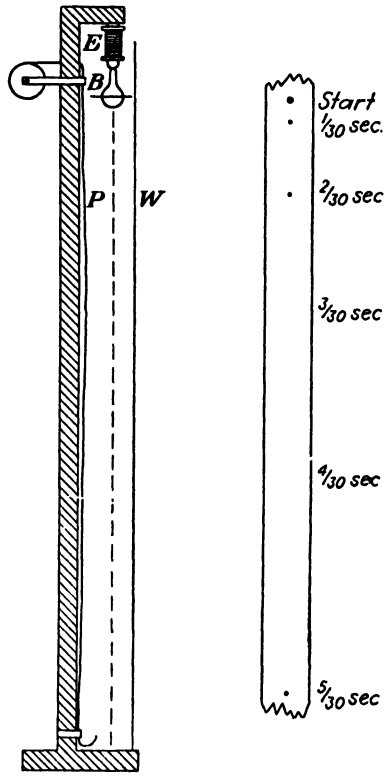


FIG. 35. *Diagram of falling-body apparatus, and part of paper strip with record of spark marks made every thirtieth of a second during bob's fall. E, electromagnet, which releases bob B when a switch is closed. Bob falls along dotted line, and sparks jump from wire, W, through bob to another wire behind paper strip P.*

Instead of dirtying our fingers with a laboratory experiment, we might have deduced a relation between d and t by some algebraic gymnastics based on our definitions of velocity and acceleration. Let us attempt this mathematical demonstration now, and see if we can unearth an equation resembling the experimentally determined Eq. (6).

Change of velocity per unit time: this was our definition of acceleration. Change of velocity is expressed as a difference of two velocities. Thus if a stone is thrown downward from a high cliff at a rate of 100 cm/sec and if

gravity accelerates it so that 2 sec later it is going 2,000 cm/sec, its velocity downward has increased by 1,900 cm/sec. Since this change in velocity has been accomplished in 2 sec, the acceleration caused by gravity is 950 cm/sec per second, or 950 cm/sec^2. If v_1 represents the initial

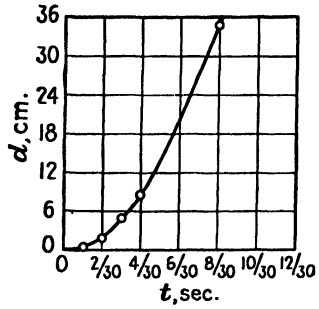


FIG. 36.

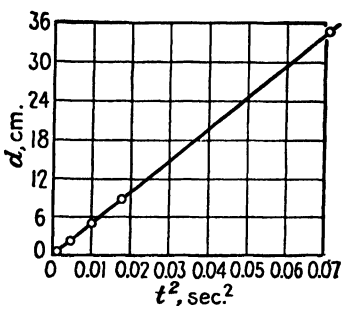


FIG. 37.

velocity (*e.g.*, 100 cm/sec) and v_2 the final velocity (*e.g.*, 2,000 cm/sec), acceleration is the difference between v_2 and v_1 divided by the time required for the velocity to change from v_1 to v_2:

$$\text{Acceleration} = \frac{\text{change in velocity}}{\text{time}} \qquad \text{or} \qquad a = \frac{v_2 - v_1}{t} \tag{7}$$

For the special case of a body falling freely from rest, Eq. (7) may be simplified by setting v_1 equal to zero, since the body starts with no initial velocity. Thus

$$a = \frac{v_2 - 0}{t} = \frac{v_2}{t} \tag{8}$$

v_2, of course, is the velocity attained at the end of t seconds. Now, *if we assume that a is constant,* the velocity will increase steadily from 0 up to this value, v_2; and, when half of the time t has elapsed, the velocity will be just half of v_2. This halfway velocity, $v_2/2$, represents the average value of the velocity over the entire interval, or $\bar{v}$. Then

$$\frac{v_2}{2} = \bar{v} \qquad \text{or} \qquad v_2 = 2\bar{v}$$

Substitute this value of v_2 in Eq. (8).

$$a = \frac{2\bar{v}}{t} \tag{9}$$

Then in this equation substitute the value of $\bar{v}$ given in Eq. (3).

$$a = \frac{2(d/t)}{t} = \frac{2d}{t^2}$$

or, by simple algebra,

$$d = \tfrac{1}{2}at^2 \tag{10}$$

This important equation shows the same proportionality between d and t^2 as does Eq. (6) and differs from Eq. (6) only in specifying that the proportionality constant is half the acceleration of gravity. *The fact that Eq.* (10), *derived by assuming the constancy of a, shows the correct (experimental) relationship between d and t is proof that the assumption is correct, or that the acceleration of gravity is constant.*

So important a constant is this particular acceleration that it is designated by a special letter g. Careful experiments have fixed its value at 980.266 cm/sec^2, or 32.16 ft/sec^2, at sea level in the latitude of New York. Although constant at any given locality, g varies somewhat at different points of the earth's surface. At London it is 981.188 cm/sec^2, at Honolulu 978.966 cm/sec^2, on Pike's Peak 978.953 cm/sec^2, at Eagle, Alaska, 982.182 cm/sec^2. It depends somewhat on latitude, somewhat on elevation, somewhat on the kind of rock immediately below the surface. Reasons for these variations will become apparent in a later chapter.

Three Rules of Proportionality

Three formal propositions concerning proportionality will occasionally be useful:

I. If x is proportional to y

then

y is also proportional to x

Thus, if $x = K_1y$, then $y = K_2x$. (Different K's simply indicate different proportions in the same discussion. Here $K_2 = 1/K_1$.)

II. If x is proportional to y

and

y is proportional to z

then

x is proportional to z

Thus if $x = K_1y$ and $y = K_2z$, then $x = K_3z$. (Here $K_3 = K_1K_2$)

III. If x is proportional to y

and

x is also proportional to z

then

x is proportional to the product yz

Thus if $x = K_1y$ and $x = K_2z$, then $x = K_3yz$. This statement gives us a means of combining several different proportions into a single equation.

Let us apply the third proposition to Newton's second law of motion (page 42, Chap. III). The statement of this law contains two proportions: (1) acceleration is directly proportional to force, or

$$a = K_1 F$$

and (2) acceleration is inversely proportional to mass, or [by Eq. (5)]

$$a = K_2 \left(\frac{1}{m}\right)$$

Combining these two proportions according to Proposition III gives

$$a = K_3 F \left(\frac{1}{m}\right) = K_3 \frac{F}{m} \tag{11}$$

By simple algebra

$$ma = K_3 F \qquad \text{or} \qquad F = \left(\frac{1}{K_3}\right) ma$$

If K_3 is constant, $1/K_3$ must also be constant, so that we may substitute K for $(1/K_3)$

$$F = Kma \tag{12}$$

Thus Newton's second law may be stated alternatively as "acceleration is proportional to the quotient of force divided by mass" [Eq. (11)], or "force is proportional to the product of mass and acceleration" [Eq. (12)].

Proportionality Constants

Having now some slight acquaintance with proportionality, let us try to find some reasonable interpretation for these ubiquitous K's which transform proportions into bona fide equations.

We had best start with a homely example. The number of apples I can buy depends on the amount of money in my pocket. With \$3 I can buy three times as many as I can with \$1, six times as many as I can with 50 cts. That is, the number of apples I can buy is proportional to the amount of money I possess, or

$$m = Kn$$

where m is the amount of money and n is the number of apples. Now this equation is a perfectly general expression of a simple proportion. It matters not if my money is in kopecks or yen or lire, or if I choose to buy expensive apples or cheap ones; the equation is true in any market. But, as it stands, the equation tells me nothing at all about hard practical facts like how many apples I can get for 75 cts. To ascertain this I shall have to

accost a grocer. Perhaps he tells me that apples are 60 cts a dozen. Now my equation is of some value: I know that when m is 60, n is 12, and I should be able to find K.

$$60 = K \times 12 \qquad \text{or} \qquad K = 5$$

This value of K is constant, by definition, so I can use it to compute the number of apples my 75 cts will buy.

$$75 = 5 \times n \qquad \text{or} \qquad n = 15$$

What is this value, 5? Obviously enough, here it is the price of one apple. *Price* is the most familiar proportionality constant in ordinary life.

Note the factors which determine price. First, K will evidently depend on what kind of apples I buy and in what store I buy them. Second, it will depend on what money units I use; thus, if the grocer had quoted me his price as \$5 a hundred instead of 60 cts a dozen, substitution would have given

$$5 = K \times 100 \qquad \text{or} \qquad K = 0.05 \text{ instead of } 5$$

That is, K would be expressed in dollars per apple, rather than cents per apple. Thus a proportionality constant remains constant only when (1) conditions of the experiment do not change, or (2) the units employed are not changed.

This example should make it clear that an equation in the form $d = Kt$ is quite useless by itself in solving problems. It tells us that under certain experimental conditions a moving object goes twice as far in 14 sec as in 7, but it gives us no hint as to what the actual distance is. To make such an equation usable in numerical problems K *must be determined by experiment.* Thus, if I find experimentally that an object goes 3 in. in 7 sec, I can substitute in the equation and find that $K = \frac{3}{7}$. Using this value of K in the equation, I can then find the distance covered in 14 sec, or in any other number of seconds. The value of $\frac{3}{7}$ represents, of course, the distance covered in each second, or the speed.

In one important circumstance it is not necessary to determine K experimentally: if the units of some quantity in the equation have not been chosen, K can be given any desired value by proper selection of these units. A good example is the equation for Newton's second law, derived at the end of the last section.

$$F = Kma \tag{12}$$

This equation may be used with F, m, and a expressed in any units we wish. For mass and acceleration we know several possible units, but how shall we express a force? The only force units we have encountered are units of weight, like the gram, pound, or ton. We might select one of these

units for F in Eq. (12) and determine the corresponding value of K experimentally; we could then use this value of K for any other problem involving the same units. But in careful scientific work weights are unsatisfactory as force units for the same reasons that weight is unsuitable as a measure of mass (page 38). So physicists find it convenient to invent arbitrarily a new force unit which will make the K in Eq. (12) equal to 1. In the metric system this is accomplished by *defining* a one-unit force as *a force which will give a mass of one gram an acceleration of one centimeter per second per second.* When $m = 1$ and $a = 1$, F is arbitrarily set equal to 1. With these numbers Eq. (12) becomes

$$1 = K \times 1 \times 1,$$

and K must also be 1. The unit force just defined is called a force of 1 *dyne* —a very small force, about equal to that which a mosquito exerts on your arm when she alights for her evening meal. A force of 10 dynes will give a 5-g. mass an acceleration of 2 cm/sec^2, or a 2-g. mass an acceleration of 5 cm/sec^2, or a 10-g. mass an acceleration of 1 cm/sec^2. Since, when F is in dynes, m in grams, and a in centimeters per second per second, the proportionality constant is 1, we may rewrite Eq. (12) in its usual form (page 43, Chap. III)

$$F = ma \tag{13}$$

Gravity is the commonest force of our acquaintance, and its nearly constant acceleration is designated by the letter g (page 55). So to express the force which gravity exerts on an object, or the weight of the object, we may substitute W (weight) for F and substitute g for a in Eq. (13).

$$W = mg \tag{14}$$

Since g is 980 cm/sec^2, the weight of a 1-g. mass in dynes is

$$W = 1 \times 980 = 980 \text{ dynes}$$

In other words, a force of 1 g. is equivalent to about 980 dynes; it varies from place to place on the earth's surface in accordance with the variation in g (page 55). *The great importance of the constant g lies in the fact that it connects weight with mass. It serves thus as a direct measure of the force of gravity.*

The Meaning of Equations

A few pages are scarcely sufficient to explain proportionality constants. Only the familiarity that comes with long practice can make fully clear their meaning and manifold uses. The chapters and problems of succeeding pages will provide occasional examples of their use and should serve to make them more understandable. But for our purposes the chief function

of such a constant is to indicate proportionality. If we were interested in arithmetic, in putting numbers into formulas and getting other numbers out, then we should concern ourselves more seriously with evaluation of the constants. But arithmetic for us here would be a waste of time. We are concerned with getting the meaning out of formulas, not with becoming adept in their practical use. The meaning of a formula, in terms of the relations between changing quantities, is usually evident without any profound knowledge about its constants.

We could make many equations seem plausible without mentioning proportionality. We would say, for instance, that "force equals mass times acceleration." Yet there are two very good reasons why we shall find it preferable in these pages to consider a formula as a combination of proportionalities.

First, we are interested here in learning something about how a scientist approaches a problem, how he thinks, how he reaches his conclusions. When a scientist sets out to find how several quantities are related and to express these relationships in mathematical symbols, he proceeds in the manner we have sketchily discussed several times in this chapter. He selects two of the quantities in question, arranges his experiment to keep the other quantities as nearly constant as possible, and then observes what values one of the two quantities assumes for various values of the second. Unless the relationship between the two is immediately obvious, he plots his experimental numbers on a graph. From the shape of the curve he can generally tell what the relationship is—whether one quantity is directly proportional to the second, inversely proportional to the second, directly proportional to the square of the second, or more complexly related. This relation he can set down at once in the form of an equation involving one or more constants. Then he selects another pair of quantities and repeats the process. At length he obtains proportions (or other relations) involving all the quantities he is studying, and by means of Proposition III, page 55, he can combine these proportions into a single equation. Nearly all the simpler equations in science are derived by this process, or mathematically from other equations derived by this process. Hence to study equations from this point of view helps to keep in mind the general way in which equations are obtained from experimental data.

Second, we examine equations from this angle in order that we may interpret them more readily. *An equation is no more than a way of summarizing observational data, and it should be read with this in mind.* For example: We shall find in Chap. VI an equation

$$F = K\frac{Mm}{d^2}$$

If we regard this equation in the usual manner, trying to figure out how a

simple force can possibly be equal to such a forbidding combination of diverse quantities, we are bogged down at once in mathematical complexities and lose all sight of the really significant fact that the equation expresses the results of observation. If, on the other hand, we read from the equation that a certain force is proportional to each of two masses and that this same force is inversely proportional to the square of a distance, the equation looks much simpler. We can see for instance that doubling either mass would double the force, that tripling the distance would make the force one-ninth as great, etc. The mystery of any equation rapidly evaporates if it is interpreted in this manner.

Numbers will be occasionally useful for illustration in subsequent chapters, simply because a mathematical formula is always clearer when its letters are given actual, tangible values. We shall find equations important, however, not for solving problems, but as shorthand expressions of observational data which can be translated into simple relationships between pairs of quantities.

Questions

1. Light from the sun requires about 8⅓ min to reach the earth, a distance of about 150,000,000 km (93,000,000 mi). What is the speed of light, in kilometers per second?
2. How would you express the following proportions by equations?
   a. The pressure of a gas P is inversely proportional to its volume V.
   b. The force F exerted by a magnetic pole is inversely proportional to the square of the distance d from that pole.
   c. The pressure P beneath a liquid surface is directly proportional to the depth beneath the surface h and to the density of the liquid D.
3. How would you read the following equations in terms of proportions between pairs of quantities?
   a. $V = KT$ (relation between volume of a gas and its absolute temperature T).
   b. $a = v_2/t$.
   c. $\bar{v}^2 = \frac{1}{2}ad$ (relation between average velocity, acceleration, and distance, for motion with constant acceleration).
   d. $F = Kmn^2r$.
4. From the data of Table IV calculate the average velocities of the steel ball between its starting point and the end of each time interval. Make a graph of these values of $\bar{v}$ plotted against t. What is the relation between $\bar{v}$ and t?
5. From the following data, show graphically that acceleration is inversely proportional to mass for a given constant force:

Acceleration (cm/sec²)	100	60	40	30	20	10
Mass (g.)	10.0	16.8	24.9	33.1	50.0	100.2

6. Suppose you were in a barrel going over Niagara Falls. How far would you fall in the first second after leaving the brink? How far in the second second? How far in the third second? How fast would you be moving at the end of the third second?

Suppose that during the descent of the barrel you drop a ball inside the barrel. Would it appear to move toward the top of the barrel or its bottom, or would it remain stationary? Assume that your initial downward velocity is 0 and that air resistance can be neglected.

7. Suppose you are trying to find the distance to water in a deep well. You drop a stone from the top and time it until you hear a splash. If the time of fall is 4 sec, how far below the surface is the water level?
8. How long would it take a stone to drop from the top to the base of a cliff 1,600 ft high?
9. What force in dynes is equal to a weight of (*a*) 1 kg? (*b*) 1 lb?
10. Express Kepler's third law (page 13) as an equation involving a proportionality constant.
11. If $y = Kx^3$, how is the value of y affected when x is doubled? When x is multiplied by 10? When x is halved?
12. If $y = K/Z^2$, how is the value of y affected when Z is doubled? When Z is tripled? When Z is divided by 4?
13. Which of the following graphs show motion with uniform velocity, and which show motion with constant acceleration?

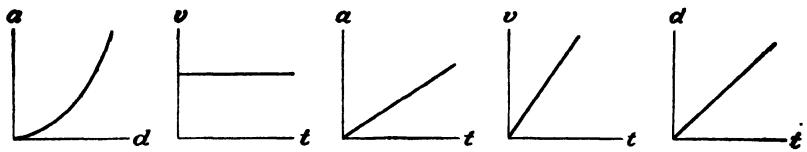


FIG. 38.

14. Sketch graphs like those in Fig. 38 to show
    a. Distance plotted against time for motion with constant acceleration.
    b. Speed plotted against time for motion with constant acceleration.
    c. Speed plotted against distance for motion with constant velocity.
    d. Acceleration plotted against distance for motion with constant acceleration.
    e. Speed plotted against time for motion with *increasing* acceleration.

CHAPTER V

Forces in Combination

IN THE ideas of force, velocity, and acceleration developed in the last two chapters, we begin to catch some glimmerings of an explanation for the complex motions of the sun's family. Each planet moves in a curved path, which means that its velocity is continually changing: this is possible only if it is acted on constantly by some force or combination of forces. On our small planet we find one force which seems to be associated with the planet itself—the force of gravity, which pulls all objects on the surface or high above the surface toward the earth's center. Perhaps we may reasonably assume that similar forces are associated with other bodies of the solar system. If this assumption is correct, may we further conjecture that these forces are the ones responsible for the curved orbits of planets and satellites?

Before we can answer this question satisfactorily, we need some further information of a more fundamental sort. Particularly, we must inquire into two problems: the problem of several forces acting together, and the problem of motion in curved paths.

Several Forces Acting on a Single Object

In previous discussions we have focused attention on single forces acting on objects which are free to move. Force problems in everyday experience are never so ideally simple; always several forces are in operation simultaneously. A falling stone, for instance, is subject not only to the force of gravity but to the resistance of the air through which it moves. The motion of an automobile depends on the force supplied to its drive shaft by the engine, on friction in its bearings, on friction between tires and pavement, on air resistance, on gravity if the car is going up- or downhill. The path of a canoe across a river depends on the force exerted by the paddler and on the force of the river's current.

Consider first the simple case of a falling stone. Gravity pulls on the stone with a constant force and, if unhindered, would cause it to move

toward the earth with steadily increasing speed. This force is represented by the arrows labeled G in Fig. 40. The opposing force of air resistance is not constant, but increases as the stone's speed increases. The increasing air resistance at different points along the stone's path is indicated by the arrows R_1, R_2, R_3 in Fig. 40. The actual force on the stone at any moment is evidently the difference between the two forces acting; we may find these actual forces for the three positions represented in Fig. 40 by laying the G and R arrows side by side and drawing a third arrow for the difference between their lengths, as indicated in the right-hand diagrams. The force difference grows smaller as the air resistance increases, until eventually, as shown in the bottom diagrams, the air resistance becomes equal to the pull of gravity. Since now the two forces are equal and directly opposite, the stone moves as if no force at all were acting on it—that is, it moves according to Newton's first law, in a straight line with constant speed. Any object falling through the air exhibits similar behavior: its speed increases until air resistance becomes equal to the pull

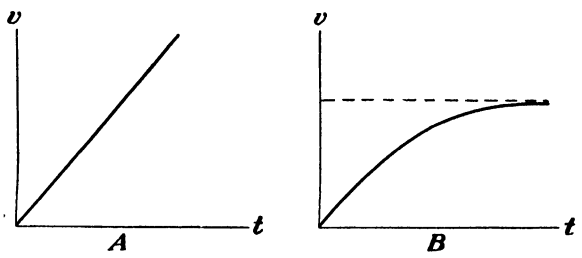


FIG. 39. *Graphs showing speed plotted against time for: A, motion of a falling body neglecting air resistance; B, actual motion of a body falling through air.*

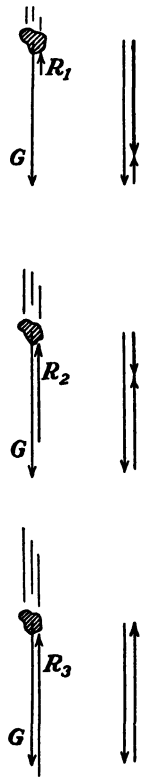


FIG. 40. *Forces on a falling stone.*

of gravity, and thereafter remains constant. Figure 39 is a graphical comparison of this sort of motion with the theoretical, constantly accelerated motion which gravity would produce if there were no air.

The forces on a car moving along a level road may be similarly represented by oppositely directed arrows (Fig. 41), except that here the resistance arrow is the sum of several parts—air resistance, friction in bearings, friction of tires, etc. If the sum of these resisting forces is equal to the driving force of the engine, the car moves with constant velocity; otherwise the car is either speeding up or slowing down. Similar examples of several forces acting in the same or opposite directions are numerous in everyday life.

More interesting are problems about forces pulling at various angles with one another. Suppose, for instance, that you set out to cross a stream in a canoe, paddling straight for the opposite shore. Suppose that you propel the boat forward against the friction of the water with an average force of 40 lb and that meanwhile the river's current drags you downstream with an average force of 20 lb. Obviously, your boat will move at an angle to the shore. In Fig. 42 the force you exert is represented by the arrow A, the force of the current by B; the boat moves as if acted on by a single force in the direction of the arrow C. If you wish the boat to move straight toward the opposite shore, you must paddle at an angle upstream, as shown in Fig. 43.

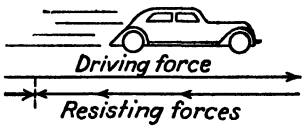


FIG. 41. *Forces on a moving car. In this case the engine's driving force is slightly greater than the resisting forces, so that the car is accelerating.*

Again, imagine an 80-kg pail of water supported by two people, each pulling at the end of a rope passed under its handle (Fig. 44). To support the pail requires a force of 80 kg pulling straight upward, shown by arrow R. This force is supplied by the combination of two oblique forces of 50 kg each, arrows P and Q. Evidently the required forces P and Q would be changed if the angle between them were changed: each would be less if

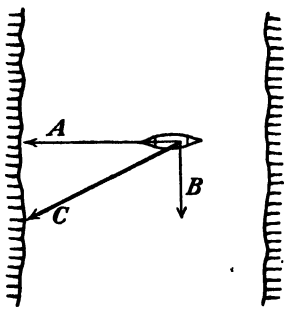


FIG. 42. *A is force on canoe exerted by paddler. B is force due to current. Canoe moves as if acted on by a single force, C.*

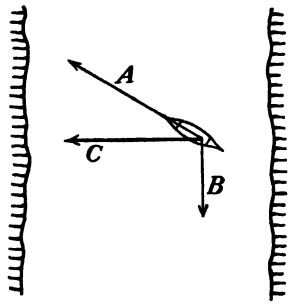


FIG. 43. *If the canoe is to move straight across, it must be paddled at an angle upstream.*

the angle were smaller, greater if the angle were larger. Or the forces might be altered by attaching another rope to the pail, thus supplying a third supporting pull. But changing the angle or the number of ropes does not alter the fundamental situation: oblique forces in combination produce the effect of a single vertical, upward force.

From these various experiments we may draw the conclusion that the effect of a single force on an object may be duplicated by several forces acting together, either along a straight line or inclined to one another.

Vectors

Each force in the diagrams of preceding paragraphs is represented by an arrow which points in the direction in which the force acts and which shows by its length how strong the force is. Arrows of this sort, representing the *direction* as well as the *size* of some quantity, are called ***vectors.*** They are useful in discussions of quantities like force, velocity, acceleration, in which direction is important. Other quantities, like mass, speed, volume, have no directional significance, and for them vectors cannot be used.

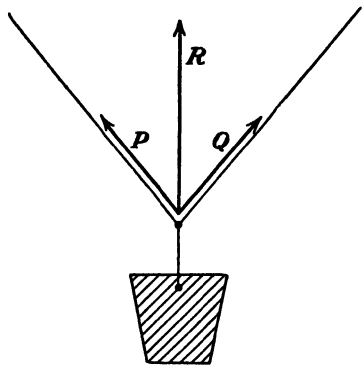


FIG. 44. *Pail supported by two oblique forces.*

Vectors are useful not only as crude representations in diagrams but also in exact mathematical computations. For instance, in the problem diagrammed in Fig. 42, two forces are given, one of 40 lb represented by vector A, the other of 20 lb represented by vector B. Intuitively we know that the resulting single force produced by their combined effects must lie somewhere between them; but how can we obtain exact numerical values for its strength and its direction? No obvious way of handling the problem suggests itself with ordinary mathematical methods, but vectors make the solution easy.

First let us try vectors with the simpler case of forces acting in a straight line. Suppose that two horses pull on a wagon, one with a force of

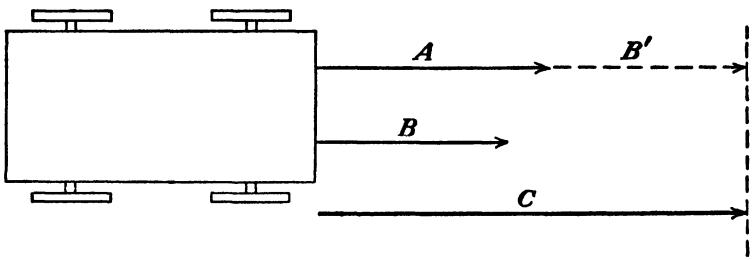


FIG. 45. *Wagon pulled by two horses. C is vector sum of A and B.*

300 lb, the other with a force of 250 lb. A diagrammatic top view of the wagon is shown in Fig. 45, with the horses represented only by vector arrows. Lengths of the arrows are determined by letting 1 cm represent 100 lb. Here obviously the total pull is the sum of the two forces, 550 lb. In terms of vectors the sum is obtained by moving B to the position B', with its tail at the tip of A, and drawing another arrow from the tail of A to the tip of B' (drawn separately as C to avoid confusing the diagram).

Vector C has a length of 5.5 cm, representing a force of 550 lb, and it points in the same direction as A and B.

Figure 40 shows a situation in which two forces are again acting in a straight line, but in opposite directions. We obtain the resulting single force by applying the rule stated in the last paragraph: lay one arrow with

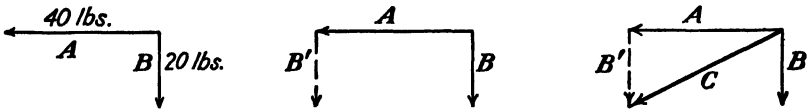


FIG. 46. *Combining vectors in the canoe problem of Fig. 42. Scale: 1 cm = 20 lb.*

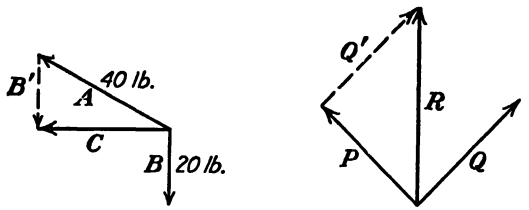


FIG. 47. *Vector addition in Figs. 43 and 44.*

its tail at the tip of the second (actually a little to one side, to avoid confusion; see diagrams at the right of Fig. 40), and draw a third vector from the tail of the second to the tip of the first (heavy-line arrows in Fig. 40).

Now let us return to the canoe problem of Fig. 42. Here the two forces are no longer in a straight line, but the procedure suggested in the last two paragraphs is still applicable. Figure 46 shows first the two force vectors redrawn from Fig. 42, with the river and canoe omitted. In the second diagram, vector B is moved to a new position B', with its tail at the tip of vector A; during this transfer, the length and direction of B remain the same. The heavy arrow of the third diagram is drawn from the tail of A to the tip of B'. This is the desired single force which represents the combined effect of A and B. Its length can be measured directly on the diagram, and the angle it makes with A or B can be found with a protractor.

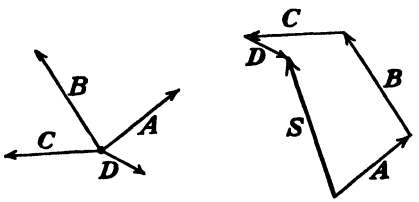


FIG. 48. *In the first diagram, four forces are shown acting at a single point. The second diagram shows how the forces may be added together to give the vector sum S.*

Figure 47 shows similar vector calculations of the resulting forces in Figs. 43 and 44. In each case the same rule is followed: place one vector with its tail at the tip of the other, keeping its length and direction unchanged, then draw a third vector from the tail of the second to the tip of the first. This process of combining vectors is called *vector addition*. It

may be applied to any number of vectors by stringing the arrows together, tip to tail, and drawing the arrow representing the vector sum from the tail of the first to the tip of the last (Fig. 48).

To show the usefulness of vectors in dealing with other quantities besides forces, let us try an example with velocities. Suppose that the speedometer of an airplane heading due north shows a speed of 120 mi/hr and that it is bucking a 40-mi/hr wind from the northwest. In what direction will it actually move and with what speed? Figure 49 shows the solution: first the velocities are drawn as vectors, then they are combined according to the usual rule. Evidently the plane will move in a direction between north and northeast, at a speed of about 100 mi/hr.

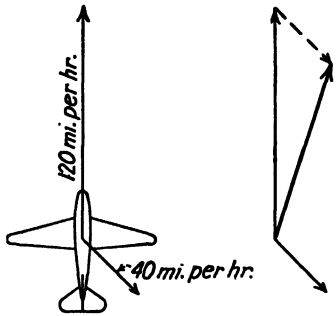


FIG. 49. *Addition of velocity vectors.*

Forces in Equilibrium

A ball suspended on a string is pulled downward by gravity, upward by tension in the string. If the forces are equal, the ball is motionless. The rope in a tug of war is pulled by two opposing forces; as long as the teams are evenly matched the rope does not move. A stone falling freely through

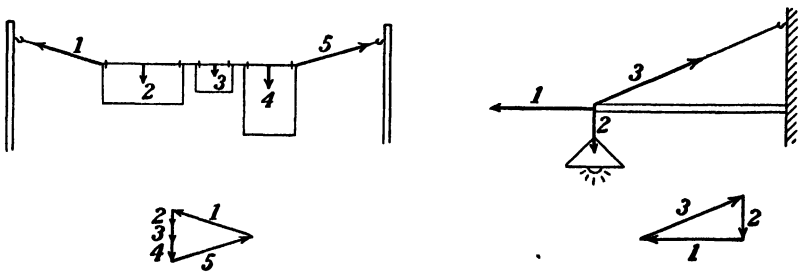


FIG. 50. *Forces in equilibrium.*

the air is acted on by two forces, gravity and air resistance. When the two are equal, the stone falls with uniform velocity (Fig. 40, third diagram). A car maintaining a uniform velocity on a level road is affected by the driving force of its engine and by the various frictional forces which oppose its motion. The weight of clothes hanging from a line is balanced by tensions in either end of the line (Fig. 50). The weight of a lamp hanging from a bracket is supported by tension in one arm of the bracket, compression in the other (Fig. 50). Forces which combine, as in these examples, to give a resulting net force of zero are said to be balanced, or *in equilibrium.*

From Newton's first law of motion it follows that an object acted on by forces in equilibrium must be either at rest or in motion with constant velocity. Whenever we find a body moving with accelerated motion—with its speed increasing or decreasing, or with its direction of motion changing—we know at once that the forces acting cannot be in equilibrium.

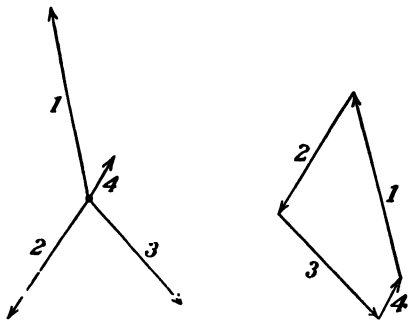


FIG. 51. *Forces in equilibrium.*

The vector sum of forces in equilibrium must, of course, be zero. This means that when the force vectors are laid tip to tail, the tip of the last arrow must meet the tail of the first. Thus a set of forces may be tested for equilibrium by laying their vectors carefully end to end and observing whether the last arrow meets the first. This test is applied in the lower diagrams of Fig. 50 to the forces shown in the upper diagrams. A more complicated example is shown in Fig. 51.

Newton's Third Law of Motion

Quite evidently, forces in the world about us generally act in combination. When we find one force at work, a bit of hunting usually brings to light several others. Let us ponder for a moment the question, is it possible for a single force to exist?

I push downward on the table. As far as I am concerned, that seems to be a single force—an elemental push, giving me the sort of sensation which I must use in any ultimate definition of force. But apparently this force is not acting alone; otherwise the table would move in response to it, with constant acceleration downward. The table resists the force, pushing upward against my hand as I push downward on its top. The harder I press down, the more stubbornly the table resists. Seemingly I cannot exert a force on the table without its exerting a force on me.

Suppose that I transfer myself and the table to the frozen surface of a lake—to a sheet of glare ice, which we will imagine to be marvelously smooth and slippery, so that it can offer no resistance to the table's sideward motion. Again I push on the table, horizontally now instead of vertically, and watch its motion accelerate, as it should under the influence of a single constant force. But again I meet difficulties: I can stick to the ice no better than the table can, and as I push it away from me I find myself starting to move in the opposite direction. Even here I cannot seem to exert a force on the table without its pushing back on me.

Considerations of this sort led Newton to his third law of motion:

For every force there is an equal and opposite force.

No force ever occurs singly; it always has a twin pushing in the opposite direction. A weight hangs on a spring balance (Fig. 52): the weight pulls downward on the spring, and the spring pulls upward on the weight. A chair pushes downward on the floor, the floor presses upward on the chair. The firing of a rifle exerts force on a bullet; the bullet simultaneously pushes backward (the recoil) on the rifle.

Sometimes the reality of the opposite force is difficult to appreciate. A book resting on a table exerts the downward force of its own weight; but just how can an inert object like the table exert a *real* upward force on the book? If the table top were made of rubber, the book would depress it, and the upward force would obviously result from the elasticity of the rubber. A similar explanation will hold for table tops of wood or metal, provided we assume the depression to be extremely small. Again, a falling apple experiences a downward pull from the earth, and by the third law must itself pull upward on the earth with an equal force. We cannot observe the effects of this force, simply because the earth is so very much larger than the apple, but we have no reason to doubt that the force exists.

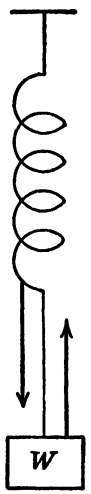


FIG. 52. *Weight pulls downward on spring, spring pulls upward on weight.*

In practice, Newton's third law is easily confused with the idea of force equilibrium outlined in the last section. It is important to remember that in cases of equilibrium (at least insofar as we have discussed them) a set of forces is acting on a *single object.* Newton's law, on the other hand, applies always to *two objects*—the force which one exerts on the second, and the opposite force which the second exerts on the first. To illustrate: forces on a stone hanging motionless at the end of a string are obviously in equilibrium. These forces are (1) gravity, pulling the stone downward; (2) tension in the string, pulling the stone upward. Gravity and tension, *both acting on the stone,* are equal and opposite—hence the equilibrium. Now by Newton's third law the string pulls upward on the stone, and the stone pulls downward on the string. Here again are equal and opposite forces, *but these act respectively upon two objects, stone and string.*

Centrifugal Force

We can now make some headway toward an analysis of motion along curved paths—the sort of motion we must understand in order to explain planetary behavior.

Rapid motion around a curve in a train or automobile produces the familiar sensation of a force pushing us toward the outside of the curve. Mud flung from a moving wheel, the pull of a stone whirled on the end of a string, are other manifestations of this force, which we call ***centrifugal*** ("center-fleeing") force. It always accompanies motion along a curve and always acts directly away from the center of curvature.

What is this force that comes into being when the direction of motion changes and that disappears when the motion is once more straight? From one point of view it is no more than another name for inertia: when our car rounds a corner, we tend to keep going in a straight line because of our inertia and so feel as if we were being pushed toward one side of the car. Centrifugal force is simply an expression of the reluctance of moving bodies to change their direction of motion.

Examined from another point of view, the idea of centrifugal force becomes a little more tangible. Let us return to an experiment which we have all performed in childhood: whirling a ball or a stone tied to the end of a string. It will help the present discussion if you will repeat the experiment in a somewhat critical mood, noting especially the forces involved. You will note that to hold the ball in its circular path requires the continual exertion of a force by your hand on the string—evidently the force responsible for the ball's continual change of velocity. Now consider the ball's motion between any two points fairly close together on the circle, such as O and Q in Fig. 53. At O the ball is moving with a velocity represented by the tangent vector a and, if unrestrained, would continue to move with this velocity along the line OP. The force exerted by your hand is directed along the string toward the center of the circle S and by Newton's second law (page 42) must produce an acceleration, or a change in velocity toward S. Now let the total change in the ball's velocity between positions O and Q be represented by the vector b. The direction of this vector is hard to represent, since it is continually changing, but we shall not be far wrong if we place it parallel to RS, halfway between OS and QS. Obviously the ball's velocity at Q is the vector sum of a (its velocity at O) and b (its total change in velocity); this vector addition is shown in the right-hand diagram of Fig. 53. If the force from S is of the

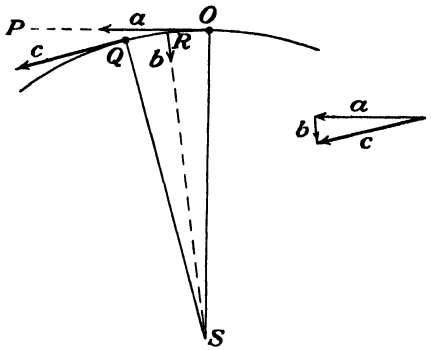


FIG. 53. *Motion of ball whirled on the end of a string. Ball tries to move along OP with velocity a; velocity b added by force toward S; hence at Q ball moves with velocity c.*

right size to keep the ball moving with uniform speed in a circle, c should have the same length as a and should be tangent to the circle at Q.

The centrally directed force which your hand, through the string, must exert on the ball is called a ***centripetal*** ("center-seeking") force. The equal and opposite force, with which (by Newton's third law) the ball pulls outward on the string, is the centrifugal force of its motion. In other words, from this point of view centrifugal force is regarded as an outward pull equal and opposite to the inward pull which is necessary to maintain motion in a curve.

Your experiment with ball and string should tell you a few other items about centrifugal force. Increasing the speed of whirling causes the force to increase rapidly. Using a heavier ball means an increase in the force. Increasing the string's length also produces an increase in force, provided that the ball makes the same number of complete revolutions per second. By making careful measurements of the force for different speeds, weights of the ball, and lengths of string, and by using a graphical analysis like the one we employed for the falling-body experiment in Chap. III, we could deduce a precise formula for centrifugal force. Rather than take the time for such an analysis, we shall simply accept the result obtained by others,

$$F = Kmn^2r \tag{15}$$

where F is the centrifugal force for circular motion, m is the mass of the moving object, n is the number of revolutions per second, r is the radius of the circle, and K is a proportionality constant.

Mathematics and Idealization

The idea of vector representation introduced in this chapter, supplementing our previous discussion of graphs and proportionality, completes our necessary store of mathematical tools. Let us pause a moment to consider critically the application of these tools to scientific problems.

We perform a series of precise measurements, say on the distances which a body falls in certain times. We plot these measurements on a graph, and discover that the graph fits an equation, $d = \frac{1}{2}at^2$. We say then that the equation describes our observations, and confidently predict that it will also describe future observations of falling bodies. This is familiar scientific procedure.

But one step here is easily overlooked. The graph whence our equation comes is a line drawn through a series of points (*cf.* Fig. 37, page 54). The line does not pass precisely through the centers of all these points; some lie above the line, some below it. No matter how carefully the experiment is performed, we cannot draw a perfectly straight line or a perfectly

smooth curve through every point. There are two good reasons for this: (1) Our measurements cannot be strictly accurate, because of imperfections in rulers and clocks; (2) small effects, due to such factors as air resistance, air currents, near-by magnets or electric charges, creep in when extreme precision is attempted. But we reason: *If we could* eliminate all extraneous effects and *if we could* procure perfect rulers and clocks, *then* all the points would lie on the curve, and the equation would accurately describe the body's fall. In other words the mathematical statement applies strictly to an ideal experiment, not to the actual one.

Application of mathematics to physical science always involves this process of *idealization*. Actual experiments, considered in all their minute details, are far too complicated, and our measuring instruments too imperfect, for mathematical formulas to be used directly. So we flee from hard reality to the idealized experiments of our imagination, in which objects and forces have simple properties which our equations can describe accurately. The exact laws and equations of preceding pages apply strictly to "balls" which are perfect spheres moving without friction, to "surfaces" which are perfectly smooth and horizontal, to "forces" which are constant and applied at a single point. In using these laws and equations in actual problems, we must add such phrases as "neglecting friction" or "neglecting air resistance."

Of course, we try to make the actual experiment in our laboratory and the ideal experiment in our imagination correspond as closely as possible. We can often modify the ideal experiment so as to make it a more accurate counterpart of the real one. For instance, suppose that we desire to investigate carefully the effect of air resistance on the motion of a falling body. Experiment shows that the speed of fall increases up to a certain point, then remains constant (*cf.* page 63). Now, if we imagine an experiment in which a perfect sphere is pulled by a constant force through a gas of uniform density, we can find a fairly simple equation to describe its motion, and this equation fits the actual observed motion better than the equation $d = \frac{1}{2}at^2$ of Chap. III. In similar fashion we might go on to consider variations in the force of gravity and in the density of air during the sphere's fall. The ideal experiment would grow more and more similar to the actual one, the equation would become ever more complicated and ever more nearly correct. But always further refinements would be at least theoretically possible.

The well-tested formulas of physical science of course describe actual experiments with sufficient accuracy for all practical purposes. But we should remind ourselves occasionally that *there is nothing inherently mathematical about the experiments themselves;* mathematics is a purely human invention, which we use as a tool, a method of describing idealized experiments which are never exact duplicates of real ones.

Questions

1. A book lies on a table. Of the following forces exerted on, or exerted by, the book, which pairs are equal and opposite? Which one of these pairs holds the book in equilibrium?
   a. The pull of gravity on the book.
   b. The upward force of the table on the book.
   c. The force exerted by the book on the table.
   d. The pull of the book on the earth.
2. Sketch a graph (*cf.* Fig. 39) of (*a*) distance against time, (*b*) acceleration against time, for a body falling through the air.
3. A 100-lb barrel is held stationary on an incline by a force of 50 lb. acting along the incline (Fig. 54). By means of vectors, find the direction and size of the single force

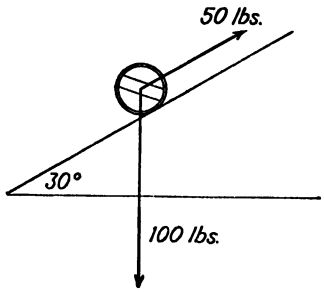


FIG. 54.

equivalent to these two forces. How would the three forces change if the incline were made steeper?
4. A bullet is fired horizontally from a high cliff at a speed of 300 m./sec. As soon as it leaves the gun it starts falling, its downward velocity increasing at the usual rate of 980 cm/sec.^{2} By means of vector diagrams find its approximate velocity after 1 sec; after 5 sec; after 10 sec. Neglect air resistance.
5. In Fig. 44, a pail is supported by two diagonal forces, P and Q. Since the pail is stationary, forces on it may be considered in equilibrium. What is the third force which pulls against P and Q? What is the relation between this third force and the force R shown in Fig. 44?
6. A 20-g. weight whirled at the end of a string 1 m. long, making 3 revolutions per second, pulls on the string with a force of approximately 75,000 dynes. From Eq. (15), what would this force become if
   a. The length of the string were doubled, keeping the weight and the number of revolutions constant?
   b. The weight were increased to 60 g., keeping the length of string and number of revolutions constant?
   c. The number of revolutions per second were increased to 6, keeping the weight and length of string constant?
7. An object at the earth's equator is affected by centrifugal force due to the earth's rotation, and at the same time by centrifugal force due to the earth's revolution about the sun. Which of these forces is the larger? (Earth's radius is approximately 4,000 mi; consider earth's orbit circular, with a radius of 93,000,000 mi.)
8. Where on the earth's surface is the centrifugal force due to the earth's rotation greatest? Where is it least?

CHAPTER VI

The Law of Gravitation

THE man who succeeded in finding a connection between Kepler's laws and the behavior of moving bodies here on the earth was Isaac Newton, the same great English scientist whose laws of motion have figured so prominently in preceding chapters (Fig. 55). The facts of Newton's life are simple and undramatic: son of an obscure farmer, his early aptitude for mechanics and dislike for agriculture led his family to send him to Cambridge; he distinguished himself as a student and a few years after graduation was appointed professor of mathematics; living quietly at Cambridge, he never married, traveled little, shunned controversy whenever possible; he held various minor administrative posts in the university and in the government; honors increased with the years, and he was at length buried with the noblest of England's dead in Westminster Abbey. In contrast with this busy but quiet life are the adventures of his far-ranging mind, adventures which cannot but amaze all who read of his work. In the sweeping law of gravitation Newton found a final solution to the problem of planetary motion, and gave science a tool for reaching out into space beyond the solar system. His formulation of the three laws of motion placed the science of mechanics on a solid foundation. By inventing calculus, Newton gave physical science a new and powerful kind of mathematics whose possibilities are still being explored. Finally, his work in optics was among the earliest systematic investigations of the properties of light.

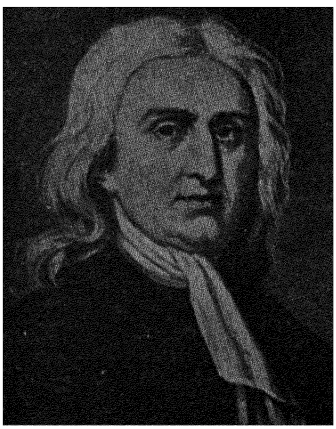


FIG. 55. *Isaac Newton (1642–1727).* (*Courtesy of Photo House.*)

Newton's great work, the *Principia*, was published in 1687, the last year of Charles II's reign. This event is one of the most important landmarks in the whole history of science.

Of all the resounding tributes that have been paid to Newton's greatness, perhaps the most elegant is that by the mathematician Lagrange: "Newton was the greatest genius who ever lived, and the most fortunate, for there cannot be more than once a system of the world to establish."

Universal Gravitation

That some force is necessary to hold the planets in their elliptical orbits had been recognized before Newton, but the nature of the force had remained a matter of vague speculation. It was Newton's great inspiration that this force might be the same as the well-known force which pulled objects to the surface of the earth. Perhaps, thought Newton, the moon revolves around the earth much as a stone on the end of a string revolves around your finger, with gravity taking the place of your pull on the string. In other words, perhaps the moon is a falling object, pulled toward the earth just as we are pulled, but moving so fast that the pull is just sufficient to keep it from flying off in a straight line away from the earth. Perhaps further, the earth and its sister planets are held in their orbits by a greater gravitational attraction from the sun.

It was a grand idea, but would it work? Galileo had discovered certain exact laws (for instance, $d = \frac{1}{2}gt^2$) expressing the relations between distance, velocity, and time for ordinary objects falling freely toward the earth; could it be shown that the moon falls according to the same laws? Kepler had set down three exact statements concerning planetary motion; could a force between planets and sun explain those three laws?

Newton's problem was to express in mathematical terms the amount of the gravitational force between different objects: was the force determined by their mass, their size, their distance apart, their rate of motion, their temperature, or by some combination of these quantities? Kepler's laws served Newton both as a guide in solving this problem and as a rigorous test for any force expression he might devise. From the second law, which states that the line joining any planet with the sun sweeps out equal areas in equal time intervals, Newton showed, with the aid of his newly invented calculus, that the force must act directly along the line between sun and planet—just as gravity on the earth pulls objects directly toward the earth's surface. Kepler's third law gave a more specific hint as to the nature of the force. This law states that the square of the time of a planet's revolution is proportional to the cube of the radius of its orbit; if T is the time of revolution and r the mean radius,

$$T^2 = kr^3 \qquad (cf.\ \text{Prob. 10, page 61})$$

T is the time, say the number of seconds, in each revolution; if we let n be the number of revolutions per second,

$$T = \frac{1}{n}$$

[Thus if an object revolves once in 3 sec ($T = 3$), in each second it completes ⅓ of a revolution ($n = \frac{1}{3}$).] Substitute $1/n$ for T.

$$\frac{1}{n^2} = kr^3$$

By an algebraic rearrangement

$$n^2 = \frac{1}{kr^3} \tag{16}$$

Now the centrifugal force due to a planet's revolution is given by

$$F = Kmn^2r \qquad \text{[}cf.\text{ Eq. (15), page 71]}$$

where m is the planet's mass, K is a proportionality constant, and n and r have the same meaning as in Eq. (16). This force must be equal to the oppositely directed centripetal pull of gravitation, so that we can use this same expression for the sun's pull on the planet. If we limit the discussion to one planet for the moment, we may regard m as constant, and the expression for the force becomes

$$F = K'n^2r$$

where $K' = Km$. In this equation substitute the value of n^2 given by Eq. (16).

$$F = \frac{K'r}{kr^3} = K''\left(\frac{1}{r^2}\right) \tag{17}$$

where $K'' = K'/k$. In other words, *Kepler's third law combined with the law of centrifugal force leads to the conclusion that the force of gravitation exerted by the sun on a planet is inversely proportional to the square of the mean distance between them.* Kepler's first law, that each planet moves in an ellipse with the sun at one focus, furnished a check for Eq. (17); with the aid of calculus, Newton showed that a planet attracted to the sun by a force inversely proportional to the square of its distance must travel in an ellipse.

Galileo's work on falling bodies gave another clue. All objects at the earth's surface fall with the same acceleration; hence the force of gravity on any object is proportional to its mass (page 44). Since, by Newton's third law of motion, the object also attracts the earth, the force between object and planet must vary also with the planet's mass. Hence the gravi-

tational force between any two bodies with masses M and m is

$$F = \frac{KMm}{r^2} \tag{18}$$

This expresses the direct proportionality between the force and each mass, and the inverse proportionality between the force and the square of the distance separating the masses. Equation (18) is adequate to account for both Kepler's laws and Galileo's experiments—to connect the force of gravity on earth with the force which holds the planets in their orbits. Evidently the gravitational attraction between two bodies depends only on their masses and their distance apart, not at all on their temperatures, compositions, or motions.

Newton was still confronted with one difficult problem: From what points in two objects should r be measured? The force between an apple and the earth is inversely proportional to the square of their distance apart; but *what* distance? From the apple to the earth's surface, to the earth's center, or to some other point in the earth's interior? Again using calculus, Newton finally succeeded in showing that for two spherical objects r is the distance between their centers—in other words, spheres behave as if their masses were concentrated at their centers. Solution of the problem is more difficult for objects of other shapes, but in general for any body a "center of gravity" can be found. With this final difficulty removed, Newton was ready to announce his law of universal gravitation:

> *Every particle in the universe attracts every other particle with a force which is directly proportional to the product of their masses and inversely proportional to the square of the distance between them.*

The important "inverse-square" relationship between force and distance may seem at first a bit difficult to grasp. By way of a numerical example, consider the earth's attraction for a ball which weighs 1 lb at sea level, as the ball is moved farther and farther from the earth. At sea level, the distance of the ball from the earth's center is approximately 4,000 mi. If this distance is multiplied by 2, the attractive force must be decreased by 2^2, or 4, times; in other words, when the ball is 8,000 mi from the earth's center, its weight is only $\frac{1}{4}$ lb. Weights of the ball at greater distances are given in Table V. Note that *as distance increases, force decreases very rapidly;* a small change in distance means a big change in the opposite direction in force.

Newton's Proof of the Inverse-square Relation

Gravitational forces between objects on the earth's surface are so exceedingly small that their measurement is very difficult. In Newton's day such measurement was impossible, so that the law of gravitation

could not be established by direct laboratory experiment. Agreement of the law with Kepler's generalizations was of course good evidence in its

TABLE V

Distance from Earth's Center	Weight of Ball		
4,000 mi (= 1 × 4,000)			1 lb
8,000 mi (= 2 × 4,000)	$\frac{1}{2^2}$	or	$\frac{1}{4}$ lb
12,000 mi (= 3 × 4,000)	$\frac{1}{3^2}$	or	$\frac{1}{9}$ lb
40,000 mi (= 10 × 4,000)	$\frac{1}{10^2}$	or	$\frac{1}{100}$ lb
240,000 mi (= 60 × 4,000)	$\frac{1}{60^2}$	or	$\frac{1}{3,600}$ lb

favor but provided no convincing connection between planetary forces and the familiar force of gravity at the earth's surface. There was one astronomical body, however, which could give Newton this necessary connection and which at the same time could furnish a proof of the gravitational law which did not depend on Kepler's laws. This body was the moon.

Freely falling objects at the earth's surface move downward with an acceleration of about 980 cm/sec^2. At the moon's distance, 240,000 mi, the pull of the earth is reduced to 1/3,600 of its value here on the surface (see Table V), so that objects should move toward the earth with an acceleration of 1/3,600 × 980, or 0.272 cm/sec^2. If the law of gravitation is correct, this means that the moon should be falling toward the earth with a velocity increasing by 0.272 cm/sec each second; in other words, that it should fall 0.136 cm in 1 sec (since $d = \frac{1}{2}at^2$), 490 cm in one minute, 10,200 km (about 6,300 mi.) in one day.

What does it mean to say that the moon "falls" toward the earth 6,300 mi in a day? Obviously our satellite is not directly approaching us at this rate. The "falling" of the moon is simply its deflection away from the straight-line path which it would follow if the earth's pull were absent. In Fig. 56, the arc TV represents part of the moon's (M) orbit around the earth E. At the instant when the moon passes M, its velocity is given by the vector A, along the tangent MR. This straight line would represent the moon's path away from M if the earth did not attract it, MS indicating the distance it would cover in one day. The earth's attraction pulls the moon away from this path, so that in one day it moves to P rather than S. The line SP represents, in effect, the distance which the moon "falls" in a day's time.

To establish the correctness of his gravitational law, Newton's problem was to show that *calculated* values for the amount of the moon's fall, like those given above, corresponded with *observed* values of its deflection

from a straight-line path. These observed values are found very simply from observations of the moon's distance and rate of motion. We know that the moon completes its nearly circular orbit once in 27.3 days, hence

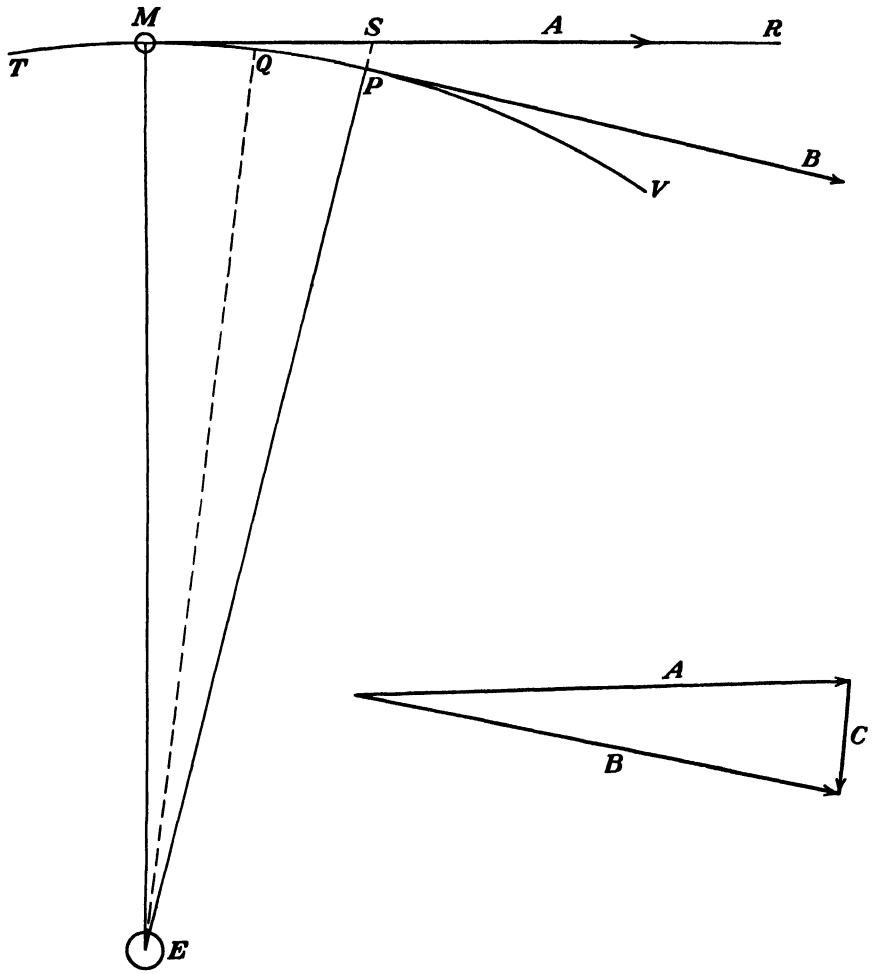


FIG. 56. *The moon as a falling body. In one day the moon "falls" a distance SP away from a straight-line path. The small triangle is a vector diagram of velocities (see Prob. 7, p. 89). Scales: Distances, 1 cm = 20,000 mi. Velocities, 1 cm = 0.1 mi/sec.*

that it moves in one day through an arc of 360/27.3, or 13.2°. We know further that its average distance from the earth is 240,000 mi. In Fig. 56, this 240,000 mi is represented by 12 cm; the angle MEP is 13.2°. By direct measurement on the diagram, the amount of the moon's deflection from a straight line SP is found to be a little over 3 mm. Since every centimeter represents 20,000 mi, 0.3 cm is equivalent to 6,000 mi. This is not far from

Obviously so small a distance as *SP* cannot be measured with any great accuracy. We shall attempt a somewhat more accurate comparison of observed and calculated values in Prob. 7 at the end of this chapter, by using a vector diagram of velocities. To make the comparison with the precision demanded in scientific work, we should need to use more accurate values for the earth's radius, for the moon's distance and time of revolution, for the acceleration of gravity; we should have to remember that the moon's orbit is actually an ellipse rather than a circle; and we should find it advisable to carry out the calculation mathematically rather than by measuring distances in a diagram. When Newton performed this computation, he found close agreement between observed values and those calculated from his gravitational law. The force which holds the moon in its orbit was proved identical with the pull of the earth on objects at its surface.

Experimental Proof of Newton's Law

Nearly a century after Newton's time another great English scientist, the eccentric Henry Cavendish, succeeded in actually measuring the force of gravitation. His method was direct and simple: two small metal spheres were attached to the ends of a light rod, which was suspended at its center from a slender wire; two heavy lead spheres were brought near the small ones, and the twist in the wire was observed. Simple as it sounds, completion of the experiment was a real feat, for the forces involved are incredibly small. By these delicate measurements, Cavendish obtained direct experimental proof of the law of gravitation.

Modern measurements with instruments similar to Cavendish's give a value of 666×10^{-10}* for the constant K in Eq. (18), when masses are expressed in grams, r in centimeters, and force in dynes. This means that two 1-g. masses placed 1 cm apart attract each other with a force of 0.0000000666 dynes. The dyne itself is too small a force for our gross senses to detect it—and this force is less than one ten-millionth of a dyne!

Even large masses on the earth's surface exert very little gravitational attraction on one another. Two 40,000-ton oceanliners, for instance, if placed ½ mi apart, attract each other with a force of about ½ ounce. But for huge aggregates of matter like the sun and planets, gravitational forces become enormous.

Consequences of Gravitation

The Shape of the Earth. We say glibly that the earth is a sphere. To be more precise we should call it an oblate spheroid, meaning that it is flattened somewhat at the poles. Is there any good reason for this particu-

* The meaning of small numbers expressed in this form is explained in Table XXIX, p. 662.

lar shape? Why should the earth not be shaped like an egg, like a pyramid, like a corkscrew?

We can answer this question by a consideration of pressures beneath the earth's surface. Pressures in water are familiar enough: swimmers can descend only a few tens of feet, submarines only a few hundred, before the pressure becomes dangerously high. These crushing forces are due simply to the weight of overlying liquid, to the earth's gravitational attraction for the upper layers of water. Less familiar are pressures in solid rocks. Most ordinary rocks are between two and three times as heavy as water—which means that pressures in rocks, due simply to their weight, should be more than twice as great as pressures at similar depths in the ocean. With increasing distance beneath the earth's surface, rock pressures quickly become enormous. Below a depth of about 12 mi the pressure is so great that solid rock will flow in response to it. We have no direct information as to what kinds of materials exist 12 mi down, but we may be practically

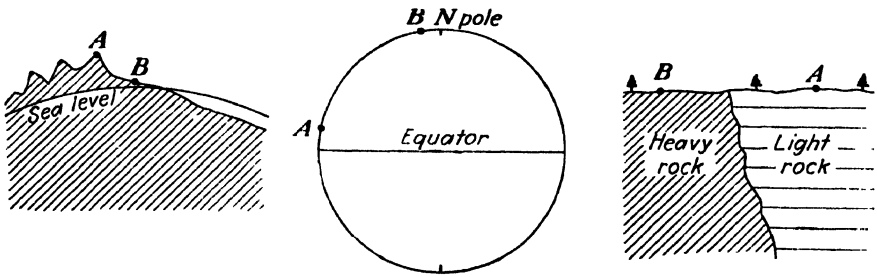


FIG. 57. *Three reasons for variation in the force of gravity. In each drawing a man would be heavier at B than at A*

certain that they behave somewhat like thick liquids in response to pressure changes. This means that one part of the earth cannot project out very much farther than other parts; if it did, pressures beneath it would be greater than under surrounding regions, and the rock beneath it would flow out to the sides until pressures were equalized.

In other words, gravity alone tends to give the earth a spherical shape, to keep all parts of its surface at the same distance from the center. Such minor irregularities as mountains and ocean basins do not greatly disturb the pressure balance, but no large protuberances can exist.

Modification of the spherical shape to spheroidal is the result of centrifugal force due to the earth's rotation. This force is evidently greatest for points on the earth's equator, where the surface is farthest from the axis of rotation. Hence the earth is somewhat expanded around its middle.

Similar reasoning explains the spheroidal shapes of other objects in the solar system. Venus rotates very slowly, hence is nearly a perfect sphere; Jupiter and Saturn show a conspicuous polar flattening from their very

rapid rotations. We may generalize for bodies even beyond the solar system: because of gravity and centrifugal force, any astronomical object above a certain size must be spherical or spheroidal. Only small objects such as meteors and small planetoids, on which gravity is too feeble to produce great pressures, can have irregular shapes.

Variations of Gravity on the Earth's Surface. A man's weight changes as he travels about over the earth. The change is never large—not more than a pound or so at most—but he would have no trouble detecting it with delicate instruments. If he is gifted with extraordinary perseverance and scientific curiosity, he will at length discover that there are three distinct types of variation in his weight (Fig. 57).

1. *Variation with altitude.* On a mountain top he will weigh less than in adjacent valleys. This observation he will explain simply as an effect of changing distance from the earth's center.

2. *Variation with latitude.* Near the equator he will find his weight a minimum, near the poles a maximum. A little reflection will show two reasons for this variation: the centrifugal force of the earth's rotation, which works against gravity, is greatest at the equator; and points along the equator are farthest from the earth's center, because of the earth's equatorial bulge.

3. *Variation with density of subsurface material.* In addition to the well-defined variations with latitude and altitude, our investigator will discover erratic minor changes in weight whose explanation is not so obvious. At least in part these minor variations are due to differences in the rocks immediately below the surface: in regions where the subsurface rocks are unusually heavy, gravity is somewhat greater than in regions underlain by light materials.

These variations in the force of gravity explain the measured variations in g mentioned in Chap. IV (page 55).

The Mass of the Earth. Using the result of Cavendish's experiment, we can weigh the earth. This sounds like a formidably difficult operation; actually, we need do no more than put a mass of 1 g. on a spring balance and read its weight. The weight, surprisingly, turns out to be 1 g., so we conclude that the earth attracts a 1-g. mass with a force of 1 g. This sounds a bit silly, but it gives us all the information we need to substitute in Newton's equation [Eq. (18)] and find M, the earth's mass. For F in this equation is simply the 1-g. force between earth and 1-g. mass; r is the distance separating their centers, or 4,000 mi; and K is Cavendish's constant, 666×10^{-10}. Before substituting, we need only change our units so that force is expressed in dynes, distance in centimeters, masses in grams.

On page 58, Chap. IV, we found that a force of 1 g. is equivalent to about 980 dynes. A distance of 4,000 mi is equal to 640,000,000 cm (since

$4{,}000 \times 5{,}280 \times 12 \times 2.54 = 640{,}000{,}000$). Hence

$$980 = \frac{0.0000000666 \times M \times 1}{(640{,}000{,}000)^2}$$

whence

$$M = 6 \times 10^{27}* \text{ g. (6 followed by 27 zeros)}$$
$$= 66 \times 10^{20} \text{ tons (66 followed by 20 zeros)}$$

Calculations of certain other astronomical masses from the law of gravitation are almost as simple. We shall undertake one of these calculations in the problems at the end of this chapter.

Weight on Other Astronomical Bodies. How much would you weigh if you could stand on the cloud-obscured surface of Venus, or on the barren deserts of Mars? To answer such a question, we need only Newton's law and data regarding planetary masses and radii.

Let is illustrate by calculating your weight on the moon. Suppose that you tip terrestrial scales at 150 lb. On the moon you will be closer to the center of mass, since the moon's radius is only about one-fourth that of the earth; hence, if the moon's mass were the same as the earth's, you should be attracted to it with a force about 16 times your earthly weight, or $16 \times 150 = 2{,}400$ lb. But the moon has a mass only about $\frac{1}{80}$ that of the earth, so that in spite of your shorter distance from the center of mass your weight will be less—about $\frac{1}{80}$ of 2,400, or 30 lb.

The calculation can be made more formally as follows. Let F_e and F_m be your weights on earth and moon respectively, M_e and M_m the masses of earth and moon, r_e and r_m the radii of earth and moon. Your mass (150 lb) is the same on both, and the constant K of course does not change. Hence

For the moon,

$$F_m = \frac{KM_m \times 150}{r_m^2}$$

For the earth,

$$F_e = \frac{KM_e \times 150}{r_e^2}$$

Divide one of these equations by the other:

$$\frac{F_m}{F_e} = \frac{M_m \times r_e^2}{M_e \times r_m^2}$$

K's and 150's cancel in the division; the ratio of the two weights depends simply on the relative masses and relative radii of earth and moon. Using $M_m = 73 \times 10^{24}$ g., $r_m = 18 \times 10^7$ cm (1,100 mi), and values of M_e and

* This form of expression for large numbers is explained in Table XXIX, p. 662.

r_e given on the last page, we substitute

$$\frac{F_m}{150} = \frac{73 \times 10^{24} \times (64 \times 10^7)^2}{6 \times 10^{27} \times (18 \times 10^7)^2}$$

whence

$$F_m = 24 \text{ lb}$$

This answer is somewhat more accurate than the 30 lb obtained in the preceding paragraph, since more accurate values for masses and radii were used.

On the moon, therefore, you and all other objects would weigh less than one-sixth as much as on earth. Without exerting yourself in the least, you could break all existing records for the high jump, the pole vault, the javelin or discus throw. High hurdles would be no more of an obstacle than pebbles to a jack rabbit. You could dive from immense heights without injury, and accomplish quadruple somersaults with ease—provided that you imported water to land in from the earth.

Discovery of Neptune and Pluto. Preceding paragraphs have given some hint as to the immense usefulness of the law of gravitation in astronomical explanations and calculations. But the most spectacular application of this law remains to be mentioned: the discovery of the two outermost members of the sun's family.

Uranus was found by accident in 1781 (page 33). Observations during the next few years enabled astronomers to work out details of the new planet's orbit and to predict its future positions in the sky. To make these predictions, it was necessary to consider not only the sun's attraction but the minor attractions of the neighboring planets Jupiter and Saturn as well. The calculations were long and tedious, but their results were accepted unquestioningly. For forty years, about half the time required for Uranus to make one complete revolution, calculated positions of the planet agreed accurately with observed positions. Then an error seemed to creep in. Little by little the planet moved away from its predicted path among the stars. The calculations were checked and rechecked, but no mistake could be found; the attractions of all known bodies had been correctly allowed for. One of two conclusions seemed necessary: either the law of gravitation, on which the calculations were based, was not strictly accurate, or else some unknown body was attracting Uranus away from its predicted path.

So firmly established was the law of gravitation that two young men, Leverrier in France and Adams in England, set themselves the prodigious task of calculating the position of an unknown body which might be responsible for the discrepancies in Uranus's position. They had little to guide them but Newton's law and observations of Uranus, but each

succeeded in finding a probable location for the disturbing body. Adams, completing his computations first, had the ill judgment to send them to England's Astronomer Royal. Busy with other matters, the Astronomer Royal put the calculations away for future checking. Meanwhile Leverrier sent his paper to a young German astronomer, Galle, who lost no time in turning his telescope to the part of the sky where the new planet should appear. Very close to the position predicted by Leverrier, Galle found a faint object which proved to be the eighth member of the sun's family. A little later the Astronomer Royal showed that Neptune's position had also been correctly given in Adams' work.

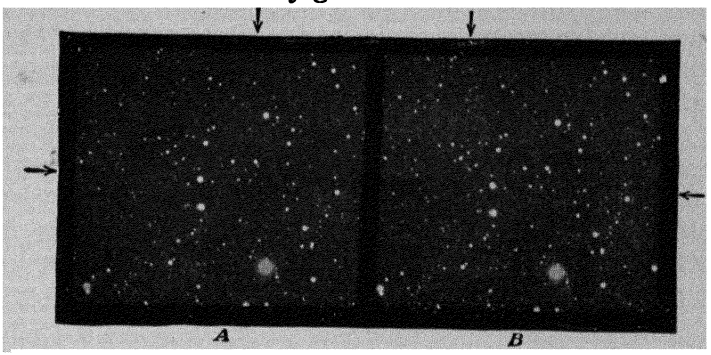


FIG. 58. *Photographs showing two positions of Pluto soon after its discovery. The faint image of the planet may be located by means of the arrows. (Photographed by Tombaugh at the Lowell Observatory.)*

When Neptune had been observed for several decades, very slight discrepancies between its observed and calculated positions led the American astronomer Lowell to predict the existence of yet another planet. The discovery of Pluto was hardly as dramatic as that of Neptune; the discrepancies in Neptune's orbit were so small that Pluto's position could not be predicted with accuracy, and the search dragged on for twenty-five years before the ninth planet was finally located in 1930 (Fig. 58).

The Scientific Method

We have made considerable progress in our attempt to explore the world and its history. At least the stage, so to speak, has been set; we now have a working basis from which to extend our inquiry into the past and farther afield into the present. In Chap. I we watched the efforts of early scholars to gain an idea of the general structure of the universe, and from their conflict of opinions we learned something of the way science progresses. Chapter II filled in details of the accepted picture of the sun's family. In Chap. III we began a search for an explanation of how this family of planets is organized, a search that took us deeply into the

nature of mass, acceleration, and force. In the present chapter this search is ended: we have found in the law of gravitation a simple rule that governs the motions of planets and their satellites. Our search has given us more than we could have anticipated, for the law of gravitation not only explains the solar system but affects the motion of every material particle.

At times in these chapters we have discussed material with apparently little relation to our primary objective. Vector diagrams of forces, for instance, had but little application to the law of gravitation or to planetary motions, and they will seldom appear in later chapters. Mathematical details and formulas introduced in the discussion of velocity and acceleration have for our purposes no intrinsic importance. But the ideas back of this seemingly extraneous material will be very necessary later. Force vectors in themselves are unimportant, but they gave us a new familiarity with the meaning and handling of forces. The equations of motion are nonessential, but they helped us to visualize the difficult concept of acceleration and served to illustrate ideas of proportionality and graphic representation which will be useful many times in the future. Still another reason for the inclusion of so much detail is the insight it gives us into the kind of reasoning which we call the *scientific method.*

A few details of this much publicized reasoning process we have glimpsed in earlier chapters. The basic assumption of the scientific method, that the simplest explanation is the best explanation, was illustrated in Chap. I. Chapter IV suggested the general procedure by which observational data are summarized in mathematical form. Two important geometric tools for scientific analysis were introduced in Chaps. IV and V: graphs and vectors. Perhaps now we have a sufficient background to go beyond these details and to describe the general method by which scientists reach their conclusions.

The scientific method contains nothing mysterious, nothing complex. Reduced to simplest terms, it may be divided into three steps: (1) observation, (2) generalization from the observed facts, and (3) checking the generalization by further observations. Thus observations of the material world are the beginning and the end of scientific reasoning. Observed facts serve both as the foundation on which a scientist builds his theories, and as the ultimate check on the correctness of the theories.

1. *Scientific observation* is carried out accurately and painstakingly, so that all pertinent facts may be collected. A considerable amount of classification and analysis of the observational data may be necessary before generalization is possible.

2. *Generalization* may be merely the statement of a rule or pattern to which the observed facts seem to conform. Or it may be a more am-

bitious attempt to explain the observations in terms of simpler rules and processes. In any case it involves the extension of results from one series of observations to similar observations in new and untried circumstances, in other words the *prediction* of results in other experiments.

3. *Checking a generalization* implies the setting up of new experiments whose results can be predicted from the generalization. If the new observations agree with the predictions, the generalization has proved its usefulness. The new observations may lead to further generalization, or to refinements of the old one, which in turn must be checked by further experiment, and so on indefinitely.

As put forward originally, a scientific generalization is commonly called a *hypothesis*. When checked and rechecked in a variety of ways, so that there is no longer any doubt of its correctness, the hypothesis becomes a *law*. The word *theory* is usually reserved for a larger logical structure, built on two or three fundamental generalizations, designed to explain a wide variety of phenomena. But there is no strict uniformity in scientific circles regarding the use of these three terms.

An appropriate example of an application of the scientific method is Galileo's work on falling bodies, for Galileo was the first to use consistently and constructively the process of generalizing from one set of facts and checking the generalization against other facts. From *observations* on balls rolling down inclined planes, Galileo collected data regarding the distances traveled in certain time intervals. Analysis of these data showed that distances and times were related according to a simple mathematical rule, and Galileo accordingly proposed the *hypothesis* that all falling bodies move so that $d = kt^2$. Further experiments with other inclined planes, and much later experiments with freely falling bodies, offered abundant *checks* for this generalization. Since Galileo's time it has been regarded as one of the *laws* of falling bodies.

For a grander and more complex example, consider the history of attempts to explain planetary motions. The original observations are innumerable records of the positions of the planets with respect to the earth and the fixed stars. From these data the Greeks made several attempts to piece together a reasonable generalization, the most successful being the Ptolemaic hypothesis. From this hypothesis future positions of the planets could be predicted, and observations checked the predictions satisfactorily. As centuries passed and observational methods were improved, it appeared that the check was not quite accurate. Some modification of the original hypothesis was necessary, but modification seemed impossible without introducing unreasonable complexities.

Then appeared a rival generalization, the hypothesis of Copernicus. Predictions from this hypothesis gave at first no better checks with observation than Ptolemy's hypothesis. Kepler refined Copernicus's

idea, basing his more accurate generalizations on the observations of Tycho Brahe. With these modifications, the Copernican system now made possible predictions of planetary positions which accurately checked observations. Because of these accurate checks and because of its greater simplicity, the Copernican hypothesis soon replaced the Ptolemaic hypothesis. Its correctness was established even more firmly by the telescopic observations of Galileo and his successors.

Behind Kepler's laws and Galileo's law of motion Newton discerned an even broader generalization, the law of gravitation. This generalization he checked by observations of the moon's motion. Newton and the physicists who followed him used his law to make one prediction after another which has been checked by observation. The most spectacular of these predictions led to the discovery of Neptune and Pluto.

For other examples of the power and usefulness of the scientific method we need look no further than the automobile, the radio, the electric light. So spectacularly successful has scientific reasoning proved in dealing with our environment that many enthusiasts would now apply it in other fields, such as ethics, education, sociology. Certainly the scientific method has some place here, but its strict limitations must not be overlooked. The scientific method is a means of discovering and organizing facts about the material world—no more than that. Observed facts are the foundation of its structure and the ultimate proof of its results. Usually the facts must be based on observations which can be repeated at will; always they must be facts which would be clear to anyone with normal senses and sufficient training to understand them. Scientists may disagree as to how observations should be interpreted, but about the observations themselves there should be no dispute. Of course, this insistence on accurate observational data is one of the great virtues of scientific reasoning. But it is likewise a severe limitation, a limitation which must be clearly recognized if the scientific method is to serve any useful purpose in fields which emphasize so-called "human values."

Questions

1. Would the acceleration of a falling body on the moon's surface be greater or less than on the earth?
2. Suggest two reasons why a man's weight should be less at the equator than in New York.
3. Is the sun's gravitational pull on the earth the same at all seasons of the year? Why or why not?
4. With how much force does a 1-ton automobile attract the earth?
5. With what acceleration does a meteor move toward the earth when it is 8,000 mi above the earth's surface?
6. From the data of Table I, page 30, calculate how much a 180-lb man would weigh on the surface of (*a*) Venus; (*b*) Mars. (Use radii in miles and relative masses as given in the table; they need not be converted into centimeters and grams.)

7. The lower right-hand diagram in Fig. 56 is a vector diagram of velocities, A and B corresponding to the moon's velocities at M and P, respectively. These vectors have the same length (on the assumption that the moon's orbit is circular) and are separated by an angle of 13.2°. Hence the vector C should represent the change in the moon's velocity in a day's time. Compute this change in velocity (*a*) by direct measurement of C and (*b*) by calculation from the laws of falling bodies (page 54). Explain how the agreement between these figures gives a check on the correctness of Newton's law of gravitation.
8. An approximate calculation of the sun's mass may be made as follows. The centripetal gravitational force exerted on the earth is balanced by the centrifugal force of the earth's revolution. The gravitational force is expressed in Eq. (18), the centrifugal force by

$$F = kmn^2r \qquad \text{[}cf.\text{ Eq. (15), p. 71]}$$

Since the two forces are equal,

$$F = kmn^2r = \frac{KMm}{r^2}$$

whence

$$M = \frac{kn^2r^3}{K}$$

K is Cavendish's constant, 666×10^{-10}, and k is $4\pi^2$ or 39.6. Using these values and appropriate figures for n and r, calculate the sun's mass. Note that it is not necessary to use the earth's mass; only its distance and time of revolution need be known. In similar manner the masses of planets may be found from the distances and times of revolution of their satellites. (This calculation is not strictly accurate, for it assumes that the sun remains stationary during the earth's revolution. Actually both objects revolve about their common center of gravity.)

CHAPTER VII

Origin of the Solar System

GRANDEST of all questions about the solar system is the problem of its origin. How did the earth and its sister planets come to be? It is a query old as history, pondered alike in primitive mythologies and in the discussions of modern science. In primitive mythologies the question was answered, but to modern science it remains an unsolved riddle.

Hitherto we have treated the sun's family as a going concern, focusing attention on its methods of operation rather than on its antecedents. For planetary behavior at present we have hit upon a simple and satisfying explanation in the law of gravitation. We expect that this same law has been true for a long time in the past and will continue to be true in the future, but it tells us precious little concerning the details of planetary history.

So we embark on a new sort of exploration, an exploration into *time*. We shall find the procedure somewhat different. No longer is direct observation possible; events of the past must be conjured up by inference from observations of the present. The only possible check on theories regarding the past is their ability to predict situations which we find in the world today. Because of this limitation, we cannot state our results, even in the best of circumstances, with the same assurance we adopted in discussing the present positions and motions of the planets. For a proof of the law of gravitation we appeal to the direct and unanswerable evidence of our senses, but a skeptic can always spoil a reconstruction of the past with the impertinent question: "Were you there?" The best we can hope for is to establish one view as more reasonable than another.

To illustrate, let us begin by disposing immediately of one eminently unreasonable hypothesis.

Perhaps the sun acquired its planetary retinue by chance encounters, picking up one planet after another simply because it happened to pass near them. This is a conceivable hypothesis: can we find any evidence for or against it?

If the solar system came into existence by such a process, we should expect from the law of gravitation that the planets would revolve about the sun in elliptical orbits, those in the largest orbits moving most slowly. This prediction agrees with observation. But we should further expect that the system would show little regularity, since the various planetary encounters would presumably occur at different speeds and different angles. The planets should move in various directions around the sun, their orbits markedly inclined to each other and to the sun's equator. This prediction is emphatically not fulfilled. All the planets revolve in the same direction; their orbits have similar shapes, and all lie nearly in the plane of the sun's equator. Add the facts that all the planets but one rotate in the same direction as their revolution and that nearly all the satellites revolve in this same direction. The order and symmetry of the system are manifest. The chance that purely accidental encounters could give so many similar motions is fantastically remote.

The Facts

Let us try the orthodox scientific approach: setting down all the facts at our disposal which might possibly have some bearing on the problem. The following tabulation is largely taken from Chap. II:

1. The solar system is isolated in space, separated from the stars by distances which are enormous relative to its own dimensions.
2. The total mass of all the planets is less than 1/700 of the sun's mass.
3. Nearly three-fourths of the total mass of the planets is concentrated in Jupiter, fifth in order of distance from the sun.
4. Planets, satellites, and planetoids, with a few exceptions, revolve and rotate in the same direction; this is likewise the direction of rotation of the sun.
5. The planets rotate at widely varying speeds; the sun's rotation is comparatively slow.
6. Orbits of planets and most satellites are inclined only a few degrees to the plane of the sun's equator.
7. Orbits of the planets are ellipses with small eccentricities—*i.e.*, nearly circular.
8. Spectroscopic examination of the sun shows the presence of most of the chemical elements found on the earth. Densities suggest that the smaller planets have compositions not far different from the earth's, but light elements are much more abundant on the larger planets.

These are facts with which any theory of the origin of the solar system must be consistent.

More valuable for present purposes are whatever facts we can muster about *progressive changes* in the motion or physical conditions of the sun, planets, or satellites. If we can discover some change which is occurring today, or better, which has been observed for several centuries, we may estimate the cumulative effect of the change over long periods of time. Reasoning backward in this manner is not altogether safe, since the rate of change may have been different in other ages, but it does give assurance that our inferences have some basis in solid fact.

For instance, we know that the sun is continually radiating immense quantities of light and heat into space. The only plausible explanation for the sun's ability to maintain this radiation is that its matter is slowly being converted into radiant energy—a process possible only at very high temperatures. This means that the sun's mass is continually growing smaller. It is possible to make a rough estimate of the rate at which the mass is disappearing and so to calculate the sun's size and temperature at different periods: the rate of change turns out to be exceedingly slow, so that even in a million years the alteration in the sun's temperature would be scarcely appreciable. This line of inquiry leads to interesting results but gives little information about planetary origins.

Another type of change which we can observe at present is the slow alteration of the earth's surface—the sort of change which goes by the name of "geological." Rivers, wind, and glaciers slowly wear down the land, volcanoes pour liquid rock onto the surface, slow movements change the level of land and sea. By carrying these changes backward, using records preserved in rocks, an expert can infer the existence of seas where is now dry land, of forests and swamps where is now barren desert. Backward nearly two billion years the record takes us; but we find, even in that remote epoch, that the earth's surface was not greatly different from its present appearance. About the planet's ultimate origin the rocks give no hint.

A progressive change which leads to somewhat more tangible results is the change in the moon's orbit produced by the tides. The subject is important enough to deserve a special section.

Tidal Friction

The tides are produced chiefly by the moon's gravitational attraction. A complete explanation is somewhat involved, but the following simple explanation accounts for the principal facts:

The moon attracts different parts of the earth with slightly different forces. For matter at A in Fig. 59 the attraction is strongest, since the distance from the moon is least; at C, the point farthest from the moon, the attraction is weakest. Because of these unequal forces, the moon tends to pull matter at A away from the rest of the earth, and to pull the

entire earth away from matter at C; in other words, the earth is somewhat bulged out at A and C. Solid rock resists the bulging effect to a large extent, but the fluid ocean responds easily. Water is heaped up on the sides of the earth facing and directly opposite the moon, and is drawn away from other parts of the earth. As the planet rotates, the water bulges are held in position by the moon. The earth, so to speak, moves under the bulges, and a given point on its surface experiences therefore two high tides and two low tides per day.

FIG. 59. *Explanation of the tides. The moon's attraction is greatest at A, least at C.*

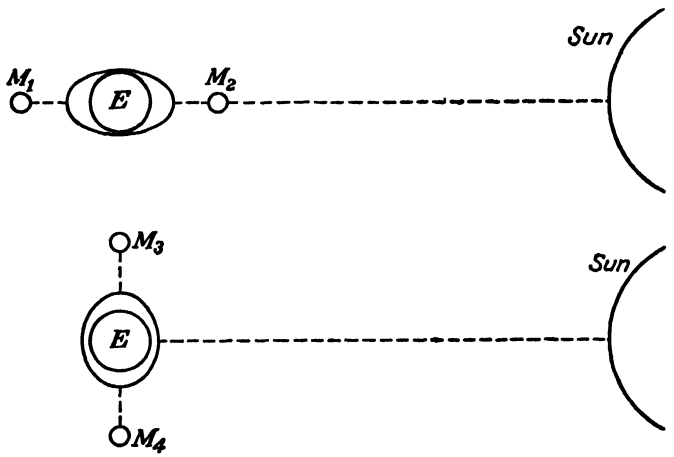


FIG. 60. *Spring tides are produced when the moon is at M_1 or M_2, neap tides when the moon is at M_3 or M_4.*

Smaller tides are produced by similar unequal attractions of the sun for different points on the earth's surface. Roughly twice a month, when sun, moon, and earth are in a straight line, solar tides are added to lunar tides to form unusually high (*spring*) tides; when the line between moon and earth is perpendicular to that between sun and earth, the tide-raising forces oppose each other and tides are unusually slight (*neap* tides) (Fig. 60).

The earth does not rotate smoothly beneath the tidal bulges, but tries to carry them around with it. The moon's attraction prevents them from being dragged very far, but the line between the bulges is somewhat inclined to the line between earth and moon (Fig. 61). In this position the

moon holds the bulges firmly, and the up-raised water drags back on the earth as it rotates. Friction between water and rotating earth is not very great in the open ocean, but along irregular coasts it may be considerable. The effect of the friction is, of course, to slow the earth's rotation; the tidal bulges act like huge but inefficient brake bands, clamped on opposite sides of the spinning planet.

In other words, *because of tidal friction the day should be slowly growing longer.* Verification of this prediction has come from records of ancient eclipses. Using the day's present length, astronomers can calculate precisely when and where eclipses have occurred in the past. These calculations do not agree with the observations recorded by ancient Egyptian and Babylonian astronomers, the discrepancies being greatest for the oldest eclipses. Calculated and observed positions, however, agree well if the slow increase in the day's length is considered. The rate of increase is very small: the time between sunrise and sunrise is longer today by 1/1,000 second than it was 100 years ago, longer by 1/50 second than in the days of Julius Caesar.

FIG. 61. *The tidal bulges are dragged ahead of the moon by the earth's rotation.*

The decrease in the earth's rotational speed is accompanied by an increase in the speed of the moon's revolution. Not only does the moon pull back on the nearer tidal bulge (Fig. 61); with an equal force the bulge pulls the moon ahead in its orbit. As the moon's speed is thus increased, the centrifugal force of its orbital motion is increased, enabling it to move outward away from the earth's gravitational attraction. That is, *our satellite is gradually becoming more and more distant.*

The rate at which the moon is receding, like the rate of the earth's slowing, seems absurdly small: about 5 ft each century. At this rate the moon will take 100,000 years to increase its distance by 1 mi. Yet over longer periods the cumulative effects of such small changes become impressive. It is calculated that some hundreds of millions of years ago (the exact time cannot be computed, since the rate of change in the past is unknown) the earth and moon were very close together, their centers a scant 9,000 mi apart; the day was only 5 hr long, and the month was scarcely longer. Beyond this point mathematical analysis becomes uncertain. It would be an attractive hypothesis to assume that the earth and moon were once even closer, so that they formed part of the same body; but against this idea there are grave objections.

Although tidal friction has apparently played a considerable role in the history of the earth-moon system, it is probably not of great importance in the development of the rest of the sun's family. No other planet, so far as we know, has the necessary distribution of fluid ocean and solid continents to provide appreciable friction.

So even the promising tidal friction idea gives us no clue about ultimate origins. In a situation of this sort, where every attempt to reason cautiously backward from observed facts leads to a hopeless impasse, a scientist has but one recourse: to abandon his beloved facts for a moment, and set up a hypothesis with little but his imagination to guide him. From the hypothesis he can predict its consequences, and if the hypothesis has any claim to validity these consequences should agree, at least in a general way, with observations. Let us examine briefly some of the flights of scientific fancy which have sought to explain the existence of the sun's family.

The Nebular Hypothesis

Imagine a great disk-shaped cloud of gas, or *nebula*, its diameter considerably larger than the present diameter of Pluto's orbit. Imagine that the cloud is slowly rotating, and that gravitation between its particles is causing the cloud to contract. As the nebula shrinks, its speed of rotation must steadily increase—just as the speed of a ball whirled on the end of a string increases when the string is shortened. Presently the outer part of the nebula is moving so fast that centrifugal force overcomes gravitation, and a ring of gas separates itself from the main mass. Contraction continues; the speed of rotation steadily increases; another ring and another separate, until ten in all are formed. The material of each ring gradually condenses into a planet, some of the condensing rings passing through an intermediate stage in which small secondary rings are detached to form satellites. The sixth ring, instead of forming a single large body, condenses into myriads of small ones—the planetoids. Meanwhile the central mass of the nebula has condensed further to form the sun.

This is the celebrated nebular hypothesis, proposed by the German philosopher Immanuel Kant, elaborated at the close of the eighteenth century by the French mathematician Laplace. Beautifully simple and reasonable, explaining neatly such peculiarities of the solar system as its flatness and the common directions of motion of planets and satellites, the nebular hypothesis remained the most widely accepted picture of the earth's origin for nearly a century.

Within the past fifty years, however, difficulties have come to light which make the hypothesis untenable. We need consider here only two outstanding ones. (1) According to the nebular hypothesis, the sun should be rotating very rapidly, since it represents the remainder of the nebula

whose steadily increasing speed of rotation had shaken off the nine planets in succession. Actually, the sun is rotating relatively slowly. In more technical terms, this difficulty is expressed by saying that the sun possesses far less of the total angular momentum of the solar system than it should have according to Laplace's idea. (2) It is extremely doubtful that rings of gas would separate in the manner described, and even more doubtful that, once separated, they would condense into planets.

The Planetesimal, Tidal, and Collision Theories

No modification of the nebular hypothesis which can overcome its basic difficulties has been proposed. The most reputable modern theories start from an altogether different assumption: that the planets were formed as a result of the *close encounter* of the sun with a passing star.

Oldest of the "close-encounter" theories is the *planetesimal hypothesis* proposed by two Americans, Chamberlin and Moulton—one a geologist, the other an astronomer. According to Chamberlin and Moulton, the passing star missed the sun by a narrow margin. As it approached, tidal bulges formed on opposite sides of the sun, much as tidal bulges are heaped up in our oceans by the moon's attraction. At the point of closest approach, the star's attraction, aided by gas explosions within the sun, tore loose great clots of matter from the tidal bulges. Larger amounts of matter, destined to form the major planets, were hurled out on the side toward the star; smaller amounts, ejected with less force from the opposite bulge, eventually became the four inner planets. Ten "bolts" in all were erupted from the sun in rapid succession, and pulled into orbits around the sun as the star receded. The material of the bolts cooled quickly to form myriads of small solid bodies, or planetesimals. The larger planetesimals in each bolt presently coalesced to form a nucleus, to which the smaller planetesimals gradually attached themselves. The planets have been built up to their present sizes by the slow addition of solid planetesimals to the original nuclei; some of the meteors which blaze across our skies today are planetesimals which the earth is only now picking up. Satellites were formed as secondary nuclei, growing like the planets by accretion of planetesimals.

Since the star in its passage would pull all the bolts in the same direction, and since they would be ejected nearly in the plane of the star's orbit, this hypothesis accounts as well as Laplace's for the outstanding regularities of the solar system. No condensation of nebulous rings is required, and the planets could be given high orbital speeds by the star's attraction without a corresponding increase in the sun's rate of turning; hence the planetesimal hypothesis overcomes two of the outstanding difficulties in the older view. In addition, it offers an explanation for the distribution and spacing of large and small planets.

The planetesimal hypothesis has been attacked on many counts, some geological and some astronomical. Chamberlin believed that one of his theory's strongest points was its insistence that the earth has been a solid body since its beginning as a small nucleus. Other geologists prefer to regard the infant earth as a molten globe, which has slowly cooled through long ages, and may even now be not entirely solidified. This is the picture given by the nebular hypothesis, and by all other recent theories.

Astronomically, the planetesimal idea was first questioned because of doubts that scattered planetesimals would actually collect to form planets, and because solar "gas eruptions" of the necessary size are improbable. These difficulties were patched up in the *tidal hypothesis* of two English physicists, Jeffreys and Jeans. The passing star was assumed almost to graze the sun, so that enormous tidal forces were produced, of themselves sufficient to cause separation of matter from the sun and perhaps from the star as well. No matter was ejected from the opposite side of the sun; the matter between sun and star was pulled out, as the star receded, into an elongated ribbon, thickest in the middle and tapering toward each end. Such a ribbon would be unstable, and would break up into separate masses which the star's attraction would set in motion around the sun. These masses, at first fluid and very hot, would cool to form the present planets.

The tidal theory is obviously similar in general outline to the Chamberlin-Moulton idea. In detail it accords somewhat better with current scientific opinion, but difficulties are by no means absent. One serious stumbling block, which applies also to the planetesimal hypothesis, is the speed with which the planets rotate. According to Chamberlin and Moulton, the planets were set in rotation by the unequal infall of planetesimals; according to Jeans and Jeffreys, by the falling back of matter disrupted from the fluid planets by a close approach to the sun. Neither cause is sufficient, without extravagant assumptions, to produce the observed motions. To remedy this defect, Jeffreys has recently proposed a third variant of the stellar-encounter hypothesis: suppose that star and sun actually collided, a bit off-center, so that their outer parts momentarily intermingled. A hot ribbon would be drawn out when the star receded, just as in the tidal theory, but now its parts would be in violent rotatory motion because of friction during the wild turmoil of the collision.

Even the *collision hypothesis* is far from settling all the difficulties to everyone's satisfaction. To enumerate the grave objections which have been suggested would take us too far afield; we have seen enough to be convinced that the earth's origin remains cloaked in mystery. One brilliant hypothesis after another fails when its predictions are matched with observed details of motion in the sun's complex family.

On one point and one only do the various theories agree: the earth's origin is bound up with the origin of the solar system. The materials which make up our small planet should be the same, except for minor variations in amount, as the materials of other planets and their satellites.

Suggestions for Further Reading—Part I

On the sun and planets:

DUNCAN, J. C.: *Astronomy*, Harper & Brothers, New York, 4th ed., 1946. A standard elementary text.

JEANS, J. H., *Through Space and Time*, The Macmillan Company, New York, 1929. A brief, popularly written book by a noted British physicist.

On force, motion, and gravitation:

STEWART, O. M.: *Physics*, Ginn and Company, Boston, 1939. A standard elementary text.

SAUNDERS, F. A.: *A Survey of Physics*, Henry Holt & Company, New York, 1936. A standard text, somewhat more elementary than Stewart.

LEMON, H. B.: *From Galileo to the Nuclear Age*, University of Chicago Press, Chicago, 1946. A lucid and entertaining account of the principal ideas of elementary physics. Less complete and less difficult than the preceding books.

On the origin of the solar system:

RUSSELL, H. N.: *The Solar System and Its Origin*, The Macmillan Company, New York, 1935. Clear and nontechnical but requires a speaking acquaintance with elementary physics and chemistry.

On the history of astronomy and physics:

MAYER, J.: *The Seven Seals of Science*, D. Appleton-Century Company, Inc., New York, 1937. A brief, popularly written history of science.

SEDGWICK, W. T., H. W. TYLER, and R. P. BIGELOW: *A Short History of Science*, The Macmillan Company, New York, 1939. More complete and more accurate than Mayer but not as easy to read.

MOULTON, F. R., and J. J. SCHIFFERES: *The Autobiography of Science*, Doubleday & Company, Inc., New York, 1945. Quotations from the original works of Copernicus, Galileo, and Newton, translated into English.

PART II

MATTER AND ENERGY

GALILEO'S emphasis on the importance of accurate observation and experiment was something new in the world. Others before him had made careful observations, others had paid lip service to the experimental method, but Galileo was the first to show clearly that the painstaking collection of facts could lead to useful results. The finding of relations among facts, the prediction of new facts from these relations, the verification of the new facts by observation and experiment—this method of reasoning, with its insistence on observable facts rather than prejudice or imagination as a basis of judgment, is as clear in Galileo's work as in the science of today. The history of modern science is a record of the application of this procedure to one part of the world after another.

We have witnessed the first great triumph of the scientific method, Newton's explanation of planetary motion. We have followed Newton's reasoning from Galileo's laws of falling bodies and from Kepler's laws of planetary orbits to his own inspiration that these laws might be combined into a single grander generalization, which would encompass both the movements of the planets and the falling of objects toward the earth. We have seen how Newton established his law by making use of observations on the speed of the moon's motion in its orbit.

The basic physical concept in this discussion was the idea of *force*. Galileo understood the action of forces clearly enough, but Newton was the first to set down the precise relations between force and motion. These force relations are fundamental to Newton's picture of the solar system, in which the sun and its family are regarded simply as objects whose motion is determined by the forces acting between them.

Forces and their effects on motion are important in everyday life as well as in explaining planetary behavior. But obviously much in ordinary experience cannot be accounted for in such simple terms: the boiling of water, for instance, or the explosion of dynamite, or the transmission of electricity along a wire. Conceivably forces on a minute scale might be

sufficient to explain these phenomena, but the forces are so far from obvious that their investigation demands new techniques and new fundamental concepts.

In succeeding chapters we set out to explore some of these other parts of experience by applying to them the same scientific method which has proved so successful in dealing with force and motion.

CHAPTER VIII

Energy

ENERGY is another word of everyday speech, like force and velocity, to which science gives a more precise meaning. We say commonly that energy is needed to climb a hill, that a moving train has energy, that certain foods are energy rich, that the earth receives radiant energy from the sun, that an uncomfortably active person is a "whirlwind of energy." Such examples show the universal, many-sided nature of the energy concept but help little in answering the simple question, What is energy? An intangible something which produces or accompanies activity: vague terms like these give the only possible answer in nonscientific language. To approach a more rigorous general definition, we shall first consider energy of motion, by extending our previous study of mass and velocity.

Kinetic Energy

The energy which an object possesses because it is moving, or its ***kinetic energy,*** is described formally by the expression

$$\text{K.E.} = \tfrac{1}{2}mv^2 \tag{19}$$

in which m is the mass of the object and v its speed. That mass and speed should determine kinetic energy seems reasonable enough: a train going 60 mi/hr should have more energy than a hummingbird traveling at the same rate, and likewise more energy than a similar train going 10 mi/hr. But it is not obvious why the m and v should be combined in this peculiar fashion. Why is the v squared, and why does $\frac{1}{2}$ appear in the formula? Why would not the simple product mv do as well?

The simpler combination mv is in fact an important quantity in physics, called the "momentum" of a moving object, but it does not satisfactorily represent kinetic energy. The reason involves mathematical comparisons of kinetic energy with other kinds of energy and will be partially explained later in this chapter. In the early days of physics, even until after the time of Newton, the distinction between momentum and

kinetic energy was as troublesome to scientists as it often is to beginning students today.

For the present we can find some justification for the squared v in a simple example. If two balls are thrown vertically upward, with the starting speed of one just twice that of the other, the former will rise not twice as far but *four times* as far as the latter. (This conclusion follows simply from the laws of falling bodies developed in Chap. IV.) Thus there is some property of the motion which depends on the square of the speed—and this property we choose to call kinetic energy.

The squared v term means that kinetic energy increases very rapidly with increasing speed. At 90 mi/hr a car has nine times as much kinetic energy as at 30 mi/hr—and requires nine times as much energy to bring to a stop. A meteor entering the earth's atmosphere at 50 km/sec and slowed by friction to 5 km/sec thereby has its kinetic energy reduced a hundredfold. The variation with mass is, of course, less spectacular: a 2-ton car going 50 mi/hr has just twice the kinetic energy of a 1-ton car at the same speed.

One common scientific unit of kinetic energy (and of other kinds of energy) is the *erg*, which may be defined as the kinetic energy of a 2-g. mass moving with a speed of 1 cm/sec. This unit is so ridiculously small—about equal to the energy of a mosquito flying 1 mi/hr—that for ordinary purposes a larger unit called the *joule*, equal to 10 million (10^7) ergs, is frequently employed. Thus the kinetic energy of a 2-g. bullet moving 300 m./sec is

$$\tfrac{1}{2}mv^2 = \tfrac{1}{2} \times 2 \times (300 \times 100)^2 = 900{,}000{,}000 \text{ ergs} = 90 \text{ joules}$$

Work and Energy

So at least one form of energy can be discussed in terms of simpler physical quantities and can be easily handled mathematically—facts which the scientist, of course, finds most satisfying. But the notion of energy is somehow more fundamental: we do not need a mathematical formula to tell us that a train moving 30 mi/hr is dangerous if we stand in its path, while a leaf blown at the same speed is not; or that the electric motor of a vacuum cleaner cannot supply enough energy to drive a 5-ton truck. From earliest childhood experience teaches us to evaluate the kinetic energies of moving objects, so that presently we learn to judge them all but unconsciously. On what basis do we make these seemingly instinctive judgments? What is there about a moving object that enables us at a glance to estimate its kinetic energy?

We think of the amount of shock our bodies would feel if they collided with the moving object—how much our hand and arm must yield to a baseball, for instance, to deaden the shock of catching it. We think of the

destruction the moving object might cause—say how completely a train would demolish a car in its path. We think of the useful work the object can accomplish—how far, for instance, a moving hammer head can drive a nail. Always the idea of accomplishing something, of producing activity in another object. This idea of causing activity, of "getting something done," is fundamental in the concept of energy—not only of kinetic energy, but of energy in general. The physicist simply refines his phraseology to read: ***energy is the capacity to do work.***

"Work," in this sense, does not have quite the same meaning as in everyday life. The physicist limits work to a process involving *motion*, motion produced by the exertion of a force. I may "work," in the usual sense, by merely supporting a heavy box, or by sitting quietly and puzzling over a problem in algebra. But in the physical sense I do no work unless I cause an object to move by exerting a force on it. How much work I do depends on two things: how much *force* I exert, and through what *distance* the object moves. To put it more precisely: Work is the product of a force and the distance through which the force acts.

$$\text{Work} = F \times d \qquad (20)$$

The force F must act in the direction of the motion produced; otherwise F refers only to that part of the force which acts in this direction.

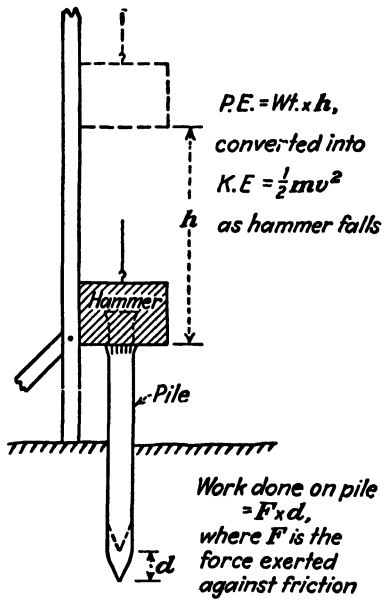


FIG. 62. *The pile driver.*

Now obviously any moving object has the capacity to do work. By striking another object which is free to move, the moving object can exert a force and cause the second object to shift its position. It is not necessary that the moving object actually do work: it may keep on moving, or friction may slowly bring it to a stop. But while it is moving it has the *capacity* for doing work; therefore we say that the object has *energy*, and, since the capacity for work depends entirely on its motion, we describe its energy as *kinetic energy*.

A good example of the relation between work and kinetic energy is furnished by the pile driver (Fig. 62), a simple contrivance which lifts a heavy weight (the "hammer") and allows it to fall on the head of a pile, thereby with successive blows driving the pile into the ground. Just before the pile is struck, the hammer possesses considerable kinetic energy,

which it loses by exerting a force on the pile. During the fraction of a second while the force acts, the pile moves a short distance against the frictional resistance of the ground; hence work has been accomplished by the energy of the falling weight.

Potential Energy

The statement that energy is the capacity to do work is not restricted to kinetic energy but is a perfectly general definition. Obviously some objects need not be in motion to possess energy; a magnet, for instance, can do work on bits of iron in its vicinity whether it is moving or not.

Many objects possess energy simply by reason of their position. Consider the pile driver of the preceding section: when the hammer has been lifted to the top of its support, it need only be released to fall and do work on the pile. The capacity for doing work is present in the hammer as soon as it has been lifted, simply because of its position several feet above the ground. The actual work on the pile is done at the expense of kinetic energy gained during the hammer's fall, but the capacity for working is present before the fall starts. Energy of this sort, depending merely on the position of an object, is called ***potential energy.***

Examples of objects possessing potential energy are literally innumerable. A book on a table has potential energy, since it can fall to the floor; a skier poised at the top of a slide, water at the brink of a cataract, a car at the top of a hill, anything capable of moving toward the earth under the influence of gravity has energy because of its position. Nor is the earth's gravity necessary: a planet has potential energy with respect to the sun, since it could do work in falling toward the sun; a nail placed near a magnet has potential energy, since it could do work in moving to the magnet.

How can potential energy be measured? For a concrete example, let us return briefly to the pile driver. The effectiveness of the hammer in driving the pile should depend on two things: (1) its weight and (2) the height to which it is lifted. Simple analysis would show that the effectiveness of the hammer is directly proportional to these quantities, so that its potential energy (with proper choice of units) may be expressed as a product of weight and height.

$$\text{P.E.} = W \times h \tag{21}$$

Since weight is a force exerted by gravity, potential energy like work is a product of a force and a distance. It is, in fact, precisely equal to the work which would be necessary to move the hammer from the end of its fall back to its starting point.

This expression, $W \times h$, is the most generally useful one for potential energy, since potential energy usually refers to the ability of an object to

move toward the earth. Other forces and distances must be used, of course, for potential energies referring to the sun, a magnet, or some other body.

The same units defined for kinetic energy—the erg and the joule—may be used for potential energy, but it is often more convenient to use simple combinations of weights and heights. For instance, a 10-g. weight lifted 30 cm from the ground has a potential energy of 30 × 10 or 300 gram-centimeters (g.-cm); the car shown in Fig. 63 has a potential energy of 2,500 × 100 or 250,000 foot-pounds (ft-lb). Note that h is always the vertical distance (*not* necessarily the total distance traveled) between the starting point and the lowest point of fall.

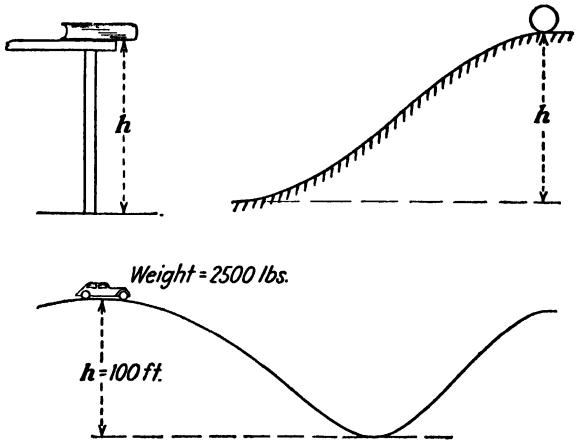


FIG. 63. *Examples of objects possessing potential energy.*

When an object starts to fall, it gains kinetic energy as its motion is accelerated, and it loses potential energy as its height above the earth decreases. Potential energy is said to be *converted into* kinetic energy. When the bottom of the fall is reached, potential energy has been entirely changed to kinetic; barring effects of friction, velocity at this point should reach its maximum value. The reverse change, from kinetic to potential energy, occurs if the object is moving from the end of its fall back to its starting point.

If friction is neglected, kinetic energy at the end of the fall should be just sufficient to carry the object back to its starting point (or to an equal height) of its own accord. For instance, the car of Fig. 63, if its wheels are frictionless, should be able to coast from the top of one hill to the top of the other, its energy being converted from potential to kinetic, then back to potential. Changes of a similar nature, from kinetic energy to potential and vice versa, are well shown in the motion of a pendulum (Fig. 64) and in the motion of a planet following an elliptical orbit around the sun (Fig. 65).

Since potential energy can be changed directly into kinetic energy, it should be possible to express both kinds by the same mathematical formula. To return once more to our overworked pile driver, the kinetic energy of the hammer as it strikes the pile should not only be

$$\text{K.E.} = \tfrac{1}{2}mv^2$$

but should also be accurately given by the expression for potential energy

$$\text{K.E. (at end of fall)} = \text{P.E. (at start of fall)} = W \times h$$

If this second statement is true, it should be possible to derive one expression for kinetic energy from the other.

To accomplish this derivation, we resort to some formulas of preceding chapters. The weight of the falling mass is a force given by the product of the mass and its acceleration, in

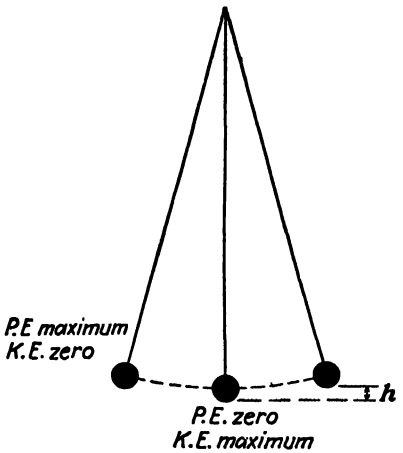


FIG. 64. *Energy changes during the motion of a pendulum.*

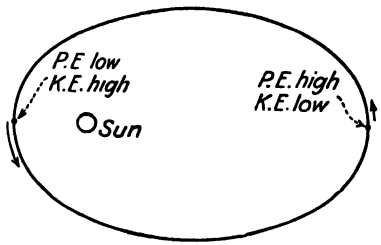


FIG. 65. *Energy changes during the revolution of a planet around the sun.* (*Cf. Fig. 7, p. 13.*)

this case the acceleration of gravity [Eq. (14), page 58].

$$W = mg$$

The vertical distance through which it will fall in a time t is given by the equation for freely falling bodies (page 55)

$$h = \tfrac{1}{2}gt^2$$

Now substitute these symbols in the second expression for kinetic energy.

$$\text{K.E.} = W \times h = mg \times \tfrac{1}{2}gt^2 = \tfrac{1}{2}m(gt)^2$$

We need one further substitution: for a freely falling object starting from rest, acceleration is equal to its velocity at any moment divided by the time during which it has fallen [Eq. (8), page 54].

$$g = \frac{v}{t}, \qquad \text{whence} \qquad gt = v$$

So v can be inserted for the gt of the last equation.

$$\text{K.E.} = \tfrac{1}{2}m(gt)^2 = \tfrac{1}{2}mv^2$$

In words, this derivation means that the weight of an object multiplied by the height through which it falls gives a number equal to half the product of its mass and the square of its speed at the end of its fall.

The derivation shows one good reason for using $\frac{1}{2}mv^2$ as an expression for kinetic energy: in processes involving changes from kinetic energy into potential (and vice versa), no other combination of m and v would give equal values for the two kinds of energy.

Energy in Other Forms

The work of our civilization is accomplished by energy in a variety of forms. The energy of moving machinery and the energy of wind and waterfalls are examples of the two forms we have just been discussing, kinetic and potential energy—often described together as *mechanical energy*. The *chemical energy* of gasoline is used to drive our automobiles; the chemical energy of food enables our bodies and the bodies of domestic animals to perform work. *Heat energy* from burning coal or oil is used to form the steam which drives locomotives. *Electrical energy* and *magnetic energy* turn motors in home and factory. *Radiant energy* from the sun, though man has yet to learn how to harness it directly, performs very necessary work in lifting water from the earth's surface into clouds, in producing inequalities in atmospheric temperatures which cause winds, in making possible chemical reactions in plants which produce foods.

Just as kinetic energy may be converted into potential energy, and potential into kinetic, so may other sorts of energy be readily transformed. In the cylinder of an automobile engine, for instance, chemical energy stored in gasoline and air is first changed to heat energy when the mixture is ignited, then to mechanical energy as the expanding gases push down on the piston. This mechanical energy is in large part transmitted to the drive shaft, but some is used to turn the generator and thus produce electrical energy for charging the battery, and some is changed to heat by friction in bearings. For another example, the principal energy changes in a hydroelectric plant are: potential energy of water above the dam → kinetic energy of moving water → kinetic energy of rotating turbine → kinetic energy of generator → electrical energy (Fig. 66); some heat energy is produced by friction. Natural processes too involve obvious energy transformations: in an animal's body the chemical energy of food is changed into mechanical and heat energy; the kinetic energy of a meteor is changed by friction in the earth's atmosphere to heat energy and light energy; in the growth of a plant radiant energy from sunlight is

changed into the chemical energy of substances produced in leaves and fruit. Energy transformations go on constantly, all about us.

If energy changes are followed backward into the past, it becomes apparent that most of the energy available to us on the earth today has come ultimately from a single source—the sun. Light and heat reach us directly from the sun; food and wood owe their chemical energy to sunlight falling on plants; waterpower exists because the sun's heat evaporates water constantly from the oceans. Coal and petroleum were formed from plants and animals which lived and stored energy from sunlight millions of years ago.

Modern civilization owes its spectacular development in large measure to the discovery of new sources of energy and to the development of new

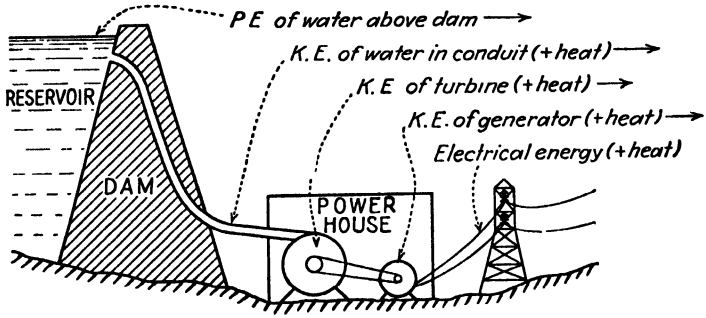


FIG. 66. *Energy transformations in a hydroelectric plant.*

methods for storing and transforming energy. Within less than 200 years man has learned to use the chemical energy of coal and petroleum, to change heat into useful mechanical energy, to store chemical energy in explosives, to get electrical energy from moving water, to use electrical energy for heating, lighting, mechanical work, communication. In the development of atomic bombs a new energy source has been tapped—the energy stored in the interior of atoms. Still other possible sources, as yet unused by industry, are the energy of tides and radiant energy direct from the sun.

Heat

Heat as a form of energy deserves special attention. For heat is invariably one of the products of an energy transformation, appearing whether we wish it to or no. In some transformations heat is almost the only product—as when electric energy is used in a heater, when wood burns in an open fire, when sunlight is absorbed at the earth's surface. More often heat appears as a minor product accompanying another form of energy, due for instance to friction in mechanical processes, to resistances in electric circuits, to chemical reactions in dynamite explosions. Sometimes the amount of heat produced is very small, but its presence cannot be

altogether avoided. We shall find later a good reason for this ubiquitous character of heat, but let us for the present try merely to gain some acquaintance with heat as a form of energy.

Heat and the related term *temperature* are often confused. We say commonly, and usually correctly, that an object possesses more heat the higher its temperature. But we cannot say with any pretense at correctness that one object possesses more heat than another because its temperature is higher. The filament of an electric light bulb, for instance, has a considerably higher temperature than a piece of burning wood; yet we should hardly choose to warm ourselves on a cold day by a light bulb in preference to a wood fire. A cup of boiling water is hotter than a pailful of lukewarm water, but the lukewarm water would melt a much larger quantity of ice. The cupful of boiling water evidently contains a smaller amount of heat despite its higher temperature.

Temperature, like force, is a fundamental quantity of which we get direct information from our sense organs, and it is equally difficult to define. I touch an object, and know at once its approximate temperature, but just how I get this knowledge is difficult to frame in words. I know that a hot and a cold object when placed together will presently reach the same temperature, and I might say therefore that one object has a higher temperature than another when heat will flow from the first to the second. But this circumlocution scarcely gives a definition of temperature itself. For the moment we shall accept our instinctive notion of temperature, and attempt a definition later.

Temperature is measured with thermometers, the familiar type of thermometer consisting of a liquid (mercury or alcohol) sealed in a slender glass tube. The liquid expands on heating, contracts on cooling, so that the position of its top in the tube gives an accurate measure of the surrounding temperature. Unhappily there are two temperature scales in common use, one devised by Fahrenheit in 1724 and named after him, the other suggested by Celsius in 1742 and called the centigrade scale. Fahrenheit chose for the zero point of his scale the lowest temperature he could obtain with an ice-salt mixture, and arbitrarily picked 96° as the temperature of the human body (his measurement was inaccurate; today we regard 98.6° as normal body temperature). Such temperatures are difficult to reproduce accurately, so that today we mark off the Fahrenheit scale by calling 32° the freezing point of water and 212° its boiling point. More sensibly, Celsius chose these easily determined temperatures of freezing and boiling water as the 0 and 100° marks on his scale. The Fahrenheit scale is an awkward makeshift, and remains in use only because England and America cling to it with the same obstinacy that preserves the British system of weights and measures. Most other civilized countries, and scientists throughout the world, use the more convenient centigrade scale. A comparison of the two is shown in Fig. 67.

The *heat* possessed by an object depends not only on its temperature, but on the amount and kind of material in it. A cup of water at 100°C has less heat than a pailful at 30°C simply because less water is present. To heat the cupful on a stove from 0 to 100° would require less gas than to heat the pailful from 0 to 30°. To include quantity of material as well as temperature in the idea of heat, we define the unit of heat, the *calorie*, as *the amount of heat required to raise the temperature of one gram of water one degree centigrade.* Raising the temperature of 1 g. of water from 20 to 25°C would require 5 calories (cal); raising the temperature of 5 g. through the same range would require 25 cal, and so on.

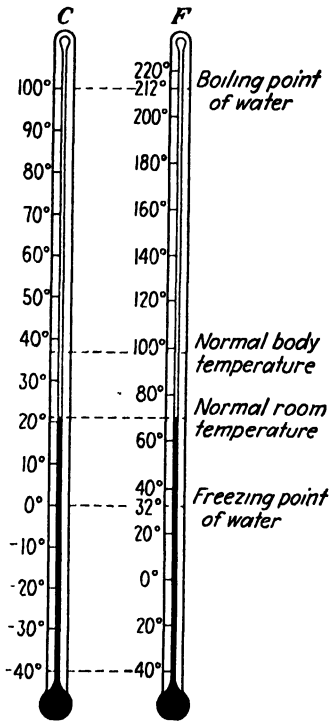


FIG. 67. *Comparison of centigrade and Fahrenheit temperatures.*

It would take us too far afield to pursue the measurement of heat further, to consider the heating of other materials besides water, to study the freezing and boiling of liquids. We shall examine such questions in later chapters, but the important idea at the moment is that heat, like other forms of energy, can be measured easily and accurately. The calorie is an energy unit, akin to the erg, the joule, the foot-pound, and the gram-centimeter.

For only about a hundred years has it been possible to state so confidently that heat is a form of energy. At the end of the eighteenth century scientists generally regarded heat as a substance, called "caloric," which an object absorbed when its temperature rose, and which escaped into the atmosphere when the temperature fell. Since the weight of an object is not affected by heating, caloric was regarded as a weightless substance, or an "imponderable"—one of several imponderables which figured prominently in the science of that time. The idea of heat as a substance worked satisfactorily for materials heated over a flame, but did not furnish any very obvious explanation for the heat generated by friction. One of the first to appreciate this difficulty was an adventurous American, born plain Benjamin Thompson in 1763, who fled this country during the Revolution and was made Count Rumford during a spectacular career in Europe. One of Rumford's many occupations was supervising the boring of cannon for a German prince, and he was struck by the large amounts of heat evolved by friction in the boring process. He showed that the heat could be used to

boil water, and that heat could be produced again and again from the same piece of metal. If heat was a substance, perhaps friction could drive it out of a metal; but that this same substance could be driven out again and again, in seemingly inexhaustible quantities, was too great a stretch for Rumford's imagination. He concluded that heat was "motion" (of an unspecified kind) rather than a substance.

The matter was finally settled by the classic experiment (1843) of an English brewer named Joule, whose name is given to one of our units of energy. By arranging a small paddle wheel to be turned in a measured quantity of water, Joule was able to measure not only the amount of heat supplied to the water by friction, but the amount of mechanical energy expended in producing the friction. He found that a given amount of mechanical energy always produced the same amount of heat. Not only was heat intimately associated with motion, as Rumford found; the *amount of motion* against friction precisely determined the *amount of heat* produced. This was a clear demonstration that heat was energy and not matter.

The amount of mechanical energy which produces one calorie of heat, according to modern experiments similar to Joule's, is 4.18 joules (4.18×10^7 ergs, or 42,600 g.-cm). This figure is often called the *mechanical equivalent of heat.*

Conservation of Energy

Joule's work led to one of the basic laws of physical science. His experiments showed clearly that in the transformation of mechanical energy into heat no energy is lost; nor is any new energy created. One kind of energy is simply converted into another, and every bit of mechanical energy expended reappears as heat energy. In further experiments on other changes of mechanical energy into heat, and in conversions of electrical, magnetic, and radiant energy into heat, always the amount of heat produced was shown to be precisely equal to the amount of some other kind of energy used. In other types of energy transformation, from potential to kinetic for instance, some energy at first glance does seem to disappear: thus the hammer of a pile driver never has quite as much kinetic energy when it strikes the pile as it had of potential energy before it fell, because some energy is lost in friction. But if the heat energy produced by the friction is added in with the kinetic energy, the sum of the two is precisely equal to the original potential energy. For all ordinary transformations of energy which have been studied in detail a similar statement holds: if all the different sorts of energy which go into a transformation are added together, and if all the energy produced is accounted for, the two sums are precisely equal. In other words, so far as we know,* *energy cannot*

* We shall find later that energy changes in the nuclei of atoms require some modification of this statement.

be created or destroyed. This sweeping generalization is the law of conservation of energy.

Just how far-reaching this law is can perhaps be appreciated only by a physicist, who has applied it to a great variety of situations and seen it time and again bring order into tangled mazes of observational data. It is probably the broadest exact principle in all science, applying with equal force to remote stars and to the intricate biological processes in living cells. In the practical world, it enables technicians to calculate accurately the amount of energy obtainable from a machine, and it has shown clearly the fallacy in the age-old dream of perpetual motion.

Questions

1. What sources of energy were used by men of the Stone Age? What sources were used by men in colonial America?
2. Can you suggest any kinds of energy found on the earth which do not have their ultimate source in the sun's radiation?
3. Is the following statement true: all changes in the physical world involve energy transformations of some sort? Why or why not?
4. Describe the energy transformations which occur
   a. When a stick of dynamite explodes.
   b. When a pendulum swings back and forth.
   c. When an electric light is turned on.
   d. When a slice of bread is toasted in an electric toaster.
   e. When a man climbs a hill.
5. One car descends a mountain by a short, steep road, the other by a longer and more gradual incline. Which, if either, loses more potential energy in the descent?
6. Which has the greater kinetic energy in its motion around the sun, the earth or Mars? The earth or Jupiter? (see Table I, page 30).
7. How much work is done in lifting a book weighing 1 kg to the top of a table 1 m. high? (Work, like potential energy, can be most easily expressed in such units as gram-centimeters and foot-pounds, which are combinations of force units and distance units.)
8. How much potential energy does the book of Question 7 possess when it is lying on the table? If it is dropped from the table to the floor, how much kinetic energy would it have just as it struck the floor? What becomes of this kinetic energy when the book comes to rest on the floor?
9. How many calories of heat are required to raise the temperature of 10 g. of water from the freezing point to the boiling point? How much mechanical energy would be necessary to produce this much heat? Express the mechanical energy in (*a*) ergs, (*b*) joules, (*c*) gram-centimeters.
10. How much energy does a man weighing 80 kg use in climbing a stairway 9 m. high? To how many calories of heat does this energy correspond?
11. In what part of its orbit (see Fig. 65) is the earth's potential energy greatest (with respect to the sun)? In what part of its orbit is its velocity greatest? Kepler reached an answer to this second question from observations which had nothing to do with energy. In which one of his three laws (page 13) is this answer expressed?
12. How high must a waterfall be in order that the energy of the falling water, when converted into heat on striking the bottom of the fall, should raise the temperature of the water by 1°C?

CHAPTER IX

Solids, Liquids, and Gases

WE LEAVE energy now, to focus attention on something more tangible: the materials of which our planet is constructed. But we shall find that energy cannot long be neglected. The characteristics of matter depend in large measure on the energy which it contains and on the effects which other sources of energy exert upon it.

Three Forms of Matter

The distinction between solid, liquid, and gaseous matter we all learn at a tender age. A moment's reflection will enable us to frame the distinction in the precise language which science requires.

A lump of sugar placed in an empty teacup is quite unaffected by its new surroundings; its shape and size have not suffered the slightest change in its transfer from the sugar bowl. Tea poured in the cup, however, quickly adjusts its shape to fit the curving bottom. No change in the volume of tea occurs during the transfer from teapot to cup, but the shape of the tea changes continuously. The gaseous substance responsible for the odor of the tea is not confined by the cup, but spreads quickly to every corner of the room; the "shape" of the gas is determined solely by the shape of the room, and its size by the size of the room. The *solid* sugar maintains its size and shape, in whatever container it is placed; the *liquid* tea has a definite size or volume, but its shape is determined by the shape of its container; the fragrant *gas* maintains neither size nor shape, but expands to fill every crevice of its container.

Or the distinction may be put differently: A solid differs from a liquid or gas in its lack of the ability to *flow*, which means that a solid can offer resistance to a deforming force which would readily alter the shape of a liquid or gas. Because of their property of flowing, liquids and gases are collectively called *fluids*, the two kinds of fluid being distinguished by the ability of liquids to maintain a fixed volume.

The dividing line between solids and liquids is not perfectly clean cut. Pitch, for instance, is hard, brittle, to all appearance quite capable of

maintaining its shape. Yet if allowed to stand long enough, a piece of pitch will flatten and spread out in the manner of very viscous liquids.

Gases

Gases have become familiar articles of commerce: the light gases hydrogen and helium, the heavy gas carbon dioxide, the evil-smelling, yellow-green, poisonous gas chlorine. Surrounding us always, necessary to our existence, is the mixture of colorless and odorless gases which we call air. The odors of flowers, foods, perfumes are all due to gaseous materials. Yet the ways of gases remain mysterious to the layman, because without special technique he cannot isolate and handle them as he can liquids and solids. He often finds it startling even to learn that gases have weight (the air in a good-sized room, for instance, weighs close to a hundred pounds), since they spread so easily outward and upward in apparent defiance of gravity.

To the scientist, on the other hand, gases are by far the best understood of the three forms of matter. He knows that all gases have certain important characteristics which are much alike, and he has learned simple laws which describe their properties. We shall presently discuss some of these important laws, but for the moment let us simply note a few of the more obvious facts about gas behavior.

The defining property of gases is their ability to expand indefinitely. Coupled with this is their extreme *compressibility*, even a small increase in pressure causing a marked reduction in the volume of a gas. A gas will expand either into a vacuum or into another gas, the second gas serving merely to slow down the rate of expansion. The spreading, or *diffusion*, of one gas into another may be well shown by opening a bottle of ammonia in one corner of a room: the irritating odor of ammonia gas spreads quickly through the air into all parts of the room. Nor is there any limit to the amount of one gas which can diffuse into another: gases are *miscible* ("mixable") with each other in all proportions. Finally, gases, unless highly compressed, are characterized by extreme *lightness* in comparison with liquids and solids.

Liquids

Water is the only common naturally occurring liquid. Petroleum, molten lava, and liquids produced by plants and animals occur in small amounts. Artificially a great variety of liquids has been prepared, but no other liquid is remotely comparable to water in abundance or in multiplicity of uses.

In contrast with gases, liquids are all but *incompressible*, the highest pressures obtainable with modern laboratory equipment serving to squeeze water into about three-fourths of its original volume. Some

liquids, water and alcohol for instance, are *miscible* with each other in all proportions, while others, like water and oil, are immiscible. Two immiscible liquids shaken together give a cloudy mixture with tiny globules of one liquid suspended in the other; when the shaking stops the globules coalesce and the liquids separate into distinct layers. If two miscible liquids are placed in contact without stirring, one will *diffuse* into the other, but the process is very much slower than diffusion in gases. If a heavy gas (like carbon dioxide) is placed in the bottom of a container and a light gas (hydrogen) in the top, after a few hours the two will be completely mixed; but if the lighter of two miscible liquids (*e.g.*, alcohol) is poured carefully on top of the heavier (*e.g.*, water), months must pass before diffusion can make the mixing complete.

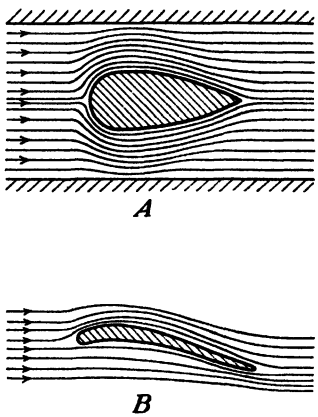


FIG. 68. *Flow lines in fluid motion around obstacles. A, flow lines around a stream-lined object; B, motion of air past the wing of an airplane.*

Like gases, liquids flow readily, but the rate of flow is much smaller. A thick liquid like honey, particularly when full of tiny bubbles or solid specks, shows well the details of liquid flow. As honey creeps slowly down an inclined surface, the bubbles gather in lines along the direction of flow, the distinctness of the lines showing roughly the amount of liquid movement in their vicinity. Lines are best formed at the top of the flow, scarcely detectable at its bottom. If the motion of the honey is confined to channels of different shapes, the bubble lines form graceful curves around obstacles and irregularities in the channel walls (Fig. 68). Always the most distinct lines are at the surface of the liquid far from the confining walls. These observations suggest that the liquid moves by the sliding of one layer over another, layers near the bottom and sides of the channel clinging to the solid surfaces and scarcely moving at all (Fig. 69). How fast the liquid flows depends on how freely layers within the liquid can slide past each other. In honey there is evidently considerable resistance to the sliding motion, in water or alcohol much less resistance. This inner resistance of liquids to flowing movement is called *viscosity*. Honey is a viscous liquid; water and alcohol are less viscous. Gases show some slight resistance to fluid motion, but their viscosities are extremely low compared with liquid viscosities.

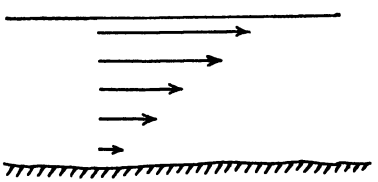


FIG. 69. *Relative motion at different levels in a slowly moving liquid.*

In rapidly flowing liquids and gases the mechanism of flow is not as simple as that just described. Flow lines are not straight or in smooth curves, but follow complex, ever changing loops and eddies. Turbulent flow of this sort is well shown in a swift mountain brook.

Some curious properties of liquids are associated with their free surfaces. Steel, for instance, is heavier than water, but a steel needle can be made to float if lowered very carefully to a water surface. The needle rests in a slight depression of the surface, the water showing no tendency to "wet" the metal (Fig. 70). Disturb the needle so that any part cuts through the surface, and it sinks immediately to the bottom. Touch the

FIG. 70. *Enlarged cross section of needle on water surface.*

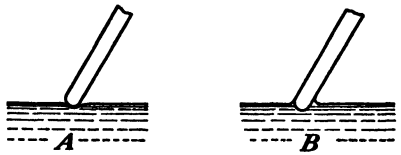


FIG. 71. *Glass rod (A) before breaking water surface and (B) after breaking surface.*

water surface very carefully with the end of a glass rod: at first contact the glass, like the needle, simply depresses the surface without being "wet" by the liquid. Lower the rod ever so slightly; of a sudden the surface breaks, and a little ridge of water climbs up around the rod (Fig. 71). Lift the rod out; its end is now "wet," covered with a thin film of water. The water surface, after a momentary disturbance, becomes smooth and flat as before. In experiments like these the surface layer of water seems to behave differently from the rest of the liquid; water will "wet" both glass and steel, but the surface layer must apparently be broken before liquid and solid can make intimate contact. It is almost as if the water were covered with a thin, stretched sheet of rubber, which must be ruptured before the water can make contact with metal or glass. Unlike rubber, the surface layer has the capacity for repairing itself: no matter how many times the surface is broken, it shows the same behavior to each new disturbance. The apparent stretched, elastic character of liquid surfaces is shown also by the spherical shape of raindrops and by the spherical shape of globules formed when immiscible liquids are shaken together. Here each liquid fragment has a free surface on all sides, and the surface contracts to give the drop a shape with the least possible surface area—the sphere. The amount of this force tending to decrease the area of a liquid surface is a property of the liquid called its *surface tension.*

Solids

Solids surround us in bewildering variety. To such naturally occurring substances as rocks, minerals, products of plant and animal life, man has

added an endless list of artificial materials. Jelly, iron, glass, feathers, and salt are all solids, since all will retain their shapes indefinitely regardless of the shapes of their containers.

Solids are practically incapable of *diffusion*, although experiments are reported in which two metals show very slight mixing after long contact. Like liquids, solids are nearly *incompressible* for all ordinary pressures. When a solid is bent, stretched, or squeezed, it cannot flow as would a fluid. If the deforming force is small, the solid changes shape only slightly and springs back to its original form when the force is removed. This behavior we describe with the statement that solids are *elastic*, some of course being much more elastic than others. Subjected to a greater force (how much greater depends on the solid), a solid may be permanently deformed. Thus an iron wire, if bent slightly, springs back to its original shape but, if subjected to greater force, is permanently bent. Brittle solids, like glass, can undergo scarcely any permanent deformation without breaking; other solids, like some metals, can be deformed almost indefinitely—gold, for instance, can be hammered into marvelously thin, translucent foil, and platinum can be drawn into wires so fine that they are scarcely visible.

Some solids occur in the beautiful geometric shapes called *crystals*. The formation of a crystalline solid is an awe-inspiring sight, if one but takes the time to watch it carefully. When water freezes to ice, for example, slender ice needles radiate out over the liquid surface, branching and interlocking, each new-formed bit of solid seeming to know by some uncanny instinct just where it should appear to preserve the shapes and pattern of the interlaced needles. One can hardly miss the conclusion that the shape of an ice needle, or the shape of any other growing crystal, reflects something about the inner structure of the solid, and that perhaps the secret of this inner structure might be revealed by a painstaking study of the crystal.

Effects of Pressure

We have already noted the great compressibility of gases, the near incompressibility of liquids and solids. Let us examine a little more closely the influence of pressure on various materials.

Definition of Pressure. First of all, what is pressure? What do we mean, for example, by the statement that air pressure in an automobile tire is 30 pounds per square inch (lb/sq in.)? Consider a piston-and-cylinder arrangement like that of Fig. 72, with an ideal sort of piston at once gastight and freely movable. Suppose that the piston weighs 100 g. and has a cross-sectional area of 50 square centimeters (sq cm). If now a fluid is confined in the cylinder by the weight of the piston alone, the total force pressing downward on the fluid is 100 g. This force is exerted over an

area of 50 sq cm, so that the force on each square centimeter of fluid is $^{100}/_{50}$, or 2 g. This *force per unit area* is called the ***pressure*** acting on the fluid. In symbols

$$P = \frac{F}{A} \tag{22}$$

where P is pressure, F is force, and A is area. Pressures are commonly expressed in grams per square centimeter, pounds per square inch, etc.

Fluid Pressure. In a fluid, since gravity pulls downward on all parts of it, the bottom layers are under pressure from the weight of overlying fluid, whether external pressure is acting or not. At the bottom of the cylinder just described, for instance, the pressure must be slightly greater than 2 grams per square centimeter (g./sq cm), since the bottom layer must support not only the piston's weight but a considerable weight of fluid as well. In ordinary laboratory experiments with gases, pressures due simply to the weights of the gases involved are usually negligible, but in experiments with liquids such pressures are important.

Except for the steady downward increase due to gravity, pressures in all parts of a fluid must be the same. It would be unthinkable, for instance, to have the fluid confined in the cylinder of Fig. 72 under a pressure of 40 g./sq cm in the upper part of the cylinder and under a pressure of 2 g./sq cm in the lower part. Since the fluid is capable of flowing, it would move from the region of high pressure to the region of low pressure until the pressure was equalized. If the air pressure in a tire is 30 lb/sq in., then the air exerts an outward force of 30 lb on every square inch of the tire's inner surface. Different parts of a solid, on the other hand, may be under very different pressures, since the solid cannot flow. One end of a timber may be placed under terrific pressure in a vise, while the other end is under no pressure except that of the surrounding atmosphere.

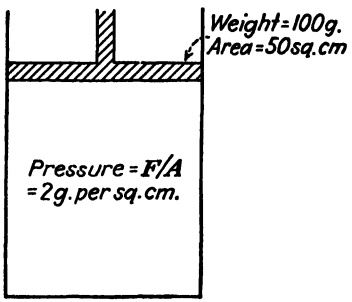


FIG. 72. *Piston and cylinder.*

Also because a fluid can flow, the pressure at any one point within it must be the same in all directions. Suppose that a piece of paper can be supported horizontally in the middle of the cylinder of Fig. 72. If the downward pressure on its upper surface is 2 g./sq cm, then the upward pressure on its lower surface must be 2 g./sq cm. If the two pressures were different, the paper would move toward the direction of lower pressure—in other words, the fluid would flow until the pressure at each point becomes the same in all directions. This characteristic of fluid pressure is well shown by the device of Fig. 73, which consists of a funnel capped with

a thin rubber sheet and connected with an instrument for measuring pressure. If the funnel is held with the center of the membrane at any one level in a liquid or gas, turning the funnel in different directions has no effect on the indicated pressure.

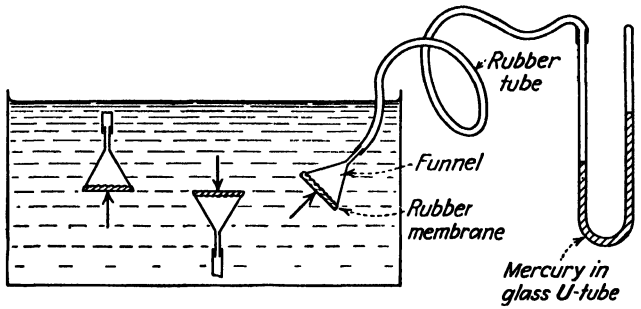


FIG. 73. *Pressure is the same in all directions at the same level in a fluid.*

Pressures in Air, Oceans, and Rocks. Since we live at the base of the atmosphere, we are under pressure from the weight of gas above us—a pressure amounting to nearly 15 lb/sq in. at sea level, sufficient to crush a container from which air has been removed unless its walls are fairly stout. We are not conscious of the 15-lb force pushing inward on every square inch of our bodies, simply because our bodies are sufficiently permeable to air so that pressures inside are maintained equal to those without. Atmospheric pressures are measured with instruments called *barometers*, of which the commonest type is a closed tube of mercury about 1 m. long inverted in a dish of mercury (Fig. 74). The upper part of the tube is evacuated, either directly with a vacuum pump, or by taking care that no air enters the tube when it is inverted in the dish. At a point *A* in the tube, the downward force is equal to the weight of the mercury column. This must be balanced by an upward force, else the mercury would flow out of the tube. The upward force is maintained by the downward push of the atmosphere on the mercury surface in the dish. Thus the pressure of 100 miles or so of atmosphere is balanced by the pressure of 30 in. of mercury. The pressure of the atmosphere at any one time is measured by the height of the mercury column, which fluctuates with varying weather conditions over a range of about 5 in.

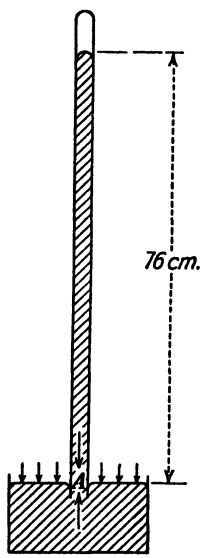


FIG. 74. *The barometer.*

Normal atmospheric pressure at sea level corresponds to a barometric height of 76 cm (about 30 in.). This pressure (nearly 15 lb/sq in.) is

frequently used as a pressure unit, called the *atmosphere.* Thus a pressure of 2 atmospheres (atm) would correspond to 152 cm of mercury, or about 30 lb/sq in. Smaller pressures, since they are often measured with mercury columns, are commonly described in terms of mercury heights: thus a pressure of "5 cm of mercury" is equal to about 0.066 atm, or roughly 1 lb/sq in.

Pressures in liquids become rapidly greater with increasing depth, as anyone knows who has swum even a few feet below the surface of a lake or pool. The stoutest submarine must be wary of venturing more than a few hundred feet down, for fear of collapsing. At a depth of 6 mi in the ocean the pressure is about 8 tons/sq in., or roughly 1,000 atm—sufficient to compress water by some 3 per cent of its volume. Fish which inhabit these depths can withstand such enormous pressures for the same reason that we can endure the pressures at the bottom of our ocean of air: pressures inside their bodies are kept constantly equal to pressures outside. When brought quickly to the surface, deep-ocean fish often explode because of their high internal pressures.

Pressures in solids cannot be measured or discussed with the same ease as pressures in fluids, since the pressure at one point in a solid may be quite different from the pressure at a neighboring point. In general, however, pressures in the earth's rock crust increase downward rapidly, just as do pressures in the ocean, because of the weight of overlying material. At depths of a few miles, pressures become greater than any we can duplicate in the laboratory, and the exact behavior of rocks at such depths must remain a matter of conjecture. From the experimental work that has been done and from geological evidence, we know that deep burial profoundly alters the ordinary properties of rocks, causing them to lose their brittleness and flow like viscous liquids. We have used this fact before (page 81) to account for the spheroidal shape of the earth.

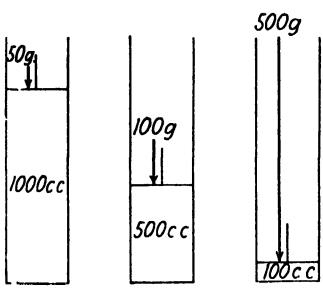


FIG. 75. *Boyle's law.*

Boyle's Law. The pressures and volumes of gases are related by a simple expression which has no counterpart for liquids and solids. Suppose that 1,000 cc (1 liter or 1 l.) of gas is confined in the cylinder of Fig. 72 under the weight of the piston, 50 g. If a 50-g. weight is added to the piston, so that the pressure on the gas is doubled, the piston will compress the gas until its volume is only 500 cc, just half the original. If the total pressure is made ten times as great, the piston will move down until the gas occupies only 100 cc (Fig. 75). *Pressure is evidently inversely proportional* (page 51) *to volume.*

$$P = \frac{K}{V} \quad (K = \text{proportionality constant}) \quad (23)$$

or alternatively

$$\frac{P_1}{P_2} = \frac{V_2}{V_1}$$

In honor of the English physicist who discovered it, this inverse proportion is called Boyle's law. It is approximately true for any gas, except at very high pressures, provided the temperature remains constant.

In contrast to this simple statement which holds for all gases, pressure-volume relationships for liquids and solids are exceedingly complex, each separate substance behaving differently.

Effects of Temperature

Nearly all properties of matter are affected in some degree by temperature changes. Viscosity and surface tension of most liquids decrease when the temperature rises, as do the elasticity, hardness, and breaking strength of most solids. Almost all substances expand on heating, gases as a rule more than liquids and solids. To allow for expansion in warm weather, concrete pavements are cut at intervals by cracks filled with tar, and steel rails are laid with their ends not quite touching. The expansion of liquids like alcohol and mercury is utilized in constructing thermometers. The rapid expansion and rise of heated air on summer days produce the atmospheric disturbances called thunderstorms. If normal expansion with temperature is hindered, considerable pressures may develop, as is shown by the buckling of poorly constructed pavements on hot summer days and by the increase in pressure in automobile tires during long drives on hot roads.

Just as gas volumes are simply related to pressure changes, so do they change uniformly with temperature. If a gas is cooled steadily, starting at 0°C, while its pressure is maintained constant, its volume decreases by $\frac{1}{273}$ of its amount at 0° for every degree the temperature falls. If the gas is heated, its volume increases by the same fraction. Thus if a gas occupies 1,000 cc at 0° under a pressure of 1 atm, it will occupy

$$1{,}000 + (\tfrac{1}{273} \times 1{,}000) \text{ or } 1{,}003.7 \text{ cc at } 1°\text{C}$$
$$1{,}000 + (\tfrac{10}{273} \times 1{,}000) \text{ or } 1{,}036.6 \text{ cc at } 10°\text{C}$$
$$1{,}000 + (\tfrac{273}{273} \times 1{,}000) \text{ or } 2{,}000 \text{ cc at } 273°\text{C}$$
$$1{,}000 - (\tfrac{100}{273} \times 1{,}000) \text{ or } 634.0 \text{ cc at } -100°\text{C}$$

—all provided that the pressure remains 1 atm, and that the gas is not near its condensation point at any of the temperatures specified (Fig. 76). If volume rather than pressure is kept fixed, the pressure increases with

rising temperature and decreases with falling temperature, again by the fraction $\frac{1}{273}$ of its 0° value for every degree change.

These figures suggest an awkward question: what would happen to a gas if we could lower its temperature to −273°C? If we should try to maintain constant volume, the pressure at this temperature would fall to zero; if the pressure remained constant, the volume should fall to zero. It is hardly probable, however, that our experiment would have so startling a consummation. In the first place we should find great difficulty in attaining so low a temperature, and in the second place all known gases liquefy before that temperature is reached. Nevertheless, this temperature, −273°C, has apparently some special significance, a significance which will become clearer in the next chapter. It is called ***absolute zero.***

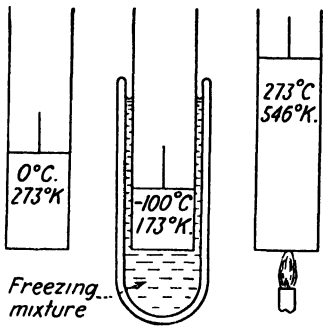


FIG. 76. *Charles' law.*

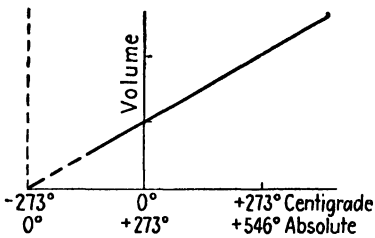


FIG. 77. *Graphic representation of Charles' law.*

For some scientific purposes it is more convenient to reckon temperatures from absolute zero than from the freezing point of water. Temperatures on such a scale, given as degrees centigrade above absolute zero, are called ***absolute temperatures.*** Thus the freezing point of water is 273° absolute, written 273°K in honor of the English physicist Lord Kelvin, and the boiling point of water is 373°K. Any centigrade temperature may be changed to degrees absolute by adding 273.

Using the absolute scale, we may express the above relationship between gas volumes and temperatures very simply: *the volume of a gas is directly proportional to its absolute temperature* (Fig. 77). This relation should be obvious from the figures on the last page. It may be expressed mathematically

$$V = KT \qquad \text{or} \qquad \frac{V_1}{V_2} = \frac{T_1}{T_2} \tag{24}$$

where the V's are volumes, the T's are absolute temperatures, and the K is a proportionality constant. Discovered by two eighteenth-century French physicists, Charles and Gay-Lussac, this relation is commonly

known as Charles's law. Like Boyle's law, it holds to a fair approximation for all gases at ordinary pressures, but becomes inaccurate at high pressures.

Changes of State

An important effect brought about by increasing and decreasing temperatures is the change from one form of matter to another. All gases can be liquefied, and all liquids frozen, if the temperature is made sufficiently low. Nearly all solids can be liquefied and vaporized, except for those like paper and coal which decompose below their melting points. Since nearly every separate substance can exist as a solid, liquid, and gas, these three are known as different *states* or forms of matter rather than different kinds, and changes from one to another are called *changes of state.*

Let us consider the temperature changes accompanying the melting of ice and the vaporization of water under ordinary conditions (normal atmospheric pressure). If we start with a dish of chipped ice below its melting point, the first effect of supplying heat is to raise the temperature. The rise continues steadily up to 0°C, then stops abruptly as the ice begins to melt. During the melting the temperature remains at 0°, no matter how much heat we supply, provided the ice-water mixture is kept well stirred. If heating continues after the last bit of ice disappears, the temperature again starts upward and rises continuously to 100°. During this rise we find vapor given off in increasing quantities. Water vapor itself, or steam, is a colorless gas and therefore completely invisible, but above the heated dish it cools and condenses into the white clouds of tiny liquid droplets popularly called "steam." If we had had the proper means of detecting it, we should have found that vapor was given off during the entire heating process, from the cold ice as well as the heated water; rising temperature simply increases the rate of vaporization. At 100° the liberation of vapor becomes so rapid that it no longer takes place from the surface alone: bubbles of vapor form all through the liquid, and we say that the water is boiling. At the boiling point as at the melting point the temperature remains constant. We may use as hot a flame as we like, or as many burners—the thermometer will remain stubbornly at 100° until the water disappears. Thereafter, if we arrange to heat the gaseous water in a closed container, its temperature will rise indefinitely as more heat is supplied. This sequence of changes is shown graphically in Fig. 78.

When steam (under 1 atm pressure) is cooled to 100°, it condenses to liquid water, and the liquid freezes to ice at 0°; that is, the sequence of changes shown in Fig. 78 is exactly reversed by decreasing temperatures. The condensation point of steam and the freezing point of water are identical, respectively, with the boiling point of water and the melting

point of ice. Because these temperatures are the same from whichever side approached, and because they remain constant while heat is added or removed, they are eminently suitable for fixing the points on a thermometer scale—as Celsius discovered 200 years ago (page 109). The only factors which influence the temperatures of freezing and boiling are the surrounding pressures and the purity of the water. The boiling point is particularly susceptible to pressure changes: the decrease in atmospheric pressure in going from sea level to a high mountain, for instance, lowers the boiling point so markedly that eggs and vegetables to be thoroughly cooked must be boiled considerably longer. The advantage of a pressure cooker lies in the fact that increased pressure raises the boiling point, hence decreases the time necessary for cooking.

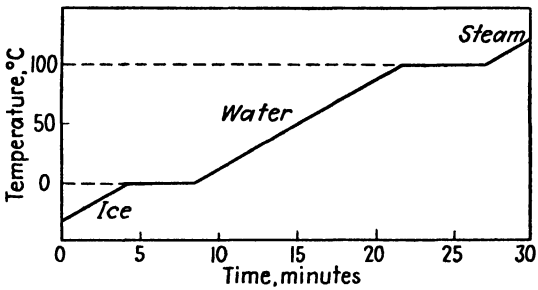


FIG. 78. *Changes in temperature when heat is supplied to a piece of ice below its melting point.*

Note that in the melting of ice and the vaporization of water the material *absorbs heat without becoming hotter*. Just to melt 1 g. of ice, with the temperature constantly at 0°, requires the addition of 80 cal of heat—enough to raise the temperature of 1 g. of liquid water from room temperature to the boiling point. To vaporize 1 g. of water at its boiling point, without any change in temperature, requires the addition of considerably more heat—nearly 540 cal. These two figures are called, respectively, the *latent heat of fusion* (*melting*) and the *latent heat of vaporization* of water.

From the law of conservation of energy, it follows that the condensation of 1 g. of steam must liberate 540 cal, and that the freezing of 1 g. of water must liberate 80 cal. If, for instance, freezing liberated a smaller quantity, say 70 cal, then every time 1 g. of ice was melted and refrozen, 10 cal of energy would disappear—which contradicts the conservation law. That steam does liberate heat when it condenses is attested by the severe burns which live steam can produce. The liberation of heat when lakes freeze in winter is sufficient to keep their shores at a somewhat higher temperature than that of surrounding regions.

Changes of state in other substances are similar to those we have just discussed for water. Most substances have definite melting points and

definite boiling points and have characteristic latent heats of fusion and vaporization. A few materials, like glass, have no sharp melting point, but gradually soften when heated. Some solids cannot be melted at ordinary pressures, but pass directly into the gaseous state; iodine, dry ice (solid carbon dioxide), camphor are familiar examples. These substances may be liquefied if the pressure is sufficiently increased. The direct formation of vapor from a solid, called *sublimation*, is not a peculiar property of these few substances—a light snowfall, for example, gradually disappears by sublimation even on a very cold day—but for most solids the amount of vaporization is so slight that we do not ordinarily observe it.

In this long but far from exhaustive survey of matter and its properties we have concerned ourselves chiefly with *description*, attempting to gain an acquaintance on more intimate terms with the characteristics of familiar substances. We find that gases, though elusive and difficult to handle, behave more uniformly and according to simpler rules than liquids and solids. We find peculiar characteristics in liquid surfaces, as if the material in them were different from the body of the liquid. Solids ordinarily cannot flow like fluids, but we find that they behave like very viscous liquids if pressures are high. In many solids we find the regular geometric shapes of crystals. We find that the changes from solid to liquid and from liquid to gas involve absorption and liberation of heat energy without accompanying temperature changes.

To the scientist, a collection of facts like this demands explanation. Why are gases so strikingly uniform in behavior, why are liquid surfaces so peculiar, why do solids form crystals? He seeks an answer in a theory about the intimate structure of matter, a theory which with a few simple assumptions will account for the differences between solids, liquids, and gases, and which will lead him on to the discovery of new and unsuspected properties. We shall discuss in the next chapter a theory which accomplishes these ends.

Questions

1. Describe an experiment to show that gases have weight.
2. Name three gases of commercial importance besides those mentioned on page 114.
3. Name two pairs of miscible liquids and two pairs of immiscible liquids.
4. Lead shot is often prepared by allowing drops of molten lead to fall some distance through the air into water. Why does this ensure that the shot will have an approximately spherical shape?
5. What is the pressure on a gas confined in a cylinder 20 cm in diameter, if the piston weighs 500 g.?
6. The accompanying diagram shows a liquid compressed by two pistons, one of much larger cross section than the other. Suppose that the small one has a surface area of 20 sq cm and the large one an area of 200 sq cm. Suppose that a 1,000-g. weight is placed on the small piston. How much is the pressure which it exerts

increased? How much does this increase the pressure which the liquid exerts on the large piston? How much does this increase the *total force* which the liquid exerts on the large piston? (Note that a small force on the small piston can produce a relatively great force on the large one. This is the principle of the hydraulic lift.)

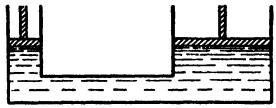


FIG. 79.

7. What is the pressure at the base of a column of water 10 cm high and 1 sq cm in cross section? What would the pressure be if the column had a cross section of 5 sq cm but the same height? What would the pressure be at a depth of 1 m. in a pool? Can you derive a relation connecting pressure with depth in water? Would the same relation hold for other liquids?
8. Mercury is 13.6 times as heavy as water. If water were used in a barometer instead of mercury, how high would the column be?
9. A certain amount of hydrogen occupies 500 cc at 0°C and 1 atm pressure. What will its volume be if the temperature is kept constant while the pressure is increased to 5 atm? If the pressure remains constant (at 1 atm) while the temperature is increased to 273°C?
10. Derive a relationship between the pressure and temperature of a gas when the volume is kept constant.
11. Why is a piece of ice (at 0°C) more effective in cooling a drink than the same weight of cold water (at 0°C) would be?
12. How much heat would be required to change 50 g. of ice at 0°C into water at 20°C?

CHAPTER X

The Kinetic Theory

SUPPOSE there were no limit to the power of our microscopes, so that we could examine a drop of water under stronger and stronger lenses indefinitely. What sort of a microscopic world would we discover when the drop was enlarged, say, a million times? Would we still see structureless, transparent, liquid water? Or would we perhaps see distinct particles, the building blocks, as it were, of the substance which to our gross senses is structureless, transparent, and liquid? These are questions as old as civilization, posed whenever men have speculated deeply on the nature of things. Our first record of an intelligent approach to the problem, as to so many other problems in science, dates from the fifth century B.C. in Greece. Of the Greeks we remember particularly Democritus, since he championed our modern view that matter consists of individual particles. Four centuries afterward, in the Rome of Caesar and Cicero, Democritus's views were elaborated by the great scholar-poet Lucretius.

Lucretius and Democritus used their hypothesis that matter is made up of tiny particles chiefly for philosophical speculation and tried only superficially to connect it with actual observations. Physicists of the nineteenth century, less interested in philosophy than in factual knowledge, found in the 2,000-year-old idea a powerful tool for explaining and correlating a great variety of observations and simple experiments like those described in the last chapter. This extension of the idea requires several assumptions regarding the nature of the tiny particles, the assumptions being known collectively as the *kinetic theory*. Perhaps because it has been so long accepted, or perhaps because it deals so largely with homely, familiar phenomena like the expansion of gases, pressure in fluids, the freezing and boiling of water, the theory has never appealed to the imaginations of radio listeners and Sunday supplement readers as have the more melodramatic theory of relativity, electronic theory, and quantum theory, but it deserves a place with them as one of physical science's most fruitful generalizations.

Molecules

Like any other scientific theory, the kinetic theory is a group of assumptions. The assumptions were made originally as guesses, intelligent guesses of course, describing the peculiarities of the particles supposed to make up gases, liquids, and solids. Physicists then tried to show that the familiar characteristics of matter follow from an application of ordinary physical laws to particles with these assumed peculiarities. This attempt was highly successful: the behavior of matter could be described accurately, or "explained," in terms of the assumptions. Furthermore, use of the assumptions enabled scientists to predict unsuspected quirks in the behavior of matter, and these predictions could be checked by experiment. As the original guesses proved themselves capable of explaining more and more known experiments and of predicting the results of new ones, they were accepted ever more widely as probable facts rather than as mere assumptions. In what follows, we shall state the various "guesses" of the kinetic theory as assumptions, and then see how their agreement with observed facts justifies their acceptance as true statements.

The basic assumptions of the kinetic theory are simply that (1) *matter is composed of tiny discrete particles called molecules* and (2) *these molecules are in constant motion.* Rather than set down here a complete list of the further assumptions which make up the theory, we shall discuss them piecemeal, showing how the specific assumptions for gas molecules make possible an explanation of the behavior of gases, the specific assumptions for liquid molecules an explanation of the properties of liquids, and so on. Before going on to these specific assumptions, let us get a little better acquainted with molecules in general.

A ***molecule*** may be defined as *the smallest portion of a substance which retains the essential characteristics of the substance.* Like most definitions in science, this statement is not rigorous without several paragraphs of explanation and qualification, but it will serve for the present.

Today we have much information about the actual sizes, speeds, even shapes of the molecules in various kinds of matter, which of course was not available when the kinetic theory was formulated. Our information is indirectly obtained, since even today we cannot directly measure the dimensions of a molecule; but it is confirmed in so many different ways that we have every reason to believe it accurate. For example, a molecule of nitrogen, the chief constituent of air, has a diameter of about 18 billionths of a centimeter (1.8×10^{-8} cm) and weighs 4.8×10^{-24} g. (48 preceded by 23 zeros and a decimal point). It travels (at 0°C) with an average speed of 50,000 cm/sec (1,500 ft/sec, or about the speed of a rifle bullet), and in each second collides with over 5 billion other molecules. Of similar dimensions and moving with similar speeds in each cubic centi-

meter of air are some 27 quintillion (27×10^{18}) other molecules. To visualize what this fantastic number means, suppose that you and 100 of your friends had started to count the molecules in a cubic centimeter of air (about a thimbleful) at the time when the earth was ejected from the sun: if each of you had counted three a second faithfully for ten hours every day, you would now be about half finished with your job.

The extreme smallness of ordinary molecules will probably make it impossible for us ever actually to see them, because the resolving power of our microscopes is limited by the nature of light* (page 258). But particles not many times larger than molecules are visible in the highest powered microscopes, and these particles are small enough to move in response to the blows of swiftly traveling molecules against their sides. Molecular motions are highly erratic, so that the visible particles are buffeted about in irregular, zigzag paths (Fig. 80), like fat men caught in a crowd of Christmas shoppers. The smallest visible particles move rapidly under the molecular bombardment, darting this way and that, now brought to a stop, now starting out in a new direction. Larger particles show only a slight jiggling motion, or do not move at all, since so many molecules strike them that at any one instant the forces on all sides are nearly the same. This motion of microscopic particles, called the *Brownian movement*, is well shown by particles of certain dyes and by tiny oil droplets suspended in water, and by smoke particles suspended in air. It is the most direct and convincing evidence we have of the reality of molecules and their motions.

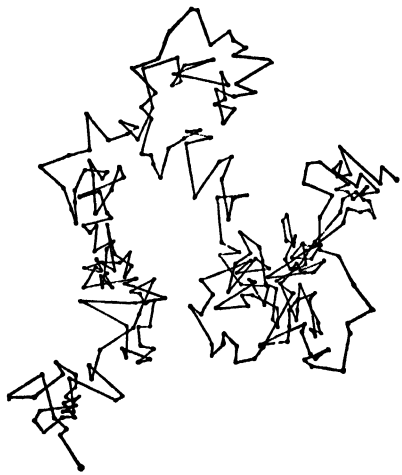


FIG. 80. *The zigzag path of a particle bombarded by molecules. The line connects the positions of a single particle observed at 10-sec intervals.*

The Kinetic Theory of Gases

We go on now to discuss the specific assumptions of the kinetic theory regarding gas molecules. We begin with gases, because the uniformity in

* The giant molecules of certain proteins have recently been made visible with the electron microscope. This instrument, which uses a stream of electrons rather than visible light, has a resolving power many times greater than that of an ordinary microscope. Even the electron microscope, however, has little chance of making visible the tiny molecules of ordinary substances.

To leave the earth an air molecule would have to reach a speed of 7 mi/sec in the cold upper atmosphere. Attaining this speed at lower elevations would, of course, do no good, for collisions would immediately reduce it. The number of molecules with speeds so far above the average in the upper atmosphere is practically zero, so that our planet is in no danger of losing its valuable air blanket. Even molecules of hydrogen, lightest and fastest of all gas molecules, probably cannot now escape the earth's attraction, although if the earth was hotter in the remote past they may have then escaped in large numbers.

Thus a planet's ability to hold gas molecules depends primarily on its gravitational pull, to a small extent on its temperature. Mercury, small and hot, is as barren as the moon. Venus, nearly the size of the earth, has an atmosphere as dense as ours. Mars is heavy enough to hold a thin atmosphere, but the swift molecules of hydrogen, helium, and neon must have escaped from it long since. The four giant planets have enormously thick atmospheres containing abundant hydrogen.

Liquids and Solids

The kinetic theory grew from the efforts of several nineteenth-century physicists to find an explanation for the behavior of gases. The greatest triumphs of the theory have come from work on gases, but in recent years it has been applied with much success to the vastly more complicated inner structures of liquids and solids.

Assumptions. The incompressibility of liquids and solids suggests the assumption that *their molecules are close together, probably in actual contact.* The Brownian movement in liquids is ample basis for assuming that *liquid molecules, like gas molecules, are in constant, random motion.* The assumption that *particles in solids are held more or less rigidly in fixed positions* follows from the definite shapes which solids maintain and from their inability to diffuse.

Strong attractive forces are assumed to be holding together the particles of liquids and solids. These intermolecular attractions are the same as the very slight ones between gas molecules, which are responsible for deviations from Boyle's law at high pressures. In liquids and solids the forces are simply more conspicuous because the molecules are so much closer together. The nature of the attractive forces is not completely understood, but they are in large part of electrical origin. So great are the forces between adjacent molecules in many solids and liquids that the molecules lose their identity as individual particles. Some of these substances are made up of electrically charged molecular fragments, others of larger particles formed by the joining together of molecules in twos and threes. Because solids and liquids are thus not all constructed of bona fide

with top and bottom, hence will strike only half as often; the horizontally moving molecules must spread their blows over twice as great an area, hence on each square centimeter the number of blows will be cut in half. Thus the pressure in all parts of the cylinder is exactly halved, as Boyle's law would predict. To extend this reasoning to a real gas, with molecules moving in all directions, requires a plunge into mathematics which would take us too far afield.

The third assumption mentioned above is not strictly accurate: gas molecules do exert slight attractions on one another. These attractions become conspicuous when a gas is greatly compressed, and account (in part) for the fact that Boyle's law does not hold at high pressures.

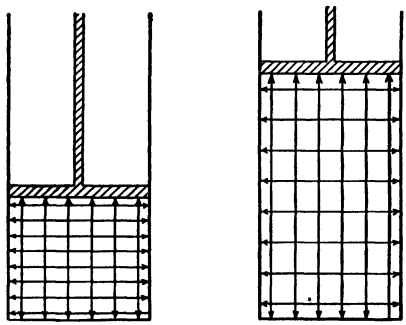


FIG. 81. *Boyle's law.* (*Molecular motions assumed vertical and horizontal only.*)

Temperature. To explain the effects of temperature requires one further assumption: (4) *The absolute temperature of a gas is directly proportional to the average kinetic energy of its molecules.* Just why the assumption is stated in this particular manner is not obvious without a mathematical discussion. But the fact that temperature should be closely related to molecular speeds, and hence to molecular energies, follows from the simple observation that the pressure of a confined gas increases as its temperature rises (p. 121): increase in pressure must mean that the molecules are striking their confining walls more forcefully and so must be moving faster.

In the last chapter we learned that the pressure of a gas approaches zero as its temperature falls toward −273°C. For pressure to become zero, molecular bombardment must cease. Thus absolute zero finds a logical explanation in terms of the kinetic theory, as the temperature at which gas molecules would lose their kinetic energies completely. There can be no lower temperature simply because there can be no smaller amount of energy than none at all. The regular increase of gas pressure with absolute temperature if the volume is kept constant, and the increase of volume if the pressure is constant (Charles's law), are understandable from this definition of absolute zero, although the precise mathematical demonstration of direct proportionality is a bit complicated.

As a prediction from our temperature assumption, we might guess that compressing a gas in a cylinder should cause its temperature to rise. For while the piston is moving down, molecules rebound from it with increased energy (Fig. 82), just as a baseball rebounds with increased energy from a moving bat. Hence the average kinetic energy of the gas molecules

should be raised, and the temperature of the gas should increase. To verify this prediction, one need only pump up a bicycle tire, and observe how hot the pump becomes after the air in it has been compressed a few times. If, on the other hand, a gas expands by pushing a piston outward, its temperature should fall, since each molecule which strikes the retreating piston gives up some of its kinetic energy. In a steam engine, for instance, compressed steam at a temperature well above the boiling point of water is cooled nearly to its condensation point as it expands by pushing against the piston of the engine.

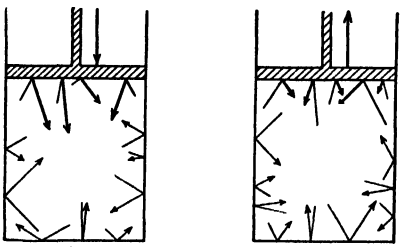


FIG. 82. *A, piston moving down, gas molecules rebound with increased energy; B, gas expanding, pushing piston up; molecules rebound with decreased energy.*

Assumption 4 may be stated in slightly more general fashion as a definition of temperature: *The temperature of a gas is a measure of its average molecular kinetic energy.* It follows that all gases at the same temperature have the same average molecular kinetic energies.

From this fact about temperature we might venture another prediction: heavy molecules should move more slowly than light molecules at the same temperature. For the average kinetic energy of a molecule is

$$\text{K.E.} = \tfrac{1}{2}m\bar{v}^2$$

where m is the molecule's mass and $\bar{v}$ its average speed; a light molecule (small m) must have a greater speed (bigger $\bar{v}$) than a heavy one to keep the product $\tfrac{1}{2}m\bar{v}^2$ the same for both. One easy way to test this prediction is to measure the rates at which two gases leak out through small openings. The gas carbon dioxide, for instance, has molecules over twenty times heavier than molecules of hydrogen (from chemical evidence); through small openings hydrogen escapes much more rapidly, the difference in rates (about 4.7 to 1) being just sufficient to make the average kinetic energies of the two kinds of molecules the same.

Heat. As we have mentioned repeatedly, heat is a form of energy. What meaning can we attach now to the heat energy of a gas, in the language of the molecular theory? Since supplying heat to a gas raises its temperature, thereby increasing the kinetic energy of its molecules, we might reasonably guess that the heat energy of a gas is precisely this energy of molecular motion. But such a guess must be examined carefully.

Up to now, "energy of molecular motion" has implied energy associated with the movement of molecules from place to place, or *translational* movement. If heat energy is merely the kinetic energy of translational

motion, then a given amount of heat ought to affect the temperatures of all gases alike. But this is emphatically not true: 1 cal of heat supplied to 1 g. of hydrogen raises its temperature by about 0.3°C, but 1 cal supplied to 1 g. of carbon dioxide raises its temperature nearly 5°. Evidently heat energy absorbed by a gas is not all used in making its molecules move around more quickly.

Studies of the effects of heat on different gases, coupled with results of other research, have shown that complex molecules may absorb heat energy in three principal ways. (1) Their translational motion may be increased. (2) One part of a molecule may be set to vibrating with respect to another part, as if the parts were connected by a spring. (3) Each molecule may be set in rotation (Fig. 83). *When heat is supplied to a gas, it becomes molecular energy in one or more of these different forms; but only the energy which goes into one particular form, kinetic energy of translation, affects the temperature of the gas.* Ten calories of heat given to 1 g. of any

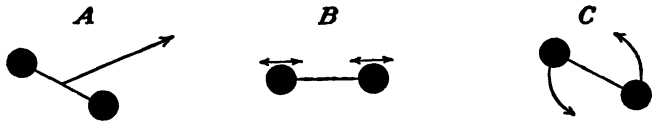


FIG. 83. *Types of motion possible for a simple molecule shaped like a dumbbell (e.g., hydrogen, oxygen, chlorine). A, translation; B, vibration; C, rotation.*

gas causes its temperature to rise, but for some gases the rise is less than for others because more of the 10 cal is consumed in making their molecules vibrate and rotate.

Count Rumford (page 110) and several abler scientists before him had surmised that heat is a form of motion. The kinetic theory finally presented, for gases at least, a definite picture of just what kind of motion heat is—a complex, disordered motion of tiny particles, spinning around, vibrating back and forth, flying about in all directions.

This view of heat answers an awkward question which bothers many a novice in physics: What keeps the molecules moving? We picture them as something like tiny billiard balls, moving rapidly, colliding with each other many times a second. Real billiard balls, after a few seconds of such motion, would lose their energy in the friction of collisions and come to rest at the bottom of their container. All other motions of our experience, save perhaps the motion of stars and planets, are similarly brought to a halt by friction unless some outside force maintains them. Molecular motions are maintained by no outside force, yet continue indefinitely with no sign of diminishing speed. Why is it that friction does not bring these tiny particles to rest, as it does other moving objects?

The answer is simply that friction implies a transformation of mechanical energy into heat energy, which is molecular energy. Friction between molecules would mean simply a transformation of molecular energy into

molecular energy—which is not a transformation at all, except in the sense that one molecule can give its energy to another. Molecules keep moving, in accordance with Newton's first law, because there is nothing to make them stop. The question of the last paragraph is, in fact, nearly meaningless, appearing to make sense simply because we are so accustomed to thinking of heat and motion as distinct concepts. Motion in the molecular world cannot produce heat, as it does in the larger world of everyday life, because molecular motion *is* heat.

Perhaps a warning is appropriate here. We shall presently undertake the dissection of molecules and peer into a realm of particles so tiny that molecules seem gigantic in comparison. As we examine this world of the inconceivably small, we must be wary lest concepts based on impressions from our larger world change their aspect completely, as our idea of heat changes in the realm of molecules.

Escape of Atmospheres

We digress a moment to see what light the kinetic theory of gases can shed on a question from an earlier chapter. Why do the moon and all other small objects in the solar system possess no atmospheres?

The answer, or at least one very likely answer, lies in the rapid motion of gas molecules. A planet holds a moving molecule, as it holds any other object, by gravitational attraction. To escape from a planet's attraction a molecule need only acquire sufficient speed in an outward direction. The necessary speed, called the *escape velocity*, is high for a large planet like Jupiter, low for a small body like the moon. The earth's escape velocity is not quite 7 mi/sec: this is the speed which an interplanetary rocket must possess in order to leave the earth behind, and likewise the speed which a molecule at the top of the atmosphere must attain before it can wander off into space. The moon's escape velocity is much smaller, about 1.5 mi/sec. Now air molecules at ordinary temperatures move with average speeds of less than ½ mi/sec, so their chances of escaping from either earth or moon seem pretty slim. But remember, this ½ mi/sec is an *average speed;* many air molecules at any given instant are moving more slowly, many others considerably faster. Remember too that temperatures on the moon's surface at midday approach 100°C, a temperature at which the average molecular speed is somewhat higher. If a sample of air could be placed on the moon today, a few of its faster molecules would attain speeds over 1.5 mi/sec and drift off into space. As these escaped, collisions in the remaining gas would presently give a few more the necessary speed. The air would slowly vanish, molecule by molecule. Probably the moon has been for ages without an atmosphere because the original gas surrounding it escaped by just this process.

To leave the earth an air molecule would have to reach a speed of 7 mi/sec in the cold upper atmosphere. Attaining this speed at lower elevations would, of course, do no good, for collisions would immediately reduce it. The number of molecules with speeds so far above the average in the upper atmosphere is practically zero, so that our planet is in no danger of losing its valuable air blanket. Even molecules of hydrogen, lightest and fastest of all gas molecules, probably cannot now escape the earth's attraction, although if the earth was hotter in the remote past they may have then escaped in large numbers.

Thus a planet's ability to hold gas molecules depends primarily on its gravitational pull, to a small extent on its temperature. Mercury, small and hot, is as barren as the moon. Venus, nearly the size of the earth, has an atmosphere as dense as ours. Mars is heavy enough to hold a thin atmosphere, but the swift molecules of hydrogen, helium, and neon must have escaped from it long since. The four giant planets have enormously thick atmospheres containing abundant hydrogen.

Liquids and Solids

The kinetic theory grew from the efforts of several nineteenth-century physicists to find an explanation for the behavior of gases. The greatest triumphs of the theory have come from work on gases, but in recent years it has been applied with much success to the vastly more complicated inner structures of liquids and solids.

Assumptions. The incompressibility of liquids and solids suggests the assumption that *their molecules are close together, probably in actual contact.* The Brownian movement in liquids is ample basis for assuming that *liquid molecules, like gas molecules, are in constant, random motion.* The assumption that *particles in solids are held more or less rigidly in fixed positions* follows from the definite shapes which solids maintain and from their inability to diffuse.

Strong attractive forces are assumed to be holding together the particles of liquids and solids. These intermolecular attractions are the same as the very slight ones between gas molecules, which are responsible for deviations from Boyle's law at high pressures. In liquids and solids the forces are simply more conspicuous because the molecules are so much closer together. The nature of the attractive forces is not completely understood, but they are in large part of electrical origin. So great are the forces between adjacent molecules in many solids and liquids that the molecules lose their identity as individual particles. Some of these substances are made up of electrically charged molecular fragments, others of larger particles formed by the joining together of molecules in twos and threes. Because solids and liquids are thus not all constructed of bona fide

molecules, we shall use for their tiny building blocks the more general term "particle."

To explain the effects of temperature on liquids and solids, we may keep the assumption which was so useful for gases, that *absolute temperature is proportional to average kinetic energies of moving particles.* (This is not quite accurate for solids.) In liquids a rise in temperature means an increase in average speed of translational motion, while in solids a rise in temperature means an increase in the vibrational motion of its particles about their fixed positions. The almost universal slight expansion of liquids and solids on heating is then merely an effort of their particles to make room for faster motion. As for gases, we further assume that liquid and solid particles can take up heat energy in other ways, so that absorption of 1 cal of heat by 1 g. of different materials may produce different increases in temperature.

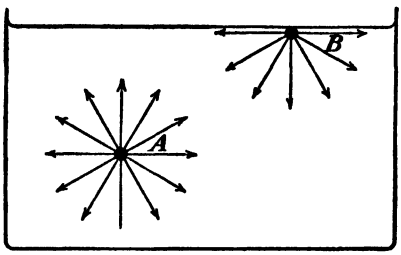


FIG. 84. *Forces on a molecule within a liquid are the same in all directions (A), but forces on a molecule at the surface are unbalanced (B).*

Liquids. If a gas resembles a swarm of angry bees, the particles of a liquid may fairly be likened to bees in a hive, crawling over one another incessantly, each one continually in contact with its neighbors. The slowness of liquid diffusion is explained by the difficulties of motion when each particle is closely surrounded by others. Liquids can flow because their molecules slide easily over one another, but they are more viscous than gases because intermolecular attractions impede the motion. The viscosity of a liquid is a rough measure of how strongly its particles attract each other. Surface tension is due to unbalanced attractions on the surface molecules (Fig. 84): while a molecule in the interior is attracted equally by its neighbors on all sides, a molecule at the surface has no upward force on it to balance the downward pull of the molecules beneath. Hence the surface tends always to contract, to confine the liquid within as small a volume as possible.

Solids. Attractive forces between the particles of solids are stronger than in liquids, so strong that the particles are no longer free to move about. They are far from motionless, however; held in position as if by springs attached to its neighbors, each particle oscillates back and forth rapidly and continuously. A solid is elastic because its particles spring back into their normal positions after being stretched apart or pushed together. A solid breaks, or is permanently deformed, when subjected to a force larger than the forces of attraction between its particles. In a brittle solid rupture takes place suddenly, along a single surface, and the particles

are pulled so far apart that healing is impossible. In a solid which can yield by slow deformation to excessive forces, we may imagine tiny fractures developing all through the solid, each fracture healing itself as the particles slide to new positions and find new partners for their attractive forces.

Crystal form and crystal growth suggest arrangements of particles in patterns of equally spaced rows and planes, like soldiers on parade (Fig. 85). Each particle, so to speak, faces the same way, and the distances to its neighbors on all sides are determined by its attractive forces in various directions. The smooth faces and sharp angles of the crystal are then the outer expression of this inner regularity. The microscope and X rays prove convincingly that many solids which do not show crystal faces have nevertheless a regular pattern in their inner structures. Solids of this sort, whether or not they occur in well-shaped crystals, are called ***crystalline solids;*** salt, diamond, quartz, and most metals are familiar examples. Solids whose particles have no regularity of arrangement (for instance, glass and rubber), and which of course never show crystal forms, are called ***amorphous solids.***

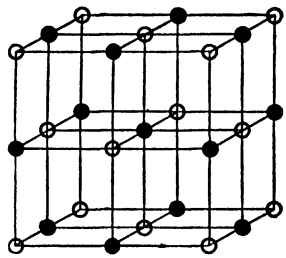


FIG. 85. *Arrangement of particles in a crystal of ordinary salt. (This diagram shows arrangement only; sizes of the particles should be greater compared with their distance apart.)*

So for liquids and solids, as well as for gases, a few well-chosen assumptions about their constituent particles lead us to reasonable explanations for a variety of observational material. Here the explanations are more often in general terms, seldom in precise mathematical language, for the intermolecular attractions which so largely determine the properties of liquids and solids are difficult to express mathematically. Nevertheless, the assumptions of preceding paragraphs have been so thoroughly justified by their agreement with observation and by their fruitfulness in predicting new experimental results that today they have almost the standing of established facts.

Changes of State

To illustrate once more how neatly molecules and their motions account for the ordinary properties of matter, let us consider the peculiar temperature-energy relationships involved in changes of state.

Vaporization. Suppose that two liquids, water and ether, are placed in open dishes. Particles in each are moving in all directions, with a variety of speeds. At any instant some particles are moving fast enough upward to break through the liquid surface and escape into the air, in spite of the attractions of their slower neighbors. By this loss of its faster

molecules each liquid gradually evaporates; since the molecules remaining behind are the slower ones, evaporation makes the liquids cool. The ether evaporates more quickly (or is more *volatile*), and cools itself more noticeably, because the attraction of its particles for each other is smaller and a greater number can escape.

Now suppose that a tight cover is placed on each dish. Molecules of vapor can no longer escape completely: by collisions with each other and with the cover they are sooner or later knocked back into the liquid. So the fast particles which leave each liquid return to it, and it no longer is cooled. Evaporation goes on as before but is balanced now by the reverse process of condensation. How many molecules are present in the space above each liquid at any given instant is determined by how rapidly they leave its surface; over the ether considerably more vapor is present than over the water, since it evaporates more easily. The amount of vapor is expressed most easily in terms of its pressure: thus we say that the ***vapor pressure*** of the ether is higher than that of water. More formally, we may define the vapor pressure of any liquid at a given temperature as *the pressure which its vapor exerts when confined above the liquid.* The vapor pressure of a substance is a measure of how readily its particles escape from its surface.

Let us now remove the covers from our dishes of water and ether and heat each liquid slowly. We aid the process of evaporation by supplying heat, giving more and more particles the energy necessary to escape. Vapor rises from each dish in steadily increasing quantities; if we stop the heating at intervals, cover each dish, and measure the vapor pressures, we find that these pressures are growing rapidly larger. At a temperature of 35° even the particles of average speed in the ether dish apparently gain sufficient energy to vaporize, for bubbles of vapor begin to form all through the liquid. We say now that the liquid is *boiling.* Its vapor pressure has become equal to the pressure of the surrounding atmosphere: particles within the liquid form bubbles, because now the pressure which they exert as vapor is sufficient to overcome the downward pressure of the atmosphere on the liquid surface. At this temperature the vapor pressure of water is still only a small fraction of an atmosphere (4.2 cm of mercury); only when the temperature approaches 100°C does its vapor pressure also become equal to atmospheric pressure, so that the liquid can boil. We may define the ***boiling point*** of a liquid as *the temperature at which its vapor pressure becomes equal to the surrounding pressure.* Standard boiling points listed in tables are the temperatures at which vapor pressures become equal to normal atmospheric pressure, 76 cm of mercury (Figs. 86, 87).

Whether evaporation takes place spontaneously from an open dish or is aided by heating, the formation of vapor from a liquid requires energy. In the one case energy is supplied from the heat energy of the liquid itself

(since the liquid grows cooler), in the other case from the external source of heat. For water at its boiling point, 540 cal (the latent heat of vaporization, page 124) are used up in changing each gram of liquid into vapor. Here there is no difference in temperature between liquid and vapor, hence no

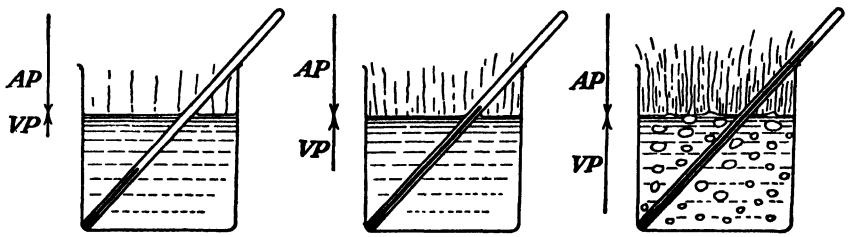


FIG. 86. *A liquid boils when the temperature becomes high enough to make its vapor pressure (VP) equal to atmospheric pressure (AP).*

difference in their average molecular kinetic energies. If not into kinetic energy, into what form of molecular energy do the 540 cal of heat energy go?

Intermolecular forces suggest an answer. In the liquid these forces are strong, because the molecules are close together. To tear the molecules apart, to separate them by the wide distances which exist in the vapor, requires that these strong forces be overcome. Each molecule must be moved against the attraction of its neighbors, moved to a new position in which their attraction for it is very small. Just as a stone thrown upward against the earth's attraction acquires potential energy, so molecules moved apart in this fashion acquire potential energy—*potential energy with reference to intermolecular forces.* So heat supplied in evaporating a liquid goes into another form of molecular energy besides the translational, vibrational, and rotational energies we have considered: molecular potential energy. When a vapor condenses to a liquid, its molecules "fall" toward each other under the influence of their mutual attractions, and their potential energy must be taken up as heat energy by the surroundings.

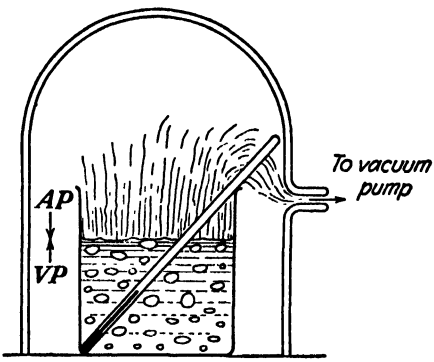


FIG. 87. *A liquid may be made to boil at a low temperature by reducing the external pressure (AP).*

Melting. Temperature changes accompanying melting are quite different for crystalline and amorphous solids and afford one easy experi-

mental method for distinguishing the two types of material. A crystalline solid like ice melts sharply, at one definite temperature, and requires the addition of a certain quantity of energy at this temperature simply to effect the change from solid to liquid. An amorphous solid like glass softens gradually on heating, so that no one temperature can be given as its "melting point."

In terms of the kinetic theory this difference in behavior is a consequence of differences in inner structure. Particles of a crystalline solid are arranged in a definite pattern, each one oriented so that the forces binding it to its neighbors on all sides are as large as possible. To overcome these forces and give the particles the disorderly arrangement of a liquid structure requires that they gain potential energy, just as liquid particles must gain potential energy during evaporation. This potential energy is the latent heat of fusion (page 124), which must be supplied to melt any crystalline solid and which is given out when the liquid crystallizes again. Particles of amorphous solids, on the other hand, have no definite pattern but are already in the random, disorderly arrangement characteristic of liquids. Melting involves merely a gradual loosening of the ties between adjacent particles, without any marked increase in potential energy at a certain temperature.

Thus, without any new assumptions, the kinetic theory offers a rational interpretation of vaporization and melting in terms of the motions, the potential energies, and the arrangements of tiny particles.

Dissipation of Energy

Different kinds of energy, we learned in an earlier discussion, can be transformed one into another. But heat energy is peculiar, in that it cannot be transformed into other kinds *efficiently*. From the heat energy of burning coal we may obtain mechanical energy, chemical energy, or electrical energy, but always in these transformations a large fraction of the heat energy is wasted. There is no escape from this waste; the transformations will not take place without it. This odd characteristic of heat energy was discovered early in the nineteenth century, at the beginning of the Industrial Revolution, as a result of attempts to improve the recently invented steam engine. Attacked both by practical men trying to get as much mechanical work as possible out of a ton of coal, and by scientists more interested in the peculiarities of heat as a form of energy, the problem of why transformations involving heat should be so wasteful was finally solved by the kinetic theory's picture of heat as random, disorderly motion of molecules.

The only practical method ever suggested for obtaining mechanical energy from heat, the method used in both steam and compressed air engines, is to supply heat to a compressed gas and let it expand against a

piston or the vanes of a turbine. Without bothering about the valves, flywheels, and other gadgets of a real engine, we may picture the fundamental process involved as follows: Suppose that a gas under pressure in a cylinder (Fig. 88) is heated and allowed to expand by pushing the piston outward. Heat energy is thereby converted into mechanical energy of the piston, which may be used for turning a dynamo, running a pump, or for any purpose we wish. Now when expansion has reduced the pressure of the heated gas to atmospheric pressure, or when the piston reaches the end of the cylinder, further heating accomplishes little. To give the piston more mechanical energy, we must somehow recompress the gas and let it expand again. Recompressing the gas, however, makes necessary the application of mechanical energy from outside. If the gas is compressed while hot, just as much energy will be needed as the gas has given out by expansion, so that the net gain will be zero. But if we cool the gas first, we find that less energy is required to compress it than we gain from the expansion. So to make the engine do useful work, we must arrange to compress the gas while it is cold, then heat it and allow it to expand while hot, then cool it again for the next compression.

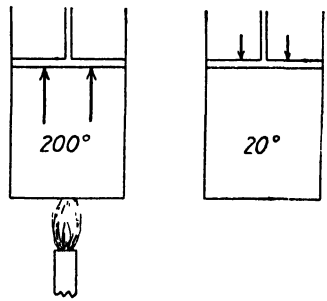


FIG. 88. *A gas at 200° gives out more energy in expanding than is required to compress the gas at 20°.*

Note carefully what happens to the heat supplied to the gas: a part of it is used to drive the piston, but some is deliberately allowed to escape when the gas is cooled before compression. For the engine to run, we need both a source of heat and something to which the gas can give part of its heat—usually the surrounding atmosphere. In effect, *heat flows through the engine from the heat source to the atmosphere, and during the flow we manage to change some of the heat into mechanical energy.* All heat engines work on this principle, taking advantage of the flow of heat from a hot object to a cold object in order to recover some of it as mechanical energy. A *difference of temperature* between two objects or two places is essential.

In the molecular motions of the atmosphere is a vast quantity of heat energy, but we cannot recover it because no cold object is available to which the heat can flow. Of course, we might set up a refrigerating machine to maintain a low temperature, but we should find that more energy is required to run the refrigerator than we could get by using it as the cold side of a heat engine. A refrigerator, in fact, is the reverse of a heat engine: it employs mechanical energy to force heat from a cold object to a warmer object, while a heat engine uses the natural flow of heat from hot objects to cold as a means of obtaining mechanical energy. Setting up

a refrigerator to maintain a low temperature for running a heat engine is just the old perpetual-motion dream in a new guise.

We are strictly limited, then, in our ability to turn heat energy into mechanical energy. We can obtain mechanical energy only by letting heat flow from a region of high temperature to a region of low temperature, and in the best of circumstances the conversion is incomplete. This limitation is not due to friction, which saps the energy of all kinds of motors, but is a limitation imposed by the fundamental nature of heat. Why should heat energy in this respect be so different from other forms?

The kinetic theory answers that heat is energy of random, disorganized motion. When heat is supplied to a gas, its molecules increase their speeds in all directions. We use only the increase in speed in a particular direction, the direction in which the piston moves; but to obtain an increase in this direction, we must speed up motions in all other directions as well. If we could line up the molecules and fire them all, like tiny bullets, straight at the piston, then all the energy we give them would go to make the piston move. In a real gas most of our bullets go astray, only a few of them giving their excess energy to the piston. From energy of random, disorderly motion we can extract only a fraction as the energy of ordered motion in definite directions which we need to run the world's machinery.

In contrast to the difficulty in obtaining mechanical energy efficiently from heat is the ease with which other forms of energy may be converted into heat energy. When coal burns, its chemical energy changes directly to heat. When a pendulum swings, kinetic energy is transformed to potential and back again, but with each swing friction removes some energy as heat. Electrical energy in a light filament is changed partly to heat, partly to light; the light, falling on walls and furniture, is absorbed and converted into heat energy. In all the energy transformations of our acquaintance some, and more often all, of the original energy becomes eventually energy of disordered molecular motion.

Once in the form of heat, energy cannot easily be restored to its original form. A little may be recovered, as we have seen, by making the heat flow to a cooler object, but even to recover this little is possible only for a brief time after the heat is produced. For heat energy spreads quickly into its surroundings, and the temperature difference necessary to make a heat engine operate soon disappears. A stone falls to the pavement, and kinetic energy turns into heat; for a few moments particles in the stone and concrete move a little faster, the faster movement affects neighboring molecules, and the heat energy spreads in ever-widening circles. Coal burning in a fireplace warms the air of a room; when the fire goes out, the warmth persists for a little while, but spreads gradually to the walls, to adjoining rooms, to the outside air, and presently is dis-

tributed so widely through the surroundings that we can no longer detect it. From any hot object heat flows in like manner to cooler objects about it, spreading indefinitely until it becomes a part of the general molecular motion of earth and atmosphere. From this reservoir of heat we cannot recover even a fraction of the original energy, for a perceptible temperature difference no longer exists. The energy has not disappeared, but it is no longer available for conversion into other forms.

We may summarize these observations on heat by the statement: *In every energy transformation, some of the original energy is always changed into heat energy which is not available for further transformations.* This statement is the law of dissipation of energy (called in more erudite circles the second law of thermodynamics). The law is merely a scientific expression of the everyday observations that other forms of energy commonly become heat energy, and that heat spreads out, or is dissipated, into its surroundings.

So far as we know, this law of the wastage of energy applies quite as universally as the law of conservation of energy. The radiant energy of stars, the mechanical energy of planetary motions, the chemical energy of food, all are being steadily changed into the energy of disordered molecular motion. The law seems to imply that the universe in the past had more radiant energy, more chemical energy, more mechanical energy than at present. It seems to imply also that the distant future will bring a time when there is no energy but heat energy, heat energy evenly distributed so that no part of the universe is warmer than another.

Questions

1. Why does not bombardment by air molecules produce the Brownian movement in large objects like chairs and books?
2. Account for the ability of gases to leak rapidly through very small openings.
3. How many nitrogen molecules could you place side by side along a line 1 mm long? (See page 128.)
4. How could you tell experimentally whether a fragment of clear, colorless material was glass or a crystalline solid?
5. Each molecule of the gas sulfur dioxide has a mass almost exactly twice that of an oxygen molecule. If both gases are at the same temperature, which will have the higher average molecular speed? If the speed of oxygen molecules for this temperature averages 40,000 cm/sec, what is the average speed of the sulfur dioxide molecules?
6. Suggest an explanation in terms of molecular motions for the expansion of liquids and solids on heating.
7. How does perspiration give the body a means of cooling itself? Why is moving air apparently cooler than still air?
8. The table on page 144 shows the vapor pressures of alcohol and water for various temperatures. Which is the more volatile liquid? What is the normal boiling point of alcohol? At what temperature would water boil on a mountain top where atmospheric pressure was 60 cm of mercury? How low must the pres-

sure be reduced to make water boil at 0°C? How could you make alcohol boil at 100°C?

TABLE VI

Temperature, °C	*Vapor pressure of water, cm of mercury*	*Vapor pressure of alcohol, cm of mercury*
0	0.46	1.27
20	1.74	4.45
40	5.49	13.37
60	14.89	35.02
70	23.33	54.11
75	28.88	66.55
80	35 49	81.29
85	43.32	98 64
90	52.55	118.93
95	63.37	142 51
100	76.00	169.95

9. A shallow dish containing ether is placed on a moist table top, and air is blown across the surface of the ether. Presently the dish becomes frozen to the table. Explain.
10. Suggest two processes in which molecular kinetic energy is in part converted into molecular potential energy.
11. What are the two chief factors which determine the amount and kind of atmosphere surrounding a planet? Explain briefly.
12. On the surface of Mars would a barometer show a lower or higher reading than on earth? Would water boil at a lower or higher temperature?

CHAPTER XI

Chemical Change

Our discussions of matter up to this point have been concerned chiefly with its general characteristics. We have sought to distinguish the three states of matter, focusing attention on characteristics common to all gases, all liquids, and all solids. Only incidentally have we mentioned differences in the specific characteristics of various gases, liquids, and solids—the heaviness of mercury as compared with water, the volatility of ether and alcohol, the amorphous character of glass, the difference in molecular speeds between hydrogen and carbon dioxide. To go more deeply into the study of matter requires that we turn our attention to these specific characteristics of different materials.

Properties

The essential characteristics of a substance, which enable us to distinguish it from other substances, are called its ***properties.*** We describe water as a colorless, odorless, tasteless material; liquid at ordinary temperatures, freezing at 0°C, and boiling at 100°C; noninflammable; capable of dissolving sugar, salt, and many other materials; miscible with alcohol in all proportions, immiscible with oils: these are all properties of the substance water.

The list of properties by which we describe a substance ordinarily includes such features as its color, taste, odor, whether it is solid or liquid or gaseous, its relative heaviness compared with other substances, to what extent it will dissolve other materials or be dissolved by them, how easily it will burn. For a liquid we might add its viscosity and its miscibility with other liquids; for a solid we should describe its hardness, its brittleness, its crystal form. We could go on to mention the behavior of a substance toward a magnet, the ease with which it conducts heat and electricity, its physiological effects, its behavior when mixed with all manner of different substances. Often in scientific work it is useful to express some of these properties *quantitatively:* thus instead of describing alcohol as a volatile

liquid, we may say more precisely that alcohol boils at 78.3°C, freezes at about −117°, and has a vapor pressure of 4.45 cm of mercury at 20°C.

The property called "relative heaviness" in the last paragraph needs a moment's attention. We say loosely that lead is heavier than aluminum, that mercury is heavier than water; in more precise language we mean that a certain volume of lead weighs more than the same volume of aluminum, a certain volume of mercury more than the same volume of water. To make such comparisons conveniently and precisely, we use a property called ***density,*** which for any substance is defined as its *mass per unit volume.* Density is commonly expressed as the number of grams in a cubic centimeter, or the number of pounds in a cubic foot. Thus 1 cc of water weighs 1 g., so its density is 1 g./cc; 1 cu ft of water weighs 62.5 lb, so its density may also be expressed as 62.5 lb/cu ft. In symbols, for any substance,

$$D \text{ (density)} = \frac{m \text{ (mass)}}{V \text{ (volume)}} \tag{25}$$

The density of lead (at ordinary temperatures) is 11.37 g./cc, which means that every cubic centimeter of lead weighs 11.37 g. The density of mercury is 13.6 g./cc; that of alcohol, a liquid lighter than water, is 0.79 g./cc. The lightest gas, hydrogen, has a density of only 0.00009 g./cc (at 0°C and 1 atm pressure); the heavier gas carbon dioxide has a density of 0.0020 g./cc, about twenty times as great.

A term similar to density and often used in technical work is *specific gravity,* defined as the number of times heavier a substance is than an equal volume of water. Thus 1 cc of lead is 11.37 times as heavy as 1 cc of water, so its specific gravity is 11.37; 1 cc of hydrogen is 9/100,000 as heavy as 1 cc of water, so its specific gravity is 0.00009. In metric units specific gravity and density are numerically equal.

Properties of materials undergo profound changes in many familiar processes. In preceding chapters we have discussed at length the changes in properties brought about by changes in molecular motions and distances when solids melt and liquids vaporize. Other processes, like the rusting of iron, the burning of wood, the explosion of dynamite, involve more drastic changes in properties and require for their explanation an alteration in the molecules themselves. To processes of this sort, called *chemical changes* or *chemical reactions,* we shall devote this and the next several chapters.

Details of a Chemical Reaction

Suppose we mix a little zinc metal, in finely powdered form, with a somewhat greater amount of powdered sulfur, set the mixture on a sheet of asbestos, and touch it with the flame of a Bunsen burner. We are

rewarded with a most impressive explosion, accompanied by much heat and light. When the fireworks have died down and the excess sulfur has burned away, we find in place of the zinc-sulfur mixture a brittle, white substance resembling neither of the original materials. What has happened?

Further experiment would quickly show that neither zinc nor sulfur alone gives any such reaction on heating; that explosion of the mixture takes place as well in a vacuum as in air; that a metal or porcelain surface may be substituted for the asbestos without influencing the reaction. In other words, the process requires the presence of both zinc and sulfur, but no other materials. Unless we choose to believe that zinc or sulfur or both have vanished from the face of the earth, we are forced to conclude that the two substances have *combined* to form the new material. To convince the most skeptical, we could find ways to change the white, brittle product back into zinc and sulfur, but to accomplish this change would require considerable time and resourcefulness.

This is a simple example of the kind of alteration in the properties of matter which we call a ***chemical change.*** Let us analyze it in true pedagogic fashion by examining in detail the properties of the three substances concerned. Sulfur is a yellow solid, rather soft, with a low density, melting at about 114°C, not dissolved by water or acid but easily dissolved by a liquid called carbon disulfide. Every grain of the yellow powder exhibits these properties, and if we should crush the grains their fragments would still show the same properties. Because every particle of sulfur is like every other particle, we call it a ***homogeneous*** substance. Zinc too is homogeneous, with such characteristic properties as a light gray color, fairly high density, a melting point of 419.4°C, solubility in dilute acids, and insolubility in carbon disulfide. The brittle, white product of reaction is a third homogeneous substance, called zinc sulfide, with properties different from those of either zinc or sulfur: a very high melting point, a fairly low density, insolubility in either very dilute acids or carbon disulfide.

Now suppose we prepare two identical zinc-sulfur mixtures, heating one until it reacts but leaving the second unheated. Both the resulting zinc sulfide and the unheated mixture contain zinc and sulfur, and both differ from pure zinc and sulfur in such properties as color and density. But the mixture is a ***heterogeneous*** material: its properties change from one particle to the next. With a microscope and a needle we can separate particles of zinc and particles of sulfur. Either carbon disulfide or dilute acid will dissolve only a part of the mixture. Properties of individual particles in the mixture have not changed at all. In the zinc sulfide, on the other hand, every particle has the same properties as every other particle, and these properties are quite different from those of zinc and

sulfur particles. Two substances may thus be mixed, often very intimately, simply by stirring them together, but their properties are radically altered only if they undergo a chemical reaction.

Most familiar chemical changes do not involve so spectacular a liberation of energy as the zinc-sulfur reaction. Rust, for instance, is the product of a slow combination of iron with certain gases of the atmosphere. Silverware slowly tarnishes because the metal combines with small amounts of sulfur contained in certain foods and in compounds in the air. Cooking involves many complex but hardly spectacular reactions. In the growth and decay of living things occur even more complex types of slow chemical change. Among commonplace chemical processes only the reactions involved in burning result in rapid liberation of energy.

To interpret a chemical change, for instance the zinc-sulfur reaction, in terms of the kinetic theory, we might guess tentatively that it involves a combination of the ultimate particles of zinc and sulfur to form particles of zinc sulfide. This explanation is a bit too simple and will need modification later, but at least the conclusion is safe that the original particles of zinc and sulfur have disappeared and a new type of particle has been formed. Just how the change from one type of particle to another comes about, in this reaction and in others, is the central problem of chemistry. We can best approach the problem by a brief review of older ideas.

Early Ideas of Chemical Change

Chemistry cannot, like physics, claim descent from disinterested speculations about the planets and stars. Man's earliest interest in chemistry was a practical one: chemical change supplied him with artificial heat and light, with much of his food and various other necessities of existence. Chemistry has remained a more practical science than physics, and its history is a record not only of man's developing ideas but of his avarice and chicanery as well.

Almost as soon as he learned to write, man discovered how to obtain metals from certain rocks, how to extract dyes from plants, how to make alcoholic drinks from fermenting grains and fruits. These were useful arts which satisfied his immediate needs, and they blossomed into flourishing industries. But centuries of practice brought amazingly little improvement in technique. Formulas and procedures passed from one generation to the next, altered only rarely by accidental discoveries. Until the heyday of Greece, chemistry was not a science, but a group of lucrative and unrelated trades.

Ideas of the Greek philosophers regarding the constitution of matter eventually turned chemical thought into new channels. Most influential of the Greek theorists was Aristotle, who taught that all matter is composed of four "elements," earth, water, air, and fire, mingled in various

proportions. Aristotle and his followers are not altogether clear as to the meaning of these elements. Evidently they were regarded not as actual substances, but as qualities or attributes which, added to a mysterious "essence," accounted for the properties of different materials. Chemical change accordingly consisted in divesting a substance of certain attributes and substituting others. This doctrine remained a harmless speculation in Greece, but the more practical Egyptians of second-century Alexandria, about the time of the astronomer Ptolemy, found it an incentive for widening the scope of their chemical industries.

If matter is as simple as Aristotle imagined it, reasoned the Egyptians, men should be able not only to direct known chemical changes toward desired ends, but to predict and discover new and useful chemical processes. For instance, men had learned centuries earlier that a reddish rock could be changed to iron by heating it with charcoal in a clay furnace. The rock lost its qualities of redness and brittleness and gained qualities of grayness and toughness. Why should it not be possible to discover a process whereby iron might be shorn of its grayness and its hardness and given the softness and yellowness of gold? The inquiry was reasonable enough and might have led in a happier age to a thorough testing of the Greek theories. But Alexandria in the second and third centuries was a part of Rome's disintegrating empire, and could scarcely support or appreciate scientific research. The idea that transformations of matter might be predicted and controlled served only to incite craftsmen and petty scholars to an undignified search for ways of simulating gold and silver with combinations of cheaper substances.

Enough of Alexandrian science survived the downfall of Rome and the fanaticism of early Christianity to blossom forth again under the Arabic civilization which, in the eighth and ninth centuries, spread from Spain to India. Among the Arabs the search for a way to change, or "transmute," base metals into gold acquired a new name, *alchemy*, and a mystic language of symbols and philosophical terms which only the initiated could understand. Evolving into a bizarre combination of experimental science, philosophic speculation, and secret rituals, alchemy found its way into Europe in the early Middle Ages. Here it prospered as never before, attracting a motley crowd of charlatans and fanatics who soon convinced the credulous that they had discovered not only the transmutation of base metals into gold but the secret of everlasting youth as well. Some of the abler alchemists, to be sure, were dispassionately seeking knowledge about matter and its transformations, but they were too few to save alchemy from falling gradually into disrepute.

The general caliber of men whom alchemy attracted is perhaps best indicated by the notable scarcity of significant discoveries in its annals. A handful of new metals, a few of our familiar acids, several drugs, a few

improvements in technique: this is the net accomplishment of a thousand-year-long search which led men to a study of everything they could lay their hands on in the animal, vegetable, and mineral kingdoms. Alchemists proposed no theories significantly different from those of Aristotle; chemical change, in their eyes, remained a mysterious process of adding and subtracting "qualities" or "elements" from "principles" and "quintessences."

Paracelsus, in the sixteenth century, turned alchemy away from the fruitless search for gold toward the preparation of medicines and drugs. In the following century men here and there in Europe (John Mayow and Robert Boyle in England, Jean Rey in France, Georg Wilhelm Stahl in Germany) began a realistic inquiry into the properties of matter and their changes during chemical reaction. From the work of these men has sprung our modern science of chemistry. By turning their backs on the search for gold, seeking primarily information concerning the ultimate nature of matter and its alterations, testing each hypothesis rigorously, scientists have gained a control over chemical change which has given us, not a means for transmuting base metals into gold, but much of the richness of our industrial civilization.

Elements

From the long, disheartening search of the alchemists emerged at last an idea that certain materials like iron, mercury, gold, sulfur were simple substances which could be neither decomposed nor transformed into one another. Such simple substances were called ***elements.*** Belief slowly grew that the earth contains only a limited number of these elements and that all other materials are combinations of them in various proportions. *The formation of one substance from others by chemical change, then, is possible only if its elements are present in the other substances.* Never expressed as a law, this statement is nevertheless the fundamental axiom which distinguishes chemistry from alchemy.

These modern elements are concrete, tangible substances, not mere qualities like the elements of Aristotle. Our belief that they are the building materials of all matter is founded on the cold fact that every other variety of matter which we can bring into the laboratory can be broken down into two or more of them. Scientists recognize ninety-odd naturally occurring elements and have reason to believe that not more than two or three remain to be discovered. Of the known elements ten are gases, two liquids, and the remaining seventy-odd solids at ordinary temperatures and pressures. Hydrogen, oxygen, chlorine, neon are familiar gaseous elements; bromine and mercury are the two liquids; iron, zinc, tin, aluminum, copper, lead, silver, gold, carbon, sulfur are solid elements.

To define an element precisely in simple terms is extraordinarily difficult. A few years ago we could have said glibly, "An element is a sub-

stance which cannot, by ordinary chemical means, be decomposed into simpler substances." Modern physics has proved this statement inexact, but the more rigorous definition involves physical concepts which we have not yet discussed. For the present we may use the older statement, remembering that it will need drastic revision later.

Compounds and Solutions

Elements are put together to form the other materials of the earth in a variety of ways. Some materials contain two or more elements united in a chemical ***compound,*** as zinc and sulfur are united in the compound zinc sulfide. Other materials consist of ***mixtures*** of elements, or mixtures of compounds. The distinction between compounds and mixtures is of fundamental importance, since chemical reactions always involve the formation or the breaking down of one or more compounds.

Let us consider first the kinds of material which can be formed from two elements only. The elements may form a *heterogeneous mixture,* like the mixture of zinc and sulfur we considered on a previous page; in such a mixture each element retains its own distinguishing properties, and small fragments of each may be separated mechanically from the mixture. The elements may be mixed more intimately to form a *homogeneous mixture,* or *solution:* thus the gases hydrogen and oxygen when placed in the same container will mix so thoroughly that no ordinary means will show that any one part of the mixture is different from any other part. Finally, the elements may form a *compound:* if an electric spark is allowed to jump through a mixture of hydrogen and oxygen, the two react violently and liquid droplets of the compound water are formed.

The distinction between compounds and heterogeneous mixtures is easy, since the separate elements are still recognizable in the mixtures. It is not always a simple matter, however, to tell whether two elements have formed a solution or have undergone a chemical reaction to produce a compound. We may say in general that their properties are more profoundly altered if they have united chemically, and we may put the matter to experimental test in one of the three following ways:

1. Measure the freezing point or boiling point of the material. For a compound the freezing point or boiling point is a constant temperature, but for a solution the temperature changes during both boiling and freezing. Thus water, a *compound* of hydrogen and oxygen, boils at precisely 100°C (if the pressure is 1 atm); liquid air, a homogeneous *mixture* of nitrogen and oxygen, starts to boil at about −192°C and boils at various temperatures up to −182°C as vaporization continues.

2. See whether the material can be separated into its elements by boiling or freezing. Ordinarily the composition of a compound is not altered by a change of state, while a solution is wholly or partially separated into its constituents. Water shows no tendency to decompose

into its elements at 100°C or even far above this temperature; when liquid air boils, however, the vapor which comes off first is largely nitrogen, the vapor which comes off in the last stages largely oxygen.

3. Add more of one of the constituent elements to the material, and see whether the material remains homogeneous. Elements in a compound are combined in a definite, invariable proportion, while the composition of a solution or mixture is variable. In water every gram of oxygen is combined with precisely 0.126 g. of hydrogen; if more oxygen or more hydrogen is added, it does not mix with the water but forms a heterogeneous mixture of gas and liquid. With liquid or gaseous air, on the other hand, more nitrogen or oxygen will readily mix, and the material remains homogeneous.

Experiments of this sort will determine whether any unknown material is a compound or a solution. The experiments, of course, reflect the fact that in a compound the elements have combined to form a new substance, with characteristic properties of its own, while in a mixture each element retains its identity.

The Classification of Matter

Materials made up of more than two elements may be compounds, or may be mixtures of various sorts. The distinction between heterogeneous mixtures and homogeneous mixtures holds as well for mixtures of compounds as for mixtures of elements. Thus the common rock granite is a heterogeneous mixture of several compounds, the various compounds being easily distinguishable by their differences in color, in crystal form, in hardness, etc. Whole milk is a heterogeneous mixture which separates into milk and cream on standing. A solution of salt in water, on the other hand, is a homogeneous mixture of two compounds; the boiling point of the solution is not constant, the salt may be separated by boiling off the water, more salt or more water may be added to the solution without destroying its homogeneity. Ordinary soda water is a homogeneous mixture of carbon dioxide dissolved under pressure in water; when the pressure is released, it becomes a heterogeneous mixture of liquid water and gas bubbles.

All the earth's materials may be classified as elements, compounds, homogeneous mixtures, and heterogeneous mixtures. Formally, the classification may be expressed as follows:

Heterogeneous matter (mixtures of compounds, elements, or both)
Homogeneous matter, including
  Elements } Pure substances
  Compounds }
  Solutions (homogeneous mixtures)

In terms of the molecular theory, a heterogeneous material contains large aggregates of dissimilar molecules. A homogeneous material is either a substance containing molecules of only one kind or a solution containing two or more molecular types intimately mixed. Because all the molecules in an element or compound are very nearly alike, these two are often classed together as *pure substances*, in contrast with solutions.

This classification of matter needs some critical comment. "Homogeneous" and "heterogeneous" apply to matter as we see and feel it with our senses or with the aid of a microscope, not to its ultimate particles. Any kind of matter is, of course, heterogeneous if we consider it as made up of molecules and empty spaces. Also a gradual change in properties from place to place does not destroy the homogeneity of a material: thus air is considered homogeneous, although air on a mountaintop differs markedly in density and composition from air at the seacoast. Pure substances we have defined as materials with molecules very nearly alike; the adverbs "very nearly" are necessary, as we shall see later, to make the statement conform with modern physical research. The distinction between compounds and solutions is not always as sharp as we could wish. Some compounds are decomposed on boiling; certain solutions have constant boiling points; the compositions of some compounds, notably minerals, vary within small limits.

Despite these qualifications, the above classification is of basic importance in chemistry. It brings a semblance of order into the multitude of dissimilar materials around us, showing which ones may be studied most profitably. Since the properties of mixtures should depend directly on the properties of their constituent pure substances, it is to these pure substances that we turn first for an understanding of the earth's materials.

Physics and Chemistry

Before going further into chemistry, let us consider for a moment the relations of this branch of physical science to other branches which we have studied and will study.

Any attempt to set up boundaries between the sciences can only emphasize their unity. We say broadly that the physical sciences deal with inorganic matter, the biological sciences with living matter. We consider that physics relates to one type of phenomenon, chemistry to another. These are arbitrary separations, dictated solely by convenience, for in nature there is no fine distinction between one set of materials and processes and another set. Between any two of our separate sciences is a borderland where the two overlap.

Astronomy and geology can be defined without much ambiguity. Geology, the "earth science," deals with all the naturally occurring inorganic materials and phenomena of the earth, embracing those parts of

physics and chemistry and biology which apply to inorganic terrestrial matter. Astronomy considers all things and processes beyond the earth, and the earth as it is related to other objects in space. Since the earth in its motions and history is a part of the solar system, astronomy oversteps somewhat the bounds of geology. Scientific information about celestial bodies until recent years concerned primarily their masses and motions, their distribution, their radiation of energy, so that astronomy is pretty largely physics on a grand scale. In the future, as modern instruments reveal more and more about the forms and changes of matter in stars and planets, astronomy will be linked ever more closely with chemistry.

The distinction between physics and chemistry is more subtle, depending on different points of view rather than on differences in subject matter. A physicist would analyze a dynamite explosion, for instance, in terms of the gas pressure produced, the strength of the rocks rent apart, the amount of heat lost to the rocks and the surrounding air. A chemist, observing the same explosion, would consider the elements and compounds involved in the reaction which converts dynamite into gas, the source of the heat liberated, the possible changes produced in the rock material by heat and pressure. Physics deals with the more general properties of matter, its density, tensile strength, compressibility, its motion in response to forces, its ability to absorb and radiate heat; chemistry deals rather with specific properties of different kinds of matter and with changes from one kind of matter to another. Physics is concerned, too, with energy of all varieties, its source, its measurement, its changes from one form to another; chemistry is concerned with energy only insofar as it influences or is produced by chemical reactions.

Early physics was an offspring of astronomy, early chemistry an outgrowth of misguided efforts to accomplish impossible changes in the form of matter. Inevitably, analytical minds in both sciences sought an explanation of phenomena in terms of the ultimate structure of matter, and researches into this field cannot properly be assigned to one science or the other. Modern inquiries have so broadened our knowledge of the structure of matter that this borderland field might well be considered a science in its own right.

Pointless though any attempt to separate the two sciences must be, certain of the fundamental concepts of each may be profitably differentiated. The *physical properties* of a substance, for instance, are those with which physics is primarily concerned—color, density, hardness, boiling and melting points; its *chemical properties* describe its capacity for reacting with other substances. Again, *physical change* refers to a change like that from solid to liquid or liquid to gas, or the change in form produced by crushing a solid or emulsifying a liquid; *chemical change* involves a change from one compound to another, or from elements to compounds.

We say that "the components of a solution or a heterogeneous material can be separated *by physical means*," indicating thereby such processes as distillation, settling, filtration. *Chemical means*, on the other hand, are necessary to effect separation of the elements in a compound. Since the boundary between physics and chemistry is so hazy, terms like physical and chemical properties, physical and chemical change, cannot be precisely defined but serve a purpose in rough descriptions of matter and its alterations.

The difference in viewpoint between physics and chemistry can hardly be appreciated without a considerable background in both sciences. Such a background in physics we have acquired in preceding chapters; in the immediately following chapters we shall try to get the chemist's slant on things.

Questions

1. List as many properties as you can of ice and of water. Why do we call the change from water to ice a physical change?
2. List the properties of iron and of rust. Why do we call the change from iron to rust a chemical change?
3. Arrange the following in order of decreasing density: water, gold, air, aluminum, helium, ice, iron.
4. A cube of iron 10 cm. on a side weighs 7,800 g. What is the density of iron?
5. If 32 g. of oxygen occupy 22.4 l. at 0°C under 1 atm pressure, what is the density of oxygen under these conditions?
6. How is the density of a gas changed if it is compressed to one-tenth its former volume (temperature remaining constant)? Derive an algebraic relationship between the density and volume of a given mass of gas when temperature is constant.
7. Could iron be prepared from rust? Sulfur from zinc sulfide? Lead from iron? Hydrogen from oxygen?
8. How can you show that water is a compound rather than a homogeneous mixture of hydrogen and oxygen?
9. Air was long considered an element. How could you show that this idea is false?

CHAPTER XII

Weight Relations in Chemical Reactions

THE first chemical change to be studied intensively by methods which we can call "modern" was the process of burning, or *combustion*. The transformation from wood to smoke and ashes, with its accompanying heat and dancing flames, is by all odds the most spectacular chemical change with which men of earlier times had immediate contact, and it quite naturally has piqued the curiosity of thoughtful individuals from remotest antiquity to the present. Primitive man based his explanation on ever-present demons and spirits. In the religions of many advanced civilizations the fire god has a respected place. The more scientifically minded Greeks gave the first rational explanation in nonsupernatural terms, recorded in the writings of Aristotle: every inflammable material was supposed to contain the elements "earth" and "fire," the latter escaping while the material burned, the "earth" (ashes) remaining behind. In various guises this explanation of Aristotle's persisted through the centuries of alchemy down even to the time of the French Revolution. The history of the overthrow of this hoary idea and the establishment of modern conceptions of chemical change is an impressive chapter in the development of human thought.

The Phlogiston Hypothesis

The particular form which Aristotle's explanation took in the eighteenth century was the phlogiston hypothesis developed by two Germans, Becher and Stahl. Instead of "fire," Becher and Stahl called the substance which supposedly escaped during combustion by the more esoteric name *phlogiston*. The precise nature of phlogiston was never clearly described: sometimes it was regarded as a definite substance, with color, odor, and weight; sometimes as one of the weightless, intangible "essences" which

appear so frequently in alchemical writings. In spite of, or perhaps because of, its indefinite character phlogiston proved capable of explaining most of the chemical facts known in the eighteenth century.

Many metals when heated in air change slowly to soft powders: zinc and tin give white powders, mercury a reddish powder, iron a black scaly material. These changes, as well as the changes in ordinary burning, were ascribed to the escape of phlogiston. A pure metal can often be recovered from its powder by heating the powder with charcoal; in terms of the phlogiston hypothesis, this means that the charcoal is nearly pure phlogiston, which on heating reunites with the powder to form the metal. A similar explanation accounted for the recovery of metals from their ores by heating with charcoal. The observation that air is necessary for burning was explained by the assumption that phlogiston could leave a substance only if air was present to absorb it. The further observation that wood heated in a limited quantity of air is only partially burned meant that the capacity of air to absorb phlogiston is limited.

These explanations and many more like them seemed straightforward enough and established the phlogiston hypothesis on a firm footing. They illustrate well how attractive a wholly false hypothesis can be if it is not too critically examined. A false hypothesis may even be useful in advancing scientific knowledge: thus predictions based on the phlogiston hypothesis could be verified and hence new facts about chemical change discovered.

The great difficulty which phlogiston encountered was in explaining changes in weight. When wood burns its ashes weigh less than the original wood, and the decrease in weight may be reasonably interpreted as the weight of phlogiston which has escaped. When a metal is heated, however, the resulting powder weighs *more* than the original metal. To overcome this difficulty, supporters of the phlogiston theory had to assume that phlogiston could have a negative weight, so that its escape would make a substance heavier. Phlogiston's weight seemed to be variable, changing capriciously from one reaction to the next. Like the heat substance, caloric, phlogiston was regarded as an "imponderable"—but a more versatile imponderable, which could show positive, negative, or zero weight as each situation required.

That so fantastic a hypothesis could have flourished for well over a century, that it could have been defended by some of the ablest scientists of the time, seems incredible from our modern viewpoint. We must remember that chemistry in the early eighteenth century had only recently emerged from alchemy. Color, taste, odor were regarded as more important attributes of a substance than weight. Good balances were not available. Gases were imperfectly understood; they were commonly believed to have zero or negative weights. With chemical knowledge in so

rudimentary a state, the existence of a substance with weights varying from positive to negative did not seem unreasonable.

Lavoisier and Priestley

By 1770 chemistry had made considerable progress. The balance had been improved, new methods for handling gases had been developed, the role of gases in chemical reactions was better understood. With the new knowledge came more and more assumptions to bolster up the phlogiston idea, so many assumptions that a few scientists at last ventured to question the whole unwieldy hypothesis.

FIG. 89. *Antoine Laurent Lavoisier (1743–1794).*

Foremost among these was the great French scientist Antoine Laurent Lavoisier, whose clearheaded reorganization of the chemical knowledge of his time paved the way for modern chemistry (Fig. 89). Son of a wealthy lawyer, Lavoisier was given a good education and had more than ample means for carrying on his scientific work. For many years of his busy life he served as a public official and showed himself keenly aware of the acute social problems which France was facing. But neither an immense scientific reputation nor

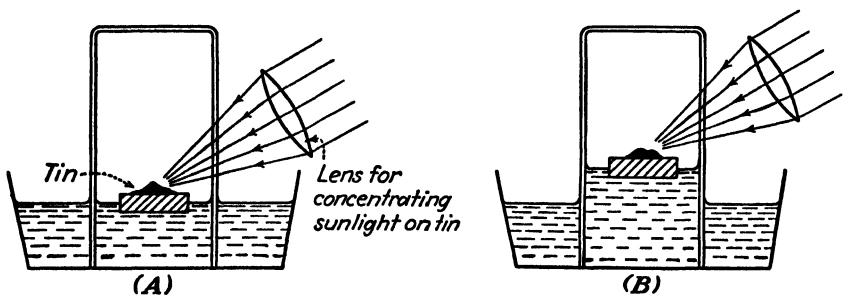


FIG. 90. *Lavoisier's demonstration that tin on heating combines with a gas from the air. A, before heating; B, after heating.*

long public service could save him during the Revolution; denounced by Marat, he was sent to the guillotine in 1794.

Lavoisier demonstrated the falsity of the phlogiston hypothesis most clearly by a series of experiments on tin, a metal which is readily con-

verted to a white powder by heating. He repeated the simple experiment of weighing tin before and after heating, to show that the powder weighed more than the metal. He placed a little tin on a wooden block floating in water, covered the block with a glass jar, and heated the tin by focusing the sun's rays upon it with a magnifying glass—a common method of heating before gas burners and electric heaters were invented (Fig. 90). Part of the tin was converted to white powder, and the water level rose in the jar until only four-fifths as much air was left as at the start; further heating caused no detectable change. (This experiment was not original.) Finally tin was heated in a sealed flask until as much as possible was converted to powder. The flask was accurately weighed before and after heating, and the two weights proved to be identical. Then the flask was opened, and air rushed in. With the additional air, the weight of the flask was greater than at the start.

Now it would be possible, of course, to explain these results by supposing that the air had absorbed a substance with negative weight, provided that some assumptions are added about the effect of this absorption on the air's volume. But to Lavoisier such roundabout reasoning seemed stupid when the experiments suggested so clearly that the tin had absorbed a gas from the air. We need only imagine that one-fifth of the air consists of a gas which can combine with tin: then the powder is a compound of this gas with the metal, and the increase in weight is the weight of the gas; water rises in the jar to take the place of this gas which has been removed; when the sealed flask is broken, air rushes in to replace that which the tin has absorbed. This explanation is simple, direct, involving only substances which have definite weights.

At about the time these experiments were completed, Lavoisier heard that an English experimenter, Joseph Priestley, had obtained by strongly heating a reddish powder the metal mercury and a new gas with strange properties. A lighted candle placed in the gas flared up brilliantly, glowing charcoal burst into flame, a mouse kept in a closed jar of the gas lived longer than one in a jar of air. The red powder, Lavoisier knew, could be prepared by heating mercury in air; if mercury, like tin, formed the powder by combining with a gas from the air, then Priestley, by further heating the powder, must have succeeded in preparing the pure gas which combined with the metals. Isolation of the pure gas was all that Lavoisier needed to complete the proof that metals when heated do not give up phlogiston but combine with a substance from the air. He repeated Priestley's experiment, starting this time with mercury, heating it until a red powder had formed, then by stronger heating obtaining the same active gas. Here then, prepared directly from the air, was a specific, tangible substance which accounted for the behavior of heated metals.

Joseph Priestley, discoverer of the new gas which we today call *oxygen*, was a man of altogether different stamp from the brilliant Frenchman. A retiring, all but poverty-stricken minister of a small church, Priestley had only his spare time and severely limited funds to devote to his beloved experiments. No theoretical genius or organizer like Lavoisier, he had a great flair for careful experimentation. He contributed particularly to the study of gases, discovering several besides oxygen and studying their reactions. Like Lavoisier he was a victim of the French Revolution: he did not lose his head, but suffered so much persecution in England for his outspoken views in favor of the revolutionists that he was forced near the end of his life to seek refuge in America.

Priestley, a firm believer in phlogiston, named his new gas "dephlogisticated air," thinking that it aided burning because it contained no phlogiston and therefore could absorb more than ordinary air. He made no attempt to prove or disprove this idea, nor did he use the new gas for any purpose except to try out its effects on various materials. Lavoisier gave the gas its modern name, oxygen, and found it useful not only for explaining the changes in metals on heating but for explaining the processes of combustion as well. The burning of candles, wood, coal, according to Lavoisier, involves a combination of their materials with oxygen. They appear to lose weight because the products of the reaction are gaseous; actually, as he showed by experiment, the gaseous products weigh more than the original material. Thus burning and the reactions of metals in air both received a rational explanation in terms of real substances, and the necessity for assuming the mysterious phlogiston disappeared.

Oxygen

The element oxygen will play so important a role in future discussions that we had best pause a moment to examine its more conspicuous properties in the light of modern knowledge.

Oxygen under ordinary conditions is a colorless, odorless, tasteless gas, most easily distinguished from air by experiments like those performed by Priestley. If a splinter of wood, for instance, is allowed to burn until one end is glowing charcoal, and if the flame is then blown out and the splinter plunged into oxygen, the flame will rekindle spontaneously and the wood will burn brightly until the oxygen is consumed (Fig. 91). The pale flame of burning sulfur becomes a brilliant blue in oxygen, and phosphorus blazes up with a dazzling yellow light. Several metals will burn brightly in oxygen; even a piece of iron wire will burn if it is heated strongly and plunged into oxygen. We may express this behavior with the statements that oxygen *supports combustion* and is *chemically active.*

If cooled sufficiently, oxygen condenses to a clear blue liquid with the strange property of being strongly attracted to a magnet. The boiling

point of liquid oxygen is −183°C and its freezing point is about −225°C. Oxygen is chemically active even in the liquid state, the closer packing of its molecules offsetting the very low temperature. Iron, for instance, will burn so vigorously in liquid oxygen that the metal is melted by the heat of the reaction, in spite of the intensely cold liquid surrounding it.

Oxygen is by far the most abundant of the elements which make up the earth's crust, its total amount (by weight) being nearly equal to that of all the rest put together. Most of the oxygen is in compounds, compounds which are the chief constituents of rocks, soil, and living things. Water is a compound of hydrogen and oxygen. The free element is one of the important constituents of the atmosphere.

Air owes its ability to support combustion to the free oxygen which it contains; it cannot support combustion as well as pure oxygen because the element is so diluted with inactive gases. (Air is about one-fifth oxygen, four-fifths nitrogen, with small amounts of other gases.) That air is a *mixture* or *solution* of oxygen and nitrogen rather than a compound may be readily shown by liquefying it and allowing the liquid to boil. Boiling commences near the boiling point of nitrogen (−196°C), and the vapor which comes off first consists mostly of this element; as boiling proceeds the temperature rises, and toward the end practically pure oxygen is left in the liquid. The ease with which oxygen can be separated from nitrogen by letting liquid air boil makes this a convenient method for preparing pure oxygen for commercial use.

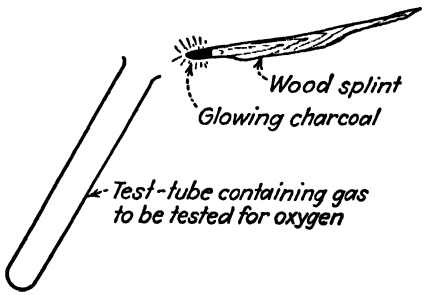


FIG. 91. *A test for oxygen.*

When oxygen combines chemically with another substance, the process is called ***oxidation*** and the other substance is said to be ***oxidized.*** (We shall find later that these terms can be extended to include other types of reaction.) Rapid oxidation accompanied by the liberation of noticeable heat and light is the process of combustion, or burning. In the experiments of Lavoisier tin and mercury oxidized slowly when heated; a lighted candle oxidizes rapidly in air, still more rapidly in pure oxygen. Slow oxidation is involved in many familiar processes, such as the rusting of iron, the decay of wood, the hardening of paint. The energy to maintain life comes from the slow oxidation of food in our bodies, by oxygen breathed in through the lungs and transported by the bloodstream.

A substance formed by the union of another element with oxygen is called an *oxide.* The white powder which Lavoisier obtained by heating tin is tin oxide; the red powder of Priestley's experiment was mercuric oxide.

Rust is largely iron oxide. In general, oxides of metals are solids. Oxides of other elements may be solid, liquid, or gaseous: the oxide of sulfur is an evil-smelling gas (the odor of burning sulfur) called sulfur dioxide; carbon forms two gaseous oxides, called carbon monoxide and carbon dioxide; the oxide of silicon is found in nature as the solid called quartz, the chief constituent of ordinary sand; the oxide of hydrogen is water. Oxides of nearly all the elements can be prepared, most of them by simply heating the elements with oxygen. A few oxides (mercuric oxide, lead oxide, barium peroxide) are easily decomposed by heating, giving one convenient laboratory method for preparing oxygen, while other oxides, such as lime (calcium oxide), are not decomposed even at the temperature of the electric arc, 3000°C.

Conservation of Mass

Lavoisier's discovery of the true nature of combustion was made possible by his use of the balance and by his insistence on the importance of weights in studying chemical reactions. This emphasis on weights marked a profound change in viewpoint, and is one of Lavoisier's great contributions to chemistry. From his day to ours the balance has remained the chemist's most valuable tool.

Not only did the balance enable Lavoisier to overthrow the phlogiston hypothesis; it led him also to a generalization as fundamental to modern chemistry as the law of energy conservation is to physics. Accurate weighing of many substances before and after undergoing chemical reaction convinced him that *the total mass of the products of a chemical reaction is always the same as the total mass of the original materials*, no matter how startling the chemical change may be. This is the *law of conservation of mass*. It may be stated more tersely: *matter can be neither created nor destroyed*. When wood burns, mass seems to disappear because some of the products of reaction are gaseous; if the mass of the original wood is added to the mass of the oxygen which combines with it, and if the mass of the resulting ash is added to the mass of the gaseous products, the two sums will turn out exactly equal. Iron increases in weight on rusting because it combines with gases from the air, and the increase in weight is exactly equal to the weight of gas consumed. In the thousands of reactions which have been tested with accurate chemical balances, no deviation from the law has ever been found.

Modern physics has made necessary some modification in our ideas about the conservation of mass and energy. We can no longer believe that matter and energy are indestructible, for there is good evidence that some processes, notably those taking place in the "atom smashers" of modern physical laboratories and at high temperatures in the interiors of the stars, actually involve the transformation of matter into energy. To take

account of these processes, we must for strict accuracy combine the two conservation laws in the single statement, *the total amount of energy plus matter in the universe is constant.* We shall discuss transformations of matter into energy in future chapters, but for the present our concern is with ordinary processes in which the older conservation laws hold rigidly.

The explanation of combustion, emphasis on the use of the balance, and the law of conservation of mass by no means tell the full story of Lavoisier's achievements. Two other contributions were especially valuable in the development of chemistry: a revision of chemical nomenclature (undertaken with a committee of French scientists), which finally cleared away the complex and mystifying jargon inherited from alchemy; and a reemphasis on the importance of elements as the fundamental chemical substances. One curious member of Lavoisier's list of elements is the heat substance, caloric: while he discredited the phlogiston imponderable, Lavoisier still felt it necessary to use an imponderable for the explanation of heat.

Although comparisons between scientists in different fields are never safe, the thoroughgoing reorganization and systematization which Lavoisier gave to chemistry are reminiscent of the changes which Isaac Newton had wrought in physics a hundred years earlier.

The Law of Definite Proportions

In the years immediately following Lavoisier's death, the balance was called upon to settle an important question regarding the composition of chemical compounds. Were the elements in a compound combined in a definite, constant ratio by weight, or could the ratio vary somewhat? Take the familiar compound water, for example. Analysis had shown that in water 8 g. of oxygen were combined with each gram of hydrogen. Did this mean that *every* sample of water contained oxygen and hydrogen in the ratio 8:1, or could the ratio sometimes be 7:1 or 7.5:1? The problem was a difficult one in those days, for analytical procedures were not well developed and substances used in analysis were often impure.

Two eminent French scientists took opposite sides of the question: Berthollet, a chemical engineer, friend and adviser of Napoleon; and Proust, in his early years an apothecary, later given a magnificent laboratory in Madrid by King Charles IV of Spain. Berthollet's careful analyses convinced him that the composition of a compound could vary within certain limits, while the equally careful analyses of Proust indicated that compositions were strictly constant. Proust finally had the better of several years of argument; by repeating Berthollet's work, he showed that the apparent variations in composition could be explained by impurities in the substances used for analysis.

Thus Proust could venture the general statement that *the elements which make up a chemical compound are combined in a definite proportion by weight.* This generalization is called the *law of definite proportions* (Fig. 92).* It means, for example, that if one sample of water contains 8 g. of oxygen

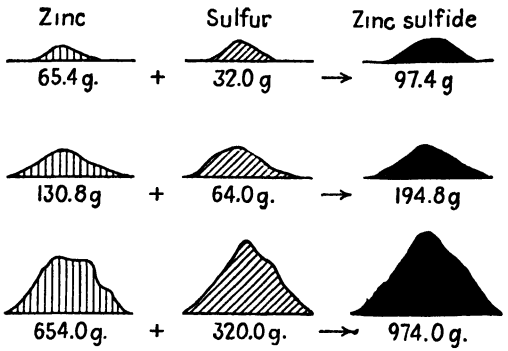


FIG. 92. *The law of definite proportions.*

to every gram of hydrogen, then any other sample of water will show the same ratio. Take water from the ocean, from rain, from a melting glacier, or prepare it in the laboratory: invariably it will contain eight parts by weight of oxygen to one part by weight of hydrogen (more precisely, 8 to 1.008). If 3 g. of hydrogen and 24 g. of oxygen are mixed and a spark is passed through the mixture, the gases will be entirely used up to form 27 g. of water, since they are present in a 1:8 ratio; but if 4 g. of hydrogen are mixed with 24 g. of oxygen and ignited, the result is a heterogeneous mixture of 27 g. of water and 1 g. of unused hydrogen gas.

The law of definite proportions appeared implicitly in the last chapter, as one means of distinguishing compounds from solutions. Its discovery at the beginning of the nineteenth century made possible finally a clear recognition of the differences between elements, compounds, and mixtures and thus added one more step to the systematization which Lavoisier had begun. But it did more than simply help to organize existing knowledge: together with Lavoisier's conservation law, it pointed the way to a theory designed to explain just what happens when chemical changes take place.

Questions

1. When a large jar is placed over a lighted candle, the candle burns for a few minutes and then goes out. Explain this experiment in terms of phlogiston and in terms of oxygen.
2. The gas ammonia is a compound of the two elements nitrogen and hydrogen. Analysis shows that for every 3 g. of hydrogen in ammonia 14 g. of nitrogen are

*As with so many of the basic laws of physical science, modern research has shown that this one is not quite accurate in the simple form given above. The simple statement of the law, however, is satisfactory for present purposes.

present. If a given sample of ammonia contains 12 g. of hydrogen, how much nitrogen must it contain? How many grams of nitrogen would 34 g. of ammonia contain? How many grams of ammonia could be prepared from 100 g. of hydrogen?

3. The most common ore of iron is iron oxide, or hematite, which contains 7 g. of iron to every 3 g. of oxygen. How much iron could be recovered from 100 kg of pure iron oxide ore? How much from 100 tons?
4. When water is decomposed into its elements, the *volume* of hydrogen produced is twice as great as the volume of oxygen, but its *weight* is only one-eighth the weight of oxygen. If the density of hydrogen is 0.00009 g./cc, what is the density of oxygen (both gases at 0°C and 1 atm pressure)?
5. Nitrogen and oxygen, the principal constituents of air, are present in a practically constant ratio of 4 g. of nitrogen to every 1 g. of oxygen. The two elements can also be made to combine chemically, forming the compound nitric oxide, which contains 7 g. of nitrogen to every 8 g. of oxygen. Both air and nitric oxide are colorless gases. How could you show experimentally that one is a compound and the other a mixture?
6. When carbon (coke or charcoal) burns in air, either carbon monoxide or carbon dioxide may be formed, depending on conditions of burning. In carbon monoxide every 3 g. of carbon is combined with 4 g. of oxygen; in carbon dioxide every 3 g. of carbon is combined with 8 g. of oxygen. How much oxygen would be required to burn a kilogram of coke to form carbon monoxide? How much to burn a kilogram of coke to form carbon dioxide? If the burning were done in air rather than in pure oxygen, how much air would be required in these experiments?
7. State the phlogiston hypothesis of combustion, indicate the observational evidence on which it was based, and suggest two predictions which can be made from it. Similarly describe Lavoisier's explanation of combustion, the observational evidence behind it, two predictions that can be made from it, and simple experiments to check the predictions. Why is Lavoisier's explanation better than the phlogiston hypothesis?

CHAPTER XIII

The Atomic Theory

LAVOISIER had given chemistry a new point of view and a rational nomenclature; Proust had made precise the idea of a chemical compound; other workers were discovering new substances and new reactions in great numbers. As the nineteenth century opened, the time was again ripe for a bold scientific theory which could show the underlying simplicity in an accumulation of observational facts.

John Dalton

The man who laid down the outlines of this theory was an English Quaker named John Dalton, an awkward, colorless individual whose outward life was the humdrum existence of a poorly paid teacher (Fig. 93). By training and by talent Dalton seemed singularly unfitted for a brilliant career in science. His formal education stopped at an early age, and what he knew of physics and chemistry he learned himself; he was a slow thinker; he was a notoriously poor experimenter. For scientific work his only assets seemed to be a stubborn perseverance and an unshakable habit of taking notes about everything he saw—about the weather, about his experiments, about his travels, about his income and expenditures. Perhaps we should list also as assets his meager education and his plodding, literal mind: for the one led him to seek new explanations uninfluenced by the misconceptions of his contemporaries, and the other made necessary persistent efforts to portray complex subjects with simple diagrams.

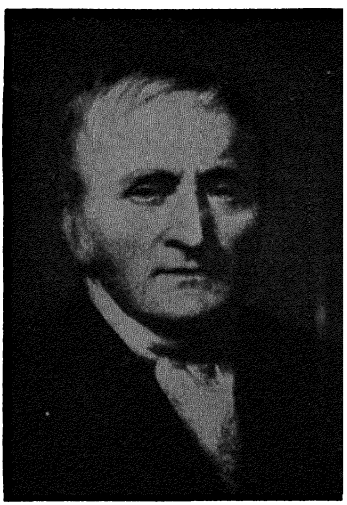
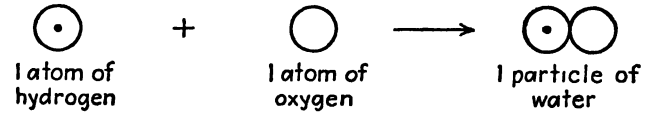


FIG. 93. *John Dalton (1766–1844).*

The atomic theory emerged from Dalton's crude attempts to picture to himself the ultimate particles of gases. Like a few physicists before him and many after, he had found in the "atoms" of Democritus and Lucretius a ready explanation for the more obvious characteristics of gases. His interest inclined more to chemical behavior than to physical properties, and he was seeking a way to express chemical reactions of gases in terms of ultimate particles. The simplest idea seemed to be that a combination of two elements represented a union of their "atoms," as Dalton christened the ultimate particles of an element. Dalton imagined the atoms to be tiny spheres and represented them in pictures by circles with various designs to distinguish different elements. Thus the reaction between hydrogen and oxygen to form water could be drawn



The decomposition of mercuric oxide was represented by

If the atoms are further imagined to have unchangeable weights, the law of conservation of mass follows immediately, since a reaction involves no change in the number of atoms but merely a joining together or separation.

Dalton had not been toying long with these diagrams when Proust completed his demonstration of the law of definite proportions. This changed the atomic idea at once from an idle speculation to a promising hypothesis; for definite compositions were precisely what one would expect if chemical combination took place as Dalton imagined. *If* every particle of mercuric oxide, for instance, is composed of one atom of mercury and one of oxygen, and *if* a mercury atom weighs, say, twelve times as much as an oxygen atom, *then no matter how much mercuric oxide is present, there must be just twelve times as much mercury as oxygen.* Furthermore, mercuric oxide from different localities, or prepared in different ways, should have the same composition, since its properties depend on its having one mercury atom joined with every oxygen atom. Fired by his success in explaining this quantitative law, Dalton went on to apply his theory to other numerical relationships among chemical compounds. In almost every case predictions from the theory agreed with observed facts: atoms seemed capable of explaining far more than he had anticipated. Dalton's crude diagrams had led him to a magnificent discovery.

Assumptions of the Atomic Theory

These are the principal assumptions which make up Dalton's atomic theory (some are not quite in the form which he used):

1. Every element consists of tiny particles called atoms.
2. The atoms of any one element are all exactly alike.
3. Atoms are indestructible: they cannot be divided, created, or destroyed.
4. When two or more elements unite to form a compound, their atoms join together to form molecules of the compound.
5. In general, atoms combine in simple ratios: thus one atom of an element *A* may combine with one atom of *B*, or with two atoms, or with three atoms, but not with large numbers of atoms.
6. If two elements can unite to form more than one compound, molecules of the most stable and abundant compound contain only one atom of each element. (This assumption is now known to be wrong.)

The first four assumptions are sufficient to explain the laws of definite proportions and conservation of mass, as we have seen in the preceding section. The fifth assumption is justified by the simple predictions which can be made from it regarding weight relations. The sixth assumption was a guess, justified only by convenience. Dalton's reasons for these two assumptions can be made clear by a few simple examples.

TABLE VII

	Carbon	*Hydrogen*	*Sulfur*	*Oxygen*	*Atomic ratios (Dalton)*	*Atomic ratios (modern)*
Carbon monoxide	3 g.			4 g.	(1:1)	1:1
Water		1 g.		8 g.	(1:1)	2:1
Methane	3 g.	1 g.			1:2	1:4
Hydrogen sulfide		1 g.	16 g.		(1:1)	2:1
Sulfur dioxide			1 g.	1 g.	1:2	1:2
Carbon disulfide	3 g.		16 g.		1:2	1:2

The first four columns of Table VII show the *experimentally determined* compositions of six compounds, each containing only two elements. Thus carbon monoxide contains 3 g. of carbon to every 4 g. of oxygen; methane is a compound of carbon and hydrogen containing 3 g. of carbon for every 1 g. of hydrogen. (We are using approximately correct numbers according to modern standards, not the inaccurate weights which Dalton used.) According to Dalton's theory, we can analyze these figures as follows. We *assume* for some of the compounds that each of their molecules contains

only one atom of each constituent element (by assumption 6, above.) Following Dalton, we make this guess for carbon monoxide, water, and hydrogen sulfide. Now if the oxygen in water weighs 8 times as much as the hydrogen (by experiment), and if every molecule contains one atom of oxygen and one atom of hydrogen (by assumption), then an oxygen atom must weigh 8 times as much as a hydrogen atom. Similarly, since the carbon in carbon monoxide weighs only $\frac{3}{4}$ as much as the oxygen, our assumption suggests that a carbon atom weighs $\frac{3}{4}$ as much as an oxygen atom. A hydrogen atom weighs $\frac{1}{8}$ as much as an oxygen atom, a carbon atom $\frac{3}{4}$ as much: hence a carbon atom must weigh 6 times as much as a hydrogen atom ($6 \times \frac{1}{8} = \frac{3}{4}$). Now methane, a compound of carbon and hydrogen, contains 3 times as much carbon as hydrogen (experimentally) —which suggests that every methane molecule contains two atoms of hydrogen and one of carbon. This is a simple ratio (2:1), in agreement with assumption 5.

Using now the assumption of a 1:1 atomic ratio for hydrogen sulfide, we find from the weights of hydrogen and sulfur that a sulfur atom must weigh sixteen times as much as a hydrogen atom. An oxygen atom, remember, weighs eight times as much, so that a sulfur atom must be twice as heavy as an oxygen atom. In sulfur dioxide these two elements are combined in equal proportions by weight; hence a molecule of sulfur dioxide must contain two atoms of oxygen to one of sulfur—again the simple ratio 2:1. Exactly similar reasoning applied to carbon disulfide shows that here again the atoms are united in a simple 2:1 ratio.

TABLE VIII

	Carbon	*Oxygen*	*Hydrogen*	*Nitrogen*	*Atomic ratios (Dalton)*	*Atomic ratios (modern)*
Carbon monoxide	3 g.	4 g.			1:1	1:1
Carbon dioxide	3 g.	8 g.			1:2	1:2
Ethylene	6 g.		1 g.		1:1	2:4
Methane	3 g.		1 g.		1:2	1:4
Nitrous oxide		4 g.		7 g.	2:1	2:1
Nitric oxide		8 g.		7 g.	(1:1)	1:1
Nitrogen dioxide		16 g.		7 g.	1:2	1:2
Nitrogen trioxide		12 g.		7 g.		2:3
Nitrogen pentoxide		20 g.		7 g.		2:5

These atomic ratios are tabulated in the fifth column of Table VII, parentheses indicating those used as assumptions. Today we know that

two of the assumptions are wrong, those for water and hydrogen sulfide. Use of modern figures for these two gives the values in the sixth column. In spite of different initial assumptions, the numbers of the sixth column are still simple ratios of small whole numbers, as assumption 5 predicts they should be.

Table VIII exhibits in even clearer fashion the strikingly simple relationships among the weights of elements in various compounds. Three groups of compounds are represented, each group containing compounds of a single pair of elements. Again the first four columns show *experimentally determined* compositions. The fifth column contains the atomic ratios as calculated by Dalton from the assumptions listed in Table VII and from the one additional assumption that molecules of nitric oxide contain one oxygen atom and one nitrogen atom apiece. In the sixth column are atomic ratios found by modern methods. Again molecules of these compounds show simple ratios among their constituent atoms.

Examples like these, which could be multiplied indefinitely, show the reasons back of the fifth assumption on page 168. The weights of elements in different compounds are related in a strangely simple manner, which is neatly explained by simple combinations of Dalton's indestructible atoms. The sixth assumption, however, is not essential. Dalton might have assumed for any of his "stable and abundant" compounds a ratio of 1:2, or 2:3; this would have changed some of the numbers in other ratios but would not have disturbed their simplicity. Dalton realized this, of course, but since he had no way of knowing the true ratios he chose the simplest possible assumption. Today we can establish these ratios by a variety of methods which were unknown in Dalton's time.

It was numbers, relations among weights, which led to the atomic theory. In our day we have other evidence for the existence of atoms, but Dalton rested his case solely on weights of elements and compounds which could be measured with a balance. Comparison of these weights seemed to show a numerical order in the structure of matter: the materials of the earth were not put together haphazardly, but according to mathematical rules as rigid and precise as the rules which Kepler had found in the motions of the planets. Dalton was wise enough, or fortunate enough, or both, to see that this amazing numerical orderliness betrayed the presence of tiny, indestructible atoms, joined together in simple patterns, as the building blocks of all matter.

Avogadro's Law

Successful as the atomic theory appeared to be, it encountered grave difficulties which delayed its universal acceptance for many years. Two of these difficulties were outstanding.

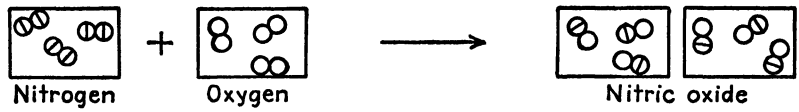
First, there was the awkward assumption which we labeled 6 on page 168. Here, at the heart of the theory, was a statement with no justification beyond Dalton's wish for simplicity. Until this assumption could be proved or disproved the actual number of atoms present in a molecule was anybody's guess.

Second, there was trouble with volume relations in reactions among gases. A French scientist, Gay-Lussac, had shown that when gases are consumed or produced in a chemical reaction, the volumes of the gases are related by simple whole numbers. Thus when water is decomposed into its elements, the volume of hydrogen is just *twice* the volume of oxygen; when nitrogen and oxygen unite to form nitric oxide (a colorless gas), *one* volume of nitrogen and *one* volume of oxygen produce *two* volumes of nitric oxide. Here were some simple numerical relations which the atomic theory ought to explain but did not. Nitric oxide, for instance, according to Dalton should contain one atom of nitrogen and one of oxygen in each molecule. This meant that the nitric oxide molecule should be larger than the atom of either element. But the particles of a gas were presumably far apart, and the volume of a gas should be determined primarily by the distances between its particles rather than by their sizes. When the atoms in a given volume of oxygen paired off with nitrogen atoms to form the compound, the number of particles of the compound should be the same as the number of the original oxygen atoms; the volume which these particles would occupy might be slightly greater than that of the oxygen, but why so much greater? In particular, why *exactly twice* as great? Baffled by such questions, Dalton ungenerously concluded that the Frenchman's measurements were at fault—though Gay-Lussac was one of the most skillful experimenters of his time.

Both of these difficulties with Dalton's theory were settled in one brilliant stroke by a young Italian physicist, Avogadro—or would have been settled, if the chemists of Europe had listened to him. Impressed with the quantitative similarities in the behavior of different gases (Boyle's law and Charles's law, for instance) and with the simple volume relations discovered by Gay-Lussac, Avogadro made the happy guess that *equal volumes of all gases under the same conditions of temperature and pressure contain the same number of molecules.* Avogadro suggested further that in some elements his molecules might be identical with atoms, but that *in other elements, as well as in compounds, each molecule might consist of several atoms.*

Avogadro's idea not only opened the possibility of molecules containing several atoms for elements, but, coupled with Gay-Lussac's work, showed how the number of atoms per molecule could be found. Consider, for instance, the formation of nitric oxide, in which two volumes of

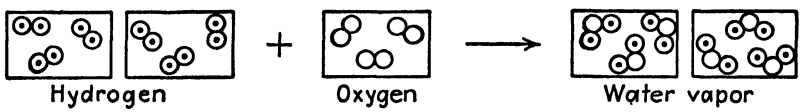
product are obtained from one volume each of nitrogen and oxygen:



Since each volume represented contains the same number of molecules (Avogadro's hypothesis), a given number of oxygen molecules must react with the *same* number of nitrogen molecules to give *twice* as many molecules of nitric oxide. A thousand oxygen molecules would give 2,000 nitric oxide molecules; 10 oxygen molecules would give 20 of nitric oxide; 1 oxygen molecule would give 2 of nitric oxide. Now each nitric oxide molecule contains some oxygen; hence the oxygen molecule must have split, half of it going to each molecule of product. Avogadro interpreted this deduction to mean that an *oxygen molecule consists of at least two atoms.*

Similar reasoning applied to other reactions shows that oxygen molecules apparently often split into two parts, but *never into more than two.* We may safely infer that each molecule has no more than a pair of atoms. Also for the other common gaseous elements—nitrogen, hydrogen, chlorine—Gay-Lussac's volume relationships suggest molecules made up of two atoms apiece.

To see how Avogadro's hypothesis makes possible the determination of atomic ratios for Dalton's "stable and abundant" compounds, consider the formation of water from its elements. Gay-Lussac had shown that at temperatures above 100°C two volumes of hydrogen reacting with one volume of oxygen produce two volumes of water vapor:



In terms of molecules, a given number of hydrogen molecules react with half as many oxygen molecules to give the same number of water molecules. Two hydrogen molecules plus one oxygen molecule give two water molecules. Now each of the original molecules contains two atoms apiece; hence four hydrogen atoms and two oxygen atoms go into the building of the two water molecules. Accordingly each water molecule must contain two hydrogen atoms and one oxygen atom, a 2:1 ratio instead of the 1:1 ratio which Dalton assumed.

So Avogadro's simple idea about the structure of molecules cleared away the outstanding difficulties in Dalton's theory, and showed the relationship between the molecules of physics and the atoms of chemistry. But Dalton, strangely, did not see the importance of this addition

to his theory. Stubbornly he clung to his indivisible atoms as the ultimate particles of gases. Dalton was famous, Avogadro all but unknown; science for once followed reputation rather than reason, and the younger man's work was forgotten. Not until 1860, fifty years after Avogadro had made his suggestion, did chemists open their eyes to its value. In those fifty years, difficulties in the atomic theory had grown so formidable that even the existence of Dalton's atoms was seriously questioned. But once the distinction between molecules and atoms became clearly recognized, discrepancies in the theory vanished as if by magic, and atoms were henceforth accepted almost universally as a fundamental part of chemistry.

Avogadro's idea that the volume of a gas at any given temperature and pressure is determined only by the number of molecules present, in his day merely a daring hypothesis, has been so well confirmed in recent years that today we regard it as an established law.

Atomic Weights

As we found in the discussion of weight relations on pages 168 to 170, the composition of a compound enables us to state how much heavier one kind of atom is than another, provided we know how many atoms of each make up a molecule of the compound. Thus the composition of carbon monoxide given in Table VII, together with the assumption that its atoms are combined in a 1:1 ratio, indicates that a carbon atom weighs three-fourths as much as an oxygen atom. Dalton's 1:1 ratio for the hydrogen and oxygen atoms of water suggests that an oxygen atom weighs eight times as much as a hydrogen atom; but our deduction from Avogadro's law that two hydrogen atoms are present for every oxygen atom shows that an oxygen atom is actually sixteen times as heavy as a hydrogen atom. Similarly from the figures of Table VII (using the "modern" atomic ratios) we may conclude that an atom of carbon weighs twelve times as much as a hydrogen atom, a sulfur atom thirty-two times as much.

Relative weights of this sort, *numbers expressing how much heavier one atom is than another*, are called ***atomic weights.*** They are useful in characterizing different elements, particularly in discussions where the comparative lightness or heaviness of different particles is important. They tell us nothing, of course, about the *actual* weights of the atoms; these actual weights are exceedingly small numbers, not generally useful because of the awkwardness of working with long decimals.

In Chap. II (page 30) we used a similar scheme for dealing with the weights of the planets. Here the actual weights are awkwardly large instead of awkwardly small. For purposes of comparison, we described the weight of a planet as so many times heavier or lighter than the earth's

weight. In Table I, for instance, the number 318 for Jupiter is not a real mass or weight, but means that Jupiter is 318 times as heavy as the earth; the number 0.8 for Venus indicates that Venus has four-fifths the weight of the earth, and so on.

For expressing the weights of the planets, the earth's weight is a convenient standard. To express atomic weights conveniently, we must similarly select some one element as a standard for comparison. Following Dalton, we might choose hydrogen, assuming for its atom the weight 1; then the atomic weight of oxygen must be 16, that of carbon 12, that of sulfur 32—these numbers not indicating a certain number of grams or pounds, but simply that an oxygen atom weighs sixteen times as much as a hydrogen atom, etc. For several reasons, however, the modern atomic weight scale uses oxygen as a standard rather than hydrogen, giving it a weight of exactly 16. Accurate determinations show that the hydrogen atom weighs a little more than one-sixteenth as much as the oxygen atom, so on the oxygen scale the atomic weight of hydrogen is not 1 but 1.0081.

Atomic weights of the known elements on this scale are given in Table IX. They range from 1.0081 for hydrogen to 239 for the heavy metal plutonium. One curious fact shown by this table, whose significance we shall discuss on a later page, is that *atomic weights for many elements are very nearly whole numbers.*

Experimentally the atomic weight of an element may be determined from the composition by weight of its compound with oxygen, hydrogen, or some other common element, provided that the ratio in which the atoms combine is known. For gaseous elements a simpler method is based on Avogadro's law: since equal volumes of two gases under the same conditions have the same number of molecules, a comparison of the weights of equal volumes gives a direct comparison of the weights of their molecules. For example, 1 l. of nitrogen weighs seven-eighths as much as 1 l. of oxygen under the same conditions, so that a nitrogen molecule must be seven-eighths as heavy as an oxygen molecule; each molecule contains two atoms, so a nitrogen atom must be likewise seven-eighths as heavy as an oxygen atom. Since oxygen has an atomic weight of 16, the atomic weight of nitrogen must be 14.

The *molecular weight* of an element or compound is the weight of a molecule on the atomic-weight scale—*i.e.*, a number expressing how much heavier or lighter the molecule is than an oxygen atom. Since an oxygen molecule contains two atoms, the molecular weight of oxygen is 32; the molecular weight of water, with two atoms of hydrogen and one of oxygen per molecule, is 1.008 + 1.008 + 16 or 18.016; the molecular weight of carbon monoxide is 12 + 16 or 28. Experimentally one easy method of determining molecular weights is by a comparison of the

TABLE IX. ATOMIC WEIGHTS

	Symbol	*Atomic weight*	*Atomic number*		*Symbol*	*Atomic weight*	*Atomic number*
Aluminum	Al	26.97	13	Neodymium	Nd	144 27	60
Antimony	Sb	121.76	51	Neon	Ne	20 183	10
Argon	A	39.944	18	Neptunium	Np	239	93
Arsenic	As	74.91	33	Nickel	Ni	58 69	28
Barium	Ba	137.36	56	Nitrogen	N	14 008	7
Beryllium	Be	9.02	4	Osmium	Os	190 2	76
Bismuth	Bi	209 00	83	Oxygen	O	16 000	8
Boron	B	10 82	5	Palladium	Pd	106 7	46
Bromine	Br	79.916	35	Phosphorus	P	30 98	15
Cadmium	Cd	112.41	48	Platinum	Pt	195 23	78
Calcium	Ca	40 08	20	Plutonium	Pu	239	94
Carbon	C	12.010	6	Potassium	K	39 096	19
Cerium	Ce	140.13	58	Praseodymium	Pr	140 92	59
Cesium	Cs	132.91	55	Protactinium	Pa	231	91
Chlorine	Cl	35.457	17	Radium	Ra	226 05	88
Chromium	Cr	52 01	24	Radon	Rn	222	86
Cobalt	Co	58 94	27	Rhenium	Re	186 31	75
Columbium	Cb	92.91	41	Rhodium	Rh	102 91	45
Copper	Cu	63 57	29	Rubidium	Rb	85 48	37
Dysprosium	Dy	162 46	66	Ruthenium	Ru	101.7	44
Erbium	Er	167.2	68	Samarium	Sm	150 43	62
Europium	Eu	152.0	63	Scandium	Sc	45 10	21
Fluorine	F	19.00	9	Selenium	Se	78 96	34
Gadolinium	Gd	156.9	64	Silicon	Si	28.06	14
Gallium	Ga	69.72	31	Silver	Ag	107.880	47
Germanium	Ge	72.60	32	Sodium	Na	22.997	11
Gold	Au	197.2	79	Strontium	Sr	87.63	38
Hafnium	Hf	178 6	72	Sulfur	S	32.06	16
Helium	He	4 003	2	Tantalum	Ta	180 88	73
Holmium	Ho	164 94	67	Tellurium	Te	127 61	52
Hydrogen	H	1 0081	1	Terbium	Tb	159.2	65
Indium	In	114 76	49	Thallium	Tl	204 39	81
Iodine	I	126.92	53	Thorium	Th	232.12	90
Iridium	Ir	193 1	77	Thulium	Tm	169.4	69
Iron	Fe	55 85	26	Tin	Sn	118 70	50
Krypton	Kr	83.7	36	Titanium	Ti	47 90	22
Lanthanum	La	138 92	57	Tungsten	W	183 92	74
Lead	Pb	207 21	82	Uranium	U	238 07	92
Lithium	Li	6.940	3	Vanadium	V	50.95	23
Lutecium	Lu	174 99	71	Xenon	Xe	131.3	54
Magnesium	Mg	24 32	12	Ytterbium	Yb	173 04	70
Manganese	Mn	54.93	25	Yttrium	Y	88 92	39
Mercury	Hg	200 61	80	Zinc	Zn	65.38	30
Molybdenum	Mo	95 95	42	Zirconium	Zr	91.22	40

weights of equal volumes of gases, like that described in the last paragraph: thus a given volume of carbon dioxide weighs $1\frac{1}{8}$ as much as an equal volume of oxygen, so that its molecular weight must be $1\frac{1}{8} \times 32$ or 44.

So the atomic theory makes possible a determination of the comparative weights of the tiny particles of matter, both atoms and molecules, from weights and volumes which we can measure in the laboratory.

Symbols and Formulas

To make the atomic theory useful and readily understandable, some means was needed for representing pictorially the atoms and their combinations. Dalton filled this need in a measure with his little patterned circles, but his diagrams for complex molecules were cumbersome and confusing. A far better system (although Dalton never admitted its merits) was invented by the Swedish chemist Berzelius, born thirteen years later than Dalton, to whom chemistry also owes the first accurate determination of atomic weights.

In Berzelius's scheme, which is in common use today, an atom of an element is represented by an abbreviation of the element's name. For many elements the first letter is used: an atom of oxygen is O, an atom of hydrogen H, an atom of carbon C. When the names of two elements begin with the same letter, two letters are used in the abbreviation for one or both: Cl stands for an atom of chlorine, He for helium, Zn for zinc. For some elements abbreviations of Latin names are used: a copper atom is Cu (cuprum), an iron atom Fe (ferrum), a mercury atom Hg (hydrargyrum). These abbreviations are called *symbols* of the elements.

Two or more atoms joined to form a molecule Berzelius represented by writing their symbols side by side: a carbon monoxide molecule is CO, a zinc sulfide molecule ZnS, a mercuric oxide molecule HgO. When a molecule contains two or more atoms of the same kind, a small subscript indicates the number present: the familiar expression H_2O means that a molecule of water contains two H atoms and one O atom; a molecule of oxygen, containing two O atoms, is written O_2; a molecule of carbon tetrachloride (CCl_4) contains one C atom and four Cl atoms; a molecule of nitrogen pentoxide (N_2O_5) contains two N atoms and five O atoms. Each subscript applies only to the symbol immediately before it. These expressions for molecules are called *formulas*.

Symbols and formulas give the chemist a convenient shorthand for expressing the joining together and separating of atoms which, according to Dalton's theory, are the fundamental processes of chemical change.

Modern Ideas about Atoms

Like any good scientific theory, the atomic theory has undergone

We have already noted the changes made necessary by the resurrection of Avogadro's hypothesis, and the improvement in pictorial representation introduced by Berzelius. More modern changes in viewpoint strike at the basic assumptions of the theory. Here we can only list the revisions which modern research has made necessary; their full discussion must wait until we have gained some acquaintance with electricity.

We no longer believe that the atoms of an element are necessarily all alike, for some elements have been separated into substances (isotopes) with slightly different properties. We no longer consider atoms indestructible; some heavy atoms disintegrate into lighter atoms spontaneously (radioactivity), and in modern "atom smashers" changes from one atom to another can be produced artificially. We no longer believe that combination of atoms in simple ratios is universally necessary, for ratios are anything but simple in the complex molecules of organic chemistry. These refinements of viewpoint, dealing with materials and processes outside the realm of elementary chemistry, cast no reflections, of course, on the usefulness and near correctness of Dalton's assumptions for *ordinary processes* and *ordinary materials*. We may safely ignore them for the present, simply keeping them in the backs of our minds for future reference.

Changes of opinion in modern times regarding the composition of ordinary liquids and solids cannot be so easily ignored. The units of structure in liquids and solids are not always, as Dalton believed, individual atoms or molecules, but may be aggregates of several molecules or electrified fragments of molecules. We cannot say correctly, for instance, that an atom of zinc reacts with an atom of sulfur to give a molecule of zinc sulfide; a "molecule of zinc sulfide" does not exist in the solid state, for crystalline zinc sulfide consists of a network of electrically charged zinc and sulfur particles. We had best examine the three states of matter in some detail, to see in each case with what sort of particles we must deal, and what meaning we can attach to Berzelius's formulas.

For *gases*, we believe with Avogadro that the ultimate particles are molecules. The molecules may consist of one or more atoms: in mercury vapor each "molecule" is a single atom, and we write the formula of mercury vapor accordingly Hg; in hydrogen each molecule has two atoms, and its formula is H_2; the formula CO_2 for carbon dioxide represents a molecule with one atom of carbon and two of oxygen. In gases each formula is, so to speak, a diagram of a kind of particle which we believe actually to be present.

Some *liquids* are made up of discrete molecules; for example, the particles of carbon tetrachloride are molecules, and its formula (CCl_4) represents a molecule. Other liquids, like molten salt and molten metals, consist of electrified particles. Still others, like water and alcohol, contain aggregates whose size varies with the temperature. The structure of

many liquids is yet imperfectly known. Obviously, the assignment of formulas to these complex liquids presents difficulties. Rather than try to make the formulas fit the structures, chemists generally dodge the problem by using *the formulas of the corresponding vapors.* The vapor of ordinary salt consists of molecules with the formula NaCl (Na is an atom of the metal sodium), so NaCl is used also for molten salt; the formula Hg for mercury vapor is retained for the liquid; the formula for water vapor (H_2O) is applied to liquid water as well.

In some *solids* (solid carbon dioxide, sugar) molecules may be regarded as the structural units, and their formulas (CO_2, $C_{12}H_{22}O_{11}$) refer to distinct particles in Berzelius's original sense. A few solid elements, like sulfur and carbon (diamond), consist of crystal lattices made up of atoms, and their formulas are therefore the symbols of the elements (S for sulfur, C for diamond). Many other solids (salts of all kinds, metals, metallic oxides) have crystalline structures with electrically charged particles as their building blocks. For these we assign formulas on the same basis as for liquids, if the formulas of the vapors are known: ordinary table salt has the formula NaCl, solid mercury the formula Hg, solid zinc the formula Zn. If the formula of the vapor cannot be determined, the formula of the solid merely represents the relative number of atoms which have combined to form it. Thus mercuric oxide decomposes before vaporizing appreciably, so we have no direct means of determining its formula; we assign to it the formula HgO simply because we know that mercury and oxygen combine atom for atom to produce it.

The formula for a liquid or solid, therefore, *may or may not represent a distinct kind of particle;* the only definite information it gives is the relative number of atoms which have combined to make the substance. The formula for a gas, on the other hand, *always represents the composition of an actual molecule.*

The atomic theory makes possible all manner of predictions about the amounts of substances which will react together or which can be produced by reaction, and it gives us a deep insight into the processes of chemical change. Some important questions, however, it leaves wholly untouched. The matter of energy, for instance: why should some reactions liberate so much energy spontaneously, while others do not take place unless we supply energy? A related question concerns the nature of the forces which hold atoms together in a compound. Again, why are atoms so particular about which others they will combine with—why, for instance, will zinc combine so much more readily with sulfur than with nitrogen or with carbon? And why the peculiar numerical relations in which atoms combine—why should four atoms of hydrogen combine with one of carbon, two atoms of hydrogen with one of oxygen, one atom of hydrogen with

one of chlorine? In attempting to answer such questions we shall find Dalton's theory of little assistance.

Questions

1. Starting with the "modern" atomic ratios for water and carbon monoxide, show from the data of Table VII that a methane molecule contains four hydrogen atoms and one carbon atom. What is the formula of methane gas?
2. Starting with the "modern" atomic ratios for carbon monoxide and sulfur dioxide, show that a carbon disulfide molecule contains two atoms of sulfur and one of carbon. What is the formula of carbon disulfide gas?
3. Find the atomic weight of nitrogen from the data of Table VIII.
4. Analysis of ammonia, a compound of nitrogen and hydrogen, shows that 28 g. of nitrogen are combined with every 6 g. of hydrogen. Each nitrogen atom is 14 (approximately) times as heavy as a hydrogen atom (from atomic weight table). What is the ratio of nitrogen to hydrogen atoms in ammonia?
5. One volume of hydrogen reacts with one volume of chlorine gas to give two volumes of a gas called hydrogen chloride. Show from these figures that each hydrogen molecule must contain at least two atoms.
6. What is the molecular weight of a gas if 1 l. of it weighs nine-sixteenths as much as 1 l. of oxygen (weighed under similar conditions)?
7. From the atomic weight table and Table VII, find the molecular weight of methane, sulfur dioxide, and carbon disulfide (use "modern" ratios).
8. How many molecules of water vapor could be made from two molecules of hydrogen? How many liters of water vapor from 2 l. of hydrogen (both at 200°C and 1 atm pressure)? How many grams of water from 2 g. of hydrogen?
9. How many atoms of carbon must be burned to make eleven molecules of carbon dioxide? How many grams of carbon to make 11 g. of carbon dioxide?
10. What is the symbol for hydrogen? What is the formula for hydrogen gas? What does each represent?
11. The following are formulas of gases. State in words what each one means.

$$HCl \quad N_2O \quad CH_3Cl \quad H_2S \quad CH_2O \quad C_2N_2 \quad C_3H_8 \quad Cl_2$$

(EXAMPLE: HCl represents a molecule containing one atom of hydrogen and one atom of chlorine.)

CHAPTER XIV

The Language of Chemistry

BEGINNING the study of chemistry is like entering a foreign land: a land where the scenery consists of strange objects called atoms and molecules, and where the King's English is distorted into a new language of unfamiliar words and symbols. We now turn briefly from the scenery of chemistry to its language, for here, as in any other foreign country, one feels truly at home only when he can speak to the inhabitants in their own tongue.

In physics also we learned a new language, but it was a language of familiar words given new meanings. In the language of chemistry the words themselves are unfamiliar. Such an ordinary material as salt becomes sodium chloride; lime becomes calcium oxide; alum becomes hydrated potassium aluminum sulfate. There is no intent to mystify in this relabeling of commonplace substances. Rather, the chemist feels that these artificial words make his subject infinitely clearer than it could possibly be otherwise. He must deal with a vast multitude of different materials, and he tries to describe them with words which will give some hint as to their composition and relationships.

The most familiar of the chemist's terms are his names for the elements: iron, sulfur, mercury, oxygen are words with which everyone has some acquaintance. We shall seek first to give a more precise meaning to some of these names of elements, by describing briefly the properties of the elements. Next we shall see how these names are altered and put together to make the names of compounds. Finally we shall discuss the shorthand methods by which results of chemical changes may be concisely described.

Five Common Elements

In Chap. XII the element oxygen was briefly described. Here we set down similar thumbnail sketches of five other elements, whose compounds will figure prominently in later discussions.

Hydrogen. Lightest of all substances: density 0.00009 g./cc at 0°C and 1 atm pressure. A colorless, tasteless, odorless gas. Boiling point −253°C (20°K), freezing point −259°C (14°K). Atomic weight 1.008. Symbol H, formula at ordinary temperatures H_2.

Hydrogen is a fairly abundant element, making up about 1 per cent by weight of the earth's crust. Most of it is combined with oxygen in water. In compounds with carbon and oxygen, hydrogen is present in all animal and vegetable tissue. Free hydrogen, uncombined with other elements, is very scarce; it sometimes occurs as a minor constituent of volcanic gases and of natural gas.

In the laboratory, hydrogen is commonly prepared (1) by the reaction between certain metals and water or acids and (2) by the electrolysis of water. A convenient apparatus for the first method is shown in Fig. 94;

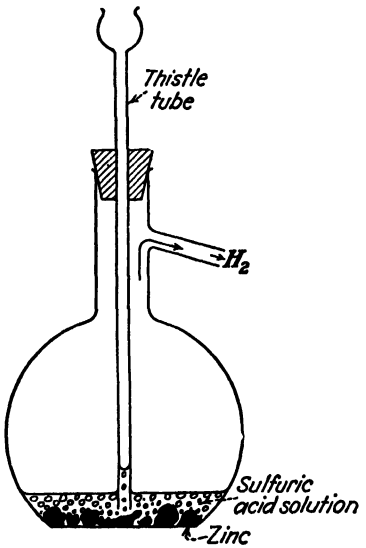


FIG. 94. *Preparation of hydrogen by the reaction between a metal and an acid.*

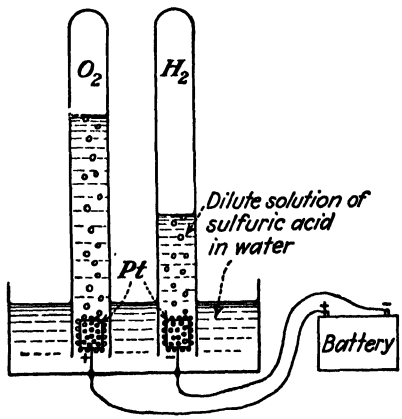


FIG. 95. *Electrolysis of water.*

acid is poured down the thistle tube onto pieces of metal in the flask, and the gas bubbles off steadily as long as both acid and metal are present. Zinc is often used as the metal and sulfuric acid as the acid; the reaction between these two may be written:

$$\text{Zinc} + \text{sulfuric acid (sol'n)} \rightarrow \text{hydrogen} + \text{zinc sulfate (sol'n)}$$

In words, this statement is read, "zinc added to a solution of sulfuric acid gives hydrogen and a solution of zinc sulfate." After reaction the zinc sulfate is not visible, but may be obtained as a white, crystalline solid by evaporating the remaining liquid. The second method for pre-

paring hydrogen, *electrolysis*, implies the passage of an electric current through a fluid, with resulting decomposition of the fluid (Fig. 95). Here the fluid is water, made a conductor by addition of a little acid or alkali, and the current passes between small platinum plates (marked "Pt") connected to a battery or generator. Hydrogen bubbles collect at one plate, oxygen bubbles collect at the other, and each gas rises to the top of its tube. As the gases accumulate, the volume of hydrogen remains always twice as great as the volume of oxygen (page 171). This reaction may be summarized

$$\text{Water} \rightarrow \text{hydrogen} + \text{oxygen}$$

Such a statement says nothing about how the reaction was carried out but merely describes the chemical change which has occurred.

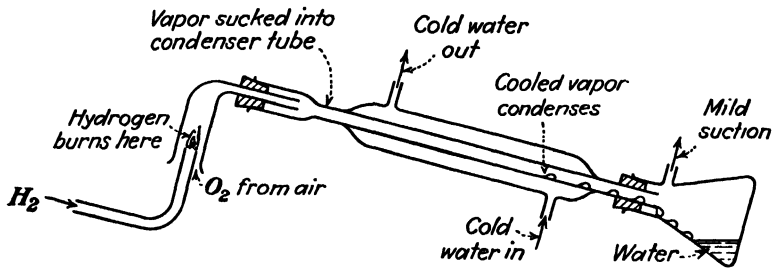


FIG. 96. *Experiment to demonstrate that water is formed when hydrogen burns in air.*

Hydrogen burns readily in air or oxygen with a hot, colorless flame. The gas produced in the flame is water vapor, as may be shown by condensing some of it (Fig. 96). This reaction is the reverse of the decomposition of water by electrolysis:

$$\text{Hydrogen} + \text{oxygen} \rightarrow \text{water}$$

A mixture of hydrogen and oxygen will not react at ordinary temperatures, but once the mixture is *ignited* by the heat of a flame or an electric spark, the reaction generates sufficient heat to keep itself going. If the gases are mixed in certain proportions, an explosion results, owing to the sudden expansion of the mixture as it is heated by the reaction. Miniature explosions produced by bringing a flame near hydrogen-air mixtures in a test-tube give a convenient method of testing for the element.

One other compound of hydrogen and oxygen may be prepared indirectly. This is *hydrogen peroxide* (H_2O_2), an explosive liquid when pure, commonly used in dilute solution as a disinfectant and bleaching agent. Its formula is written H_2O_2 rather than HO because its molecular weight, determined by comparison of the density of its vapor with that of oxygen (and in other ways), is 34 and not 17.

Hydrogen is an active element at moderately high temperatures, combining directly with a number of other nonmetallic elements besides oxygen and less readily with several of the metals. Thus with nitrogen it forms *ammonia* (NH_3); with chlorine, *hydrogen chloride* (HCl); with calcium, *calcium hydride* (CaH_2). With many metallic oxides hydrogen reacts to form water and the free metal; thus

Copper oxide (CuO) + hydrogen (H_2) → copper (Cu) + water (H_2O)

In a reaction of this sort, where oxygen is removed from combination with a metal, the oxide is said to be *reduced*. Reduction is the opposite process to oxidation.

The low density of hydrogen makes it useful for filling balloons and dirigibles, although its inflammability is a constant source of danger. The intense heat produced when hydrogen burns in oxygen is made use of in cutting and welding metals with the oxyhydrogen blowtorch. Hydrogen is a principal constituent of artificial gas fuels. The hardening of oils to form solid fats and the production of synthetic ammonia and wood alcohol are among the other commercial uses of the element.

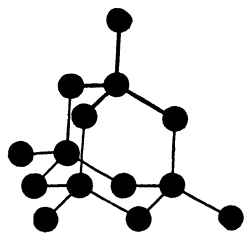


FIG. 97. *The arrangement of carbon atoms in a diamond crystal. The unit of structure is a tetrahedron consisting of four atoms around a central atom; one of these units is shown by heavy lines.*

Carbon. (Atomic weight 12. Symbol C, formula C.) Diamond and graphite are two naturally occurring forms of the pure element carbon. Diamond is the hardest known natural substance, clear and colorless when strictly pure, not breaking easily in any direction, a very poor conductor of electricity. Graphite is soft, opaque, steel gray to black, occurring in tiny flakes which split apart easily, a fairly good electrical conductor. Ordinary carbon in the form of coke, soot, and charcoal is highly impure graphite in minute crystals. Both diamond and graphite are extremely resistant to heat, vaporizing appreciably only at temperatures near 3500°C. Carbon has never been liquefied.

That two materials as different in properties as diamond and graphite can be forms of a single element seems at first incredible. The fact may be proved by burning each in oxygen, at temperatures above 700°C; both give carbon dioxide gas as the only product. The differences between the two arise from a difference in crystal structure: the carbon atoms of diamond are arranged in a compact framework in which each atom is surrounded by four others at the corners of a tetrahedron (Fig. 97); the carbon atoms of graphite lie in parallel planes, each plane made up of hexagonal rings (Fig. 98). Distances between the planes in a graphite crystal are greater than distances between the atoms within each plane,

so that the crystal splits easily in a direction parallel to the planes. In diamond the atoms are closer together (as shown by its greater density—3.5, compared with 2.2 for graphite), and each atom is separated from the four that surround it by equal distances; hence diamond is not easily split in any direction.

Many other solids besides carbon, both elements and compounds, can exist in two or more different forms, the properties of each depending on the particular arrangement of particles in its crystal lattice.

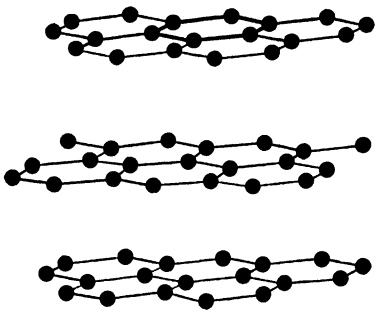


FIG. 98. *The arrangement of carbon atoms in the crystal lattice of graphite. The unit of structure is a hexagonal ring of six atoms; one of these units is shown by heavy lines.*

Inactive at ordinary temperatures, carbon at high temperatures reacts readily with many substances. It burns in air or oxygen, forming the gas *carbon dioxide* if abundant oxygen is available, the gas *carbon monoxide* if the oxygen supply is limited. Carbon combines slowly with a few other nonmetallic elements: with sulfur to form the volatile, inflammable liquid carbon disulfide (CS_2), with hydrogen to form the gas methane (CH_4) and related compounds. Some metals react with carbon to form solids called carbides, such as calcium carbide (CaC_2) and iron carbide (Fe_3C). Like hydrogen, hot carbon reduces the oxides of many metals; for instance,

Zinc oxide (ZnO) + carbon (C) → zinc (Zn) + carbon monoxide (CO)

The two oxides of carbon, carbon monoxide and carbon dioxide, are colorless, odorless, tasteless gases. One or the other, depending on the amount of oxygen available, is formed not only from the burning of carbon but from the burning of any carbon compound. Slow oxidation of carbon compounds in the bodies of animals produces carbon dioxide, which is exhaled from the lungs. The monoxide is inflammable and forms an important constituent of artificial gas fuels; it is a deadly poison, especially dangerous because it has no odor. Carbon dioxide is heavy, nonpoisonous, neither inflammable nor able to support combustion. Dissolved in water under pressure, it forms ordinary "soda water." It is the gas responsible for the "rising" of bread and cakes during baking. Solid carbon dioxide, "dry ice," is used extensively in refrigeration.

Carbon is a relatively scarce element in the earth's crust, making up only about 0.03 per cent by weight of the crust's materials. But its importance to humanity is out of all proportion to its abundance. The free element in the form of coal is our most important industrial fuel. Coal and

coke are indispensable in winning many of the common metals from their ores. The compound carbon dioxide is an all-important minor constituent of air, since plant growth depends on its presence. Combined with calcium and oxygen in calcium carbonate ($CaCO_3$), carbon is a constituent of the common and useful rock limestone. Compounds of carbon with hydrogen (hydrocarbons) make up natural gas, gasoline, lubricating oils. More complex carbon compounds are the chief constituents of our bodies, of the food we eat, of the clothes we wear, of the wood from which our houses are built. Artificially produced carbon compounds include an endless variety of dyes, perfumes, explosives, drugs, plastics. In the number, variety, and importance of its compounds carbon outranks all other elements.

Sulfur. The "brimstone" of ancient times. A yellow solid, odorless and tasteless. Melts at 114.5°C, boils at 448.5°C. Density about 2 g./cc. Atomic weight 32.06. Symbol S, formula of solid S.

The common occurrence of sulfur near active volcanoes, together with the blue flame and sharp odor produced when it burns, probably explains its long literary association with the subterranean abode of deceased sinners. The free element is also found among the deposits of volcanoes long extinct and in marine deposits associated with gypsum and rock salt. Gypsum is a familiar naturally occurring sulfur compound ($CaSO_4.2H_2O$); the sulfides pyrite (FeS_2), galena (PbS), cinnabar (HgS) are others.

Liquid sulfur has the peculiar property of becoming highly viscous as its temperature is raised toward 200°C, then less viscous as the boiling point is approached. When sulfur near its boiling point is quickly cooled by pouring into water, it solidifies to a brown, pliable, elastic material called amorphous sulfur, quite different from the familiar yellow, crystalline form.

Chemically sulfur is a moderately active element. It burns readily in air or oxygen to form *sulfur dioxide* (SO_2), a gas whose odor is that often described as "the odor of sulfur." Sulfur unites with many metals on heating to form compounds called *sulfides*, for example, copper sulfide (CuS) and silver sulfide (Ag_2S). With active metals this reaction may liberate considerable energy, as in the zinc-sulfur reaction described on page 147. With hydrogen, sulfur combines to form the "rotten-egg gas," hydrogen sulfide (H_2S). Other common sulfur compounds contain both metals and oxygen: "Epsom salts," magnesium sulfate ($MgSO_4.7H_2O$); "blue vitriol," copper sulfate ($CuSO_4.5H_2O$); gypsum, calcium sulfate ($CaSO_4.2H_2O$). (The "H_2O's" in these formulas indicate water, loosely held in the solid crystals, which may be removed on mild heating.)

Industrially the most important sulfur compound is sulfuric acid, H_2SO_4, a heavy, colorless, viscous liquid, highly corrosive, dissolving readily in water with evolution of much heat. A solution of the acid is a

common laboratory reagent; among its countless industrial uses the most important are in petroleum refining and in the manufacture of fertilizers and explosives.

Chlorine. A greenish-yellow gas with a disagreeable odor. Poisonous. Boiling point −35°C, freezing point −102°C. Density at 0°C and 1 atm, 0.0032 g./cc. Atomic weight 35.46. Symbol Cl, formula at ordinary temperatures Cl_2.

Chlorine is far too active to exist free in nature. Its most abundant compound is sodium chloride (NaCl), familiarly known as ordinary salt, which occurs in solution in the ocean and in salt lakes, and in solid form as deposits of rock salt. In the laboratory chlorine is commonly prepared from sodium chloride by heating it with sulfuric acid and a black powder called manganese dioxide; industrially the gas is prepared by electrolysis of a concentrated sodium chloride solution.

Chemically chlorine is one of the most active of all elements, combining directly with all metallic elements and many nonmetallic elements. Oxygen is one of the few with which it will not react, but unstable oxides can be prepared indirectly. The more active metals burn brilliantly in chlorine, just as they do in oxygen, forming *chlorides* instead of oxides. Thus copper foil burns to form *copper chloride* ($CuCl_2$) and sodium burns with an intense yellow flame to form white clouds of tiny *sodium chloride* crystals. This spectacular reaction between the active gas chlorine and the active metal sodium, both highly poisonous, to produce the harmless substance salt is one of the most impressive examples of the profound changes in properties which chemical reaction can bring about.

A jet of hydrogen burns in chlorine as readily as in oxygen, forming a colorless gas with a sharp, unpleasant odor called *hydrogen chloride* (HCl). A mixture of hydrogen and chlorine will react at room temperature, provided that the mixture is exposed to light; if the mixture is prepared in darkness and then brightly illuminated, the reaction is explosive. For either this *photochemical* ("caused by light") reaction or the burning of hydrogen, the chemical change may be summarized

$$\text{Hydrogen} + \text{chlorine} \rightarrow \text{hydrogen chloride}$$

A solution of hydrogen chloride in water is a common strong acid called hydrochloric acid (the "muriatic acid" of commerce).

Chlorine reacts with many colored compounds of carbon, changing them to colorless compounds. This property accounts for the extensive use of the element in bleaching. It is used also as a disinfectant and in the preparation of dyes, explosives, and poison gases.

Sodium. A silver-gray solid, tarnishing quickly on exposure to air. Melting point 97.5°C, boiling point 880°C. Density 0.97 g./cc. Atomic weight 23.00. Symbol Na, formula of solid Na.

Sodium is one of the more abundant elements, making up nearly 2.5 per cent of the earth's crust, but its extreme chemical activity prevents it from occurring free in nature. Its compounds are widely distributed in rocks, soil, and in solution in bodies of water.

Sodium is so soft it can be cut like cheese, so light it will float on water, so easily corroded by air and water that it must be kept under oil. By ordinary standards these properties suggest that sodium should be called anything but a metal, yet to the chemist sodium is a metal par excellence. It does have a silvery luster on freshly cut surfaces, and it is an excellent conductor of electricity—two characteristically metallic properties. Its chemical behavior shows in an exaggerated form certain properties common to all the more active metals. Among these chemical properties are: (1) its ability to burn brightly, with evolution of much heat, in both oxygen [forming *sodium peroxide* (Na_2O_2)] and chlorine. Some nonmetallic elements will burn in oxygen or chlorine, a few in both, but the energy liberated is in general not as great as for the metals. (2) Its ability to liberate hydrogen from acids. Many common metals, for instance zinc (see page 181), iron, and aluminum, liberate hydrogen slowly from acids, but with sodium the reaction is violent and the liberation of gas exceedingly rapid. (3) Its ability to liberate hydrogen from water. This reaction produces, in addition to hydrogen, a solution of a compound called *sodium hydroxide* (NaOH).

$$\text{Sodium} + \text{water} \rightarrow \text{hydrogen} + \text{sodium hydroxide (sol'n)}$$

Several of the commoner metals, even iron, can be made to react slowly with hot water, but a little chunk of sodium need only be dropped on cold water to start a reaction which sets it skimming over the surface and generates enough heat to melt the metal.

Sodium combines readily with most of the nonmetallic elements, but not in general with other metals. Like hydrogen and carbon, it reduces many metallic oxides.

Of the simple compounds of sodium, the more familiar are ordinary salt (NaCl), washing soda or sodium carbonate (Na_2CO_3), baking soda or sodium bicarbonate ($NaHCO_3$), caustic soda or sodium hydroxide (NaOH).

How Compounds Are Named

With this descriptive background for six common elements, and a speaking acquaintance with several more, we turn now to find the meaning behind those confusing names which the chemist uses so freely—sulfate, chloride, nitrate, hydroxide, and all the rest. We shall see at the same time how formulas of compounds are built up from the symbols of their elements. To discuss the names of compounds we follow in the

footsteps of Lavoisier, to discuss their formulas in the footsteps of Berzelius.*

Order. For a compound containing a metal and one or more nonmetals, the name of the metal stands first and its symbol appears first in the formula. Thus salt is sodium chloride, and its formula is NaCl rather than ClNa. For compounds containing only nonmetallic elements, there is no simple rule. If carbon or hydrogen is present, it usually stands first: carbon dioxide (CO_2); hydrogen sulfide (H_2S); carbon tetrachloride (CCl_4). An exception, justified only by long habit, is ammonia (NH_3). If oxygen or chlorine is present, it is usually written last: carbon monoxide (CO); phosphorus trichloride (PCl_3); sulfur dioxide $(SO)_2$. In most ordinary compounds containing more than two elements, oxygen is one of the constituents; it commonly does not appear explicitly in the name, and in the formula stands last: calcium carbonate ($CaCO_3$); sodium sulfate (Na_2SO_4). An exception is certain compounds containing both hydrogen and oxygen, called hydroxides, in whose formulas hydrogen comes last: sodium hydroxide (NaOH).

Compounds with Two Elements Only. Names of these compounds always end in *-ide*, used as a suffix to the name of the second element. Thus: HCl, hydrogen chlor*ide;* $ZnCl_2$, zinc chlor*ide;* $AlCl_3$, aluminum chlor*ide;* H_2S, hydrogen sulf*ide;* Fe_3C, iron carb*ide.* In only one other type of compound, the hydroxides, does this ending appear.

When two or more compounds contain the same pair of elements, they are distinguished by one of two methods.

1. A prefix (*mono-*, *di-*, *tri-*, etc.) may be added to the name of the second element in each, indicating the number of its atoms per molecule.

Carbon *mono*xide (CO)	Phosphorus *tri*chloride (PCl_3)
Carbon *di*oxide (CO_2)	Phosphorus *penta*chloride (PCl_5)

2. The suffixes *-ic* and *-ous* may be added to the name of the first element, the *-ic* referring to a compound containing more atoms of the second element relative to the first.

Ferr*ic* chloride ($FeCl_3$)	(For iron, the suffixes are added to
Ferr*ous* chloride ($FeCl_2$)	the Latin name, *ferrum*)
Mercur*ic* oxide (HgO)	Cupr*ic* sulfide (CuS)
Mercur*ous* oxide (Hg_2O)	Cupr*ous* sulfide (Cu_2S)

In general, the second method is used for compounds of a metal with a nonmetal, the first for compounds of two nonmetals, but this rule has several exceptions.

* We omit here discussion of the special rules for naming complex compounds of carbon.

The term "peroxide," often applied to dioxides, refers properly only to a group of dioxides with special properties. The only common peroxides are hydrogen peroxide (H_2O_2), sodium peroxide (Na_2O_2), and barium peroxide (BaO_2).

Compounds with Three Elements. The majority of these compounds contain oxygen, another nonmetallic element, and a metal. Of this majority the greater number are named by adding *-ate* to the name of the nonmetal.

Sodium sulf*ate* (Na_2SO_4)	Potassium nitr*ate* (KNO_3)
Calcium sulf*ate* ($CaSO_4$)	Magnesium carbon*ate* ($MgCO_3$)

A great many compounds contain the group of atoms SO_4 united with a metal. In addition to the two listed above, common ones are potassium sulfate (K_2SO_4), copper sulfate ($CuSO_4$), zinc sulfate ($ZnSO_4$), magnesium sulfate ($MgSO_4$). Collectively these compounds are referred to as the "sulfates," and the atom group SO_4 is called the "sulfate group." Through many reactions this group remains intact, appearing simply to sever connections with one metal and join up with another. Thus

$$\underset{MgSO_4\ (sol'n)}{\text{Magnesium sulfate}} + \underset{BaCl_2\ (sol'n)}{\text{barium chloride}} \rightarrow \underset{BaSO_4}{\text{barium sulfate}} + \underset{MgCl_2\ (sol'n)}{\text{magnesium chloride}}$$

Other atom groups which remain intact in chemical reactions and which appear as part of many compounds are the nitrate group (NO_3), the carbonate group (CO_3), and the hydroxide group (OH). When two or more groups of a single kind appear in the molecule of a compound, the formula is written with parentheses around the group: thus calcium nitrate is $Ca(NO_3)_2$ rather than CaN_2O_6, and ferric hydroxide is $Fe(OH)_3$ rather than FeO_3H_3.

Acids. We shall not attempt a rigorous definition of this important class of compounds until later. In general, acids are characterized experimentally by the facts that (1) their solutions have a sour taste and (2) their solutions will change the color of certain dyes—for example, they will turn a blue litmus solution red. Their formulas are characterized by the presence of hydrogen combined with one or more nonmetallic elements. If only one other element besides hydrogen is present, the acid is named by adding the prefix *hydro-* and the suffix *-ic* to the name of the second element: HCl may be called either hydrogen chloride or hydrochloric acid; H_2S may be called either hydrogen sulfide or hydrosulfuric acid. If the acid contains oxygen in addition to another nonmetal, the prefix *hydro-* is omitted from the name: H_2SO_4 is hydrogen sulfate or sulfuric acid, HNO_3 is hydrogen nitrate or nitric acid, H_2CO_3 is hydrogen carbonate or carbonic acid.

Bases. The hydroxides violate the rules of naming set up for other compounds and are exceptional also in their chemical properties. Soluble hydroxides of metals are characterized by (1) their bitter taste and (2) their ability to reverse the changes in the color of dyes brought about by acids—for instance, a soluble hydroxide will turn a red litmus solution blue. In many respects opposite in behavior to acids, the hydroxides of metals are called collectively *bases*, but the term *base* does not appear in their names.

These rules will enable us to name most of the compounds which we shall study in future chapters. Note that the rules enable us to name a compound if its formula is given but are insufficient to tell us the formula of a compound from its name. Concerning the formula of aluminum chloride, for instance, we could guess that it would contain Al followed by Cl, but we would not know whether to write AlCl, $AlCl_2$, Al_2Cl, or some other combination. In the next chapter we shall find a means of determining the necessary subscripts.

Formulas and Equations

To complete our study of the chemist's language, we glance finally at his shorthand methods.

Chemical formulas we have used for three purposes, as explained in the last chapter.

1. To show the kinds of atoms present in a compound.
2. To show the ratio of the numbers of atoms of different kinds.
3. For gases and for some solids and liquids, to show the numbers of different kinds of atoms in a molecule.

We shall find formulas increasingly useful for a fourth purpose.

4. To stand as abbreviations for the names of compounds.

Thus "HCl" is often more convenient to write than "hydrochloric acid," "KOH" than "potassium hydroxide," etc. For the sake of completeness, two additional purposes of chemical formulas should be named.

5. To show the ratio by weight of the elements in a compound.
6. For gases and for some solids and liquids, to show the molecular weight of a compound.

These last purposes, important to a professional chemist, need not concern us here.

For a shorthand method of expressing the results of a chemical change, Berzelius suggested that the formulas of the substances involved be combined into a *chemical equation.* An equation includes the formulas of all the substances entering the reaction on the left-hand side, formulas

of all the products on the right-hand side. The formulas may be written in any order and are connected by + signs; between the two sides of the equation is placed either an arrow or an equality sign. Thus when carbon burns, the two substances which react are carbon (C) and oxygen (O_2), and the only product is carbon dioxide (CO_2).

$$C + O_2 \rightarrow CO_2$$

This equation means, in words: "carbon reacts with oxygen to form carbon dioxide."

Equations are very like the "summary statements" used earlier in this chapter, with one important added provision: the number of atoms of any one kind must be the same on both sides of the equation. This provision means that the law of conservation of mass is expressed in every chemical equation. For example, the summary statement for the decomposition of water by electrolysis is

$$\text{Water} \rightarrow \text{hydrogen} + \text{oxygen}$$

Substitution of formulas gives

$$H_2O \rightarrow H_2 + O_2$$

Here two atoms of oxygen are shown on the right-hand side, only one atom on the left; in chemical terms, the equation is "unbalanced." We cannot help matters by simply writing O instead of O_2 on the right, for we know that oxygen has the formula O_2. Nor is it legitimate to write a subscript "2" under the O in H_2O, for H_2O_2 is the formula of hydrogen peroxide, not water. The remedy is to show two units of H_2O on the left, giving two molecules of hydrogen and one of oxygen:

$$2H_2O \rightarrow 2H_2 + O_2$$

Now the equation is "balanced," for there are two O atoms and four H atoms on each side. Note that a number placed before a formula multiplies everything in the formula, while a subscript applies only to the atom immediately before it.

"Balancing an equation" consists of making the number of atoms of each kind the same on both sides, by writing the proper numbers in front of various formulas. For simple equations balancing involves no more than careful inspection. We shall consider three examples:

1. Mercury heated in oxygen is changed to red mercuric oxide.

Summary statement:

$$\text{Mercury} + \text{oxygen} \rightarrow \text{mercuric oxide}$$

Unbalanced equation:

$$Hg + O_2 \rightarrow HgO$$

Balanced equation

$$2Hg + O_2 \rightarrow 2HgO$$

2. Zinc added to a solution of hydrochloric acid liberates hydrogen and forms a solution of zinc chloride.
Summary statement:

Zinc + hydrochloric acid (sol'n) → hydrogen + zinc chloride (sol'n)

Unbalanced equation:

$$Zn + HCl\ (sol'n) \rightarrow H_2 + ZnCl_2\ (sol'n)$$

Balanced equation:

$$Zn + 2HCl\ (sol'n) \rightarrow H_2 + ZnCl_2\ (sol'n)$$

3. When hydrogen sulfide is burned in oxygen, water and sulfur dioxide are formed.
Summary statement:

Hydrogen sulfide + oxygen → water + sulfur dioxide

Unbalanced equation

$$H_2S + O_2 \rightarrow H_2O + SO_2$$

Balanced equation

$$2H_2S + 3O_2 \rightarrow 2H_2O + 2SO_2$$

Two facts about equations need emphasis. First, an equation shows simply what chemical change has taken place; it tells nothing about the conditions of temperature, pressure, or illumination which are necessary to bring about the change. Second, an equation is not a means for predicting what chemical change *will take place* but is a concise summary of a change which *has taken place.* To write an equation, the formulas of all products as well as those of all the substances which react must be known.

An equation in itself is not a means for predicting chemical changes; but, if a chemist is familiar enough with the chemical behavior of an element, he can often express by means of an equation his prediction as to the reaction of the element in a new set of circumstances. We shall learn in the next chapter his bases for such predictions.

Questions

1. At the temperature of liquid air (about −185°C) is hydrogen a gas, liquid, or solid? Answer the same question for chlorine and oxygen.
2. On what property or properties of graphite does each of the following uses depend: (*a*) for mixing with clay and wax to form the "lead" of lead pencils; (*b*) for terminals ("electrodes") placed in melted sodium chloride, between which elec-

tricity is passed during electrolysis; (*c*) for making crucibles used to heat metals to high temperatures?

3. In what ways is the chemical behavior of sodium similar to that of hydrogen? In what ways does oxygen resemble chlorine?
4. Which of the following (*a*) will float on water, (*b*) melt within 20° of the boiling point of water: sulfur, sodium, diamond, graphite?
5. If a flame is brought near a test tube containing pure hydrogen, the gas burns quietly, but if the tube contains a mixture of air and hydrogen, an explosion results. Suggest an explanation.
6. Name the following compounds:

$Ca(OH)_2$	$MgCO_3$	H_2CO_3	KOH	$HgSO_4$
$AgNO_3$	AgCl	Na_3N	CaO	K_2CO_3
$Al_2(SO_4)_3$	$Zn(NO_3)_2$	H_3PO_4	NaH	HNO_3

Which of these substances are acids and which bases?

7. Name (*a*) the two oxides of phosphorus, P_2O_3 and P_2O_5; (*b*) the two chlorides of mercury, Hg_2Cl_2 and $HgCl_2$; (*c*) the two hydroxides of iron, $Fe(OH)_2$ and $Fe(OH)_3$.
8. Interpret the following equations in terms of molecules:

$2CO + O_2 \rightarrow 2CO_2$ $\qquad$ $4P + 5O_2 \rightarrow 2P_2O_5$

$3H_2 + N_2 \rightarrow 2NH_3$ $\qquad$ $C + H_2O$ (steam) $\rightarrow CO + H_2$

(EXAMPLE: the first equation means that two molecules of carbon monoxide react with one molecule of oxygen to give two molecules of carbon dioxide.)

9. Which of the following equations (*a*) are balanced, (*b*) show a reaction between two gases, (*c*) show a reaction involving an acid, (*d*) show the production of a colorless gas, (*e*) show reduction of an oxide?

$Zn + H_2SO_4$ (sol'n) $\rightarrow H_2 + ZnSO_4$ (sol'n)

$Al + 3O_2 \rightarrow Al_2O_3$ $\qquad$ $N_2 + H_2 \rightarrow 2NH_3$

$H_2CO_3 \rightarrow H_2O + CO_2$ $\qquad$ $6Na + Fe_2O_3 \rightarrow 2Fe + 3Na_2O$

$3CO + Fe_2O_3 \rightarrow 3CO_2 + 2Fe$ $\qquad$ $MnO_2 + 4HCl \rightarrow MnCl_2 + 2H_2O + Cl_2$

10. Write balanced equations for the following reactions:

   a. Hydrogen and oxygen combine to form water.
   b. Zinc and sulfur combine to form zinc sulfide (ZnS).
   c. Copper oxide (CuO) is reduced by hydrogen, with the formation of copper and water vapor.
   d. Carbon burns in air to form carbon monoxide.
   e. Sulfur trioxide (SO_3) combines with water to form sulfuric acid.
   f. Sulfur dioxide combines with oxygen to form sulfur trioxide.
   g. Sodium burns in chlorine to form sodium chloride.
   h. Aluminum reacts with hydrochloric acid solution to liberate hydrogen and produce a solution of aluminum chloride ($AlCl_3$).
   i. Sodium liberates hydrogen from water, forming a solution of sodium hydroxide.

CHAPTER XV

The Periodic Law

SCIENCE progresses by seeking relationships among things and processes. Relationships can be interpreted by generalizations, in the form of laws, theories, and hypotheses. Generalizations, once established, lead to a search for new relationships, and on these in turn further generalizations are based. At each stage in this progress the relationships become more fundamental, and the laws and theories wider in scope. So, as scientific thought develops, explanations of the natural world become possible in simpler and simpler terms.

This is an old theme, but an important one. The myriad details of scientific knowledge must not obscure for us the underlying process by which the knowledge has been obtained. The guiding motive of this process, the mainspring of the scientific method, is the search for relationships. Always, when faced with new phenomena, the scientist unconsciously asks: How are these phenomena related, among themselves and to other phenomena of my experience? In what ways are they similar, in what ways dissimilar? What mathematical rules connect them?

Kepler sought relationships in the observations of Tycho Brahe on planetary motions, and found the three generalizations which we call Kepler's laws. Newton sought a connection between these generalizations and Galileo's laws of falling bodies, and found the law of gravitation. Rumford and Joule tried to find the relation between mechanical energy and heat; from the rule connecting these two forms of energy and similar rules for others came the conservation law. Boyle, Charles, and Gay-Lussac searched for relationships among gas volumes at various pressures and temperatures and discovered the simple proportionality laws. Relationships between these laws and other regularities in gas behavior led to the kinetic theory.

When a large number of objects are under investigation, the search for relationships becomes primarily a problem of classification. Any

classification implies relationships, for objects are grouped according to their similarities and differences. From a classification, other relationships and generalizations often emerge.

Thus chemistry in its beginnings faced a formidable number of different materials. Relationships among these materials suggested finally their classification into elements, compounds, and mixtures. This classification focused attention on relations among the compositions of different compounds, and from these relations Dalton built the atomic theory.

By Dalton's time chemists were convinced that the earth's materials were constructed from a handful of elementary substances, therefore from a limited number of different kinds of atoms. But the handful was a big one, and steadily growing bigger: by the middle of the nineteenth century about sixty different elements had been discovered. Was it possible that the elements themselves might be interrelated, that their atoms might show resemblances and differences which would reveal some even more fundamental principle of chemistry? Here again was a problem primarily of classification.

Attacked by several chemists in the 1860's, this problem was brilliantly solved by a Russian, Mendelyeev. To understand Mendelyeev's work, let us inquire a bit further into the properties of the elements, paying special heed to similarities and dissimilarities which might serve as a basis for their classification.

Metals and Nonmetals

Among the more obvious distinctions between different kinds of elements is that which divides metals from nonmetals. So generally familiar is the idea of a metal that we have used this distinction frequently without trying to make it precise. Iron, mercury, gold, aluminum, sodium, tin are examples of metallic elements; carbon, sulfur, hydrogen, chlorine, helium are nonmetals.

The outstanding physical properties which differentiate metals from other substances are (1) their characteristic sheen, or "metallic luster," and (2) the ease with which they conduct heat and electricity. Instinctively we associate also the qualities of hardness and toughness with metals, but a moment's glance at the properties of gold, lead, and sodium shows that these are not general characteristics. Nonmetals in the solid state are usually brittle materials without metallic luster (graphite and one form of silicon are exceptions) and are very poor conductors of heat and electricity (graphite is an exception, but its conductivity is small compared with that of most metals). In some other physical properties nonmetals have an extreme range: in melting point from helium (−269°C, or 4°K) to carbon (above 3500°C), and in hardness from diamond to soft white phosphorus.

In chemical behavior metals show considerable differences among themselves. Sodium, for instance, is extremely active, while gold and platinum are highly resistant to chemical change. In general, (1) metals combine with nonmetals much more readily than with each other. All metals combine directly with fluorine and chlorine and most combine directly with oxygen. Many metals mix readily to form alloys, but definite compounds between metals are few and unstable. (2) All the more active metals react with dilute acids to liberate hydrogen, and very active metals liberate hydrogen from water. (3) Oxides of the more active metals react with water to form bases. Thus

$$Na_2O + H_2O \rightarrow 2NaOH \text{ (sol'n)}$$

The slaking of lime is a similar combination of a metallic oxide (CaO) with water to form a hydroxide.

$$CaO + H_2O \rightarrow Ca(OH)_2$$

Nonmetals show an even greater variety of chemical properties than do the metals. Some (argon, helium, neon) form no compounds at all, while others (chlorine, fluorine) are highly active. In general, (1) nonmetals (except for the argon group) combine readily with active metals, somewhat less readily with each other. Thus chlorine and fluorine react violently with active metals but will not combine directly with oxygen. Sulfur and phosphorus, on the other hand, burn brightly in oxygen. (2) The nonmetals do not react with dilute acids. Several are attacked by bases, but the reactions do not follow a single pattern. (3) Oxides of nonmetals, if soluble, combine with water to form acids. Thus the slightly sour taste of soda water is due to carbonic acid, formed from the dissolved oxide of carbon.

$$CO_2 + H_2O \rightarrow H_2CO_3$$

Sulfuric acid is manufactured by dissolving SO_3 in water.

$$SO_3 + H_2O \rightarrow H_2SO_4$$

The nonmetal hydrogen is unique: although distinctly nonmetallic in its physical properties and much of its chemical behavior, some of its reactions suggest similarities with the metals. For instance, hydrogen combines more readily and more violently with nonmetals than with metals, and its compounds with atom groups like SO_4, CO_3, etc., are similar in formula to metallic compounds. The oxide of hydrogen, of course, forms neither an acid nor a base on combination with water.

We shall learn later a more sophisticated general definition of metals and nonmetals. Even this better definition, however, does not make the distinction sharp, for a few elements show properties in some measure characteristic of both groups. Metals far outnumber the nonmetals, only

twenty of the ninety-odd elements known today being considered definitely nonmetallic.

Active and Inactive Elements

Sodium we call an "active" metal, gold a very "inactive" one. Precisely what do these terms mean?

We remember that sodium is tarnished by a few seconds' exposure to air, while a gold ring keeps its luster after years of exposure to air and perspiration. We think of the spectacular combustion of sodium in chlorine, accompanied by much heat and light energy; gold combines sluggishly with chlorine, setting free little energy. We recall that sodium liberates hydrogen rapidly from dilute acids, even from water; gold is unaffected by ordinary acids, dissolving only in a mixture of concentrated HCl and HNO_3. In reactions like these we say that sodium exhibits its greater activity.

Let us consider in more detail certain reactions involving these metals and two others of intermediate activity, copper and iron. Suppose that a sample of each is allowed to combine with chlorine, and that the number of calories of heat liberated in each of the four reactions is measured. To make the comparison fair, suppose that the same amount of chlorine, 35.5 g., is used in each case, with enough of the metal to consume it completely. From this experiment we should find that the amounts of energy liberated are 5800 cal for gold, 25,700 cal for copper, 32,100 cal for iron, and 98,400 cal for sodium. Evidently, then, *one method for comparing activities numerically is by measuring amounts of energy given out in similar reactions.*

Or we might start with the four metal chlorides, and see how easily each metal could be separated from the nonmetal. We should find that gold chloride is decomposed by heating to about 300°C, copper chloride by heating to somewhat above 1000°C, iron chloride and sodium chloride only by heating to much higher temperatures. Gold chloride is said to be a relatively *unstable* compound, while sodium chloride is a very *stable* compound. In general, the more active an element is, the more difficult are its compounds to decompose. *Relative stabilities of similar compounds, then, give us a second method for comparing activities.* Roughly quantitative measurements of relative stabilities may be obtained by determinations of decomposition temperatures, or better by the amounts of energy which must be supplied to bring about decomposition.

If samples of the four metals are placed in test tubes and dilute HCl solution is poured over them, hydrogen bubbles off rapidly in the tube containing sodium, more slowly in the tube containing iron, not at all from the copper and gold. Thus *the rate at which hydrogen is evolved from dilute acids is a third, somewhat cruder, method for comparing activities of metals.*

For different nonmetals a comparison of activities is more difficult than for metals, since a nonmetal which is only moderately active in reactions with metals (for example, phosphorus) may be highly active in reactions with other nonmetals. In general, however, the term "active nonmetal" applied to an element refers to its behavior in reactions with metals and metallic compounds. With this understanding, the activities of two nonmetals may be compared by methods similar to those which apply to metals: the heat liberated in their reactions with equal quantities of the same metal may be measured, or the stabilities of similar compounds may be determined. Thus chlorine is a more active nonmetal than oxygen, since it combines directly with more metals than does oxygen and since in most of its reactions with metals it liberates more energy. Sulfur is less active than oxygen, since sulfides are as a rule more easily decomposed than oxides.

By using the results of several different methods, both metals and nonmetals may be arranged in continuous series showing the order of their activities. In the following partial lists active elements are placed at the beginning, inactive ones at the end:

Metals		*Nonmetals*
Potassium	Iron	Fluorine
Sodium	Lead	Chlorine
Calcium	Copper	Bromine
Magnesium	Mercury	Oxygen
Aluminum	Silver	Iodine
Zinc	Gold	Sulfur

Thus a study of the varying activities of metals and nonmetals points the way to a possible grouping of elements within the two major classes.

Families of Elements

Among some elements resemblances are so striking that they seem to form natural groups or families. As examples of families of elements, we shall discuss a group of active nonmetals, a group of active metals, and the group of inert gases.

The Halogens. These curious elements, fluorine, chlorine, bromine, and iodine, are responsible for some of the vilest odors and some of the most brilliant colors in an elementary chemistry laboratory. Their family name means "salt former," in token of the white, crystalline solids which they form by combination with many metals.

Atomic weights increase in the order F (19), Cl (35.46), Br (79.92), I (126.92). Melting points and boiling points increase in this same order: fluorine is a gas (boiling point −187°C); chlorine, a gas (boiling point −35°C); bromine, a volatile liquid (boiling point 59°C); iodine, a volatile solid (boiling point 184°C). All have colored vapors: fluorine, pale

yellow; chlorine, greenish yellow; bromine, reddish brown; iodine, reddish violet. All are soluble in carbon tetrachloride, giving solutions colored like their vapors. Fluorine reacts with water to liberate oxygen; the others are slightly soluble in water, the solubility decreasing from chlorine to iodine. All have two atoms per molecule at ordinary temperatures: F_2, Cl_2, Br_2, I_2.

In chemical behavior the other halogens closely resemble chlorine (page 186). All are active nonmetals, combining directly and violently with active metals. The resulting compounds have strikingly similar formulas.

NaF	CaF_2	ZnF_2	AlF_3
NaCl	$CaCl_2$	$ZnCl_2$	$AlCl_3$
NaBr	$CaBr_2$	$ZnBr_2$	$AlBr_3$
NaI	CaI_2	ZnI_2	AlI_3

Note that in all compounds with any one metal, the halogens have the same subscript, that is, the same number of atoms combined with each metal atom. All the halogens react directly with hydrogen, although for bromine and iodine the reaction is slow at ordinary temperatures. The hydrogen compounds HF, HBr, and HI, like HCl, dissolve in water to form acids.

Although the halogens are all active elements their activity declines markedly with increasing atomic weight. Fluorine is the most active of all nonmetals, so active that it is difficult to prepare, difficult to keep, dangerous to work with. Chlorine is somewhat less active, bromine and iodine still less. All metals combine directly with fluorine and chlorine, but the less active ones are not affected by bromine and iodine. Amounts of energy liberated by the reactions with potassium are (for 39.1 g. of K):

F_2 (forming KF)	Cl_2 (forming KCl)	Br_2 (forming KBr)	I_2 (forming KI)
118,000 cal	104,300 cal	95,100 cal	80,100 cal

Thus the halogens form a group of elements with many similar properties, some of which change progressively as atomic weight increases. The following summary shows in the first column the outstanding similarities, in the second the properties which change from element to element. Properties which increase with increasing atomic weight are designated "inc," properties which decrease by "dec."

Similar Properties	*Properties Which Change Progressively*
Strong odors	Boiling point (inc)
Molecules have two atoms	Melting point (inc)
Soluble in CCl_4	Solubility in water (dec)
Slightly soluble in H_2O (F_2 liberates O_2)	Chemical activity (dec), shown by
Active as nonmetals	Speed of reaction with H_2 (dec)
Hydrogen compounds soluble, form acids	Energy liberated by reactions with metals (dec)
Compounds have similar formulas	

The Alkali Metals. These are all soft, light, highly active metals like the most familiar member of the group, sodium. A few of their physical properties are shown in Table X. Note that densities increase steadily (except K) with atomic weights, while melting points and boiling points decrease. Cesium has so low a melting point that it would be liquefied if held in the hand—but this experiment is not advisable, because cesium's intense chemical activity makes contact with it dangerous.

TABLE X. PROPERTIES OF THE ALKALI METALS

Name	*Symbol*	*Formula*	*Atomic weight*	*Density*	*Melting point, °C*	*Boiling point, °C*
Lithium	Li	Li	6.94	0 53	186	1200
Sodium	Na	Na	23.00	0 97	97 5	880
Potassium	K	K	39 10	0.86	62 5	760
Rubidium	Rb	Rb	85.45	1.53	38 5	700
Cesium	Cs	Cs	132.81	1.90	26.5	670

Like sodium, the other alkali metals tarnish quickly in air, liberate hydrogen from water and dilute acids, combine energetically with active nonmetals to form very stable compounds, and form oxides which combine with water to make bases. Formulas of their compounds are strikingly similar:

Bromides:	LiBr	NaBr	KBr	RbBr	CsBr
Sulfides:	Li_2S	Na_2S	K_2S	Rb_2S	Cs_2S
Hydroxides:	LiOH	NaOH	KOH	RbOH	CsOH

In general, chemical activity increases as atomic weight increases. The three heavier metals liberate so much energy in their reactions with cold water that the hydrogen produced ignites spontaneously, while from lithium and sodium hydrogen is evolved without burning. Cesium forms the most stable compounds with chlorine and bromine, lithium the least stable compounds.

Like the halogens, the alkali metals show striking similarities in their chemical and physical properties, but many of the properties change progressively with increasing atomic weights.

The Inert Gases. In strong contrast to the active elements of the alkali metal and halogen groups, these gases are extremely inactive, so inactive that they form no compounds with other elements. Atoms of the individual gases do not even join together in pairs to form molecules, as do those of other gaseous elements. All the inert gases except radon are minor constituents of the atmosphere; argon makes up nearly 1 per cent of air, the others much less. Their scarcity and inactivity prevented

their discovery until the last decade of the nineteenth century—nearly thirty years after Mendelyeev's work.

The common physical properties of the inert gases, outlined in Table XI, show the same general similarity and regular gradations with increasing atomic weight that we have found in the other groups.

TABLE XI. PROPERTIES OF THE INERT GASES

Name	*Symbol*	*Formula*	*Atomic weight*	*Density of liquid*	*Melting point, °C*	*Boiling point, °C*
Helium	He	He	4.00			−269
Neon	Ne	Ne	20.18	1.20	−249	−246
Argon	A	A	39.91	1.40	−189	−186
Krypton	Kr	Kr	83.7	2.6	−169	−152
Xenon	Xe	Xe	130.2	3 06	−140	−109
Radon	Rn	Rn	222	4.4	− 71	− 62

Valence

The usefulness in classification of a chemical property called *valence* is well illustrated by these three families of elements.

In the chlorides of the alkali metals one chlorine atom is united with each metal atom, giving the formulas LiCl, NaCl, etc. Chlorides of some other metals contain two halogen atoms for every metal atom: $CaCl_2$, $MgCl_2$, $BaCl_2$. Aluminum and a few rare metals form chlorides with three chlorine atoms per metal atom: $AlCl_3$, $GaCl_3$. *The number of chlorine atoms per metal atom in the formula of a metal's chloride is a property of the metal called its* ***valence.*** By convention, valences of metals are labeled "positive": thus sodium has a valence of +1, calcium (from $CaCl_2$) a valence of +2, aluminum a valence of +3, etc. Hydrogen, though a nonmetal, is considered to have a valence of +1, since it forms the chloride HCl.

For nonmetals their hydrogen compounds are the basis for a similar definition of valence. *The number of hydrogen atoms per nonmetal atom* in the formula of the compound of a nonmetal with hydrogen *is called the* ***valence*** *of the nonmetal.* Valences of nonmetals are labeled "negative": the halogens (from HCl, HBr, etc.) have the valence −1, oxygen (H_2O) has the valence −2, nitrogen (NH_3) has the valence −3. For the present the terms "positive" and "negative" need be no more than convenient labels, although we shall find a good reason for them later.

In formulas of simple compounds, *the total number of positive and negative valences must be equal.* This rule follows from the two definitions just given. In Na_2O each sodium atom has a valence of +1, giving a total positive valence of 2, and the single oxygen atom supplies an equal

negative valence of 2. In $AlBr_3$, the Al supplies a total positive valence of 3, and the three Br atoms (valence of each = −1) supply an equal negative valence. By use of this rule, the valence of an unfamiliar element may be determined if the formula of any one of its simple compounds is known. Thus if we know that the rare metal scandium forms a fluoride ScF_3, we need no further information about its chemical behavior to assign it a valence of +3, since 3 positive valences are necessary to equal the 3 negative valences supplied by three fluorine atoms. Since scandium has a valence of 3, we should expect the formula of its chloride to be $ScCl_3$, of its oxide to be Sc_2O_3, and so on. *The idea of valence, therefore, supplies the information necessary to write the formula of a compound with correct subscripts when the name of the compound is given.*

Formulas of more complex compounds may be written if the idea of valence is extended to atom groups. Since the nitrate group (NO_3) appears in the compound HNO_3, the group as a whole may be assigned a valence of −1, just as Br is assigned a valence of −1 from the formula HBr. Similarly the formula H_2SO_4 suggests a valence of −2 for the sulfate group, the formula H_2CO_3 a valence of −2 for the carbonate group, the formula HOH (or H_2O) a valence of −1 for the hydroxide group. Corresponding to these valences we find the compounds $NaNO_3$, Na_2SO_4, Na_2CO_3, NaOH, $Ca(NO_3)_2$, $CaSO_4$, $Al_2(SO_4)_3$, etc. For scandium, with its valence of 3, we might predict a nitrate with the formula $Sc(NO_3)_3$, a hydroxide $Sc(OH)_3$, and so on.

The importance of valence in classification is apparent from the constant valences in the three families of similar elements just discussed: +1 for the alkali metals, −1 for the halogens, 0 for the inert gases (since they form no compounds). Evidently a characteristic valence links together elements of similar properties. As a basis for a general classification of the elements, however, valence has two drawbacks: (1) Some elements of very dissimilar properties have the same valence; for instance, the inactive metal silver shows the same valence as the alkali metals. (2) Some elements have two or more different valences; thus copper forms the two chlorides CuCl and $CuCl_2$, and iron the two chlorides $FeCl_2$ and $FeCl_3$. We shall find that valence is indeed useful in the general classification, provided that these drawbacks are given due weight.

The concept of valence will be broadened later to include compounds of nonmetals with each other as well as their compounds with metals.

The Periodic Classification

Mendelyeev was a Russian out of Siberia, long-haired, bearded, patriarchal in appearance (Fig. 99). For many years professor of chemistry at the University of St. Petersburg, he devoted himself to government service as well as to scientific work, although his outspoken liberal ideas

were frequently embarrassing to the Tsarist regime. Mendelyeev was a gifted teacher, an able experimenter, but above all a dreamer, a scientific visionary. If some of his speculations seem fantastic, for one vision at least chemistry owes him a great debt—the vision which gave him the key to the classification of the elements.

The fact that some elements have strikingly similar properties was, of course, known long before Mendelyeev's time. The grouping of elements according to valence became possible after 1860, when acceptance of Avogadro's hypothesis cleared away the difficulties in determining atomic weights and assigning formulas. What Mendelyeev saw was that *valence and other properties are related to atomic weight.* In this vision he was not alone, for a few of his contemporaries reached this conclusion independently; but Mendelyeev was the first to apply it to all the known elements, and to predict from it the existence of elements then unknown.

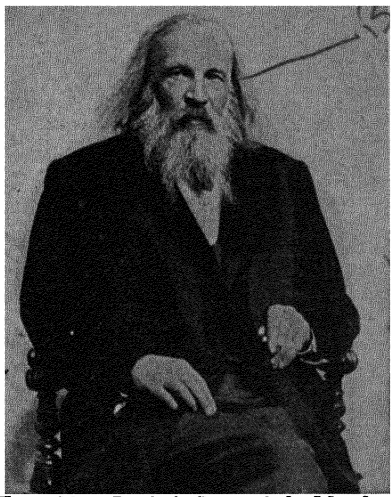


FIG. 99. *Dmitri Ivanovitch Mendeleyeev (1834–1907).*

Following Mendelyeev, but using our modern list of elements (Table IX) rather than the limited number which he knew, let us write down the elements in the order of increasing atomic weights. First is hydrogen; then the inert gas helium; then the alkali metal lithium; then a rare metal called beryllium, less active than lithium, with a valence of +2; then boron, a relatively inactive nonmetal, which forms a chloride BCl_3; then carbon, a nonmetal which forms both CCl_4 and CH_4; then nitrogen, another nonmetal; then oxygen, a more active nonmetal; and fluorine, most active of all nonmetals. From lithium to fluorine is a complete transition from a highly active metal to a highly active nonmetal, and a change in valence through positive values from +1 to +4, then through negative values from −4 to −1. After fluorine comes neon, another inert gas like helium, then sodium, an alkali metal like lithium. To suggest these resemblances, we break off the row of elements at fluorine, and start a second row with neon (Table XII). In the seven elements beyond neon we find again a transition from active metals to active nonmetals, and a change in valence like that of the first row.

After chlorine, in order of atomic weights, comes potassium, then argon, then calcium. Starting a third row with potassium would put this active metal under the inert gases helium and neon, while argon

would go beneath sodium and lithium. To avoid this obvious discrepancy, we deliberately reverse the order for once, bringing argon before potassium.

TABLE XII. PART OF THE PERIODIC TABLE

H 1.008	He 4.003	Li 6.940	Be 9.02	B 10.82	C 12.010	N 14.008	O 16.000	F 19.000		
	Ne 20.183	Na 22.997	Mg 24.32	Al 26.97	Si 28.06	P 30.98	S 32.06	Cl 35.457		
A 39.944	K 39.096	Ca 40.08	Sc 45.10	Ti 47.90	10 metals		As 74.91	Se 78.96	Br 79.916	
Kr 83.70	Rb 85.48	Sr 87.63	Y 88.92	Zr 91.22	10 metals		Sb 121.76	Te 127.61	I 126.92	
Xe 131.30	Cs 132.91	Ba 137.36								
Group	0	I	II	III	IV		V	VI	VII	
Chlorides	–	NaCl	$CaCl_2$	$AlCl_3$	CCl_4		----	----	----	
H-compounds	–	----	----	----	CH_4		NH_3	H_2O	HCl	
Valence	0	+1	+2	+3	±4		−3	−2	−1	

After calcium in the third row more difficulties appear. Scandium, the next element, has the same valence as aluminum, but differs in other important properties. Titanium (Ti) is even less like carbon and silicon. Then follow ten metals (including iron, copper, zinc), quite similar among themselves, but differing conspicuously from the nonmetals at the end of the first two rows. Only after the ten metals do three relatives of these nonmetals appear—arsenic (As), selenium (Se), and bromine. Thus between the first inert gas (He) and the second (Ne) is a sequence of eight elements; between neon and argon is another sequence of eight; but between argon and krypton the sequence includes eighteen. Beyond krypton is a second sequence of eighteen, including again a dozen metals of doubtful relationships. From xenon to the last inert gas, radon, is a yet more complex sequence of thirty-two elements.

Although the interval between similar elements changes in this peculiar manner, although a few reversals of order appear (for instance, K and A), nevertheless this arrangement of elements does show a strikingly regular repetition of properties. *If the elements are listed in the (approximate) order of their atomic weights, elements with similar properties recur at definite intervals.* This is one way of stating Mendelyeev's *periodic law.* A tabular arrangement, like Table XII, showing this recurrence of properties is called a ***periodic table.***

Most periodic tables show similar elements in vertical rows, called ***columns*** or ***groups*** of elements. The groups are numbered as shown at the bottom of Table XII. The horizontal rows, containing elements with widely different properties, are called ***periods.*** Across each period

is a sudden step from an inert gas to an active metal, then a more or less steady transition through less active metals to weakly active nonmetals and finally to highly active nonmetals. Within each column is also a steady change in properties, but a much less rapid and less conspicuous one. We have already noted that increasing atomic weight brings increasing activity in the alkali metal family and decreasing activity among the nonmetals of the halogen family. These changes are typical; chemical properties within each group change from top to bottom in the direction of increased metallic activity or decreased nonmetallic activity—which amount to the same thing.

Mendelyeev studied carefully the metals in the middle of the long periods, to see if they also could be made to fit into the table. He found it possible to arrange these elements in families among themselves, and in less satisfactory fashion, chiefly on a basis of similarities in valence, to connect some of the families with elements in the main groups. These relationships are shown in the complete periodic table (Table XIII): here elements of the main groups (*e.g.*, the halogens) are connected by solid lines, and the vaguer associations with metals in the long periods are suggested by dotted lines. Thus the elements copper, silver, and gold are arranged as a "subgroup" with the alkali metals, an arrangement justified only by the fact that the three elements in some of their compounds show a valence of +1; copper and gold more commonly have valences of two and three, respectively, and the other properties of these three elements make the relationship an exceedingly tenuous one.

In the complete table the inert gases are placed at the right rather than the left, as in Table XII. Their position is a matter of choice; the arrangement of Table XIII has the advantage of grouping them with the other nonmetals at the left of the table. The position of hydrogen is uncertain, as might be expected from the peculiarities in its chemical behavior which we have already noted. It shows usually a valence of +1, rarely a valence of −1, so in the table it is connected by dotted lines with the alkali metals and with the halogens.

Relationships shown by the periodic table are in spots exceedingly vague, but on the whole it brings together similar elements with great accuracy. Mendelyeev's achievement seems all the more remarkable if we recall that in 1869, when the periodic law was discovered, only sixty-three elements were known. This meant that Mendelyeev had to leave numerous gaps in his periods of elements in order to make similar elements fall one under the other. Sure of the correctness of his classification, he boldly predicted that these gaps represented undiscovered elements. Further, from the position of each gap, from the properties of the elements around it, and from his knowledge of the variation of these properties

TABLE XIII. PERIODIC CLASSIFICATION OF THE ELEMENTS

							1 H 1.0081									Atomic number → Atomic weight →	2 He 4.003
Group I		Group II		Group III		Group IV		Group V		Group VI		Group VII					Group 0
3 Li 6.940		4 Be 9.02		5 B 10.82		6 C 12.010		7 N 14.008		8 O 16.0000		9 F 19.000					10 Ne 20.183
11 Na 22.997		12 Mg 24.32		13 Al 26.97		14 Si 28.06		15 P 30.98		16 S 32.06		17 Cl 35.457					18 A 39.944
I a	II a	III b	IV b	V b	VI b	VII b	VIII			I b	II b	III a	IV a	V a	VI a	VII a	0
19 K 39.096	20 Ca 40.08	21 Sc 45.10	22 Ti 47.90	23 V 50.95	24 Cr 52.01	25 Mn 54.93	26 Fe 55.84	27 Co 58.94	28 Ni 58.69	29 Cu 63.57	30 Zn 65.38	31 Ga 69.72	32 Ge 72.60	33 As 74.91	34 Se 78.96	35 Br 79.916	36 Kr 83.7
37 Rb 85.48	38 Sr 87.63	39 Y 88.92	40 Zr 91.22	41 Cb 92.91	42 Mo 95.95	43	44 Ru 101.7	45 Rh 102.91	46 Pd 106.7	47 Ag 107.88	48 Cd 112.41	49 In 114.76	50 Sn 118.70	51 Sb 121.76	52 Te 127.61	53 I 126.92	54 Xe 131.3
55 Cs 132.91	56 Ba 137.36	57 to 71	72 Hf 178.6	73 Ta 180.88	74 W 183.92	75 Re 186.31	76 Os 190.2	77 Ir 193.1	78 Pt 195.23	79 Au 197.2	80 Hg 200.61	81 Tl 204.39	82 Pb 207.21	83 Bi 209.00	84 Po 210	85	86 Rn 222
87	88 Ra 226.05	89 to 96	Th?	Pa?	U?												
57 to 71 Rare earth elements		57 La 138.92	58 Ce 140.13	59 Pr 140.92	60 Nd 144.27	61	62 Sm 150.43	63 Eu 152.0	64 Gd 156.9	65 Tb 159.2	66 Dy 162.46	67 Ho 164.94	68 Er 167.2	69 Tm 169.4	70 Yb 173.04	71 Lu 174.99	
89 to 96 Actinium series		89 Ac 227.05	90 Th 232.12	91 Pa 231	92 U 238.07	93 Np 237	94 Pu 239	95 Am* 241	96 Cm* 242								

* Americium (Am) and curium (Cm) have been prepared artificially, but are not known to occur naturally.

properties of the unknown elements. His guesses included not only predictions about valence and general chemical activity, but precise numerical values for densities, melting points, etc. As the unknown elements were discovered one by one, as their properties were found to check Mendelyeev's predictions with uncanny accuracy, the correctness and usefulness of the periodic classification became firmly established. Perhaps its greatest triumph came at the end of the century, when the inert gases were discovered: here were six new elements whose existence Mendelyeev could not have predicted, but they fitted perfectly, as one more family of similar elements, into the periodic table.

Today only four gaps remain in the table, and there is good evidence that these elements are very scarce. The usefulness of the periodic law in predicting new elements is almost past, but its usefulness in coordinating chemical knowledge is greater today than in Mendelyeev's time. A chemist need not learn in detail the properties of all the elements, or even of a large proportion of them; if he knows thoroughly the chemical behavior of a few elements, and if he knows how properties vary in the periods and groups of the periodic table, then for any other element he need only glance at its position in the table to learn its chief physical and chemical characteristics.

And the periodic table has a deeper significance. Like any successful classification it not only shows the relationships on which it was based but suggests further relationships. It exhibits complex relations among different elements; these must mean complex relations among different kinds of atoms. Differences and similarities of various sorts among atoms inevitably suggest that the atoms must have *structure*, must have parts whose form and arrangement are responsible for the differences and similarities. In the search for the structure of Dalton's indivisible atoms, a search on which we embark in the next section, the periodic law plays at once the roles of guide and arbiter—pointing out how the various structures must be related, and serving as a check on any theory devised to picture them.

Questions

1. Sodium never occurs in nature as the free element, while platinum seldom occurs in combination. How are these facts related to the chemical activities of the two metals?
2. From what physical and chemical characteristics of iron do we conclude that it is a metal? From what physical and chemical characteristics of sulfur do we conclude that it is a nonmetal?
3. In the formation of 58.5 g. of sodium chloride from its elements, 98,400 cal of chemical energy are changed into heat energy. By the law of conservation of energy how many calories of heat must be converted into chemical energy to decompose this much sodium chloride?

4. Show that the positions of the first five metals in the list on page 198 are in agreement with their positions in the periodic table. Where would cesium appear in a more complete activity series? Where would platinum appear?
5. By what sort of experiments could you show that iron is a more active metal than mercury?
6. Using the periodic law and your knowledge of the halogens, make predictions concerning the following properties of the undiscovered element at the bottom of the halogen column:
   a. At ordinary temperatures, is it solid, liquid, or gaseous?
   b. At approximately what temperature does it boil?
   c. A molecule of its vapor contains how many atoms?
   d. Is it very soluble, moderately soluble, or slightly soluble in water?
   e. Write the formula of its compound with hydrogen (use *X* as the symbol of the element).
   f. Write formulas for its compounds with potassium and calcium.
   g. What is its approximate atomic weight?
   h. Does it combine more readily or less readily with hydrogen than does iodine?
   i. Is its compound with potassium more stable or less stable than KI?
7. Which is the more active metal, Ba or Ca? Sc or Al? Ca or Ti? Which is the more active nonmetal, Se or S? As or P? F or I?
8. What is the most common valence of Li, Ba, Ne, Si, S, Zn, Br?
9. What is the valence of Fe in Fe_2O_3? Hg in Hg_2O? Ba in $BaSO_4$? Cr in $Cr(OH)_3$? Sn in $SnBr_4$? Pb in $Pb(NO_3)_2$?
10. Write formulas for aluminum oxide, magnesium iodide, lithium carbonate, calcium sulfide, sodium nitride, rubidium hydroxide, potassium sulfate, barium nitrate.
11. Write balanced equations for the following reactions:
   a. Potassium and sulfur combine to form potassium sulfide.
   b. Cesium reacts with bromine to form cesium bromide.
   c. Barium reacts with water to liberate hydrogen.
   d. Aluminum reacts with ferric oxide (Fe_2O_3) to form iron and aluminum oxide.

Suggestions for Further Reading—Part II

On the kinetic theory:

LEMON, H. B.: *From Galileo to the Nuclear Age*, University of Chicago Press, Chicago, 1946. The chapters in this book on temperature, heat, and the kinetic theory are particularly good.

BRAGG, W. H.: *Concerning the Nature of Things*, Harper & Brothers, New York, 1925. A series of informal essays on the behavior and structure of matter.

On the atomic theory and the periodic law:

RICHARDSON, L. B., and SCARLETT, A. J.: *General College Chemistry*, Henry Holt & Company, Inc., New York, 1940. A standard elementary text.

DEMING, H. G.: *Fundamental Chemistry*, John Wiley & Sons, Inc., New York, 1940. A standard text, particularly valuable for its emphasis on the basic principles of chemistry and their applications.

HATCHER, W. H.: *An Introduction to Chemical Science*, John Wiley & Sons, Inc., New York, 1940. More elementary and more popularly written than either of the preceding, with particular emphasis on the industrial applications of chemistry.

On the history of physics and chemistry:

MAYER, J.: *The Seven Seals of Science*, D. Appleton-Century Company, Inc., New York, 1937. See p. 98.

SEDGWICK, W. T., TYLER, H. W., and BIGELOW, R. P.: *A Short History of Science*, The Macmillan Company, New York, 1939. See p. 98.

JAFFE, B.: *Crucibles*, Simon & Schuster, Inc., New York, 1930. A series of journalistically written biographies of the great chemists.

MOULTON, F. R., and J. J. SCHIFFERES: *The Autobiography of Science*, Doubleday & Company, Inc., New York, 1945. Quotations from the original works of Priestley, Lavoisier, Dalton, Rumford, and Joule.

PART III

THE STRUCTURE OF MATTER

THE kinetic theory assumes that all matter is made up of tiny moving particles. In gases the particles are far apart and attract each other only slightly; in liquids attraction between them is great enough to keep the particles close together but not to prevent their moving about; in solids the particles are held so firmly that their motion is restricted to vibrations about fixed positions. Increase in temperature leads to faster motion of the particles and so to a tendency for the particles to spread apart against their mutual attractions. In terms of these particles and the simple forces between them such diverse phenomena as boiling, freezing, the expansion of gases, and the flow of liquids find a complete explanation. Thus the kinetic theory introduces order and simplicity into the apparently complex behavior of ordinary materials.

Like Newton's picture of the solar system the kinetic theory is a *mechanical* explanation—an explanation depending on the motion of objects which exert simple forces on each other. In the solar system the objects are planets and satellites; in ordinary matter they are molecules. So successful were these two applications of mechanical ideas that nineteenth-century scientists looked eagerly for similar explanations in other fields. Some physicists ventured to predict that presently all natural phenomena would be interpreted by means of particles and forces acting between them—not only other phenomena in physics, but chemical reactions and even biological processes. Attainment of such a goal would satisfy the philosophical urge to find an underlying simplicity and order throughout the universe, and in the practical world would make possible the prediction and control of all manner of processes desirable for human welfare.

But in the nineteenth century the goal was still distant. In chemical phenomena certain basic principles had been recognized—that all matter is made up of a limited number of elements, that each element consists of tiny particles, that the particles or atoms of different elements are

related by the periodic law—but these principles provided no basis for mechanical explanation. Little was known about the motions of atoms and the forces which bound them together in molecules remained a subject for lively speculation. Moreover, the intricate relationships among different atoms brought out by the periodic law seemed to hint that these were not simple particles, but particles which themselves must have complex structures.

Likewise other branches of physics, in particular the phenomena of light and of electricity and magnetism, did not lend themselves readily to mechanical interpretation. These subjects had been under investigation since the time of Galileo, but rapid progress came only in the nineteenth century. Attempts to apply mechanical ideas were only partly successful and have remained so to this day.

In the chapters to follow we shall take up the study of electricity and light in some detail. We shall see how recent investigations of these subjects have unexpectedly furnished a partial mechanical explanation of chemical reactions, and we shall see further why a complete explanation in strictly mechanical terms is not possible either for chemical phenomena or for the phenomena of light and electricity.

CHAPTER XVI

Electricity and Magnetism

Comb your hair with a hard rubber comb, on a day when the air is dry: you find that your hair crackles, and its loose ends are attracted by the comb. Stand before a mirror in a darkened room and comb vigorously: you see tiny sparks jump from comb to hair. Rub a cat's fur with your hand, and touch your finger to the tip of his nose: the spark and the cat's injured expression are evidence of the electricity your rubbing has produced. Your comb rubbed on the cat's back, or more conveniently on a piece of fur no longer attached to a cat, is even more effective. Not only does it produce a sizable spark when touched to a near-by object but it attracts and picks up small bits of paper and cloth.

The first recorded investigator of such phenomena was Thales of Miletus, a Greek philosopher who lived about 600 B.C. Lacking rubber, Thales experimented with amber, in his language *electron*. We immortalize his work by saying that amber (or hard rubber) rubbed with fur possesses an *electric charge*—by which we mean simply that it is capable of producing a spark and attracting small light objects.

In Galileo's day came the first important elaboration of Thales' investigations, when Sir William Gilbert, court physician to Queen Elizabeth, proved that electric charges can be produced by rubbing together a great variety of different substances. In the seventeenth century electricity advanced slowly, many of its students seeking merely more spectacular ways of shocking and surprising audiences rather than knowledge of the phenomenon itself. Not until the time of the American genius Benjamin Franklin, in the middle of the eighteenth century, was the subject given much serious attention.

Electric Charges

Let us examine the behavior of electric charges more carefully. We begin by suspending a small pith ball from a silk thread, to serve as an

indicator of charges in its vicinity. If touched with a rubber rod which has been stroked with fur, the pith ball jerks violently away, and thereafter is strongly repelled whenever the rod is brought near (Fig. 100). We assume that the pith ball had no electric charge at the beginning of the experiment; at the instant of contact with the rod the ball acquired some of the charge on the rod, and in this charged condition is repelled by the rod.

Now bring near the same pith ball a glass rod which has been rubbed with silk. The ball is no longer repelled but strongly attracted (Fig. 100). With the ball in this condition, therefore, the charge of the rubber rod repels it, the glass rod attracts it. Now try the experiment in reverse: charge a second pith ball by touching it with the glass rod. It bounds away, evidently repelled. But the charged rubber rod attracts this second ball strongly.

We can draw only one conclusion: the charges on the two rods are somehow different. Furthermore, the kind of charge on one rod attracts the kind on the other, while each rod repels an object which has some of its own kind of charge. More simply, *like charges repel each other, unlike charges attract each other.*

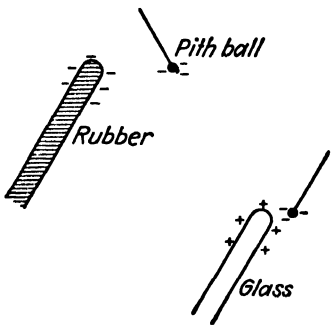


FIG. 100. *Like charges repel, unlike charges attract.*

Evidently we need names for the two kinds of electricity. In a more romantic age we might have called them male and female; or we might follow the earlier workers and label them vitreous electricity and resinous electricity. But by universal convention we follow Benjamin Franklin's suggestion, calling them positive electricity and negative electricity. A ***negative*** charge is *defined* as one similar to that produced in a rubber rod by stroking it with fur, while a ***positive*** charge is one similar to that produced in glass by rubbing it with silk.

We have concentrated our attention on the positive charge of the glass, the negative charge of the rubber. Strictly, however, we do not produce a positive charge alone by rubbing glass with silk, or a negative charge alone by stroking rubber with fur. If the fur used with the rubber is brought near a positively charged pith ball, the ball is repelled; if the fur is brought near a negatively charged ball, the ball is attracted. Thus the fur must have a positive charge. Similarly the silk used with the glass has a negative charge. In general, when electricity is produced by contact between two dissimilar objects, one acquires a positive charge and the other a negative charge, the distribution of charge depending on the nature of the two substances used.

Conductors and Insulators

Useful as a pith ball is, it can hardly pass for an accurate scientific instrument. Somewhat better is an *electroscope*, a device consisting of two leaves of thin metal foil suspended from a metal support inside a glass-walled box (Fig. 101). A charge of either kind applied to the metal support spreads itself over the leaves; charged with the same kind of

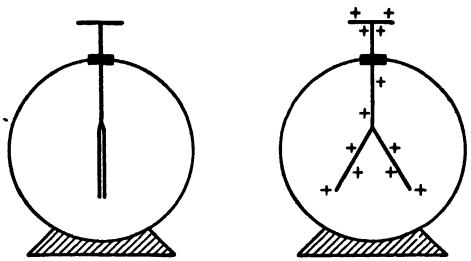


FIG. 101. *An electroscope, uncharged and charged.*

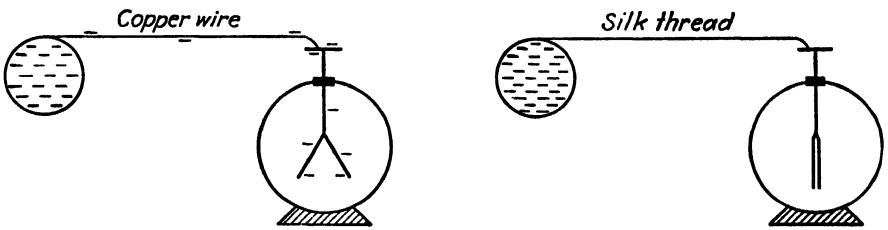


FIG. 102. *Conductor and insulator.*

electricity, they fly apart, the distance apart depending roughly on the size of the charge.

With this instrument let us study the movement of electric charges from one object to another. That charges are capable of motion, our experiments have already indicated; for example, when the pith ball was touched with an electrified rubber rod, part of the rod's charge moved onto the ball. Suppose we place an electroscope a foot or so from a charged object, say a large metal sphere, and that we make connection between the sphere and the metal of the electroscope by rods or wires of different materials. If a copper wire is laid between them, the leaves of the electroscope fly apart at once. With about equal readiness a charge is transferred from sphere to electroscope by wires of other metals. But if connection is made with a dry glass rod or a silk thread or a piece of dry wood, the electroscope is scarcely affected (Fig. 102). In other words, the ability of the charge to pass from sphere to electroscope depends on the nature of the material making the connection. Materials like iron or copper, which carry the charge readily, we call ***conductors***. Materials

like glass or dry wood or hard rubber, along which charges move with difficulty, we call ***insulators.***

A perfect conductor would be a substance along which charges could pass with no resistance to their motion, while a perfect insulator would be one through which no electric charge could be forced. No substance of either type is known. We find good conductors like copper, good insulators like rubber and silk, but these fall far short of perfection. Between good conductors and good insulators can be listed a great variety of substances along which charges move with more or less difficulty. Some of these intermediate substances may be considered either insulators or conductors, depending on circumstances; there is no sharp line between the two.

Ordinary air is a very poor conductor. If two strong opposite charges are brought close together, however, the air between them may momentarily become a good conductor. The air molecules become violently agitated, in a manner we shall describe presently, and the charges are able to leap across the gap. This type of sudden discharge is an *electric spark.* After the spark, unless the two charges are continuously renewed, the air reverts quickly to its normal nonconducting state.

The conductivity of pure water is extremely small. But traces of dissolved impurities increase the conductivity enormously; since most water which we use in daily life is somewhat impure, we come to think of it as a fairly efficient conductor. On humid days solids exposed to the atmosphere become coated with an invisible film of water, which makes even good insulators capable of conducting charges appreciably. Experiments with electric charges are difficult in humid weather, since the water films on insulators enable charges to leak away from pith balls and electroscopes; even in dry weather, experiments are improved if insulating materials are heated or rubbed with alcohol to remove adhering water molecules.

The earth as a whole, at least that part of it beneath the outer dry soil, is a fairly good conductor. Hence if a charged object is connected with the earth by a piece of metal, the charge is conducted away from the object to the earth. This convenient method of removing the charge from an object we speak of as *grounding* the object. You ground your radio by attaching the appropriate wire to a gas or water pipe (which is connected to other pipes below the surface), thereby giving electric charges in the radio a path to the ground. Ground connections for radios or lightning rods must be carefully made, using metal all the way; but for rough experiments with small electric charges, connection is sufficiently good if a person simply touches a charged object with his finger. The charge travels through his body, through the floor or the walls of the building, to water pipes or directly to the ground.

The human body is not a good conductor—a fortunate circumstance, for otherwise we should be shocked more frequently and more severely. But charges may be easily produced in the laboratory which can be very painful if allowed to move through the body to the ground. When properly insulated from the ground, however, say by standing on a plate of glass or on a platform supported by glass or rubber insulators, a person may be given a considerable charge, large enough to make separate hairs stand out away from each other and to make tiny sparks jump from his fingertips, without any unpleasant sensation. The charge is dangerous only when it can actually move through the body.

The readiness of charges to move to the ground, either through the air or through solid supports, makes exact experimentation difficult. Small charges tend to leak off slowly; even a pith ball, hung from a silk thread in dry air, will not maintain a charge indefinitely. Larger charges may find their way to the ground by sparking, unless sufficient insulation is provided.

Thus far we have said nothing about the actual nature of the electric charge. What really happens in a conductor when we say that a charge is moving from one point to another? Modern physics can answer this question fairly satisfactorily, and we shall study its answer in succeeding chapters. But the men who developed our knowledge of electricity to the point where an answer was possible worked with only a hazy notion as to what electricity actually is. Something moved, moved readily from one object to another. The only remotely comparable motion in their experience was the motion of substances like liquids and gases. So, quite naturally, they called electricity a "fluid."

At first two kinds of electric fluid were assumed. But Benjamin Franklin saw that explanations could be simplified by assuming that there was only a single fluid, that all objects normally possessed a certain amount of this fluid, and that one kind of electric charge represented an excess of this fluid, the other kind a deficiency. This was a happy suggestion, but unluckily for modern students Franklin chose the positive sign for his fluid. We know today that his choice was wrong: the electricity which usually moves through a solid from one object to another is negative electricity. This movable electricity consists of tiny particles called *electrons*, which in ordinary experiments act very much like Franklin's supposed "fluid." Our modern explanations of simple electrical phenomena, in terms of electrons, are exactly similar to Franklin's explanations with signs reversed.

Coulomb's Law

But before proceeding to the modern treatment of electricity, let us study electric charges quantitatively in an experiment which requires

no knowledge of their properties except that they attract and repel one another. The repulsive force between a rubber rod stroked with cat's fur and a negatively charged pith ball evidently depends on two things: how close the pith ball is to the rod, and how much charge each possesses. The influence of these two factors can be shown by observing (1) that the pith ball is not affected by the rod when it is some distance away, but is increasingly repelled the closer the rod is brought, and (2) that an increased charge produced by prolonged stroking of the rod with fur makes the repulsion stronger. A third factor is involved which we shall neglect for the moment—the influence of the air between the two charges. If the air were replaced by some other gas or a liquid, or were removed altogether, the repulsive force would be different.

The repulsive force between two charged objects of like sign is greater the smaller the distance between them. If the two objects are 2 cm apart, the force between them is one-fourth as great as if they are 1 cm apart, and four times as great as when they are 4 cm apart.* In other words, doubling the distance between the objects quarters the force between them, making the distance three times as large decreases the force ninefold, and so forth. Between force and distance there is evidently an inverse relationship, a relationship such that as the distance increases the force decreases as its square. In mathematical symbols

$$f = \frac{K'}{d^2}$$

We have as yet no means of stating accurately the amount of charge on an object. Suppose we select two small objects with equal charges of the same sign which, when placed 1 cm. apart, repel each other with a force of 1 dyne. Each of these charges we shall *define as a unit charge.* Suppose we increase the charge on one object until it repels the other with a force of two dynes; then we shall call its charge two units. If we increase its charge until it repels the unit charge with 5 dynes of force, its charge is five units, and so on. Similarly we could increase the other charge, leaving the first charge at one unit, and every unit increase in force would mark a unit increase in charge. By defining electric charges in this fashion, we arbitrarily make the force between them proportional to each charge; that is,

$$f = K''q_1 \qquad \text{and} \qquad f = K'''q_2$$

where q_1 and q_2 are the charges, K'' and K''' the proportionality constants. Now combine the three equations for force (Chap. IV, page 55):

* These relationships are exact only when the objects are very small compared with the distance between them.

$$f = \frac{Kq_1q_2}{d^2} \tag{26}$$

—an expression reminiscent of Newton's law of gravitation.

This formula holds as well for unlike charges as for like charges, except, of course, that the force is in the opposite direction. The formula is a universal law for all electric charges everywhere, as Newton's law of gravitation holds for all masses.* Established toward the end of the eighteenth century by a French engineer, Coulomb, it is called after him *Coulomb's law.*

Magnets

The metal iron is distinguished from most other substances by its ability to be strongly magnetized. Certain alloys, or mixtures, of other metals are also strongly magnetic; the metals cobalt and nickel, certain compounds of iron, and liquid oxygen are faintly magnetic. But for practical purposes the only materials widely used for strong magnets are iron and steel. In this limitation to a few materials† the property of magnetism is very different from the electrical properties we have been discussing.

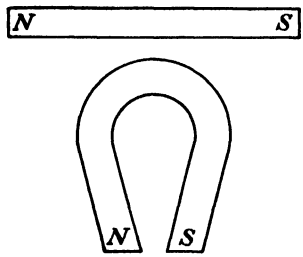


FIG. 103. *Bar magnet and horseshoe magnet.*

A bar of iron may be magnetized either by wrapping it with a coil of wire and passing a current through the wire, or by stroking it in one direction with another magnet. Magnetism can be removed from the iron by hammering it, or by heating it to a bright red heat.

A magnetized bar of iron (Fig. 103) is recognized, of course, by its ability to attract and hold other pieces of iron. Another property, illustrated by the familiar compass needle, is its ability when freely suspended to turn so that one end points north, the other south. We call the north-pointing end the ***north pole*** of the magnet, the south-pointing end the ***south pole.*** Here, near the ends, the greater part of the magnetization is concentrated, as may be easily shown by testing the attraction of various parts of the bar for small iron nails. If two magnets are brought near to each other, the two poles are found to behave quite differently. Laid end to end so that the two north poles are near together, the magnets repel each other, while if a north pole is brought near a south pole the

* With the provision, as in the case of gravitation, that the simple formula applies rigorously only to charges on spherical objects and on objects small compared with their distance apart.

† With delicate instruments, all substances can be shown to be very faintly magnetic. Some are attracted to a magnet, as is iron, while others are very slightly repelled.

two attract each other. We may formulate a simple rule analogous to that for electric charges: *like magnetic poles repel each other, unlike poles attract.*

It would be convenient for experimental purposes if we could isolate a single north pole unencumbered by the south pole at the other end of the magnet. Seemingly, isolation of the north pole should not be difficult; we need only saw the magnet through its middle. Unfortunately this method will not work. The resulting half magnets have each a north pole and a south pole, two new poles appearing miraculously where the middle of the former magnet was. We may cut the resulting magnets in two again with the same results, and continue as long as our patience holds out or as long as we have tools fine enough for the cutting, and still each magnet that we prepare, however small, will have a north pole and a south pole. There is no such thing as a free magnetic pole.

Since a magnet can be cut into smaller and smaller fragments indefinitely, each fragment acting as a small magnet, we may reasonably assume that magnetism is a property of the smallest particles of a substance. Presumably each particle of iron (or any other material which can be magnetized) behaves as if it had a north pole and a south pole. In ordinary iron the particles are haphazardly arranged, and adjacent north and south poles neutralize each other's effect. When the iron is magnetized, we imagine that many or all of the particles are aligned with their north poles in the same direction, so that the strengths of all the tiny magnets are added together (Fig. 104).

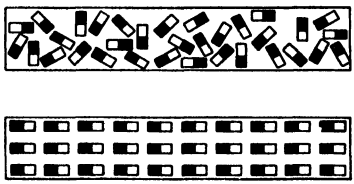


FIG. 104. *Diagram showing arrangement of particles in two bars of iron, the first one unmagnetized and the second magnetized.*

Let us try an experiment with magnets analogous to the electrical experiment with conductors and insulators. Place a magnet a foot or so from an unmagnetized piece of iron and connect the two by various materials. No substance will be found to carry the magnetism from the original magnet to the unmagnetized iron except a third piece of iron (or a magnetic alloy). If iron is used for the connection, the unmagnetized piece becomes temporarily able to attract small nails or filings; but all or nearly all of its magnetism vanishes when the connection is severed. Meanwhile the strength of the original magnet remains practically unimpaired. That is, when a magnet touches a piece of iron, its magnetism in effect extends along the iron, but is not permanently conducted through it in the sense that an electric charge is carried through a conductor.

Quantitatively, the force between two magnetic poles (like or unlike) may be expressed in a law very similar to Coulomb's law for electric

charges. Consider that each pole* is far enough from its partner so that the latter's influence is negligible. Then if the magnetic strength of one pole is p_1, of the other p_2, and if d as usual is the distance between them, the force of attraction or repulsion is

$$f = \frac{Kp_1p_2}{d^2} \tag{27}$$

The rules for the behavior of magnetic poles are thus deceptively similar to those for electric charges. Magnetism and electricity are subtly related, as we shall see presently; but magnetic poles and electric charges *at rest* have little in common and are without effect one on the other. This is an important and easily forgotten fact. *A magnet is completely uninfluenced by a stationary electric charge of either sign in its vicinity, and vice versa.*

Let us summarize a few of the resemblances and differences between magnetic poles and electric charges.

Resemblances:

1. The rules of attraction and repulsion are similar; that is, like poles repel each other, like charges repel each other, while unlike poles attract and unlike charges attract.
2. The quantitative expressions for electric and magnetic force are similar.

Differences:

1. Magnetism is exhibited noticeably only by iron, steel, and a very few other substances, while any material properly insulated may be given an electric charge.
2. Magnets, if free to move, align themselves in a north-south position, while electric charges have no such tendency to orientation.
3. An electric charge of either sign may be isolated completely, independent of the opposite charge; magnetic poles always occur in pairs in the same object, and cannot be isolated.
4. Electric charges move readily from one object to another along substances known as conductors; magnetism is not moveable.

Terrestrial Magnetism

Scheherazade, fabricating the thousand-and-one-nights' story which was to save her from the Caliph's executioner, retold the legend of the bold sailor Sinbad, whose curious adventures were brought to a close

* In this law each "pole" is a point near, but not at, the end of the magnet, where the magnetism may be regarded as concentrated.

when his ship was irresistibly drawn to a large black rock and dashed to pieces against its side. Neither Scheherazade nor the Caliph knew, presumably, that this legend was even in their day more than 1,000 years old: for the Greek Homer describes a similar calamity which befell his hero Ulysses. Again and again the old yarn appears, cropping out even in our own times when early explorers brought back from Yellowstone Park tales of a magnetic mountain which pulled the iron nails from their shoes. Responsible for these fantastic legends is the slightly magnetic black rock called *magnetite* (iron oxide, Fe_3O_4); found abundantly in Asia Minor, it was evidently known to the Greeks from very early times. But neither the ancient Greeks nor Romans nor Arabs, as far as we can tell, ever learned that magnetite, or a piece of iron stroked with magnetite, would turn to a north-south position. This fact, probably discovered in China, found its way to Europe in the twelfth century, giving navigators a much-needed means of setting their courses on the high seas.

As a nautical instrument the magnetic compass has one serious disadvantage, which the terrified sailors of Columbus discovered on their first voyage to America: the direction of the needle varies from place to place on the earth's surface, being true north at only a few points. Long after Columbus' voyage other explorers learned the reason for this variation. The north pole of a magnet is attracted, not to the geographic north pole, but to a point in northern Canada, just west of the mouth of Hudson Bay, which we call the *magnetic north pole* (actually an *S* pole, since the *N* pole of a magnet is attracted toward it). In the central part of the United States the magnetic pole is nearly in line with the geographic pole, and compasses give approximately true directions. On the Atlantic coast the needle points somewhat west of north, on the Pacific coast somewhat east of north. In far northern latitudes, as in Alaska or Greenland, the deviation of the compass from true north (its "declination") becomes very large; in the Arctic Ocean north of Canada the north end of the needle would point south.

To explain the behavior of the compass needle, we assume that the earth itself is a huge magnet, with one pole in northern Canada and the other near the edge of the Antarctic Continent, several hundred miles from the south pole. Commonly we speak of the magnetic poles as points on the surface, like the geographic poles; actually they are far beneath the surface. This is indicated by a simple experiment: suspend a compass needle so that it can move vertically as well as horizontally; the north end will dip steeply downward, in the latitude of New York making an angle of about 70° with the horizontal. This angle, the "dip," increases to the north so that the needle stands vertically over the magnetic pole. Thus the point toward which the needle is attracted apparently lies deep within the earth.

The earth's magnetism is subject to considerable variation. At times the changes are rapid and only temporary: these are the so-called "magnetic storms," probably connected with sunspot activity. Less well understood is an apparent slow movement of the magnetic poles, resulting in a change in declination in the United States of about half a degree in thirty years.

Fields of Force

Electric and magnetic forces have grown so familiar in our lives that they no longer excite especial wonder—except perhaps when displayed on a grand scale, as in artificial lightning or huge commercial electromagnets. Yet these forces, reduced even to the naked simplicity of a pith ball and a glass rod, or a small toy magnet, should perhaps be a source of deep amazement not inferior to the starry heavens. Electricity and magnetism, just in the fact of their existence, do violence to our intuitive ideas of matter, space, and the nature of force.

These forces differ in one significant respect from most other common forces: they act without contact across intervening space. I cannot move a book from the table by waving my hand toward it; my golf ball will not move from its tee, however violent my exertion, until the clubhead actually makes contact with it. But consider a charged pith ball and glass rod: the ball does not wait for contact but senses, so to speak, the presence of the rod while it is yet some distance away. Remove the air, place rod and ball in the most perfect vacuum attainable, and the force is not diminished. The two react on each other without benefit of anything that our senses can detect. For the pith ball the region near the rod is somehow different from other space, since in this space it is impelled to move. Near any electric charge is such a region of altered space. Likewise a magnet, though in a different way, changes the space around it, so that material of the proper sort is acted on by a force.

Now what do we mean by an "alteration of space"? We regard space, customarily, as emptiness, the complete absence of matter. How can *nothing* be altered? Nineteenth-century physicists avoided this philosophical difficulty by inventing the *ether*—an all-pervading "fluid" with such marvelous properties as perfect transparency and perfect rigidity, through which matter could move without friction. Modern physicists discard the ether and talk rather of the properties of space—more realistic, surely, but not very helpful in trying to visualize the presence of a force in emptiness.

A third force must be added to those of electricity and magnetism, a force so much a part of experience that we take its working for granted—the force of gravitation. Just as an electric charge or a magnet creates some sort of disturbance in the space about it, so does any object, simply

because of its mass, alter space so that other objects move toward it. These three, the only forces we know which can act from a distance, are markedly different in most respects, but similar laws connect the force in each case with the magnitude of charge, magnetic strength, or mass, and with the distance from the exciting object. Because the three laws

$$f = K\frac{m_1 m_2}{d^2}, \qquad f = K'\frac{q_1 q_2}{d^2}, \qquad f = K''\frac{p_1 p_2}{d^2}$$

each involve d^2 in the denominator, they are often called simply the *inverse square laws.*

In science as elsewhere we must sometimes cover up a fundamental mystery by giving a name to a phenomenon and describing it minutely.

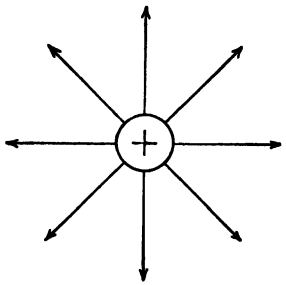


FIG. 105. *The electrostatic field around an isolated positive charge.*

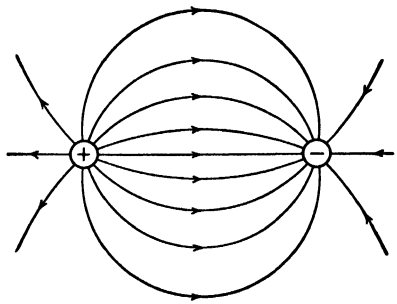


FIG. 106. *The electrostatic field around two unlike charges.*

We call the region of altered space around a mass, an electric charge, or a magnet a ***field of force,*** and in so naming it we forget our primitive wonder at its behavior. Technically a field of force extends to infinite distances in all directions, since a mass, a charge, or a magnet presumably exerts a force everywhere in the universe. But practically the force becomes negligible a relatively short distance from the exciting object, and by its "field of force" we ordinarily mean the space immediately adjacent to the object, the space in which its force is perceptible.

Fields of force are recognized by the tendency of objects in them to move, and we describe fields accordingly in terms of motions. Suppose we are to describe the electric field about a positively charged metal sphere: we ask, how would a small charge move in this field? Evidently, if the small charge is negative, it would move in a straight line toward the sphere; if it is positive, it would move directly away from the sphere. (Neglect gravity, air resistance, friction, the support of the sphere, etc.) Hence we describe paths of motion in the field as straight lines radiating out from the sphere's center (Fig. 105). Conventionally we consider the direction of motion of a *positive* charge and indicate this motion by placing outward-pointing arrows on the lines. A negatively charged sphere

would have a field of the same shape, but its lines would be directed inward.

Consider now the field about two adjacent charges, one + and the other −. Here again we ponder the question, how would a small + charge move? Between the charges it would move directly from +

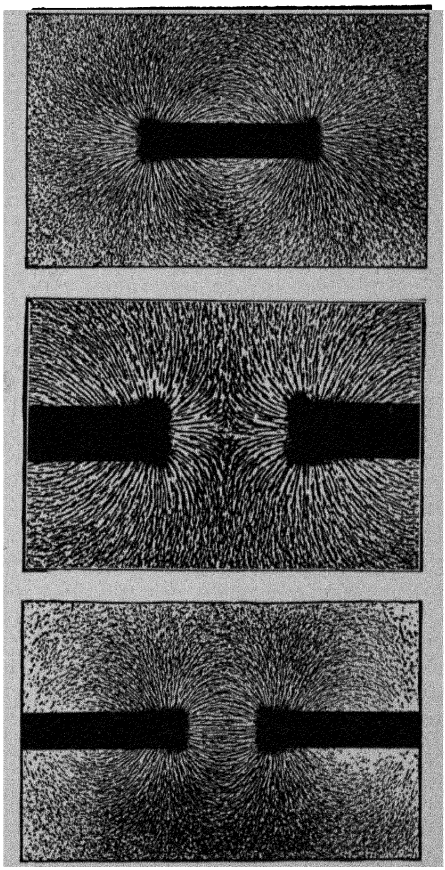


FIG. 107. *Drawings of iron filings around a single bar magnet (upper picture), around two like poles (middle picture), and around two unlike poles (lower picture). In each case the filings show the lines of force of the magnetic field. (From O. M. Stewart's Physics, by permission of Ginn and Company, publishers.)*

toward −; off to one side, repelled by one and attracted by the other, it would move in a curved path (Fig. 106).

Fields of magnetic or gravitational force may be similarly described. Magnetic fields are complicated by the continual presence of at least two poles, and by the fact that we must consider the motion of an object which does not exist—a free N or S pole. We can gain an approximate

picture of the field by using a small compass, making the reasonable assumption that the compass will point in the direction in which a free pole would move. On this principle we obtain a rough idea of a magnetic field by a simple experiment: scatter iron filings on a glass plate above a magnet, and shake them slightly. Each bit of iron, under the influence of the magnet, becomes in effect a small compass and aligns itself with the magnet's field. The result is shown in Fig. 107, and an ideal representation of the field in Fig. 122, page 242.

The lines used to describe a field—gravitational, magnetic, or electric—are called *lines of force*. Although purely fictitious, they enable us to describe concisely how an object will tend to move in any part of a field. Where many lines are close together the field is strong; where they spread apart the field is weak. Note that lines of force do not cross or branch, and theoretically can end only on one of the exciting objects.

But useful as lines of force are in advanced study of electricity and magnetism, they are purely descriptive. They tell us what will happen in the field, but not what the field is or how it can exist in empty space.

Questions

1. Describe carefully simple experiments to prove that there are two kinds of electricity.
2. Contact between road and tires often produces an appreciable electric charge in the body of an automobile. This charge is a hazard to trucks carrying inflammable materials, such as gasoline, since it may accumulate sufficiently to produce a spark. A gasoline truck usually has a metal chain attached to its frame which touches the road as the truck moves. Explain how this chain prevents dangerous sparks.
3. Name four good conductors and four good insulators.
4. Death by electric shock is sometimes caused when a person while taking a bath touches a poorly insulated light switch. Why is the shock so much more dangerous than usual under these conditions?
5. Suppose that two small, similar positive charges placed 1 cm apart repel each other with a force of 0.2 g. What would the force become if
   a. The distance is increased to 3 cm?
   b. One charge is doubled?
   c. Both charges are tripled?
   d. One charge is doubled and the distance is increased to 2 cm?
6. Which is greater, the force with which a ten-unit charge repels a one-unit charge, or the force with which the one-unit charge repels the ten-unit charge?
7. List the important differences between electric fields and gravitational fields.
8. By means of lines of force, draw diagrams of
   a. The field around two spheres carrying negative charges, with their centers 4 cm apart.
   b. The field around two small bar magnets placed in a line with the N pole of one about 2 cm from the S pole of the other.
9. Describe what would happen if a positively charged electroscope were touched with
   a. A rubber rod which has been stroked with fur.
   b. A glass rod which has been rubbed with silk.

CHAPTER XVII

The Electron

IN THE repertoire of modern physics are few experiments more spectacular than a cathode-ray demonstration. On the lecture table is supported a long glass tube, with a small metal disk sealed into each end (Fig. 108). A side tube is connected with a vacuum pump. The room is darkened and the experimenter throws a switch, giving one metal disk a strong positive charge, the other a strong negative charge. Nothing is visible yet except a faint glow at each end of the tube, since too much air sepa-

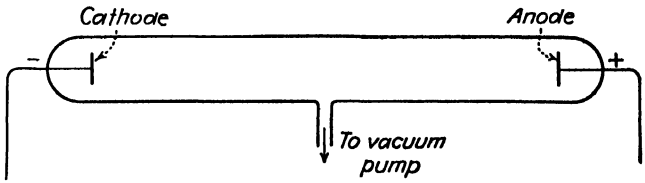


FIG. 108. *Diagram of cathode-ray tube.*

rates the charges for a spark to jump between them. Now the vacuum pump is connected, reducing slowly the air pressure in the tube. Suddenly a long spark leaps down the tube, brilliant, narrow, dancing from side to side. Widening as the air pressure falls, the spark becomes a wavering purplish column filling the tube completely. Now the pressure has dropped to less than 1 mm of mercury (less than a thousandth of normal atmospheric pressure): the purple column breaks near one end, the dark gap grows wider, and presently the tube is once more nearly invisible. Still the pump pulls traces of air from the tube, reducing its pressure to 0.001 mm or less. Now at last, on the walls of the tube opposite the negative disk, appears a greenish glow, faint at first, but spreading and brightening until much of the glass is softly luminous. The green radiance shows that the tube is filled with the once-mysterious *cathode rays.*

Electric discharges in tubes of this sort attracted the serious attention of physicists in the 1870's, but not until 1898 was the green glow of the glass at very low pressures explained. Discovery of the true nature

of cathode rays we owe to many workers, in this country and in Europe, but one name stands out above the rest—that of the great English physicist, J. J. Thomson. With Thomson's work on cathode rays begins that extraordinary scientific development which we call "modern physics."

The Nature of Cathode Rays

The two metal disks of the evacuated tube are called its *electrodes*—a general term applied to any pair of adjacent conductors on which unlike charges are maintained. The negative electrode is named the *cathode*, the positive one the *anode*. Cathode rays are so named because they stream

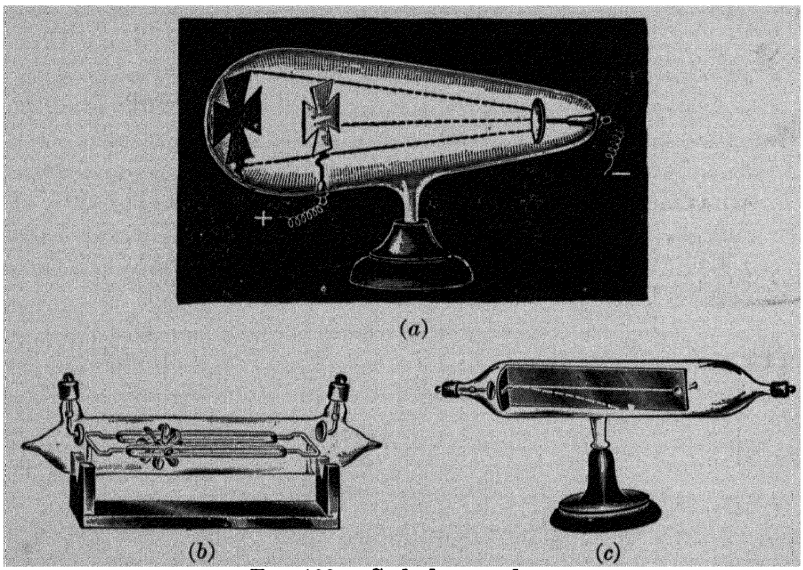


FIG. 109. *Cathode-ray tubes.*

from the negative metal plate. We can prove this fact very simply by using a tube equipped with an opaque object which can be set before the cathode: the object is silhouetted sharply against the green glow at the far end of the tube (Fig. 109*a*).

Opinion was divided in the early days of cathode-ray experimentation, one group maintaining that the radiation was a form of invisible light, the other group that the radiation consisted of material particles. Experiments designed to settle the issue brought out many properties of the mysterious rays. Under their influence not only glass but many other substances will glow softly, or *fluoresce*. Experiments showing that obstacles in the path of the rays produce sharp shadows proved that the radiation travels approximately in straight lines. Because they move in straight lines, the rays can be focused on a point by using a cathode with

a concave surface, somewhat as light is focused by a concave mirror. A piece of platinum placed at the point where the rays converge quickly becomes red hot, proving that the cathode radiation carries considerable energy. All these properties could be explained satisfactorily either by supposing that the rays consisted of particles or by supposing that they were akin to ordinary light.

But other experiments favored the particle hypothesis. If a metal plate is sealed into the side of a cathode-ray tube and given a strong negative charge, the greenish fluorescence moves from the end of the tube to the wall opposite the charge (Fig. 110). Evidently the rays are deflected away from the charge as if they possessed a negative charge themselves. No radiation like light has ever been found to possess a charge. Furthermore,

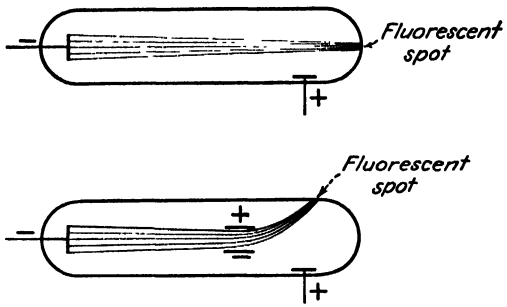


FIG. 110. *A stream of cathode rays is deflected away from a negative charge, toward a positive charge.*

if a tiny paddle wheel is placed on a runway in the tube with its vanes in the path of the rays (Fig. 109*b*), it moves rapidly away from the cathode, showing that the rays possess considerable kinetic energy—a property we associate with material particles. Light rays, we shall see presently, do possess a little kinetic energy, but not nearly enough to cause the motion of the paddle wheel. These and other experiments established beyond question that cathode rays are actually tiny material particles, each with a minute charge of negative electricity. We call these particles ***electrons.***

The mass and charge of the electron and the speeds with which it travels were ascertained in the early years of this century by experiments of two sorts: (1) measurement of the deflection of cathode rays in electric and magnetic fields and (2) measurement of the rate of movement of charged oil droplets falling through an electric field. Details of these experiments are not important for our purposes, so we shall merely summarize their results:

1. The electron is exceedingly small. Its mass is about 9.11×10^{-28} g., about 1/1,800 of the mass of a hydrogen atom.

2. The charge of the electron, measured in the units of charge defined in the last chapter, is 4.80×10^{-10}. Small as it is, this charge is enormous for an object as tiny as an electron. Thus, a gram of electrons at a distance of 1 cm from another gram of electrons would be repelled with a force of 3×10^{26} tons.
3. The speed of an electron in the cathode-ray tube is prodigious. Speeds of over 170,000 mi/sec have been measured—over nine-tenths the velocity of light.

Here is a particle smaller than the tiniest atom, bearing a negative charge enormous for its size, hurled from the cathode at an incredible speed. It makes no difference what material is used in the cathode: electrons are emitted from all substances, and the electrons are all precisely alike. Their speeds and number may differ with different substances, but not their mass* or charge. Thus we are led to the sweeping conclusion that electrons are a part of all matter—part of John Dalton's indivisible and indestructible atoms.

Electrons in Ordinary Matter

In simple experiments with solid materials, these ever-present electrons play a role similar to that which Franklin imagined for his positive electric "fluid." Franklin thought that a positively charged object had an excess of "fluid," a negatively charged object a deficiency. Today we should say rather that a positive charge indicates a deficiency of electrons and a negative charge an excess.

In the language of electrons, a conductor is a substance whose electrons are relatively free to move, while an insulator is a substance whose electrons are firmly bound in their atoms. If a copper wire is placed between a negatively charged sphere and an electroscope, some of the movable electrons in the wire, repelled by the negative charge, move onto the electroscope and give it a negative charge. Note that electrons do not necessarily move all the way from sphere to electroscope: Those on the sphere, so to speak, push electrons from the other end of the wire onto the electroscope. If silk thread is used in place of the wire, its electrons, although repelled by the charge, are so firmly fixed in their atoms that almost none move onto the electroscope. A positive charge on the sphere would draw electrons out of the wire, or out of any piece of metal connected to it. Thus, electrons move easily along a metal away from a negative charge or toward a positive charge—which is a way of repeating the statement that all metals are conductors.

* This statement needs qualification for electrons moving at very high speeds. All material objects, including electrons, show an increase in mass at speeds near the velocity of light (see p. 653).

We can also say, a little less certainly, that the charging of a rubber rod by stroking it with cat's fur means that electrons from the fur have been transferred to the rod, making the rubber negative and leaving the fur positive. Similarly a glass rod rubbed with silk acquires its + charge by giving some of its electrons to the silk. But these experiments with nonconductors are probably actually more complex, involving not only movements of electrons but also movement of positive and negative ions.

An *ion* is an atom or molecule with an electric charge. Ordinary atoms or molecules are electrically neutral: They contain electrons, but they likewise have enough positive charge to neutralize the electrons. If an atom or molecule loses one or more of its electrons, it becomes a positive ion, and if it gains an extra electron or two it becomes a negative ion. Ions are particularly important in the conduction of electricity through gases and liquids. A spark jumps through air, for instance, only when strong charges on the electrodes have changed some of the air molecules into ions; the spark itself is the disturbance created by rapid motion of positive ions toward the cathode and negative ions toward the anode. The process of electroplating, for example the plating of silver on an iron spoon, involves the movement of positive silver ions through a solution.

Thus, conduction of electricity in metals is a movement of electrons; conduction through gases or liquids is a movement of ions. Experiments with nonconductors like cat's fur, rubber, and pith balls probably involve both types of conduction. The ions here are not only ions of the substances themselves, but ions of air or water in thin invisible films on the surfaces of the objects. The complete explanation of these simple experiments becomes very complicated; we will do best to ignore the ions for the present and treat the experiments as if electrons alone were moving.

Induction

With the aid of electrons we may attempt an answer to a question we have so far carefully avoided. One sign that a body possesses an electric charge is that it causes other objects to move toward it: why this attraction for uncharged objects? Let us return to our overworked rubber rod and pith ball. If the rod is given its usual negative charge and brought near the ball, electrons in the ball, repelled by its charge, move as far away as they can—*i.e.*, to the far side of the ball (Fig. 111). The side of the ball near the rod is left with a positive charge, and the ball is accordingly attracted to the rubber. If the rod is removed without actually touching the ball, the disturbed electrons resume their normal positions and the ball is unchanged. But if contact is made, some of the rod's electrons flow onto the ball, giving the ball as

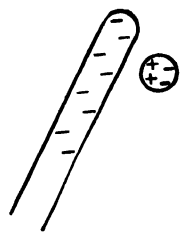


FIG. 111. *Separation of charges on a pith ball produced by induction.*

a whole a negative charge and causing the violent repulsion we have observed before.

The + and − charges on opposite sides of the pith ball, *produced without actual contact with another charge*, we call *induced charges*. The induced charges are temporary, disappearing as soon as the negative rod is removed. But a permanent charge may be fixed on the ball by induction, provided that the ball is grounded while the rubber rod is near (Fig. 112). Grounding the ball (say by touching it) makes it and the earth temporarily part of a single huge conductor; in this conductor the negative rod drives electrons as far away as possible, that is, from the ball down into the earth. If now the ground connection is broken while the rod is still held near by, the electrons have no means of returning to the ball, and it therefore has a positive charge which will remain when the

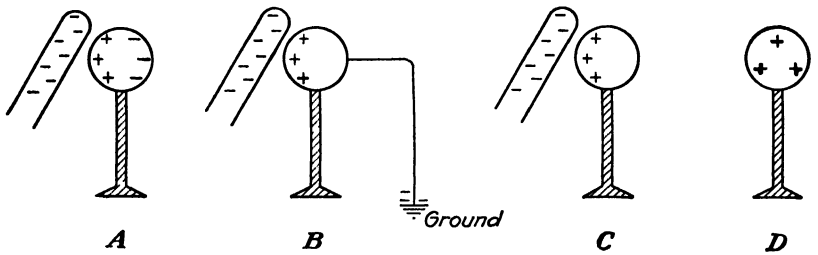


FIG. 112. *Charging an insulated sphere by induction. A, separation of charge induced by presence of rubber rod; B, sphere grounded; C, ground connection broken, + charge stranded on sphere; D, rubber rod removed.*

rod is taken away. In the same fashion a negative charge may be fixed on another pith ball: hold near it a positively charged glass rod; ground the ball, letting electrons attracted by the rod flow up from the ground into the ball; break the ground connection and remove the rod, thus leaving the excess electrons stranded on the ball. Note that the induced charge on the ball is always opposite in sign to the original charge.

In general, any object with a negative charge induces positive charges on all objects near it, simply because it repels away from itself the movable negative charges in these objects. Insulated objects near the negative charge will have a concentration of positive electricity on the side toward the charge, a concentration of negative electricity on the opposite side. Grounded objects will have lost some of their electrons to the earth. The amount of the induced charge on any object depends on the size and shape of the object and on its nearness to the negative charge. In similar fashion a positive charge induces negative charges on objects in its immediate vicinity.

We may also speak of *induced* magnetic poles produced in pieces of iron by the presence of a near-by magnet. In a nail lying near the *N* pole

of a horseshoe magnet, for instance, an S pole is induced at the end near the magnet and an N pole at the opposite end. The attraction of the nail's S pole for the magnet's N pole becomes evident if the two are moved a little closer together.

Thus ***induction*** is a general term, referring to the production of electric charges or magnetic poles by the mere presence of other charges or magnets. Actual contact is not necessary.

All machines for producing large electric charges operate by induction rather than by friction. Perhaps the simplest example of such a machine is the *electrophorus*, which consists of a flat piece of hard rubber and a metal disk of about the same size provided with an insulating handle (Fig. 113). The rubber is first given a charge by rubbing it with cat fur, and the metal disk is placed upon it. Since the metal makes actual contact with the rubber at only a few points and since the rubber is a very poor conductor, the rubber's excess electrons do not flow directly to the metal. But the metal while in this position can be given a powerful induced charge. Placing it on the rubber causes some of its electrons to be driven from its bottom face to its top face. If the metal is now momentarily grounded, these excess electrons on its top face are removed, leaving the disk as a whole with a strong positive charge. When lifted from the rubber and brought close to a grounded conductor, it is capable, under favorable conditions, of producing a spark nearly 1 cm long. The metal may be placed once more on the rubber platform and, if grounded, will acquire a new charge. The charge may be renewed in this fashion again and again, without seemingly diminishing the charge on the rubber platform at all.

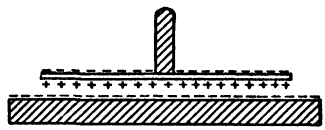


FIG. 113. *An electrophorus.*

Whence this apparently inexhaustible supply of electricity? Where does the disk gain the energy necessary to produce one fat spark after another? Consider the steps in the charging process. Little energy is required to give the rubber platform its original charge, for this is accomplished simply by contact with the cat fur. Negligible amounts of energy are involved in placing the metal disk on the rubber and in making and breaking the ground contact. But when the disk is lifted away from the rubber, two strong unlike charges are pulled apart against their attractive force—a process which evidently requires energy. It is this energy which reappears in the spark.

More complicated electric machines work on the same principle: the mechanical energy necessary for separating unlike charges is turned into electrical energy. The usual type of machine is arranged so that mechanical energy can be supplied by turning a crank and so that charges of opposite sign can be accumulated on two metal spheres (Fig. 114). When

the spheres are placed close together, only a relatively small excess of electrons on one and deficiency of electrons on the other is sufficient to make a spark pass between them. Further turning of the crank builds up a charge for a second spark. As the spheres are moved farther apart, more and more charge is necessary to produce the spark; hence a longer interval elapses between discharges. If the spheres are pulled too far apart, no amount of turning can cause a spark to jump, as the charges leak off into the air before they can attain sufficient size.

Dwarfing the largest sparks man can produce are the magnificent natural sparks we call lightning. The charges responsible for lightning accumulate by the tearing apart of water droplets and ice crystals in violent air currents, the smaller fragments being left negatively charged and the larger ones positive; further air movement concentrates negative

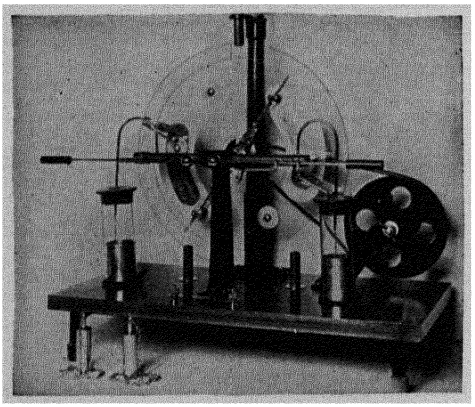


FIG. 114. *An electrostatic machine. (Courtesy of Central Scientific Company.)*

fragments in one cloud, positive fragments in another. The lightning discharge may take place between two such charged clouds, or between a cloud and a conducting object near the earth which is charged by induction. Suppose a positively charged cloud drifts near a grounded conductor, say a building or a tree or a mountaintop (Fig. 115). The strong + charge pulls electrons from the earth into the conductor, giving it a powerful induced negative charge; if the cloud approaches close enough, the charges will be neutralized in a gigantic spark. Thunder is caused by the very rapid expansion of strongly heated air along the path of the spark.

Electrons and Electricity

In the tiny charges on pith balls and in the enormous charges developed by a thunderstorm we see the same phenomenon: the piling up of electrons on one object, lack of electrons on another. Presently we shall

find in electric lights, in motors and dynamos, in telephones and waffle irons, further manifestations of the accumulation and movement of these tiny negative particles. In X-ray tubes and in the vacuum tubes of our radios we bring the electron out in the open, leaping across a vacuum, as we found it in the cathode-ray tube. Later on we shall discover that chemical reactions are ordered and simplified when explained in terms of electrons. More and more the electron enters the fabric of scientific thinking, even in geology, astronomy, and biology.

Now and again in our pursuit of the electron we must pause to ask the simple question: What is it? What manner of thing is this tiny, ever-present and ever-restless particle? For this question we shall not find an answer. To the end of our discussions we shall be talking about an object we can never see, whose shape and color and even size we cannot describe. The ablest physicists despair of gaining a satisfactory picture of the electron. Their only recourse is to describe its properties and behavior as accurately as possible; and we can but follow humbly in their footsteps, without the immense advantage of their mathematical equations.

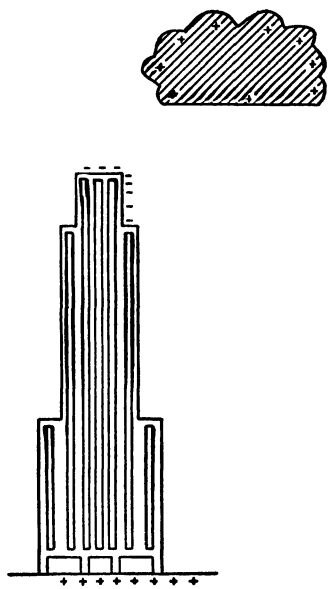


Fig. 115. *Diagram to show why lightning may strike a tall building. A high positive charge on a cloud induces a negative charge on the building.*

Two fundamental properties of the electron we have already learned: (1) It possesses mass, which means that it has inertia and should be attracted to other masses by Newton's law of gravitation; and (2) it has a negative charge, which means that it is attracted and repelled by other charges according to Coulomb's law. Now we face a subtle question: Is the electron a particle of mass possessing an electric charge, or is it an electric charge itself? That is, can an electric charge possess mass, or are the two properties separate and distinct?

The question is not as formidable as it sounds, but its answer involves a clarification of our notions about electricity—a word whose definition we have heretofore scrupulously overlooked. What is an electric charge? Dodging the issue, we can reply, "An excess or a deficiency of electrons." But then what is the charge of the electron itself? Here we can no longer dodge. An electron differs from an uncharged object of the same mass in that it will cause other charges to move: in other words it possesses energy. Perhaps this is our usual subconscious

idea of electricity: a form of energy which will cause objects to move or emit light or become hot. If we *define* electricity in this fashion, we may regard the electron as a mass having a certain quantity of electrical energy.

On the other hand, electric charges are not known apart from electrons and a few other charged particles whose acquaintance we shall make presently. Since these particles represent electricity reduced to the simplest terms we know, it would surely be reasonable to *define* electric charges as consisting of these particles. With this definition we may regard the electron as an electric charge itself.

Thus our question may be answered either way, depending on the definition we adopt for electricity. With equal correctness we may say that the smallest unit of electric charge possesses mass, or that one of the smallest particles of matter possesses electricity. The essential thing is that the two are inseparable: all ordinary matter contains electric charges, and all electric charges are associated with particles having mass.

Questions

1. Name two properties of cathode rays which suggest that they are streams of particles rather than a form of wave motion like light.
2. Account for the ability of a charged rubber rod to attract and pick up small bits of paper or cloth.
3. Suppose that a negatively charged rubber rod is held near (but not touching) an electroscope, that the electroscope is momentarily touched with the finger, and that the rod is then removed. What charge, if any, is left on the electroscope? Explain.
4. How could you charge an electroscope negatively by induction?
5. In terms of electrons explain why the production of electricity by friction must always give equal amounts of positive and negative electricity.
6. In the sequence of operations shown in Fig. 112, suppose that the charged rod is removed *before* the ground connection is broken. What charge will be left on the sphere?

CHAPTER XVIII

Electric Currents

THUS far we have concerned ourselves chiefly with the branch of electricity called *electrostatics*—the study of static, or stationary, electric charges. We have mentioned the ability of charges to move —from rubber rod to pith ball, or along a wire to an electroscope—but we have focused attention on the result of motion, rather than on the motion itself. Further, all these motions have been extremely short-lived: a charge simply moved suddenly from one object to another, then remained stationary until conditions around it were changed. Now this kind of electricity is fascinating certainly, and often spectacular; but it is not the kind of electricity which lights our houses, runs our trains, and curls our hair. The electricity of everyday use is current electricity—electric charges in continuous motion. The study of electric currents, which we shall touch briefly in this chapter, is often given the impressive title *electrodynamics*.

There is no fundamental difference, of course, between a moving charge and one at rest. A moving charge is the same sort of individual we have met in the last two chapters, attracting and repelling other charges, causing sparks if given a chance. But motion gives it a few quirks of behavior which we could not have suspected from our previous study.

Properties of Electric Currents

We shall start with the simple type of current produced when a wire is connected between the terminals of a battery. A battery is a device for maintaining a positive charge on one terminal (or electrode), a negative charge on the other. This is accomplished by a chemical reaction in the battery, chemical energy being continuously transformed into electrical energy whenever the battery is in use.

Connecting a wire between the terminals provides a path for the excess electrons at the negative electrode to move toward the positive

electrode. This flow of electrons tends to destroy the two charges, but chemical processes in the battery build up the charges as fast as they are depleted. Thus the current in the wire consists of a continuous movement of electrons from one end to the other. Note that we do not say: "The electrons carry the current," or "The motion of electrons produces a current"; the moving electrons *are* the current.

Suppose the wire is cut, and its free ends are attached to a switch: then by closing and opening the switch we may start or stop the current at will. The current flows whenever a continuous conducting path is provided between the electrodes and stops whenever the path is broken. When the conducting path is complete (*i.e.*, when the switch is closed), we say the circuit is *closed;* when the path is interrupted, we say the circuit is *open.*

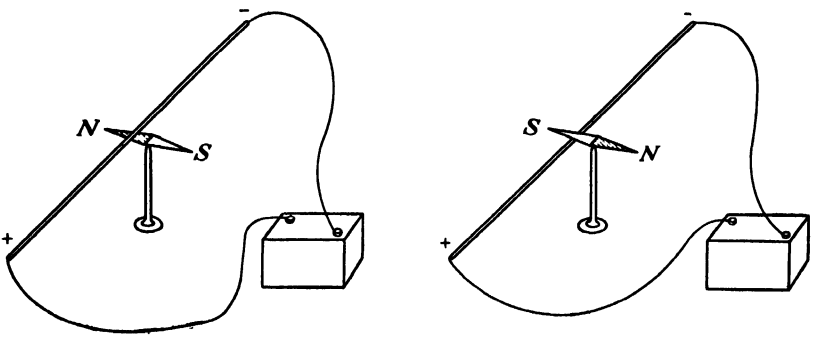


FIG. 116. *Oersted's experiment. A compass needle tends to swing into a position at right angles to a wire carrying a current.*

Given a battery, a switch, and wires of different sizes and materials, let us examine electric currents in some detail. First of all we face the question: How is an electric current recognized? A wire looks the same, whether the circuit be closed or open; how can we tell whether or not electrons are moving along it?

One way in which a current may betray its presence is the spark produced if a small air gap is left somewhere in the circuit—for instance, if the wire is detached from one terminal and its end held near but not touching the terminal. A large gap would break the circuit, but a spark will jump readily across a small air space.

Another easily detectable effect of the passage of electrons is the heating produced. The amount of heating for a given flow of electrons depends greatly on the wire used: small wires become hotter than large ones; wires of certain metals, like tungsten and iron, become much hotter than silver or copper wires of similar size. The heating is explained by the resistance which the wires offer to the movement of electrons along them —much as the heat due to friction is explained by resistance to mechanical

movement. We take advantage of the heating effect of currents in a multitude of common electrical devices. Electric light bulbs are an extreme example, in which a slender tungsten filament is made hot enough to emit light copiously.

A less familiar but tremendously important effect of an electric current is its influence on magnets in its vicinity. To repeat a famous experiment first performed over 100 years ago by the Danish physicist Oersted, let us connect a horizontal wire to a battery and hold beneath the wire a small compass needle (Fig. 116). If the circuit is closed, the needle swings into a position at right angles to the wire. Place the compass just above the wire: the needle swings completely around, until it is again perpendicular to the wire but pointing in the opposite direction.* *Evidently a wire carrying a current produces a magnetic field.*

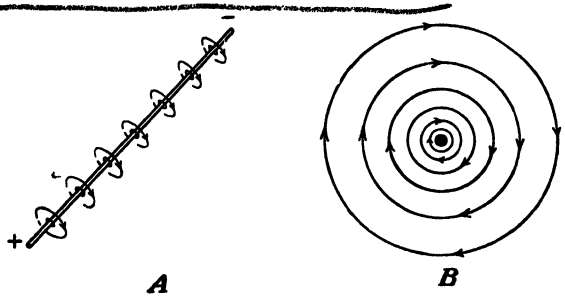


FIG. 117. *Magnetic lines of force around a wire carrying a current. In A, electrons are moving from right to left. B is a cross section of a wire in which electrons are moving out away from the page.*

Not only does the compass needle in this experiment prove the existence of a field, but its different positions suggest that the magnetic lines of force are concentric circles about the wire (Fig. 117). The direction of the lines of force (*i.e.*, the direction in which the *N* pole of the compass points) depends on the direction of flow of electrons through the wire; when one is reversed, the other reverses also. In any particular case the direction of the field may be found by encircling the wire with the fingers of the left hand, so that the extended thumb points along the wire in the direction in which the electrons move; then the finger tips point in the direction of the field. In practical electricity this rule and others connecting directions of fields and currents are naturally important, but for our purposes they are needless details. We need only remember that *the current and the field are at right angles to each other.*

Oersted's discovery was the first positive proof that a connection exists between electricity and magnetism; it was also the first demonstra-

* This experiment presupposes the absence of other magnetic fields. In the usual laboratory demonstration, the needle is affected by the magnetic field of the earth as well as that of the wire, hence swings to a position not quite at 90° with the wire.

tion of the principle on which the electric motor is based. A clear understanding of the experiment is indispensable for any intelligent appreciation of modern physics. To repeat once more: electric charges at rest do not affect a magnet, but *when in motion* produce a magnetic field. Even when moving, they do not *attract* or *repel* a magnet; they simply cause it to turn until perpendicular to the direction of their motion.

There is no simple "explanation" of this phenomenon. Oersted's experiment is a fundamental one, like the experiments which enable us to define positive and negative charges. We simply accept, as one of the basic properties of electric charges, the fact that when in motion they are surrounded by magnetic fields as well as by electric fields.

Galvanometers and Motors

Suppose that a horizontal wire connected to a battery is suspended as in Fig. 118, so that it is free to move from side to side; and suppose

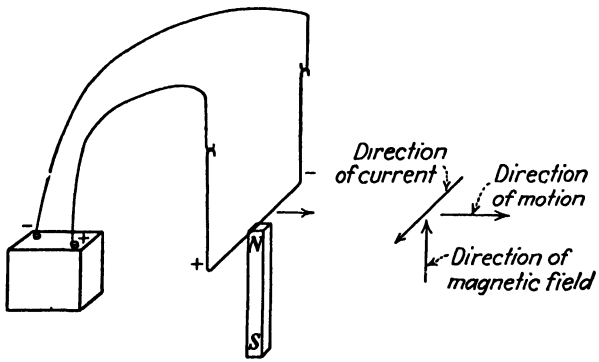


FIG. 118. *The motor effect. The wire tends to move sideward, in a direction perpendicular both to the magnetic field and to the direction of current.*

that the *N* pole of a strong bar magnet is placed directly beneath it. This setup is the reverse of Oersted's experiment: he placed a movable magnet near a wire fixed in position, while here we have a movable wire near a fixed magnet. We might predict, from Oersted's results and Newton's third law of motion, that in this case the wire will move. It fulfills the prediction, jumping out quickly to one side as soon as the circuit is closed. The direction of its motion is perpendicular to the lines of force of the bar magnet's field. Whether it jumps to one side or the other depends on the direction of flow of electrons in the wire and on which pole of the magnet is used.

The push which a magnetic field exerts on a wire carrying a current is often called the *motor effect*, since the running of an electric motor depends on this force. Keep in mind the nature of the force: it is not

attraction or repulsion, but a *sidewise push.* The wire does not move toward the magnet or away from it but *perpendicular to its field.*

The motor effect can be beautifully demonstrated with a cathode-ray tube. In this tube we have an electric current reduced to bare essentials: a flow of electrons, unencumbered by a wire, moving across a vacuum from electrode to electrode. By placing a fluorescent screen down the length of the tube, we can make the invisible electrons leave a

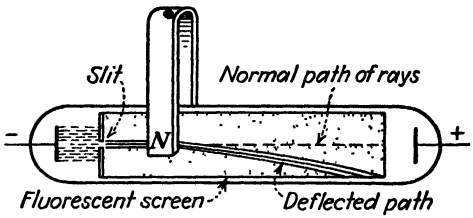


FIG. 119. *Deflection of cathode rays in a magnetic field. The direction of electron movement is from left to right; the direction of the field is into the page; the direction of deflection is at right angles to both, or downward.*

trace of their movements as a hazy purple line. If now we bring near the side of a tube so equipped one end of a bar magnet, the purple line bends either upward or downward. That is, the swiftly moving electrons are pushed in a direction at right angles to the magnetic field and so are forced out of their accustomed straight path (Fig. 119). Since the amount

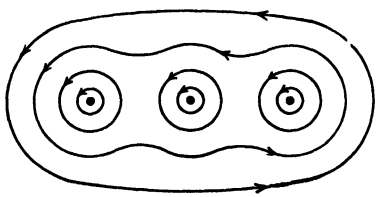


FIG. 120. *The magnetic fields of wires carrying parallel currents add together. (Each dot is the cross section of a wire perpendicular to the paper.)*

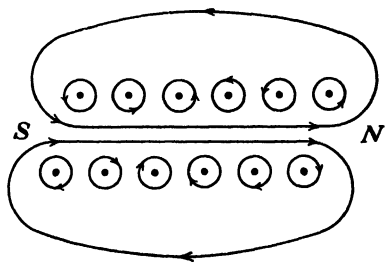


FIG. 121. *Diagrammatic cross section of a coil, showing how the fields of all loops add together to give an N pole at one end, an S pole at the other end.*

of deflection depends, among other things, on the speed, mass, and charge of the electrons, experiments of this sort give physicists valuable information about these tiny particles.

Let us return to the behavior of electrons caged in wires. As we have seen, they are deflected by a magnet just as their freely moving brethren in a cathode-ray tube are deflected, and this sidewise movement of the electrons causes the whole wire to move with them. Now suppose that we use several wires instead of one, all the wires being side by side and

all carrying a current in the same direction. Each wire builds itself a magnetic field which, except in the space very close to the wire, is in the same direction as the fields of its neighbors (Fig. 120). In effect, the fields of all the wires add together, and the sidewise movement of the group produced by the approach of a magnet is accordingly more violent than the movement of a single wire.

The simplest way to obtain the effect of a group of wires carrying parallel currents is to wrap a single wire in a coil. The current in each loop of the coil is parallel to the current in all others, and the magnetic fields of the different loops add together (Fig. 121). A little study of such a coil shows that inside the coil the lines of force are all directed toward one end of the coil, while outside they are directed to the other end. This is exactly analogous to the field of a bar magnet, as Fig. 122 shows, and we find, accordingly, that *a coil carrying a current behaves like a bar magnet.* One end is a north pole, the other a south pole; it attracts pieces of iron; and it swings to a north-south position if free to move. Which end is north and which south depends, of course, on the direction of the current and the direction of winding.

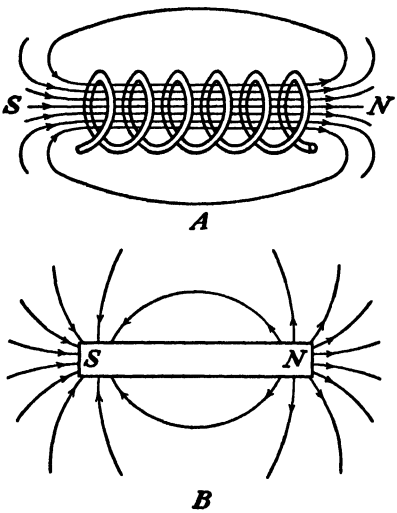


FIG. 122. *Diagram to show the similarity between the magnetic fields around (A) a coil carrying an electric current and (B) a bar magnet.*

The magnetic strength of the coil is prodigiously increased if a rod of soft iron is placed within it. This combination of coil and soft iron is called an *electromagnet.* An electromagnet exerts magnetic force only when a current flows through its coil, but by using sufficient current it can be made far more powerful than a permanent magnet of similar size. Electromagnets find manifold employment in industry, from the tiny coils which operate doorbells and telegraph relays to the huge magnets used for loading and unloading scrap iron (Fig. 123).

Suppose that a small coil is suspended between the poles of a horseshoe magnet (Fig. 124). It hangs limp until a current is passed through it; then it snaps into a position so that its *N* pole is as near as possible to the *S* pole of the horseshoe, its *S* pole near the *N* pole of the horseshoe. The thread suspending the coil resists somewhat the coil's turning, and the coil comes to rest at an intermediate position. How far the coil turns

toward a direct alignment with the magnet depends on two things: the resistance of the thread to twisting, and the strength of the coil's *N* and *S* poles. The magnetic strength of the coil depends in turn on the amount of current passing through it. Hence, for a given supporting thread, the angle through which the coil turns is a measure of the strength of the electric current applied to it. Devices built so that this angle can be measured accurately are our most convenient instruments for detecting

FIG. 123. *A large commercial electromagnet loading scrap iron.* (*Courtesy of General Electric Co.*)

and measuring electric currents. They are called *galvanometers* (after the Italian scientist Galvani). The ordinary direct-current "voltmeters" and "ammeters" used by electricians are galvanometers of this type adapted for special purposes. Extremely sensitive galvanometers are among the most useful tools in modern physical research.

A galvanometer has nearly all the essential parts of a simple electric motor. In a motor the coil must be supported on an axle rather than a delicate thread, and some device must be used for changing the direction of current through the coil every time it aligns itself with the magnet.

Changing the current reverses the poles of the coil, so that every time it swings into and a little past the position of alignment, it receives a new impulse to turn into the opposite position. Thus its motion becomes continuous. The device used for automatically changing the current direction is called a *commutator;* it may often be seen on the axle of a motor as a copper sleeve divided into two or more segments (Fig. 125).

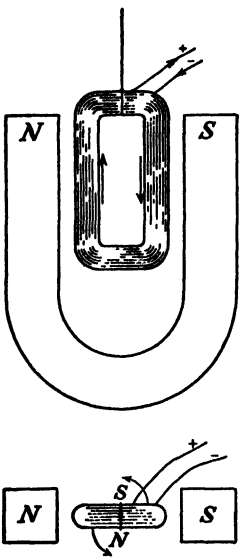


FIG. 124. *Diagram of a galvanometer. The upper picture is a side view and the lower picture a top view of a coil suspended between the poles of a magnet. Arrows in the upper picture indicate the direction of movement of electrons around the coil; arrows in the lower picture show the motion of the coil.*

Practical electric motors are not built on quite so simple a pattern. They are of many designs, adapted to different uses and different kinds of electric currents. Ordinarily electromagnets are employed rather than the simple permanent magnet of a galvanometer. In some motors the coil is fixed in position, and the magnet or magnets rotate about it. Some motors, built for alternating rather than direct current, have no need of commutators. But regardless of design, motors without exception utilize in their operation the force between a magnet and an electric current in its vicinity.

Electrical Units

Mastery of the formidable galaxy of units employed in electrical work is a task which means many sleepless nights to every beginner in physics and in practical electricity. For our present purpose such a detailed study of these units would be about as useful as an investigation of Sanskrit dialects. But a few of the units, used and misused widely in discussions of electricity, deserve at least a passing glance.

In a number of ways the flow of electricity along a wire is analogous to the flow of water in a pipe. The analogy is often helpful, because water can be visualized so much more easily than electricity. We shall find a comparison with flowing water especially helpful in this discussion of units.

Back in Chap. XVI, we defined a unit of electric charge as an amount of charge which would repel a similar charge at a distance of 1 cm with a force of 1 dyne. For practical work this "electrostatic unit" is too small; a larger unit commonly employed is the ***coulomb,*** an amount of charge equal to almost exactly 3,000 million electrostatic units. Like the smaller unit, the coulomb is simply a *quantity of charge,* the charge

carried by a certain enormous number of electrons—in this case, about 6×10^{18}. It is a quantity similar to the gallon in the measurement of water.

If electrons are flowing along a wire at such a rate that 1 coulomb of electric charge passes a given point each second, we say that the *current strength* in the wire is 1 ***ampere*** (1 amp) (after the French physicist, Ampère, who lived at the beginning of the nineteenth century). If 2 coulombs pass each second, the current strength is 2 amp, and so on. Thus the ampere is a measure of *rate of flow;* similarly, we could measure rate of flow of water in gallons per second.

Suppose that a gallon of water is poised at the brink of a waterfall. We say that it possesses here a certain potential energy, since it is

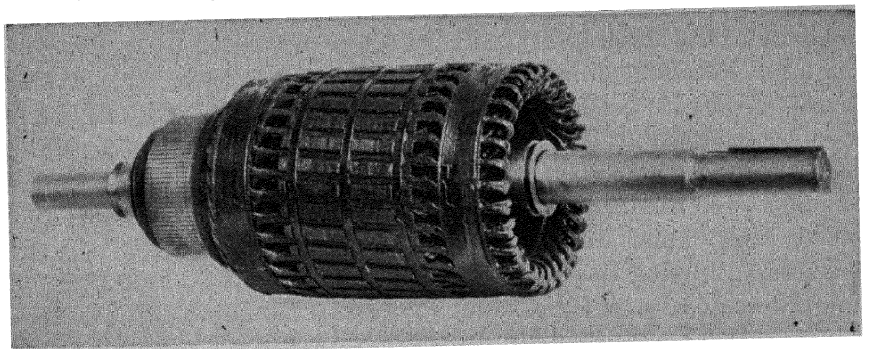


FIG. 125. *The moving coil, or armature, of a direct-current motor. The commutator is the segmented ring near the left end.* (*Courtesy of General Electric Co.*)

capable of moving under gravitational attraction. When it drops to the base of the fall, its potential energy decreases. The work obtainable from the gallon of water during its fall is measured by this decrease in potential energy—or in simpler terms, by the difference in elevation between top and bottom of the waterfall. Now consider a coulomb of electricity poised on the negative terminal of a battery. It is capable of moving, under the repulsion of adjacent negative charges and the attraction of charges on the other terminal. We say therefore that it possesses a certain potential energy, by reason of its position on the negative electrode. When it has moved along a wire to the positive terminal, its potential energy is smaller, since here it can no longer move spontaneously, except perhaps to a stronger positive charge in the neighborhood. The amount of work the coulomb can perform in flowing from negative to positive is measured by its decrease in potential energy. This decrease of potential energy brought about by the motion of 1 coulomb from negative to positive is a quantity called the *difference of potential,* or *potential difference,* between the two electrodes. It is a quantity analogous to difference of elevation in the case of water. We would measure differ-

ence of elevation in feet; we measure difference of potential in ***volts*** (named for the Italian physicist Alessandro Volta).

The maximum difference of potential between the terminals of an ordinary storage battery is about 6 volts, of a dry cell about 1.5 volts. Every coulomb of electricity at the negative terminal of the storage battery, therefore, is capable of doing four times as much work as a coulomb at the negative electrode of a dry cell—just as a gallon of water at the brink of a waterfall 600 ft high is capable of doing four times as much work as a gallon at the brink of a 150-ft fall. If a storage battery and a dry cell are connected in exactly similar circuits, the former will push four times as many electrons around its circuit in a given time as the dry cell, giving a current strength (amperes) four times as great. Very crudely, we may speak of the potential difference or voltage between two points as the amount of "push" effective in moving charges between the points. Potential difference is akin to force: in fact, the maximum potential difference between the terminals of a battery is often, but not altogether correctly, called the "electromotive force" of the battery.

Evidently the amount of work an electric current can do depends on two things: the number of coulombs available, and the potential difference, or push, which drives them. We obtain the amount of electrical work by multiplying these two factors together, that is, by multiplying volts and coulombs. A logical unit for expressing amounts of electrical work or energy would therefore be the "volt-coulomb"—a unit in common use, but usually under the name "watt-second" (see below).

The *power* of an electric current is defined as the amount of work, or the number of volt-coulombs, which it can perform per second. In mathematical form

$$\text{Power} = \frac{\text{work}}{\text{seconds}} = \frac{\text{volts} \times \text{coulombs}}{\text{seconds}}$$

Now recall the definition of an ampere: a current strength of one coulomb per second. Hence we may substitute amperes for coulombs/seconds in this formula:

$$\text{Power} = \text{volts} \times \frac{\text{coulombs}}{\text{seconds}} = \text{volts} \times \text{amperes}$$

A convenient unit of power, then, is a volt-ampere—so convenient that it is given a special name, the ***watt***. In mechanical terms a watt is equal to 1 joule (or 10 million ergs) per second.

On ordinary electric light bulbs you will find usually some such notation as: 60W, 110V. We interpret this cryptogram to mean that when the lamp is connected to the ordinary 110-volt house circuit, it will use

electrical energy at the rate of 60 watts. From the above formula it is evident that the strength of the current through the lamp is a little over ½ amp. A lamp marked 100W or 200W uses electrical energy at a proportionately greater rate. Many common electrical appliances require the expenditure of electrical energy at still higher rates: toasters, irons, motors in vacuum cleaners and electric refrigerators draw 300–600 watts.

When a 100-watt lamp burns for 1 sec, it consumes evidently 100 volt-coulombs of electrical energy. Ordinarily, however, we refer to this amount of energy as 100 *watt-seconds* rather than 100 volt-coulombs. If the lamp burns for 1 hr, it consumes 100 watt-hours of energy; if for 10 hr, 1,000 watt-hr, or 1 *kilowatt-hour* (1 kilowatt = 1,000 watts). The electric bill you pay at the end of each month is usually made out in kilowatt-hours. You pay directly for the amount of electrical energy which your various lights and motors and gadgets have consumed during the month.

Three concepts as abstruse as current strength, potential difference, and electric power can scarcely be made clear by a few pages of reading. For a time inevitably the ampere, the volt, and the watt must seem a bit obscure. But further acquaintance with these difficult terms, as they creep into future discussions, should ultimately rob them of their mystery.

The Production of Electric Currents

Energy in any of its usual forms may be converted directly into electric energy. The battery is a device for changing chemical energy to electric energy. Instruments called *thermocouples*, used in the measurement of very high and very low temperatures, convert heat directly into electric energy. Even radiant energy may produce small electric currents, as we shall see presently.

But the electric energy which is supplied so copiously to our homes and factories comes neither from light nor heat nor chemical reactions, but from mechanical energy. The great dynamos in power plants which supply electricity to cities utilize water power or steam engines. Isolated farms may have small generators operated by gasoline or Diesel motors, or more rarely by wind power. In all cases the energy which is turned directly into electricity is mechanical energy of moving machinery.

One method of converting mechanical energy to electric energy we discussed in Chap. XVII: the electrostatic machine. Turning the crank of such a machine moves insulated conductors past stationary electric charges; induced charges on the conductors are collected on spherical electrodes, and eventually an electric current in the form of a spark jumps between the electrodes. We might, of course, connect a wire between the spheres and obtain a steady current rather than a series of sparks. But the strength of the current would be disappointing. An

electrostatic machine is capable of producing enormous differences of potential (high voltages), but cannot efficiently produce a steady current of more than a tiny fraction of an ampere.

Commercial generators utilize a principle basically the same as that used in galvanometers and electric motors: the force between magnets and moving electric charges. The story goes back to some famous observations by the English physicist, Michael Faraday. Intrigued by the researches of Ampère and Oersted on the magnetic fields around electric currents, Faraday reasoned that if a current can produce a magnetic field, then somehow a magnet should be capable of generating an electric current. Now a wire placed in a magnetic field and connected to a galvanometer shows no sign of a current. But *if the wire or the magnet is moving*, Faraday discovered, *a current is produced.* As long as the wire continues to cut across lines of force of the magnetic field, the current persists; when the motion stops, the current stops. Because it is produced by motion in the presence of a magnetic field, without any direct contact with electric charges, this sort of current is called an *induced current.*

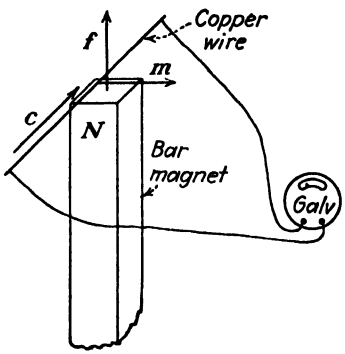


FIG. 126. *The dynamo effect. The arrow f indicates the direction of magnetic lines of force, m the direction of motion of the wire, and c the direction of the current produced by this motion.*

Let us repeat Faraday's experiment in very simple form. Suppose that the copper wire of Fig. 126 is moved back and forth across the lines of force of the bar magnet. The galvanometer will indicate a current flowing first in one direction, then the other. Note that the wire is held approximately at right angles to the lines of force; thus the motion of electrons along the wire is at right angles to the field. Which direction along the wire the induced current will flow depends on the direction of its motion and the direction of the lines of force: reverse the direction of motion, or use the opposite magnetic pole, and the current is reversed. The strength of the current depends on the rapidity of movement and on the strength of the field.

This phenomenon, often called the *dynamo effect*, is different only in appearance from the motor effect. A motor runs because electrons flowing along a wire are pushed sidewise in a magnetic field. In Faraday's experiment we again cause electrons to move through a magnetic field, but this time by moving the wire as a whole. The electrons are pushed sidewise as before and, in response to the push, move along the wire as an electric current.

To intensify the induced current produced by moving a conductor near a magnet, commercial dynamos (Fig. 127) employ a large coil rather than a single wire, and several electromagnets instead of a bar magnet. Turned rapidly between the electromagnets by steam engine or water turbine, wires of the coil cut lines of force first in one direction, then in the other. Operation of the dynamo is illustrated in simplified form in Fig. 128, where a coil is shown turning between two magnets only. Evidently during half a revolution each side of the coil cuts the field in one direction, then during the other half revolution cuts the field

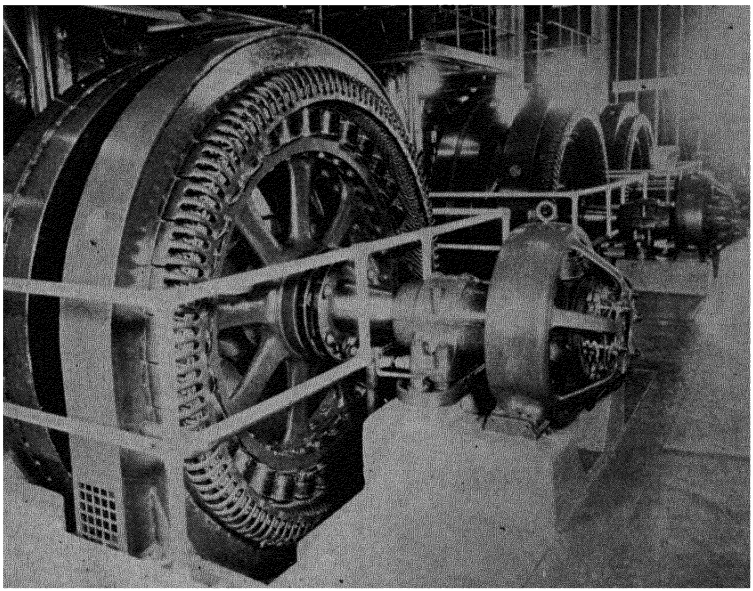


FIG. 127. *Three large generators installed in a power plant at Pierre, S. D. (Courtesy of General Electric Co.)*

in the opposite direction. Hence the induced current flows alternately one way and the other. We call such a current an *alternating current.*

Currents produced by batteries, thermocouples, and photoelectric cells are one-way, or *direct,* currents, flowing steadily in one direction unless we arbitrarily change the connections. In alternating currents electrons move first one way, then the other—not from one end of the circuit to the other, but simply back and forth over short distances. In the 60-cycle current which we ordinarily use in our homes, electrons change their direction of motion 120 times each second.

Dynamos can be constructed to produce direct current by the use of commutators similar to the commutators used on direct-current motors; but alternating current is usually produced because it is more

economical for transmission over long distances. In most household appliances, such as lights and heaters, the direction of current is immaterial, and alternating current works quite as efficiently as direct.

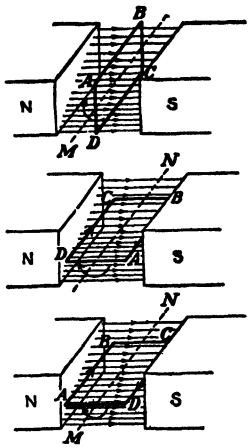


FIG. 128. *Diagram to show how a dynamo produces alternating current. In the top drawing no current is produced, since both sides of the coil ABCD are moving parallel to the magnetic field. As side AB moves down across the field (middle drawing), the current flows in one direction; half a revolution later, when AB moves up across the field, the current flows in the opposite direction. (From Elements of Physics by Smith.)*

Transformers

To generate an induced current requires that magnetic lines of force be made to move across a conductor. Two ways have been mentioned for accomplishing this motion: the wire may be moved past a magnet, or the wire may be held stationary while the magnet moves. We come now to a third, less obvious method, which involves no visible motion at all.

Suppose that coil *A* in Fig. 129 is connected to a switch and a battery, coil *B* to a galvanometer. When the switch is closed, a current flows through *A*, building up a magnetic field around it. The current and field do not start instantaneously; during a tiny fraction of a second the current increases from zero to its normal value, and its magnetic influence expands from a weak field close to the wire to a strong field perceptible at some distance. We may imagine the circular lines of force expanding, moving outward across the wires of coil *B*. This motion of the lines across coil *B* produces in it a momentary current, recorded by a sharp kick of the galvanometer needle. Once the current in *A* reaches its normal, steady value, the field becomes stationary and the induced current in *B* stops. Now suppose the switch is opened: in another small fraction of a second the current in *A* drops back to zero, and its magnetic field, so to speak, collapses back around the coil. Again lines of force cut across *B*, and the galvanometer responds with another kick, this time in the opposite direction since the motion of the lines of force has changed. Thus *starting and stopping the current in A has the same effect as moving a magnet in and out of B*. An induced current is generated whenever the switch is opened or closed.

Suppose that *A* is connected not to a battery but to a 60-cycle alternating current. Now we need no switch; automatically, 120 times each second, the current comes to a complete stop and starts off again

in the other direction. Its magnetic field expands and contracts at the same rate, and the lines of force cutting *B* first in one direction, then the other, induce an alternating current similar to that in *A*. An ordinary galvanometer will not respond to these rapid alternations, but an instrument built to measure alternating currents will show the induced current readily.

Thus an alternating current in one coil will produce an alternating current in a second coil, even though a considerable distance separates them—a fact often used to mystify the uninitiated, for instance by

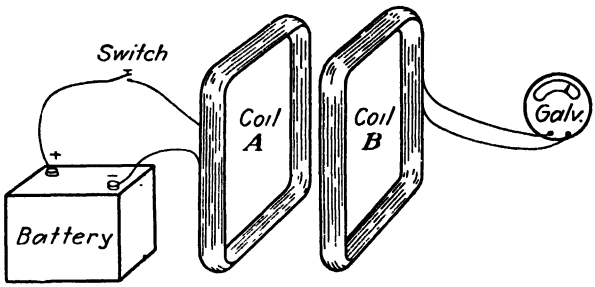


FIG. 129. *The principle of the transformer. Momentary currents are registered by the galvanometer when the current in coil A is started or stopped.*

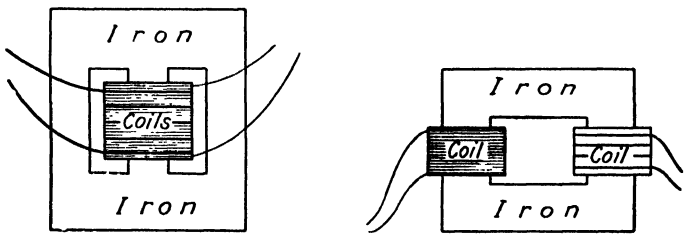


FIG. 130. *Diagrams of two kinds of transformer. Many other designs are used, but all have two coils wound on a core of soft iron.*

making a lamp burn without visible electric connections. But to generate an induced current most efficiently, the two coils should be close together and wound on a core of soft iron (Fig. 130). Such a combination of two coils and an iron core constitutes a *transformer*. The coil into which electricity is fed from an outside source is the *primary* coil, the one in which an induced current is generated is the *secondary* coil.

Transformers are useful because the voltage of the induced current can be made any desired multiple or fraction of the primary voltage by suitable winding of the coils. If the number of turns of wire in the secondary coil is the same as the number of turns in the primary, the induced voltage will be the same as the primary voltage. If the secondary has twice as many turns, its voltage is twice that of the primary; if it has one-third as many turns, its voltage is one-third that of the primary, etc.

Thus by using a suitably designed transformer, we may secure any desired voltage, high or low, from a given alternating current.

Likewise the current strength in the secondary depends simply on the number of windings in the two coils. It is greater in the secondary than in the primary if the number of turns is less, and vice versa; that is, a transformer which reduces voltage produces a current of greater amperage, one that increases voltage produces a current of smaller

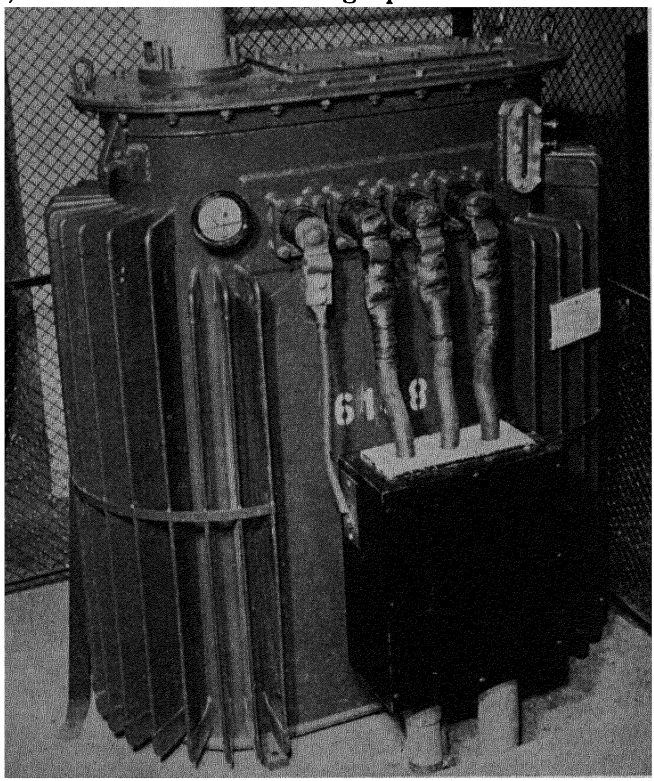


FIG. 131. *A commercial transformer. (Courtesy of General Electric Co.)*

amperage. As an illustration, suppose that a transformer has 100 turns on its primary coil, 1,000 on its secondary. Then, if a current of 110 volts and 3 amp flows through the primary, the induced current in the secondary will have a voltage of 1,100, an amperage of 0.3 (neglecting heat losses).

The relation between volts and amperes in the two coils of a transformer is a necessary consequence of the law of conservation of energy. If we neglect heat losses, the electric power generated in the secondary coil should equal that supplied to the primary. Now electric power is

the product volts × amperes; hence in a perfectly efficient transformer, volts × amperes for the primary should equal volts × amperes for the secondary. So an increased voltage in the secondary must mean a decreased current strength, and vice versa.

For a multitude of purposes, in homes and factories and laboratories, we find it necessary to change the voltage of alternating currents (Fig. 131). But perhaps the most valuable service which transformers render is in making possible long-distance transmission of power. Current strengths in long-distance transmission must be as small as possible, since large currents mean energy losses in heating the transmission wires. Hence at the powerhouse electricity from the dynamo is led into a "step-up" transformer, which "steps up" its voltage and decreases its amperage several hundred times. On high voltage or "high tension" lines this current is carried to local substations, where other transformers "step down" its voltage to make it safe for transmission along city streets.

The necessity for changing the voltage of the current at least twice between powerhouse and consumer explains our customary use of alternating currents rather than direct: no one has yet devised an instrument for changing the voltage of direct currents which can compare with the transformer in simplicity and efficiency.

Questions

1. What sort of electric current—alternating or direct—would you find in
   a. The filament of an electric light bulb in your home?
   b. The filament in a light bulb in the headlight of an automobile?
   c. The wires leading from the power source to a direct-current motor?
   d. The wires in the moving coil of a direct-current motor?
   e. The high tension lines carrying electricity from Boulder Dam to Los Angeles?
   f. The secondary coil of a small transformer?
2. In what fundamental respect are the motor effect and the dynamo effect similar?
3. A wire connected to a battery and a switch is suspended with its length vertical so that it is free to move in any direction. The *N* pole of a strong bar magnet is held near the wire at one side. Describe what, if anything, will happen when the switch is closed.
4. Describe the magnetic field around a single loop of wire carrying a direct current.
5. A long coil is suspended by a thread at its midpoint between the poles of a strong magnet. What, if anything, happens when a direct current is sent through the coil? What if alternating current is used?
6. How does the construction of an electromagnet differ from that of a transformer?
7. Given a bar magnet, a coil of wire, and a source of direct current, describe three ways by which you could make a momentary induced current in a second coil.
8. Given a coil of wire and a small light bulb, how could you tell whether the current in a second coil was direct or alternating without touching this second coil or its connecting wires?
9. How much power is consumed by a 60-watt lamp? How much energy is used by a 60-watt lamp burning for 2 hr? If the lamp is connected to a 110-volt line, what current strength in amperes does it use?

10. Why is alternating current rather than direct current used in the long-distance transmission of power?
11. In what respect is a galvanometer similar to a simple direct-current motor?
12. For what purposes are the following devices used?
    a. Transformer.
    b. Commutator.
    c. Dynamo.
    d. Galvanometer.
    e. Electromagnet.
13. The fuse box in your home is a safety valve, designed to prevent too much electricity from flowing through the house circuit. A 30-amp fuse, for instance, will burn out and automatically shut off the current whenever the current strength through it exceeds 30 amp. If your home is served by a 110-volt line, how much power in watts can you draw from the line before a 30-amp fuse will burn out? How many 100-watt light bulbs could you put in the circuit before the fuse would burn out?
14. How many coulombs of electricity move through a coil in 10 min if it is connected in a circuit with an ammeter that reads 0.5 amp? If a voltmeter in the same circuit reads 6 volts, how much electric energy moves through the coil in 10 min? Express this energy in watt-seconds and in kilowatt-hours.
15. Suppose that the dynamo of a power plant generates current at 4,000 volts and 250 amp. This current is led into the primary coil of a transformer whose secondary coil has fifty times as many turns as the primary. What is the voltage and current strength of the current in the secondary coil? Suppose that the secondary coil is connected with a high tension transmission line that carries the current to a distant city. How must the current be changed before it can be used by people in the city?

CHAPTER XIX

Light Waves

MICHAEL FARADAY (Fig. 132), son of a blacksmith, apprenticed to a bookbinder in his youth, schooled himself in chemistry and physics from the books he was learning to bind. At the age of twenty-one he obtained the humble station of bottle washer for Sir Humphry Davy, at that great man's laboratory in the Royal Institution of London. Within twenty years the blacksmith's son succeeded Davy as head of the Institution. During those twenty years Faraday's experiments, particularly in chemistry, had won him wide acclaim in scientific circles. To the later years of his life belong those more famous investigations into electricity and magnetism which we discussed briefly in the last chapter. Never an adept at mathematics, Faraday is remembered as one of physical science's greatest experimental geniuses. Like most of us, he felt the necessity of working with real, tangible things, such as the coils and magnets of his laboratory. To make real the electric and magnetic forces which he could not see or feel, Faraday invented *lines of force*—lines which do not exist, which give at best a crude picture of fields of force, but which students ever since have found useful in their early attempts to visualize the abstractions of electricity.

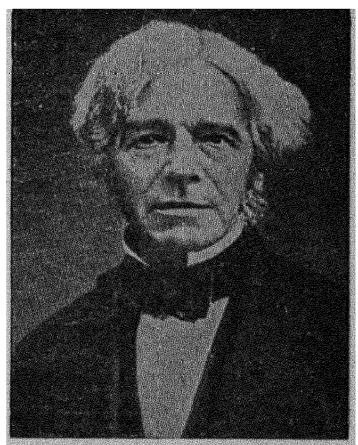


FIG. 132.—*Michael Faraday (1791–1867). (Courtesy of Gramstorff Brothers, Inc.)*

James Clerk Maxwell (Fig. 133) was born into an old and distinguished Scottish family in 1831, when Faraday was forty years old. Given the best education that England could provide, Maxwell became a precocious scientist. At fifteen he published his first paper; at twenty-

five he was made professor of physics and astronomy at Cambridge. Maxwell was gifted with extraordinary mathematical ability: for him fields of force could be better expressed by equations than in terms of Faraday's lines. With his complicated equations, based largely on Faraday's experiments, Maxwell at length not only expressed the interconnections between electricity and magnetism, but formulated a theory for the explanation of light.

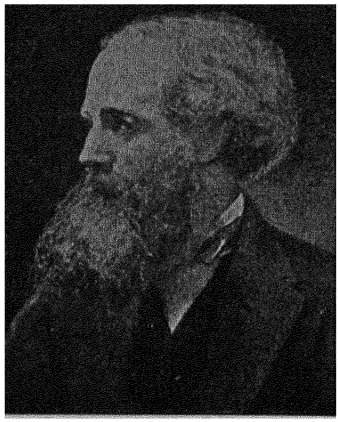


FIG. 133. *James Clerk Maxwell (1831–1879). (Courtesy of Brown Brothers.)*

Many a moral could be drawn from the lives of these two men, so different in talent and in training, who stand out above the host of nineteenth-century scientists who were seeking an explanation for the mysterious phenomena of electricity. Their work is an especially good illustration of the progress of science from experiment to broad generalization. Sometimes experimental ability and theoretical insight are combined in the same person, but more often we find these faculties in different men. As Faraday's work paved the way for Maxwell, so the experiments of Lavoisier lay behind Dalton's atomic theory, the observations of Tycho Brahe behind Kepler's laws of planetary behavior, Galileo's experiments with falling bodies behind Newton's laws of motion.

Electromagnetic Waves

One of the startling results of Maxwell's calculations was the prediction that electric charges moving with changing velocity should generate waves capable of traveling long distances through empty space.

Suppose that A, B, and C in Fig. 134 represent electrons confined in vertical wires so that their only possible motion is up and down, perpendicular to the line BC. If A moves to A_1, its repulsion for B and C causes them to move downward along their wires; if A moves to A_2, B and C move upward. In other words, the lines of force radiating outward from A all change their positions as A moves, and B and C move in response to the changes.

Now *how long* does it take a motion of A to produce a response in B and C? Will C move at the same time as B, or later because it is farther from A? Do all parts of the electric field around A change position instantaneously as A moves, or does the part of the field near A change first, the disturbance then spreading outward to greater and greater distances? Maxwell was the first to answer these questions satisfactorily. He showed that the change in all parts of the field was *not* instantaneous

as A moved, that is, that B would start to move before C. The disturbance caused by A's motion travels outward from A with the speed of light, 186,000 mi/sec (3×10^{10} cm/sec). Naively, we may think of the electric lines of force about A as long straight ropes extending outward in all directions: moving A quickly produces a kink in each rope, which travels outward along the rope much as a kink will travel along a real rope if it is jerked at one end.

During its motion A is surrounded not only by its electric field, but by the magnetic field which, as we learned in the last chapter, accompanies every moving charge. As A starts to move, this magnetic field is built up, its circular lines of force spreading outward as the motion grows faster; when A comes to a stop, the lines of force collapse back around it. As the electrons B and C are cut by the moving magnetic lines of force, they are impelled to move at right angles to the motion, that is, along their wires. Thus not only the moving electric field but the changing magnetic field of A causes a response in B and C. Maxwell's equations predict that the magnetic disturbance will travel outward from A with exactly the same speed as the electric disturbance.

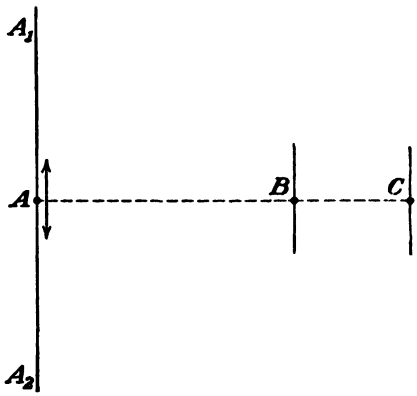


FIG. 134. *Disturbances in the electric and magnetic fields around A produced by its motion affect electrons at B and C successively.*

Now suppose that A moves continuously back and forth between A_1 and A_2. Every time it moves, a new kink is sent out along its electric lines of force; every time it speeds up and slows down, a reversal in the direction of its magnetic field travels outward with the electric kinks. In other words, a vibrating or oscillating charge of this sort sends out continuous electric and magnetic disturbances capable of affecting charges some distance away. These disturbances constitute Maxwell's ***electromagnetic waves.***

Every alternating current consists of electric charges oscillating in the manner just described, and every alternating current sets up electromagnetic waves. In a transformer, oscillating electrons in the primary coil set up waves that travel to the secondary and in it cause other electrons to oscillate. In the last chapter we said simply, "One alternating current induces another alternating current in a near-by conductor." Here we focus attention on the space between the conductors and say, "Waves set up by one alternating current travel to a near-by conductor

Electromagnetic waves in a transformer travel only a short distance from primary to secondary. If the secondary coil is moved a few inches or a few feet away from the primary, the induced current becomes very feeble; in other words, the electromagnetic waves die out. This is because the rate of oscillation is small—only 60 cycles per second in the ordinary house current. If the electrons are made to oscillate faster, the waves travel farther before dying out. When they oscillate very rapidly, say several hundred thousand times a second, Maxwell's equations predict that they will generate waves strong enough to influence other charges many miles away.

From our vantage point, secure in the knowledge which the last half-century has brought us, it seems strange indeed that the brilliant Scot did not proceed at once to his laboratory and devise means for producing these long-distance waves. Electromagnetic waves produced by rapid oscillations remained a mathematical curiosity until eight years after Maxwell's death, when a German, Heinrich Hertz, demonstrated their existence. Even Hertz seemed unaware of the possibilities of his discovery; it remained for an Italian inventor, Guglielmo Marconi, to point the way from Maxwell's mathematics to the modern radio.

Radio transmission is accomplished by means of electromagnetic waves produced by electrons oscillating several hundred thousand times a second in the aerial of the sending station. These waves strike the aerial of a receiving station and make electrons in the receiving set vibrate in unison with those of the sending set. The sending station produces irregularities in the transmitted waves timed in accordance with the vibrations of the human voice or the instruments of an orchestra. In the receiving set these irregularities are made to operate a diaphragm which simulates the sound produced by the sender.

To the practical world of high-pressure advertisers and sinking ships, explorers and music lovers, bored parties and lost airplanes, Maxwell gave the radio. To physical science he bequeathed a gift equally valuable: an explanation for the phenomena of light. Not only did Maxwell show that electromagnetic waves should travel with the speed of light, but he concluded that *light itself is a form of electromagnetic wave motion.* It was known before Maxwell's time that light shows the characteristics of wave motion, but his work gave the first intimation that light, electricity, and magnetism are intimately related. How close this relation is, and how intimately all three are connected with the innermost structure of matter, physicists have learned only in the last forty years.

Waves

We begin our study of the properties of light with a few words about waves in general.

Standing on a beach, watching the waves roll in and break one after the other, we are impressed with the ceaseless motion, an apparent motion of *something* toward the shore. As children, perhaps we naively concluded that quantities of water were rolling bodily shoreward, carrying along pebbles and shells and bits of driftwood. To most of us it comes as a surprise when parents or teachers tell us that in ordinary waves the water itself performs only an up-and-down, back-and-forth motion, as much of it moving away from shore as toward shore. So there dawns the strange idea of seemingly disembodied waves moving through water, while the water itself moves in an altogether different fashion.

What, then, does actually move shoreward? The answer is *energy*, energy which can rock great ships, batter down breakwaters, in time even undermine granite cliffs. The essence of all familiar wave motions is a transfer of energy, accomplished by a to-and-fro or up-and-down motion of material particles, but involving no bodily movement of matter in the direction of the waves.

FIG. 135. *Diagram of water waves in deep water. Each particle performs a periodic motion in a small circle. Particle 1 now is at the crest of a wave; the crest moves to the right as particles 2, 3, 4 . . . in succession reach the tops of their respective circles*

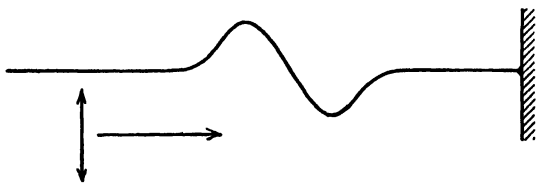


FIG. 136. *Transverse waves. The diagram represents a single wave traveling along a rope. The horizontal arrow shows the direction in which the wave is traveling, and the vertical arrow shows the motion of individual particles in the rope.*

Surface waves on water are the most familiar, but among the most complex wave motions which a physicist must unravel. Each water particle travels in a roughly circular path, rising to the top of its circle when a wave crest passes over it, falling to the bottom as the succeeding trough passes, rising again under the next crest (Fig. 135). We describe the motion as *periodic*, that is, recurring again and again over the same path. As succeeding particles perform this periodic motion, each a trifle later than the last, they produce the regularly spaced waves we find on an open lake or sea.

For our purposes a simpler wave motion is better suited. Suppose that you fasten securely one end of a long rope, then give the other end a quick jerk. A bend, or kink, in the rope travels from your hand to the

other end (Fig. 136). The rope as a whole has not moved: when the disturbance subsides, the rope is in the same position as at the start. But the energy which you supplied by jerking has traveled, as a wave, the full length of the rope. Now repeat the experiment, and fix your attention on one tiny segment of the rope. As the wave passes, you see this segment move up, then down, then up again, and thereafter remain quiet. A succession of such up-down motions makes up the wave. In other words, the motion of particles here is mostly *perpendicular to the motion of the wave itself*. Such waves we call *transverse*.

The wave motion called *sound* is quite different. Sound waves ordinarily are made up of vibrating air molecules, though, of course, sound can travel through other materials as well. The air vibrations are produced by vibrations of some solid object—a tuning fork, the taut strings of a violin or piano, the diaphragm of a loud-speaker, the vocal cords in our throats. Consider the motion of one prong of a tuning fork (Fig. 137). Air molecules on one side are first pushed violently away, then rush in to fill the vacant space when the prong moves back, then again are pushed away. Molecules farther away, pushed by those first influenced, follow this back-and-forth vibration a fraction of a second later. A similar motion would be produced in a line of people, each with his hands on his neighbor's shoulders, if someone should push and pull alternately at one end of the line: the back-and-forth motion would be transmitted from one person to the next down the line. Here the motion of each particle is a vibration *back and forth along the same line on which the wave travels*. We call such a wave *longitudinal*.

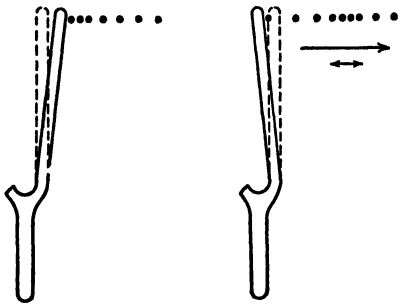


FIG. 137. *Longitudinal waves produced by vibrations of one prong of a tuning fork. Air molecules (represented by dots) are alternately pushed together and spread apart. The longer arrow shows the direction in which the waves travel, and the shorter arrow shows the motion of each particle.*

Longitudinal waves may be produced in any material, solid or fluid, since their transmission requires simply that each molecule impart a forward motion to its neighbors. But transverse waves are possible only in a rigid medium, where sideward motion of one particle drags with it adjacent particles to which it is tightly bound. In a fluid, where the molecules are unattached, sideward motion of one would simply be ignored by its neighbors.*

* Waves at the boundary between two fluids are an exception. Waves on the surface of water, for instance, involve transverse motion in part.

Three quantities essential in the description of wave motion of all kinds are the speed of the waves, the wave length, and the frequency. The idea of speed is obvious: it is simply the rate at which each wave crest appears to move. The wave length is the distance between successive crests (Fig. 138). Frequency is the number of wave crests (or troughs) which pass a given point each second. Now multiplying the number of waves which pass in a second by the length of each wave should give the speed with which the waves travel: if 10 waves, each 2 ft long, pass every second, then each wave must travel 20 ft during the second. In other words, frequency n times wave length λ gives speed v.

$$v = n\lambda \tag{28}$$

Big water waves on the open sea sometimes have wave lengths of 1,000 ft; they travel at roughly 70 ft/sec; so their frequency would be 0.07 per second. Sound in air travels at 1,100 ft/sec, or about 15 mi/sec; the tone of middle C has a frequency of 261 vibrations per second, and hence a wave length of 4.2 ft.

Electromagnetic waves, the periodic disturbances in electric and magnetic fields set up by vibrating electric charges, are transverse waves which travel with prodigious speeds—in a vacuum, 186,000 mi/sec or 3×10^8 m/sec. This, as Maxwell showed, is the same as the speed of ordinary light in a vacuum. It is probably the upper limit of possible speeds, both for the motion of particles and for the transmission of energy by waves.

Wave lengths of electromagnetic waves vary through an enormous range, from hundreds of miles down to tiny fractions of a centimeter. Waves in different wave-length regions have quite different properties, and we use a variety of instruments to produce and to detect them. The longer waves, from about 15 mi down to about 30 ft, are used in radio transmission. Shorter waves, down to lengths of about 1 cm, are used in radar and other "high-frequency" devices. Heat radiation or *infrared* radiation, the invisible wave motion which enables heat to travel across a vacuum, has wave lengths from a few hundredths of a centimeter down to 0.00007 cm. Waves with lengths between 0.00007 and 0.000035 cm are detected by the human eye as visible light. Still shorter are the invisible waves of ultraviolet light, and far shorter yet are X rays. Smaller even than these are the waves called gamma rays produced by radioactive materials, whose lengths have been measured down to 5×10^{-11} cm.

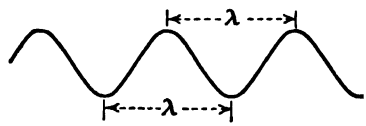


FIG. 138. *Wave length. The symbol* λ *is the Greek letter lambda.*

Frequencies of electromagnetic waves are appalling, as substitution of their speed and various wave lengths in Eq. (28) shows immediately.

Even the longer radio waves represent vibrations of several thousand times a second. Visible light has frequencies from 4×10^{14} to 8×10^{14} per second, and the frequencies of X rays and gamma rays are still higher. Just how anything can vibrate a million million times a second is quite as mysterious to a physicist as to the man in the street.

Electromagnetic waves differ from other waves not only in their unimaginable speeds and frequencies but also in the fact that their transmission does not require the vibration of material particles. Water waves mean the motion of water particles; transverse waves in a rope have no meaning without the rope; sound waves will not travel across a vacuum. But light moves readily through the best vacuum we can obtain in our laboratories, and the still better vacuum of empty space does not keep starlight from traveling billions of miles to reach our eyes. Waves of light are not waves of any material substance, but kinks in lines of force, periodic disturbances in electric and magnetic fields.

To avoid the difficulty of waves without a "waver," Maxwell and other nineteenth-century physicists imagined that electromagnetic waves were carried by vibrations in the "ether," a weightless substance supposed to pervade all space. How widely accepted this idea was a generation ago is suggested by our continued use of such expressions as "over the ether" in referring to radio broadcasts. But no positive evidence has ever come to light that the ether exists, and modern physicists prefer to describe electromagnetic waves as disturbances in fields of force rather than to accept a hypothetical substance with strange properties.

We face here, of course, the same sort of difficulty we met in our first discussion of electric and magnetic forces, back in Chap. XVI. Waves in empty space are quite as severe a strain on the imagination as is force acting across empty space. What the waves actually are, we must leave to philosophers; here we can only adopt the physicist's pragmatic view-point, conceding that the waves exist because we can perceive their effects.

Light

Four important conclusions about the nature of light have been mentioned in preceding paragraphs: (1) Light shows the characteristics of wave motion; (2) light waves are transverse; (3) light waves are electromagnetic waves of short wave length; (4) light waves travel at enormous speeds. Each of these statements can be backed up by straightforward and relatively simple experiments. But rather than undertake these experiments, let us for the moment accept the wave theory and see how we can use waves to interpret the more familiar properties of light.

In many respects the behavior of light waves is like that of ordinary water waves. If a pebble is dropped in a quiet pool, ripples spread in ever-widening circles; in similar fashion light waves spread in ever-

growing spheres around a candle or a light globe. Far from the point of disturbance ripples on water appear as a succession of nearly straight, parallel crests and troughs; so light waves far from their source, like light waves from the sun, appear to be parallel and practically straight. Just as water waves can be shown diagrammatically by a series of lines representing successive wave crests, so light waves can be shown by a similar series of lines. (In the case of light we speak of the lines as representing successive "wave fronts," rather than wave crests, since light waves properly speaking do not have crests and troughs.) The diagrams in Fig. 139 might represent ripples made by a stone in quiet water, or they might equally well suggest light waves near to and far from their source. A line like AB in Fig. 139, drawn through a series of wave fronts to show their direction of motion, is called a *ray* of light.

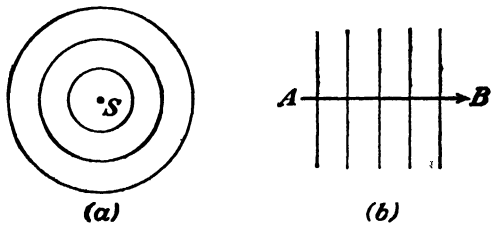


FIG. 139. *Successive wave fronts (a) near a source of light S, and (b) far enough away from the source so that the wave fronts are nearly straight.*

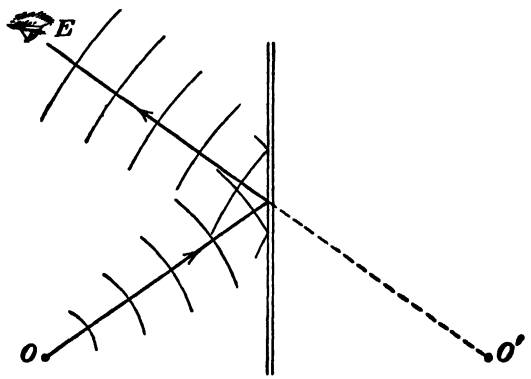


FIG. 140. *Reflection of light by a plane mirror. Light from an object at O is reflected to your eye at E, but appears to come from a point O' behind the mirror.*

Early in life we learn that "light travels in straight lines." We find support for this belief in the sharp shadows cast by objects illuminated by small light sources, and in the straight beams produced when light from small holes penetrates the recesses of a dusty basement. In terms of waves, movement in a straight line merely requires that successive wave fronts must be parallel. Wave fronts do remain parallel as long as light travels in a homogeneous medium like the air, just as water waves travel

in straight lines on the open sea. But light waves like water waves turn sharp corners when they meet obstructions, and by suitable arrangements can be made to bend gradually along a curve. The light by which we see

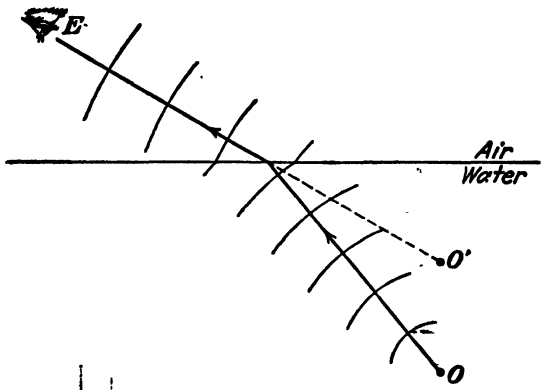


FIG. 141. *Refraction of light. Light from O is bent at the water surface, so that to an observer at E light appears to come from O'.*

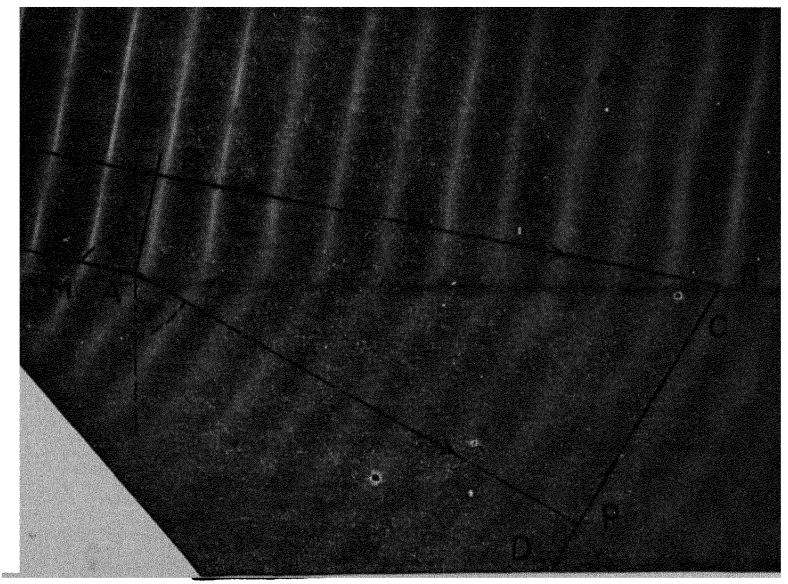


FIG. 142. *Refraction of water waves. Wave crests moving from left to right are bent at the line MN because below the line the water is shallower and the waves move more slowly. (From Webster, Farwell, and Drew's General Physics for Colleges, by permission of D. Appleton-Century Company, Inc., publishers.)*

most of the objects around us is light which has been bent sharply, or *reflected*, by striking the surfaces of these objects.

So accustomed is the human eye to find light following straight

paths that it often deceives us regarding the position of objects. Light reflected from very smooth surfaces, for instance, appears to come from behind the surfaces. Thus when you stand before a mirror, you see objects apparently behind the mirror: light rays from these objects travel to the mirror and are abruptly reflected back to you; but your eyes see the objects as if the light had followed a straight path from behind the mirror (Fig. 140).

Again, light from an object under water is bent, or *refracted*, when it enters the air. Your eye looks along the straight part of this path, so that the object appears to be above its true position (Fig. 141). In general, refraction or bending occurs when light moves obliquely from one transparent medium to another, as from air to glass, glass to vacuum, glass to water. Its explanation, not quite so evident as that for reflection, depends on the fact that light has different speeds in different materials. In Fig. 141 suppose that light is traveling obliquely upward from the object *O*. As each wave front meets the water surface, the part of it getting into the air first moves faster than the part still in the liquid and so is pulled around to a position making a steeper angle with the

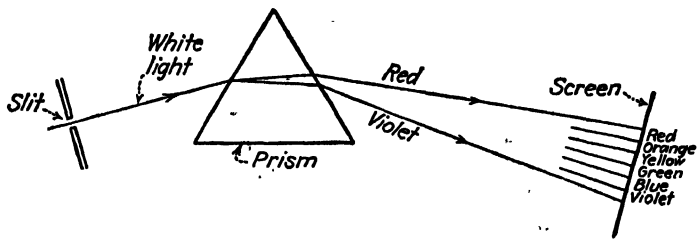


FIG. 143. *The production of a spectrum by the refraction of white light in a prism.*

liquid surface. A similar bending of wave fronts in water waves is shown in Fig. 142; here waves in the lower half of the picture move more slowly than those in the upper half because the water is shallower.

We shall have occasion later to mention a few of the manifold uses of reflection and refraction in optical instruments.

Color

The eye is amazingly sensitive to slight changes in the wave length of light. It records these changes as changes of color. As the wave length decreases through the visible range, the eye sees successively shades of red, orange, yellow, green, blue, violet. A mixture of all wave lengths is registered as white light, absence of radiation as black.

The white light we receive from the sun may be spread out into its different wave lengths by a wedge-shaped piece of glass called a *prism*. In Fig. 143, light strikes one face of a prism obliquely. The light is re-

fracted as it enters the glass, some wave lengths being deflected through a slightly greater angle than others. The color with the shortest wave length, violet, is deflected most, the red least. On leaving the glass at the other prism face the light is again refracted, and the various wave lengths are spread further apart. The brilliant band, or ***spectrum,*** of colors produced may be observed directly or projected on a screen. The natural spectrum we call the rainbow is produced in a similar manner, by refraction and reflection in myriads of water droplets.

Detailed analysis of the wave lengths present in any kind of light is accomplished by means of a spectroscope. The simplest form of this

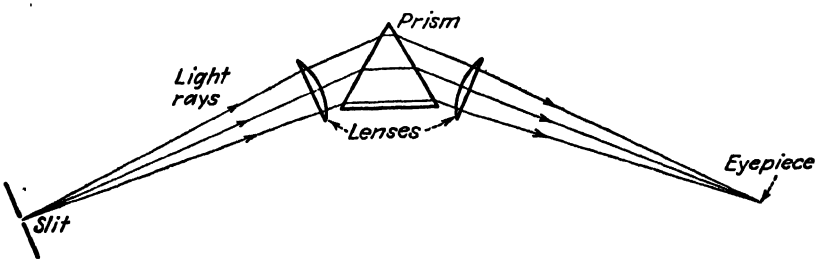


FIG. 144. *Arrangement of slit, prism, and lenses in a spectroscope.*

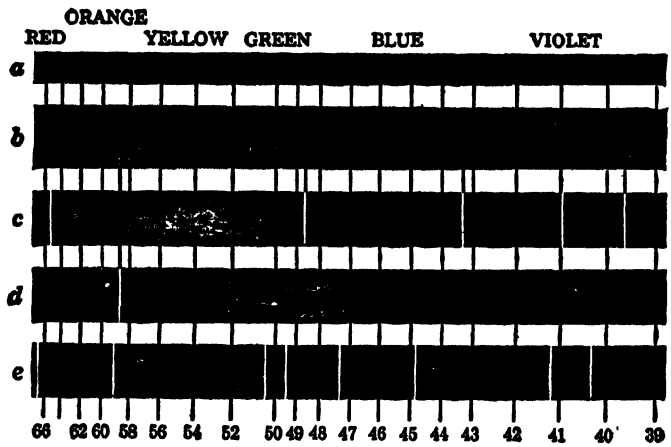


FIG. 145. *Kinds of spectra: a, continuous spectrum; b, dark-line or absorption spectrum of the sun; c, d, and e, bright-line spectra of hydrogen, sodium and helium respectively. (From An Orientation in Science by Watkeys and associates.)*

instrument consists of a prism placed between two cylindrical tubes (Fig. 144). One of the tubes has a narrow slit at one end, through which the light to be examined is admitted. The light is spread out into its separate wave lengths by the prism, and then passes down the second tube to the eyepiece. The tubes also contain a system of lenses which focus an image of the slit on the eyepiece, so that each separate wave

length appears as a colored image of the slit. If all wave lengths are present in the light, the slit images overlap and the spectrum is a continuous rainbow. If only a few isolated wave lengths are present, the spectrum consists of a few bright lines (Fig. 145).

Any solid or liquid material, or any gas if sufficiently compressed, when heated until it glows brightly gives out light of all visible wave lengths. At lower temperatures the most intense radiation is in the longer wave lengths, and the light appears red ("red heat"); as the temperature rises, the greatest intensity shifts to the middle of the spectrum, and the light appears white ("white heat"); at still higher temperatures, never reached on earth, the light becomes bluish. Because of this relation between color and temperature, astronomers can estimate the temperatures of stars from the intensity of various parts of their spectra.

Luminous gases at low or moderate pressures ordinarily produce light of only a few wave lengths. Their spectra consist of a few bright lines—*discontinuous* spectra, contrasted with the *continuous*, rainbowlike spectra of heated solids (Fig. 145). Thus the sodium-vapor lights used to illuminate foggy roads give out light largely restricted to two wave lengths in the orange-yellow part of the spectrum; the neon light that adorns our city streets consists chiefly of a few wave lengths toward the red end; mercury-vapor arcs used in "health" lamps give, besides ultraviolet radiation, a certain wave length of yellow light, one of green, and two of violet.

Colors of objects which we see by reflected light depend on the kind of light falling on them and on their composition. If an object can reflect red light but absorbs other wave lengths, it will appear red in sunlight; if it reflects chiefly green light, it will appear green in sunlight, and so on. Most objects reflect or absorb more than one color, and the color we see is a combination of those wave lengths which are reflected. Obviously an object can reflect a certain color only if that color is present in the light falling on it: thus in sodium light a red or green object would appear black. The ghastly hues which even the most carefully made up complexions assume under sodium or mercury light are due simply to the absence in these lights of wave lengths which skin and cosmetics normally reflect.

Some colors have a different origin. The blue of the sky, for instance, is due to scattering of the sun's light by molecules and dust particles in the atmosphere. The short wave lengths at the spectrum's blue end are scattered more effectively than red light; hence the sky, which we see only by scattered light, has an excess of blue, while the sun itself is a little more yellowish or reddish than it would appear if we had no atmosphere. At sunrise or sunset, when the sun's light must travel through great thicknesses of dust-laden air, the scattering of its blue light is more pronounced and the sun is often a brilliant red.

Interference

One property which light shares with other forms of wave motion is particularly important, since it can often be used, when other tests fail, to determine whether a given radiation consists of waves or of tiny material particles. This is the property of *interference.*

Suppose that two ropes AC and BC are connected to a third rope CD, as in Fig. 146. Suppose that waves are started along AC and BC by successive jerks at A and B. The effect of these waves on CD will depend on whether or not they arrive at C at the same time. If two wave crests reach C simultaneously, it and all succeeding points will move upward somewhat farther than points along AC and BC, since it is acted on by both waves. If, however, a trough on one rope reaches C simultaneously

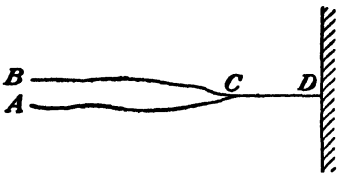


FIG. 146. *Waves started along the ropes BC and AC will interfere at C.*

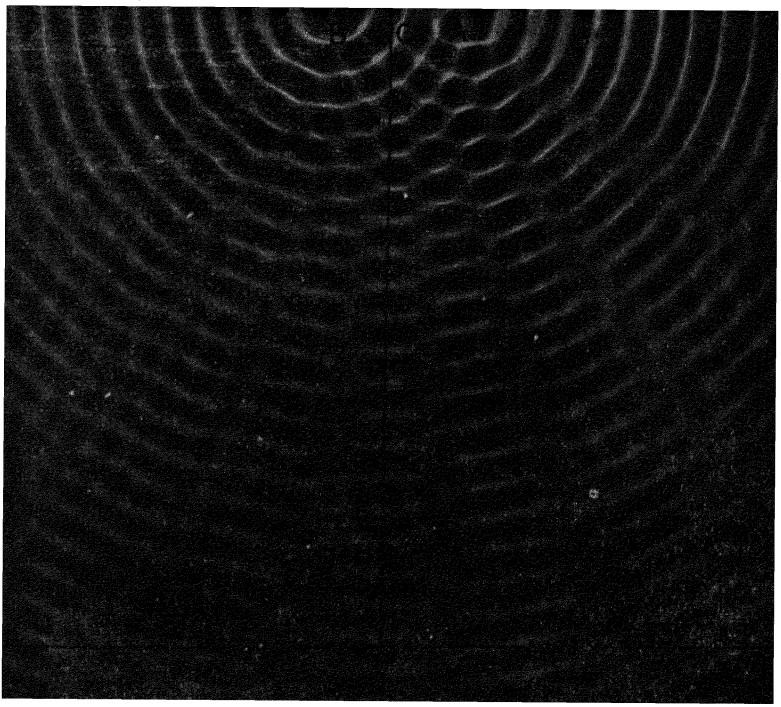


FIG. 147. *Interference of water waves. Ripples are spreading out from two sources, A and B. In some regions the ripples reinforce each other, and waves are prominent; in other regions the ripples destroy each other and waves are absent. (From Webster, Farwell, and Drew's General Physics for Colleges, by permission of D. Appleton-Century Company, Inc.)*

with a crest on the other, the two wave impulses will cancel each other and CD will not move. We might imagine further that AC and BC are ropes of different sizes, so that waves travel along them with different frequencies; then at C the waves will be sometimes in step, sometimes out of step, so that CD has waves only a part of the time.

When two trains of waves add together in this fashion to produce waves of increased or diminished strength, we say that the waves *interfere*. Interference phenomena may be beautifully demonstrated by means of water waves (Fig. 147).

Interference effects of light waves are readily shown by letting light fall obliquely on a soap film, preferably a film stretched across a wire ring (Figs. 148 and 149). Let us first use monochromatic light, that is, light of a single wave length or of a very few wave lengths close together in the spectrum—for instance, yellow light from the sodium-vapor lamp mentioned above. A soap film placed in this light shows a succession of yellow and black bands, in lines or circles or irregular figures (Fig. 148).

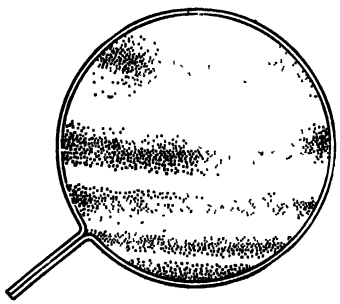


FIG. 148. *Interference bands produced by the reflection of light from a soap film.*

The bands can be explained with the aid of Fig. 150, which shows a greatly enlarged cross section of a part of the film. Light falling on the film (AO and BQ) is reflected twice, once from the upper surface and once from

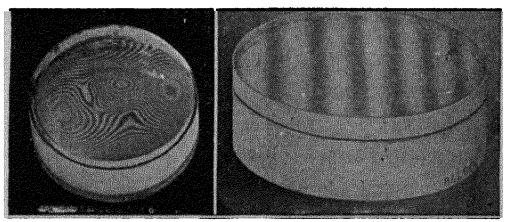


FIG. 149. *Interference of light by reflection from thin air films. In each picture the two glass plates are touching at one edge, very slightly separated at the opposite edge, leaving a wedge-shaped film of air between them. Light reflected from one surface of the film interferes with light reflected from the other surface (see Fig. 150). In the right-hand picture the surfaces are optically plane, while in the left-hand picture they are somewhat uneven. (Courtesy of Bausch and Lomb Optical Company.)*

the lower surface. Light rays from the two reflections travel upward from the film along nearly the same path, and their waves interfere. At some places on the film (as at O), wave fronts in the reflected rays are out of step, so that one cancels the effect of the other; at other places (as at Q) wave fronts are in step and reinforce each other. Where cancellation occurs, the film appears black; where reinforcement occurs, it is bright

yellow. What effect interference will have at any point on the film depends on the thickness of the film at that point, since the thickness determines how far one reflected ray lags behind the other.

When white light is used instead of monochromatic light, the reflected rays of one color will be in step at a given point while rays of other colors will not. Hence rather than white and black bands the film shows a succession of brilliant colors. These are the rainbow effects we see so often in bubbles and in oil films on water.

The interference effects produced in thin films give us the best sort of experimental evidence that light travels in waves, for no other simple kind of motion shows interference. It seems a strange bit of irony that the first man who carefully studied the colors of thin films, Sir Isaac Newton, favored another theory concerning the nature of light. Newton believed that light consists of tiny, rapidly moving material particles or *corpuscles*

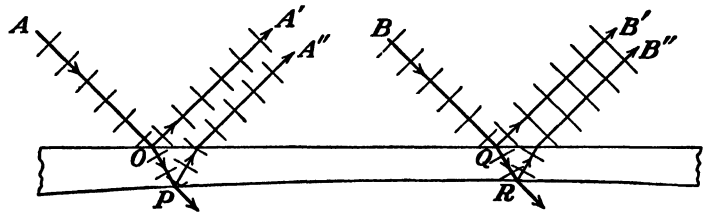


FIG. 150. *Cross section of soap film in Fig. 148, greatly enlarged. AO and BQ are two rays of monochromatic light falling on the film. AO is reflected in the two rays OA' and PA'', and BQ is reflected in the rays QB' and RB''. In each case the reflected rays interfere. OA' and PA'' cancel each other because their wave fronts are out of step, while QB' and RB'' reinforce each other.*

rather than waves. He never found a very satisfactory explanation for the colors of films in terms of these corpuscles, but the other common properties of light—straight-line transmission, reflection, refraction—could be explained with corpuscles as well as with waves. The corpuscular theory, backed by Newton's great name, found enthusiastic adherents for a century and more after his time. It was finally laid in its grave by further experimental work in the early nineteenth century.

In very recent years, as we shall see presently, physicists have in a sense resurrected the corpuscular theory, but in a transfigured form which Newton would scarcely have recognized.

Questions

1. Why is it that the electromagnetic waves set up by the alternating currents in light bulbs in your home do not produce disturbances in your neighbor's radio?
2. Would a stream of electrons moving at a steady rate along a wire produce electromagnetic waves? Why or why not?
3. If a spring is suspended in a horizontal position and one end is compressed by a sharp blow, a region of compression travels along the spring. Would a

series of compressions set up in this manner be longitudinal or transverse wave motion?

4. What is the wave length of the sound waves of C two octaves above middle C, if their frequency is 1,044 vibrations per second?
5. What is the wave length of radio waves whose frequency is 900,000 vibrations ("900 kilocycles") per second?
6. What is the frequency of X rays which have a wave length of 3×10^{-8} cm?
7. Why does another person look shorter to you when he is standing in a swimming pool than when he is standing in the air?
8. Make a diagram to show how a ray of light would be refracted on traveling obliquely from air into water. Explain the bending in terms of waves.
9. What kind of a spectrum would be given by (*a*) a piece of white-hot platinum, (*b*) molten iron, (*c*) the air in a partially evacuated tube through which an electric discharge is passing?
10. Which of three stars has the highest temperature, one whose light appears red, yellow, or blue?
11. What color would red cloth appear if illuminated by (*a*) a mercury-vapor lamp, (*b*) a sodium-vapor lamp, (*c*) the sun?
12. Why do films of oil on the surface of water often show bands of different colors?

CHAPTER XX

X Rays and Radioactivity

X RAYS are born in high-vacuum discharge tubes, the same spectacular instruments that led Thomson at Cambridge to the discovery of the electron. But the story of the discovery of X rays is scarcely a fitting parallel for Thomson's long and painstaking study of cathode rays. In his laboratory at Würzburg, in 1895, Wilhelm Roentgen happened to observe that a screen coated with a fluorescent salt glowed every time he switched on a near-by cathode-ray tube. In that moment X rays were discovered. Roentgen knew that cathode rays themselves could not escape through the glass walls of his tube—yet evidently some sort of invisible radiation was falling on the screen. The radiation was strangely powerful; thick pieces of wood, glass, even metal could be interposed between tube and screen, and still the screen glowed. At length Roentgen found that his mysterious rays would penetrate human flesh and leave a photographic record of bones beneath the flesh. With these observations Roentgen announced his discovery to the world, christening the radiations "X rays" after the algebraic symbol for an unknown quantity.

Like X rays, radioactivity was discovered by the merest accident. The year after Roentgen's announcement, Henri Becquerel, in Paris, was studying phosphorescent substances (materials which glow softly after exposure to light) to see if some of them could be made to emit X rays. Among the materials of his investigation was a yellow salt containing the heavy metal uranium. A bit of this salt was left accidentally on a photographic plate in Becquerel's darkroom; when he discovered it several days later, some obscure hunch prompted him to develop the plate. On it he found a dark spot, a rough silhouette of the pile of salt. This was extraordinary—a substance that could take a photograph of itself without benefit of X rays, cathode rays, or any other known radiation. More salt on other plates behaved similarly, no matter how carefully the plates were covered and protected from outside influences.

There could be no doubt: the uranium salt of its own accord was giving out some sort of radiation. Becquerel named this property of the material "radioactivity."

Radioactivity, X rays, and the electron, three discoveries which led directly to modern physics, were all announced within a few years of each other at the very end of the nineteenth century. It was widely believed at that time that all the really fundamental discoveries in physical science had been made: Faraday's experiments and Maxwell's equations had cleared up the mystery of electric and magnetic phenomena, light had been linked to electricity by Maxwell's equations, and John Dalton's atoms were firmly intrenched as the ultimate building blocks of the universe. To this comfortable belief the new discoveries came as a rude shock. Just how little science was expecting anything so strikingly new is shown by the fact that two of these fundamental discoveries were pure accidents.

X Rays

X rays are produced where cathode rays stop. Rapidly moving electrons must be brought to rest suddenly; the more abruptly they are stopped, the more powerful are the resulting X rays. An X-ray tube

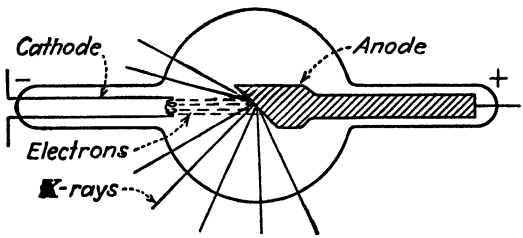


FIG. 151. *Diagram of an X-ray tube.*

is simply a cathode-ray tube designed to produce and to bring to a sudden stop a powerful beam of electrons. Usually the stopping is accomplished by the metal of the anode (Fig. 151). Crudely, we might compare the production of X rays to the stopping of a bullet by a tree: just as the bullet buries itself in the wood and gives up its kinetic energy as heat, so an electron buries itself in the anode and gives up its energy as X rays.

Several properties of X rays have been mentioned: their ability to make certain materials fluoresce, to darken a photographic plate, to pass through opaque matter. Another conspicuous property is their tendency to *ionize* air or other gases through which they pass—in other words, to disturb the gas molecules so violently that some are broken up into positively and negatively charged fragments or *ions* (page 231). An electroscope charged with either positive or negative electricity is quickly discharged when placed in such an ionized gas, since it attracts

to itself gas particles with a charge opposite to its own. This fact gives us a convenient means of detecting and measuring the intensity of X rays: we need only put a charged electroscope in their path and observe its rate of discharge (Fig. 152).

That X rays themselves carry no charge is readily shown by trying to deflect them in powerful electric and magnetic fields, as Thomson deflected cathode rays. While cathode rays are repelled by a negative charge and deflected at right angles to the lines of a magnetic field, X rays show no change in direction in either kind of field.

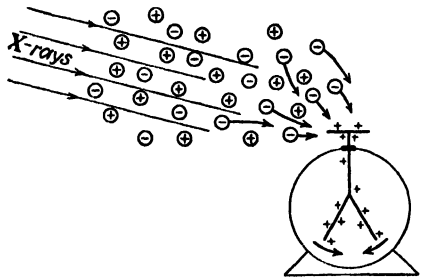


FIG. 152. *Ionization of air by X rays. Both positive and negative ions are formed. Negative ions are attracted to a positively charged electroscope, neutralizing its charge. If the electroscope were charged negatively, positive ions would be attracted.*

Since charged particles were ruled out, physicists following Roentgen had to choose one of two alternatives regarding the nature of these peculiar rays: either they were a form of wave motion, or they consisted of tiny neutral particles. Now seemingly this choice should be easy. To our gross senses, a wave and a particle are eminently dissimilar; each has its own characteristic properties and should be readily recognizable. Yet physicists bickered for seventeen years over the nature of X rays before the question was finally settled. This was but the first of several arguments in the annals of modern physics as to whether various radiations should be regarded as waves or as particles.

The difficulty in working with X rays is their penetrating power. It was useless to set up a paddle wheel in their path and expect them to demonstrate kinetic energy by pushing it, as cathode rays will; for X rays in large part go through the wheel instead of being stopped by it. Equally fruitless were early experiments designed to test their wave nature by refraction. The rays betrayed themselves at last by showing the characteristic wave property of interference—but only after a new experimental technique had been devised. Ordinary interference methods, using the soap and oil films which give such beautiful results with visible light, were tried and seemed only to support the particle hypothesis by their negative results. Finally in 1912 the German physicist Laue suggested that X rays might be made to interfere by passing them through a crystal. It occurred to Laue that X rays might be waves with such exceedingly short wave lengths that ordinary films would not be thin enough to cause interference. But in a crystal the particles are arranged in planes very close together—possibly close enough so that two adjacent

planes might act as a very thin film, X rays reflected from one plane interfering with those reflected from the other (Fig. 153). Laue's guess proved correct. The interference patterns which his students photographed by allowing X rays to pass through crystals silenced at once the physicists who supported the particle hypothesis (Fig. 154).

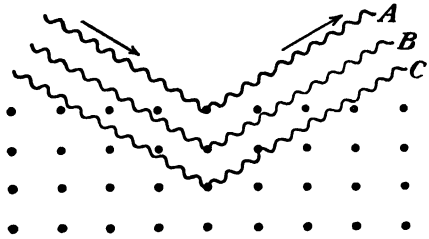


FIG. 153. *X rays are partly reflected by particles at different levels in a crystal lattice. The reflected rays from deeper layers (B, C) travel over longer paths than those from the surface layer (A). Hence waves in the different reflected rays may be either in step or out of step, depending on the angle of reflection and the spacing of particles in the lattice.*

X rays, then, are waves—disturbances in an electromagnetic field, kin to light and radio waves. Their speed is that of light, their wave length roughly a thousand times smaller. As radio waves are produced by starting and stopping electrons in a wire, so X rays result from the sudden stopping of electrons in a cathode-ray tube. The faster the electrons move before collision (*i.e.*, the higher the voltage), the more energy they give out as X rays. This means not only a greater intensity of X rays, but X rays of higher frequency and greater penetrating power.

Once the nature of X rays had been established, Laue's technique could be used in reverse; X-ray interference patterns were made to reveal the intimate structure of crystalline solids. Most of our knowledge of the arrangement of particles in crystals has come from these X-ray studies. X rays also, as we shall find later, gave physicists some important hints regarding the structure of atoms. But of course to the world at large the greatest service of Roentgen's mysterious radiations has been to make visible injured and diseased portions of the human body and to aid in curing malignant growths.

Radioactivity

Experiments following his initial discovery convinced Becquerel that the part of the salt responsible for darkening his photographic plates was the element uranium. Any compound of this metal produced a similar darkening, and the amount of darkening was roughly dependent on the amount of uranium present. Having discovered these few facts about the new phenomenon, Becquerel turned the problem over to a young woman working in the laboratories at the Sorbonne. Her name was Marie Curie.

The story of Madame Curie and her husband has been often told. Early in their work on radioactivity came the surprising discovery that a

piece of pitchblende, a black mineral from Bohemia, produced a darkening of photographic plates out of all proportion to its uranium content. This suggested that the mineral contained traces of some other element far more powerful than uranium. The Curies set themselves the fantastically difficult task of isolating this unknown element.

There followed two years of feverish labor, seeking to wrest from a ton of stubborn black ore a fraction of a gram of a substance whose properties were completely unknown. When their work was at last completed, the Curies had added not one but two new elements to the periodic table. The first to be discovered was named polonium, after Madame

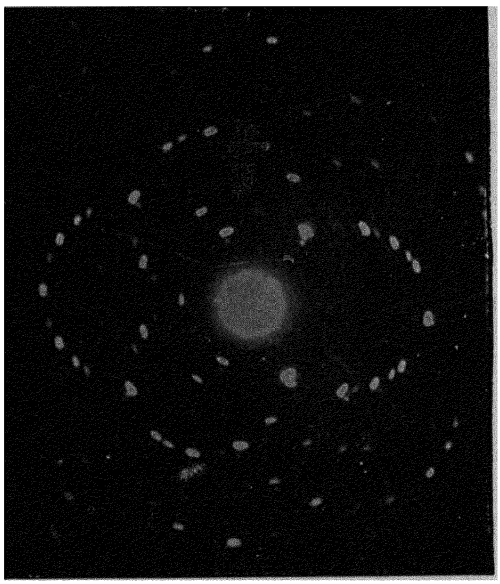


FIG. 154. *The interference pattern produced by the reflection of X rays from particles in the crystal lattice of calcite. The bright spots are the places where X rays reflected from various layers in the crystal come in step. (From Elements of Physics by Smith.)*

Curie's native Poland; the other, more abundant element was the famous metal radium.

Radium belongs in the second group of Mendelyeev's table, with calcium and barium. Like them, it is a soft, silvery metal, tarnishing rapidly in air and dissolving in water with the evolution of hydrogen. Unlike calcium and barium, radium shows an astonishing ability to produce the radiations which Becquerel associated with uranium. Thousands of times more active than uranium, radium and its salts not only blacken photographic plates but glow softly in a darkened room.

The radiation from radium and other radioactive elements remained for several years quite as mysterious as X rays. Like X rays, the radiation

caused fluorescence, darkened a photographic plate, caused ionization in gases, and at least a part of it was highly penetrating. The harshest physical or chemical treatment could not stop the radiation or even slow it down: whether the elements are free or combined in a salt, cooled in liquid air or heated in an electric arc, their ability to radiate remains unchanged. Spontaneously and continuously, year after year, these substances give out their peculiar rays. Evidently the emission is somehow associated with the atoms themselves—and with a part of each atom which is not affected by ordinary physical and chemical changes.

In electric and magnetic fields the radiation shows a complex behavior. It splits into three parts: one fraction is deflected to one side, a second fraction to the other side, and a third portion continues straight through (Fig. 155). Before their nature was known the three parts were labeled provisionally with the first three letters of the Greek alphabet: *alpha rays* were those which were deflected as if they carried a positive charge, *beta rays* those which seemed to have a negative charge, *gamma rays* those which passed undeflected through a field. Investigation has shown that gamma rays are electromagnetic waves shorter than X rays, and that beta rays are electrons. But alpha rays, by all odds the most interesting of the three, consist of particles the like of which we have not met before.

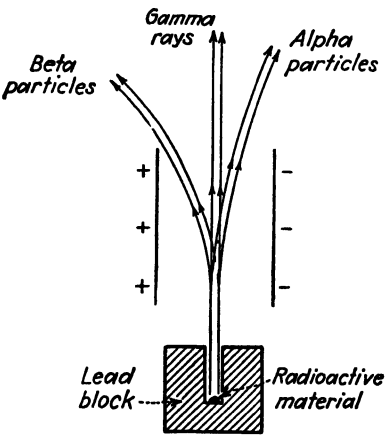


FIG. 155. *Deflection of alpha rays and beta rays in an electric field.*

They proved to be atoms of helium. More correctly, we should say *ions* of helium—atoms which have been stripped of two electrons apiece, so that they carry a double positive charge. Seven thousand times as heavy as an electron, moving much more slowly than beta or gamma rays (only about 10,000 mi/sec), the particles of alpha rays cannot penetrate as far as their brethren through gases or metals; but their relatively enormous mass makes them highly destructive to atoms which get in their way. Collisions between alpha particles and other atoms, as we shall see presently, gave physicists their best early clues in deciphering atomic structures.

One type of alpha-particle collision can be observed with astonishing simplicity. Take a watch or clock with a luminous dial into a darkened room, give your eyes time enough to become thoroughly accustomed to the darkness, and then look at the glowing paint through a low-power magnifying glass. You will find that the glow is made up of myriads

of tiny flashes, like the sparks from a rocket, each flash lasting only a moment. In the paint is a tiny bit of some radioactive substance, together with a material which will fluoresce under its influence; each flash is the disturbance created when a single alpha particle is stopped by molecules of the fluorescent substance. You do not see the helium ions themselves, but you see the effect that each one produces.

Atomic Disintegration

The fraction of a gram of paint on your watch dial, with its incredibly small amount of actual radioactive material, has been giving off helium ions in countless multitudes since the watch was made. Now it would be fairly easy to show that no helium is present in the materials of the paint when it is mixed, and equally easy to show that the helium would be produced just as rapidly if all the materials of the paint were removed except the infinitesimal speck of some radioactive element which it contains. *The helium is evidently being produced spontaneously from the atoms of the radioactive element.* With this conclusion we abandon our idea that elements are substances which cannot be broken down into simpler substances.

Naïvely we may think of the expulsion of an alpha particle from a radium atom as a sort of Lilliputian explosion: a fragment of the atom bursts from its deep interior, and part of the energy of the explosion appears as gamma radiation. The structure remaining after the explosion is no longer a radium atom; it has lost four atomic-weight units, and is accordingly an atom of an element with altogether different chemical properties. This is the element radon, a gas which forms no compounds, hence is assigned to the zero group of the periodic table. Like radium, radon is intensely radioactive. Each of its atoms emits another alpha particle, becoming in the process an atom of radium A, which likewise is radioactive. Thus the explosion of a radium atom sets afoot a long series of radioactive changes.

Radium itself is the product of a series of changes starting with uranium. Table XIV shows this complete series of radioactive elements, from uranium through radium and radon finally to lead, which is not radioactive. Each element in the series is produced by the disintegration of the one above it and produces in its turn the one below it. Note that some of the changes take place by the emission of an alpha particle, some by emission of a beta particle. Gamma radiation does not of itself produce a change from one element to another, but appears to be energy given out during alpha-particle emission.

Two other disintegration series are known, involving other atoms of high atomic weight and ending like the uranium series with lead. Elements with atomic weights less than that of lead, with two or three exceptions, are not naturally radioactive.

Radioactivity in a sense fulfills the alchemist's dream, for it is indeed transmutation from one element to another. But in a truer sense it is a bitter caricature of that dream: this transmutation proceeds not from lead to gold, but irrevocably from elements more precious than gold to the base metal lead.

TABLE XIV. RADIOACTIVE TRANSFORMATIONS IN THE SERIES URANIUM-RADIUM-LEAD

Element	*Group in periodic table*	*Atomic weight*	*Half life*	*Particle emitted during transformation*
Uranium	VI	238	4.4×10^9 years	Alpha particle
Uranium X_1	IV	234	24.5 days	Beta particle
Uranium X_2	V	234	1.14 min	Beta particle
Uranium II	VI	234	3×10^5 years	Alpha particle
Ionium	IV	230	8×10^4 years	Alpha particle
Radium	II	226	1590 years	Alpha particle
Radon	0	222	3.82 days	Alpha particle
Radium A	VI	218	3.05 min	Alpha particle
Radium B	IV	214	26.8 min	Beta particle
Radium C	V	214	19.7 min	Beta particle
Radium C′	VI	214	10^{-6} sec	Alpha particle
Radium D	IV	210	22 years	Beta particle
Radium E	V	210	4.9 days	Beta particle
Polonium	VI	210	140 days	Alpha particle
Lead	IV	206		

The number following each element in Table XIV is its so-called *half life*, the length of time necessary for half of any quantity of the element to disintegrate. Thus if you start with 1 million radium atoms, after 1,600 years 500,000 of them will have changed to radon; after another 1,600 years half the remainder will be gone. If you used instead the feebly radioactive element uranium, you would wait over 4 billion years for half of its atoms to disappear. If you worked with the extremely active element radon, you would wait less than four days. These half lives are about all the information we have concerning the time required for radioactive decay. We do not know how long it would take a million atoms of radium to change completely to radon, nor can we predict just how long an individual atom will last. We have only the sort of statistical information which enables a life insurance company to calculate accurately the death rate of a country's population, even though it can make no guess as to when an individual will die.

So precisely are the statistical laws of radioactive decay known that we shall use them presently to calculate the age of the earth. So well understood are the effects of the three kinds of radiation that we use

them confidently in medical and scientific work. But the actual mechanism of the process presents us with as nasty a set of unanswered and partly answered questions as modern physics can provide. Where, for instance, do these strange elements get their energy—energy in quantities so prodigious that radium, pound for pound, produces 250,000 times the heating effect of coal? Then why is radioactivity so rare a phenomenon? Why should it be the special prerogative of a few heavy elements at the end of the periodic table? Perhaps most perplexing of all are the problems raised by the statistical laws we have just discussed. Why should the atoms explode in so regular a sequence, one after the other, rather than all at once? So far as we know, a radium atom which will emit an alpha particle 2 minutes hence is identical with one which will endure for 10,000 years. Is it mere chance that sets one off before the other, or is each atom provided with some sort of clockwork which governs its behavior?

Questions

1. Arrange the following types of radiation in the order of decreasing wave length: X rays, red light, violet light, radio waves, gamma rays, ultraviolet light. Which of these has the greatest frequency? Which the least?
2. What is the frequency of gamma rays which have a wave length of 10^{-9} cm?
3. What is the experimental evidence for each of the following statements?
   a. X rays consist of wave motion rather than streams of tiny particles.
   b. Cathode rays consist of tiny particles rather than wave motion.
   c. Radioactive substances produce three different kinds of radiation.
   d. X rays ionize air.
4. How could you tell experimentally whether or not a chunk of rock contains an appreciable quantity of a radioactive element?
5. If 1 g. of radium is left undisturbed, how much of it will remain after 1,600 years? After 4,800 years? What becomes of the part which disintegrates?
6. What happens to the atomic weight of a radioactive element when it emits an alpha particle? When it emits a beta particle? When it emits a gamma ray? (Atomic weight of helium is 4; the weight of an electron is about $\frac{1}{1800}$ of the weight of a hydrogen atom.)
7. Make a graph showing how radium decays. Suppose you have 1 g. of radium at the start, and suppose you keep track of its decay for 16,000 years. On the vertical axis plot "grams of radium left," and on the horizontal axis plot "time in years."

CHAPTER XXI

The Atomic Nucleus

IN THE early dawn of a midsummer morning in 1945, at an out-of-the-way spot on the New Mexico desert, a small group of men waited nervously. They had reason to be nervous: For months they had worked feverishly on a new kind of explosive, more powerful than any ever dreamed of before, and this day had been appointed for the final, crucial test of the bomb they had put together. Calculations and preliminary tests had assured them that it would work, but as in every untried scientific experiment there remained the chance that something might go wrong, that some unsuspected factor had been left out of the calculations. Slowly the hands of their watches moved to the hour and minute when the bomb was to be exploded. A few more seconds of dreadful suspense and then suddenly there flashed across the desert a white glare more brilliant than the noonday sun. Ten miles away the new bomb had exploded. Even with the thickest of dark glasses the observers dared not watch the explosion directly, but photographs later showed a great ball of light, expanding rapidly and stirring up ahead of itself billowing clouds of smoke and dust. When the glare had subsided so that the last stages of the explosion could be watched, a pillar of cloud like the column of ash from a volcano was rising slowly into the sky. The watchers breathed more easily, grinned at each other, and shook hands: The atomic bomb had become a reality.

Forty years earlier the possibility of such a bomb had been foreshadowed in the work of the great German physicist Albert Einstein. Einstein's bold suggestion was that matter and energy, previously thought to be separate and distinct, could be changed one into the other. In ordinary physical and chemical reactions the amount of matter that changes into energy would be too small for detection, but some conversion of matter into energy should be observable in processes involving particles moving at very high speed. In succeeding years, as physicists devoted more and more attention to the behavior of particles shot out

from the interiors of atoms, the correctness of Einstein's prediction was repeatedly demonstrated: Matter was indeed transformed into energy, and the amount of energy formed from a very little matter was prodigious. Here, it seemed, within the atom was a new source of energy which might take the place of the coal, oil, and water power which had served mankind as energy sources for so long. Dreamers could envision liners crossing the ocean, or trains crossing continents, on the energy contained in the atoms of a cup of water. But the release of energy by destruction of matter remained for many years a curiosity of physics laboratories, taking place always on a scale too small for practical use, until the explosion on that summer day in New Mexico showed that a way had been found to produce atomic energy in large quantities.

The development of the atomic bomb is a story of fifty years of research by physicists of many countries, beginning with the discoveries of X rays, radioactivity, and electrons in the 1890's. Only the very last part of this research was directed toward the production of atomic bombs; for most of the fifty years, scientists who delved into the structure of the atom were trying simply to find out more and more about how matter is put together. Until the summer of 1945, when the first bombs suddenly made atomic energy a tremendous factor in world politics and military strategy, the study of atomic structure seemed a branch of pure science as remote from everyday experience as Kepler's study of planetary motions. The story of the half-century of research which led up to the atomic bomb is long and intricate, but it is one of the most exciting in the whole history of science.

The Tools of Atomic Research

The inner structure of atoms has been studied largely by a single kind of experiment: The atoms to be investigated are bombarded with small particles or with high-frequency electromagnetic radiation, and the products of the resulting collisions are identified. The particles used include electrons, alpha particles from radioactive elements, and several others that we shall meet soon; the electromagnetic radiation includes X rays and gamma rays. Evidently the major experimental problems are (1) the production of the bombarding particles or radiation, and (2) the detection of the products of their collisions with atoms. The story of research on atomic structure is in part a story of improvements in instruments designed to solve these two problems.

Production and Acceleration of Particles. The earliest experiments on atomic bombardment were carried out with alpha particles from radium. It was quickly realized that the bombardment could be made more effective by using smaller particles and by giving them greater speeds than natural alpha particles possess. To speed up a charged par-

ticle, the obvious method is to place it between two electrodes in an evacuated tube and then give the electrodes strong charges of opposite sign. The stronger the charge the faster the motion and the more effective the bombardment; so considerable effort was devoted to building electrostatic machines of enormous size, some able to generate potential differences of several million volts. A different approach was tried by E. O. Lawrence at the University of California: Instead of using a single tremendous charge to speed up particles, he designed an instrument that sent a stream of particles again and again through a fairly small potential difference; each time the stream passed through the electric

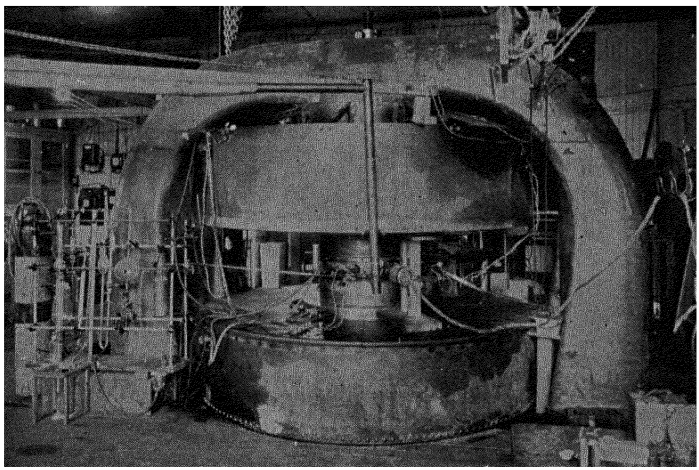


FIG. 156. *A cyclotron, one of the modern instruments used to accelerate electrified particles for atomic bombardment experiments. (Courtesy of E. O. Lawrence, University of California.)*

field its speed was increased, much as a gasoline engine is speeded up by successive fairly small explosions in its cylinders. With this device, called the *cyclotron* (Fig. 156), Lawrence and his students have succeeded in giving particles speeds that correspond to a potential difference of more than a hundred million volts. The cyclotron and related devices have become the most widely used "guns" for atomic bombardment experiments.

The production of high-frequency radiation is not as difficult a problem as the production of high-speed particles. Gamma rays from radioactive materials can be used directly, and X rays are produced abundantly by the modified cathode-ray tubes described in the last chapter.

Detection of the Products of Atomic Bombardment. When an atom is damaged by collision with particles or radiation, the products of the collision are other moving particles or more radiation or both. Like the

particles and gamma rays from radium, these products of collision (most of them, at least), have the ability to ionize gases through which they move. It is this property which is used in nearly all instruments for detecting them.

The most spectacular of these instruments is the *cloud chamber*, devised by C. T. R. Wilson at the suggestion of J. J. Thomson. The story goes that Thomson casually remarked to Wilson that he needed an instrument to photograph the paths of individual electrons moving through a gas; and Wilson, instead of laughing the matter off or concluding that Thomson had suddenly gone crazy, undertook the job in all seriousness. He succeeded with a device of amazing simplicity: It had no complex array of transformers and magnets and glowing tubes, but consisted simply of a box with small windows and a movable piston in

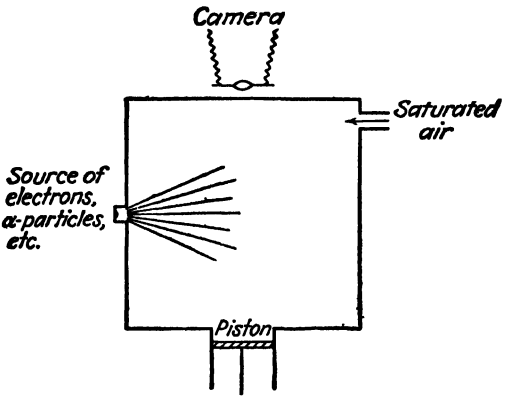


FIG. 157. *Diagram of a Wilson cloud chamber. Air saturated with water vapor is admitted to the box and caused to expand by lowering of the piston. A moment later a light flashes on to expose the plate in the camera.*

its bottom (Fig. 157). Air saturated with water vapor could be introduced into the box; moving the piston down caused the air to expand and cool, so that part of the water vapor would condense as a light fog or cloud. Now water vapor condensing into liquid droplets collects preferably around dust particles or electrically charged air molecules. Electrically charged molecules, or ions, are produced in abundance when a fast-moving electron travels through air, so if an electron is moving through the box when the air expands its wake provides a favorable place for fog droplets to condense. The ions and the droplets collecting around them move away from the precise path of the electron because of molecular collisions, so that the fog becomes spread over a strip much wider than the original path. Within a small fraction of a second, in fact, it becomes wide enough to be visible to the naked eye as a white streak in the thinner fog around it. Now if details of timing are worked out so that a light

flashes on to expose the plate in the camera just after the fog is produced, a picture will be obtained of the paths of any electrons which are moving through the box at the time. Thus Wilson succeeded in making visible the tracks of particles that are far too small to be photographed directly.

Not only electrons but any moving particles that ionize gases leave fog tracks on cloud-chamber photographs. Paths of alpha particles are particularly clear (Fig. 158). A complete record of an atomic collision is sometimes obtained—the path of the high-speed particle before collision, the sudden break in its path as the particle meets an atom, and the paths of any particle or particles that result from the collision (three collisions are shown in Fig. 158).

Another device for detecting the presence of charged particles by means of their ionizing effects is an *ionization chamber*. This instrument

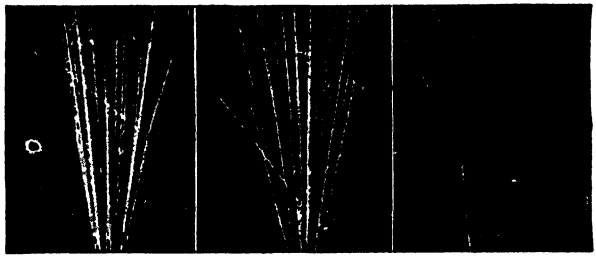


FIG. 158. *Cloud-chamber photographs of alpha-particle tracks. (Blackett.)*

consists of a small, partly evacuated box or tube, containing two metal electrodes kept at a constant difference of potential. The voltage is low enough so that normally no current flows through the gas from one electrode to the other, but if ions are produced in the chamber by a moving particle they are drawn to the electrodes and a momentary current is recorded. The current, although extremely feeble itself, can be used to control larger currents, which cause a light to flash or the needle of a meter to be deflected. Ionization chambers can be made so sensitive that they will respond to the ions produced by a single moving particle. A particularly sensitive chamber called a *Geiger-Muellcr counter*, with a voltage high enough to produce a spark discharge between the electrodes when ionization occurs, is widely used for automatically counting the number of particles passing through it.

The Emptiness of Atoms

The earliest systematic experiments using the atom-bombardment technique were carried out by Ernest Rutherford in the second decade of this century. This energetic New Zealander (Fig. 159), in later life Thomson's successor at Cambridge, set out to explore the atom's interior by observing the paths of alpha particles through gases and thin

metal sheets, keeping track of the particles at first by their flashes on a fluorescent screen (page 278), later with the vastly more convenient cloud chamber. His first astonishing conclusion was that atoms of matter are not solid spheres, but empty structures consisting of extremely minute particles separated by relatively great distances.

Rutherford based this conclusion on the straightness of alpha-particle tracks. As recorded in cloud-chamber photographs the paths through a gas are nearly always straight lines. Now of course a gas is largely empty space, simply because its molecules are far apart; but simple calculation from the kinetic theory shows that an alpha particle traversing a few centimeters of gas must nevertheless pass *through* several thousand atoms. By these encounters the swift-moving particle is at length slowed down, but is not appreciably turned from its straight path. Once in a while, however, an alpha-particle track does show a deflection, usually a very sharp deflection as if the particle had struck some massive obstacle (Fig. 158). Thus from the point of view of an alpha particle a gas behaves like a soft cushion with scattered hard pellets imbedded in it; collision with a pellet will turn the particle abruptly aside, but the pellets are so small and so far apart that its chance of hitting one before being slowed and stopped by the cushion is very slight.

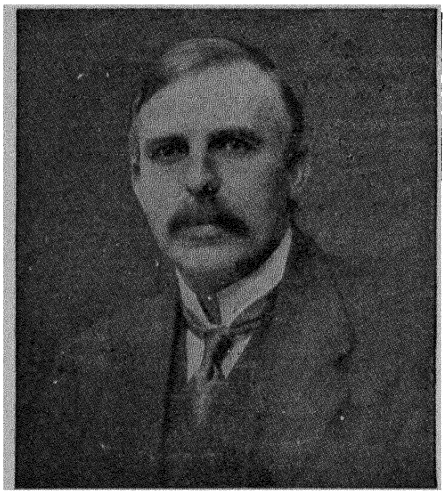


FIG. 159. *Ernest Rutherford (1871–1937).*

This behavior can be explained, according to Rutherford's calculations, if we assume that within atoms of the gas are powerful electric fields but very little actual substance except for small, heavy cores at their centers. Suppose, as a rough analogy, that a small star approaches the solar system from outer space. With sufficient speed, the chances are excellent that it will pass through unharmed, except for a slight slowing down by the gravitational fields of sun and planets; even direct collision with a small planet would not affect its motion appreciably. Only if it happened to approach closely the great central mass of the sun would its direction be radically altered. Similarly, said Rutherford, an alpha particle plows straight through the electric field of an atom, undeterred even by striking an electron now and then; only close approach to the heavy central core turns it aside.

So from Rutherford's work emerges our modern concept of the atom:

A small, heavy core, or ***nucleus,*** containing all the atom's positive charge and most of its mass, surrounded by the ***electron cloud,*** a superstructure of electrons at relatively great distances from the nucleus and from each other. Somewhat naïvely we may imagine the electrons moving around their nucleus like planets around the sun, at speeds just great enough to prevent electrical attraction from pulling them into the nucleus, just as a planet's orbital motion prevents gravity from pulling it toward the sun.

Ordinary matter, then, is mostly empty space. The solid wood of my table, the steel which supports a skyscraper, the hard rock underfoot, all are but myriads of moving electric charges, isolated from each other by greater distances, comparatively, than is the earth from its sister planets. If all the actual matter, all the electrons and nuclei, in your body could somehow be packed closely together, you would shrivel to a speck just visible with a magnifying glass.

Atomic Numbers

Rutherford's general picture of an atom as a positive nucleus surrounded by moving electrons can be used to explain many familiar observations. Since an ordinary atom is electrically neutral, the total positive charge on the nucleus must equal the negative charge of all the electrons in the electron cloud. If one of the electrons in the cloud is temporarily lost, the atom as a whole will show a positive charge; if an extra electron is temporarily added to the cloud, the atom as a whole will have a negative charge. Ions produced by high-speed particles are just such charged atoms or charged molecules that have lost or gained electrons as the particle moved through them. Ordinary positive charges on pith balls or electrodes imply a deficiency of electrons in the electron clouds of the atoms present, and negative charges imply an excess. The easy movement of electric currents through metals suggests that some electrons in the clouds of metal atoms are loosely held and can jump from one atom to the next.

Rutherford's work supplied also more precise information about the make-up of different atoms. The deflection which an alpha particle undergoes as it approaches an atomic nucleus depends on the amount of positive charge in the nucleus, so that measurements of the deflection by atoms of different elements provide a means of estimating the amounts of nuclear charge. Rutherford found that all atoms of any one element have the same nuclear charge, and that this charge is different for different elements; and further that the amount of charge increases regularly from element to element in the periodic table. The nucleus of hydrogen has a positive charge equal to the negative charge of a single electron; the nucleus of helium has a charge equal to that of two electrons; the lithium nucleus has a charge equal to that of three electrons; and so

on up to the most complex element known to occur naturally, plutonium, whose nucleus has a positive charge equal to the negative charge of 94 electrons. The number of unit positive charges on the atomic nuclei of an element is called the ***atomic number*** of the element; thus the atomic number of hydrogen is 1, of helium 2, of lithium 3, and of plutonium 94. Atomic numbers of all the elements are given in Table IX, page 175.

Since in normal atoms the positive nuclear charge must be equalized by the total negative charge of the electrons in the cloud, atomic number can also be defined as the number of electrons in the uncharged atoms of an element. For example, the atomic number of phosphorus is 15; this means that the nucleus of the phosphorus atom has a positive charge of 15 and that around this nucleus is an electron cloud of 15 electrons.

The atomic number of an element is its most fundamental property. Its atomic weight may vary slightly, as we shall see presently; atoms may have somewhat different weights and still show almost identical physical and chemical properties. But no change in atomic number is possible without a radical change in properties. The amount of positive charge on the nucleus of an atom seems to determine the fundamental nature of the atom and to distinguish it from all others. As an illustration of the significance of atomic numbers, Mendelyeev's periodic law becomes an exact law when these numbers are used in place of atomic weights: *If the elements are arranged in the order of increasing atomic number, elements with similar properties recur at regular intervals.* This statement of the law eliminates the awkward exceptions (such as the reversal in order of potassium and argon, page 204) that creep in when atomic weights are used.

Protons, Neutrons, and Other Particles

Rutherford not only established the existence of the nucleus, but took the first important step in deducing its structure.

In 1919, on a photograph of alpha-particle tracks through nitrogen, Rutherford found one track that ended in a peculiar sort of collision. From the point where the alpha particle was stopped, a thinner fog track started off in a different direction. Evidently out of the collision with a nitrogen atom had come another particle less effective in ionizing the gas than the alpha particle. Rutherford computed for this secondary particle a mass about one-fourth that of the alpha particle and a charge equal to that of the electron but positive instead of negative. The new particle was labeled the ***proton.***

Now the alpha particle is an ion of helium with an atomic weight of 4. A particle with one-fourth of this weight and a unit positive charge should be identical with the nucleus of a hydrogen atom, since hydrogen has an atomic weight of 1 and also an atomic number of 1. *Thus the proton is a*

hydrogen nucleus. A normal hydrogen atom consists of a proton with a single electron revolving around it. The two particles have equal charges, but the mass of the proton is almost 2,000 times that of the electron, so that the proton accounts for practically the entire weight of the atom. On the atomic weight scale its mass is 1.0076, nearly equal to that of the hydrogen atom itself.

Rutherford's experiment is important not only for showing that protons can be knocked out of nitrogen atoms, but also for demonstrating that it is possible artificially to change one element into another. The nitrogen nucleus from which the proton comes is no longer nitrogen, but oxygen, formed by addition of the alpha particle to the nitrogen nucleus. In symbols, the complete reaction is

$$\text{N (nucleus)} + \text{He (alpha particle)} \rightarrow \text{H (proton)} + \text{O (nucleus)}$$

This is a queer sort of chemical reaction, showing one pair of elements being formed from a different pair. It expressly violates Dalton's rule that atoms are unchangeable and indestructible. We encountered somewhat similar processes in discussing radioactivity, for example, the decomposition of radium atoms into atoms of lead and helium; but there the process is spontaneous, while here Rutherford had accomplished the change from one atom to another in a laboratory experiment. This was the first *artificial* transmutation of elements.

In later photographs of alpha-particle tracks through substances other than nitrogen Rutherford found occasional collisions showing the telltale thin fog tracks of protons. Since protons could be knocked out of many kinds of atoms, he hazarded a guess that these particles made up a part of all atomic nuclei. *Perhaps protons were the fundamental positive charges,* as electrons are the fundamental negative charges; perhaps the atomic number of an element is simply the number of protons within its nucleus.

In the late 1920's, several observers reported a different kind of collision involving alpha particles and atoms of very light elements, particularly atoms of the metal beryllium (atomic weight 9). In these collisions no fog tracks of emitted particles appeared from the actual sites of the collisions, but emission of some kind of radiation was suggested by tracks which appeared to start spontaneously in other parts of the cloud chamber and by discharges produced in near-by ionization chambers. Evidently the radiation had no ionizing effect itself, but was capable of exciting atoms at a distance so that they would produce ionization. This effect might be produced by very high frequency electromagnetic radiation—gamma rays of very short wave length—and for some time this explanation was favored. Finally in 1932, Chadwick in Rutherford's laboratory proved that the radiation consisted not of gamma rays but of tiny un-

charged particles, which the alpha particles had knocked out of beryllium nuclei. These uncharged particles, christened ***neutrons,*** have approximately the same mass as the proton but lack its positive charge. Because of their lack of charge and their extremely small size they can travel through the electron clouds of atoms without producing enough disturbance to cause appreciable ionization, but when they happen to strike a nucleus, charged particles are emitted, which cause the observed fog tracks at a distance from the original collision.

Neutrons are important not only because they are essential constituents of atomic nuclei but because they have proved more effective than

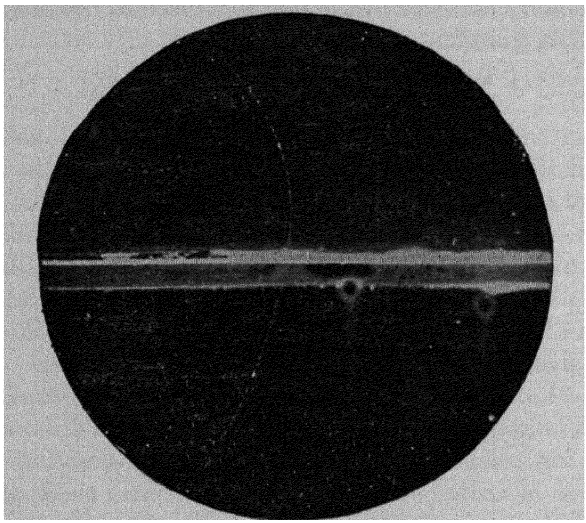


FIG. 160. *Cloud-chamber photograph of a positron track. The horizontal band across the middle of the picture is the edge of a lead sheet, through which the positron moved along a curved path. The curvature of the track is due to deflection of the particle by a magnetic field perpendicular to the plane of the figure.* (*Courtesy of Carl Anderson, California Institute of Technology.*)

charged particles as "bullets" for atomic bombardment experiments. Their lack of charge makes them difficult to control and difficult to detect, but the development of techniques for handling them was essential in the construction of atomic bombs.

Cosmic rays, high-energy radiations which come to the earth from outer space (page 315), have revealed three other particles which sometimes are emitted from atomic nuclei. The *positron,* discovered in 1932 by an American, Carl Anderson, in the debris of a cosmic-ray explosion, has since made its mark on cloud-chamber photographs of several kinds of atomic disintegrations (Fig. 160). It is a twin of the electron, possessing the same mass and a charge equal in magnitude but positive instead of negative. It differs from the proton in its much smaller mass. The pos-

itive and negative *mesons* (or *mesotrons*), particles with charges equal to that of the electron but with masses at least a hundred times greater, were also discovered on cloud-chamber photographs of cosmic-ray explosions. Positrons and mesons are exceedingly short-lived, losing their independent existence by joining together with other particles within a small fraction of a second after they are formed.

These discoveries give us a total of six *fundamental particles*, which seem to have some part in the make-up of atoms. Table XV is a summary of data regarding their masses and charges. Masses are in atomic weight units, charges in terms of the electronic charge. Very possibly, of course, the list is not yet complete.

TABLE XV. FUNDAMENTAL PARTICLES

	Charge	*Mass*
Electron	−1	0.00055
Positron	+1	0.00055
Proton	+1	1.0076
Neutron	0	1.0089
+Meson	+1	0.1 (approx.)
−Meson	−1	0.1 (approx.)

The Structure of the Nucleus

Rutherford's picture of the atom as a heavy, positive nucleus surrounded by scattered electrons poses two structural problems: (1) the arrangement of electrons in the electron cloud, and (2) the nature and arrangement of particles within the nucleus. The first problem, solved in great detail by use of spectroscopic data, will occupy us at some length later. But the nucleus, the atom's "inner citadel," is yielding its secrets only slowly to intensive study. Physicists can tell us with some confidence the kind and number of particles present in different nuclei, but details of arrangement are still unknown.

Of the six fundamental particles known to be emitted in various nuclear processes, probably only two are present in the nuclei of normal atoms. These are neutrons and protons, both of them particles with masses about the same as the hydrogen atom. Together these particles make up almost the entire mass of an atom; the electron is so much lighter that even in the heaviest atoms the combined weight of all the electrons in the cloud is only a small fraction of the weight of a single proton.

While the sum of neutrons plus protons determines the mass of a nucleus, the protons alone determine its positive charge. In other words *the atomic number of an element is the number of protons in the nuclei of its atoms, and the atomic weight is the number of protons plus neutrons.* For

example, the element fluorine has an atomic weight of 19 and an atomic number of 9; this means that a fluorine atom has a nucleus of 9 protons and 10 neutrons, surrounded by a cloud of 9 electrons.

If the nucleus ordinarily contains only protons and neutrons, we face the awkward question of how other particles can sometimes shoot out from it. Alpha particles, emitted when some radioactive atoms decompose, may be present in the original nuclei as combinations of protons and neutrons. Other particles—electrons emitted from some radioactive atoms, positrons and mesons emitted when some nuclei are disturbed by collisions—are probably formed from protons or neutrons at the instant of emission. Electrons are produced by splitting of neutrons into protons and electrons, positrons by splitting of protons into neutrons and positrons. How mesons come into existence is not known.

Isotopes

The idea that nuclei consist entirely of particles with a mass approximately equal to the mass of a hydrogen atom is reminiscent of a speculation debated among chemists over a century ago. Impressed with the fact that atomic weights of many elements are almost exact multiples of the weight of the hydrogen atom, some chemists suggested that perhaps all other atoms are built from hydrogen atoms, somehow packed very closely together. It was an attractive proposal, the sort that appeals to our instinctive notions of the simplicity which atomic structures ought to show. But it was quickly discarded as atomic weights were determined more accurately. Many elements do have atomic weights that are nearly whole numbers: thus carbon has a weight of 12.00, as if each atom contains 12 hydrogens; fluorine has a weight of 19.00, sodium 23.00, and nitrogen 14.01. But far too many elements have weights that are anything but whole numbers: for instance, chlorine 35.46, copper 63.57, and magnesium 24.32. To construct these from hydrogen atoms was out of the question.

It looks as if we face the same difficulty in building atoms out of protons and neutrons. For carbon or fluorine or sodium a combination of protons and neutrons fits the observed atomic weight splendidly. But what can we do with an element like chlorine, whose atoms apparently weigh about 35½ times as much as a hydrogen atom? No possible combination of protons and neutrons in the nucleus could give such an atomic weight. And we cannot evade the difficulty by supposing that the weight has been inaccurately determined; atomic weights have been checked and rechecked by too many careful observers to leave any doubt of their correctness.

A possible answer to this dilemma occurred to several scientists in the early 1900's: perhaps an element like chlorine is a mixture of atoms

of slightly different weights. This was a radical suggestion, completely at variance with Dalton's atomic theory. Its correctness was indicated by the results of atomic weight determinations on certain radioactive elements, but convincing proof was obtained only in 1919, through some ingenious experimental work by J. J. Thomson and another Cambridge physicist, F. W. Aston. The experiment consisted of sending positively charged atoms (positive ions) at high speeds through an electric and a magnetic field. In their passage through the fields heavy atoms were deflected somewhat less than light atoms, and the amount of the deflection provided a very accurate means of determining just how heavy different atoms are. Now when ions of chlorine were used, the ions after passing through the instrument did not all form a single beam corresponding to a mass of 35.46, but instead were divided into two beams, the stronger one corresponding to a mass of 35 and the weaker to a mass of 37. This means that natural chlorine must contain two kinds of atoms, which normally are mixed in a constant proportion giving the apparent atomic weight of 35.46.

Not only chlorine but many other elements are found to have atoms with slightly different weights when examined in Aston's instrument. These varieties of an element with different atomic weights are called *isotopes*. Extensive research has shown that fractional atomic weights always mean mixtures of isotopes; once the atoms are ionized and separated, each isotope proves to have an atomic weight very nearly a whole number. In other words, *all atoms have weights that are nearly integral multiples of the weight of the hydrogen atom.* This conclusion, of course, is consistent with the idea that all atomic nuclei are built wholly of neutrons and protons.

Atoms of two isotopes have the same number of protons in their nuclei but different numbers of neutrons. The light isotope of chlorine, for instance, has nuclei containing 17 protons and 18 neutrons, while the heavy isotope has nuclei containing 17 protons and 20 neutrons. The electron clouds, whose structure is determined by the total positive charge on the nucleus, are practically identical. In other words, the different isotopes of an element have *the same atomic number but different atomic weights.* These relations are illustrated for several elements in Table XVI.

The ordinary physical and chemical properties of an element are determined almost wholly by the electron clouds of its atoms and therefore by its atomic number. Since different isotopes have the same atomic numbers and almost identical electron structures, their properties are very nearly the same. The two isotopes of chlorine, for instance, have the same yellow color, the same suffocating odor, the same efficiency as poisons and bleaching agents, and the same readiness to combine with metals and hydrogen. Their densities, boiling points, freezing points,

and rates of diffusion depend somewhat on the masses of the atoms and thus are very slightly different. The extreme similarity in properties explains why isotopes are not appreciably separated in natural processes and why chemists failed to detect them before Aston's work.

Separation of isotopes in the laboratory is exceedingly difficult. Aston's apparatus accomplishes the separation, but only in minute quantities unless the operation is long continued. One may take advantage of the slight differences in such properties as vapor pressure, boiling

TABLE XVI. ATOMIC STRUCTURES

	Atomic number	*Atomic weight*	*Protons + neutrons in nucleus*	*Protons in nucleus*	*Neutrons in nucleus*	*Electrons in cloud*
Hydrogen	1	1	1	1	0	1
(3 isotopes)	1	2	2	1	1	1
	1	3*	3	1	2	1
Chlorine	17	35	35	17	18	17
(2 isotopes)	17	37	37	17	20	17
Lead	82	203	203	82	121	82
(6 isotopes)	82	204	204	82	122	82
	82	206	206	82	124	82
	82	207	207	82	125	82
	82	208	208	82	126	82
	82	210	210	82	128	82
Helium (no isotopes)	2	4	4	2	2	2
Alpha particle	2	4	4	2	2	0

* This isotope does not exist in ordinary hydrogen, but has been prepared artificially.

point, and rate of diffusion, but at best the process is long, tedious, and expensive. One of the major engineering accomplishments in the development of the atomic bomb was the construction of plants capable of separating the isotopes of uranium on an enormous scale.

The element hydrogen is an exception to the general rule that separation of isotopes is extraordinarily difficult. Natural hydrogen consists almost entirely of atoms containing simply a proton and an electron, but one in about every 4,000 atoms has a nucleus twice as heavy as the proton. Because the heavy isotope, called *deuterium*, is *relatively* so much heavier than the lighter isotope, the two can be separated by fairly simple processes of distillation or electrolysis. The separation can be performed so readily that deuterium oxide, or "heavy water," has become

a commonplace in scientific circles. The nucleus of deuterium (the *deuteron*), consisting of a neutron and a proton, is extensively used as a "bullet" in atomic-bombardment experiments.

The existence of isotopes makes necessary a reexamination of our ideas about chemical elements. In an earlier chapter (page 150) we defined an element as a substance not decomposable into other substances, and one of the principal assumptions of Dalton's atomic theory is that all atoms of an element are alike. Aston's work shows that both the definition and the assumption are wrong—although they remain serviceable in rough descriptions of ordinary chemical processes. On the basis of our present knowledge of atomic structure, we can frame a more rigorous definition of an element as *a substance whose atoms all have the same atomic number.*

Atomic Energy

We say that atomic weights of isotopes are "very nearly" whole numbers, or that the mass of a heavy atom is "very nearly" an integral multiple of the mass of a hydrogen atom. Why must we make these statements indefinite? Why shouldn't atomic weights be *exactly* whole numbers? If we examine the matter more closely, and add up the masses of the individual neutrons and protons (as given in Table XVI) in a complex nucleus, we find that the sum of the masses turns out very nearly equal to the mass of the nucleus, but *not quite.* If our ideas about the structure of the nucleus are correct, why should such discrepancies crop up?

For a concrete example, consider the two simplest elements, hydrogen (atomic weight 1) and helium (atomic weight 4). If we knew the proper technique, it should be possible to build a helium nucleus out of four hydrogen nuclei—in other words, to make an alpha particle by combining four protons. Since the alpha particle consists of two neutrons and two protons, two of the original protons would have to change to neutrons by losing positrons. This reaction has not been observed in the laboratory, but takes place on a large scale in the interior of the sun (page 616). If we write the reaction in symbols, and set down the exact weight of the particles concerned, we obtain

$$\underset{4 \times 1.00758}{4\text{H nuclei (protons)}} \rightarrow \underset{4.00280}{1\text{He nucleus (alpha particle)}} + \underset{2 \times 0.00055}{2 \text{ positrons}}$$

The total mass on the left side of the equation is 4.0303, on the right side 4.0039. What has become of the extra 0.0264 units of mass?

Einstein had suggested an answer to such questions in 1905, long before protons or neutrons or positrons had been heard of. His answer, based on calculations concerning experiments of an entirely different

sort, was that *the mass which disappears becomes energy.* For the amount of energy formed Einstein deduced the now famous equation

$$E = mc^2 \tag{29}$$

which means

Energy formed = mass lost × the square of the speed of light

The speed of light is an enormous number (3×10^{10} cm/sec), and its square is very much larger, so the equation implies that a huge amount of energy is produced when even a small quantity of matter disappears. In the hydrogen-helium reaction, for instance, the disappearance of 0.0264 units of mass means the liberation of $0.0264 \times (3 \times 10^{10})^2$ or about 2.3×10^{19} *ergs of energy for every* 4 *g. of helium produced.* This is equivalent to over 160,000 kw-hr of electrical energy, more than is used

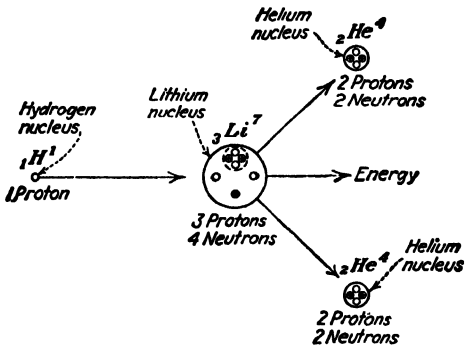


FIG. 161. *An example of transmutation by the bombardment of atomic nuclei with swiftly moving particles. A lithium nucleus struck by a proton breaks down into two nuclei of helium.*

by a thousand average homes in a year. If the reaction could be started and controlled, a small tank of hydrogen, renewed once a year, would be ample to supply a city's electrical needs.

Einstein's prediction cannot be directly checked for this reaction, but it has been verified again and again for other nuclear processes. If a lithium target is bombarded by protons, for example, a reaction takes place that is pictured in Fig. 161. The masses of the various particles are as follows:

$$\underset{1.00758}{\text{H nucleus (proton)}} + \underset{7.0165}{\text{Li nucleus}} \rightarrow \underset{2 \times 4.00280}{\text{2He nuclei (alpha particles)}}$$

The total mass on the left side is 8.0241, on the right 8.0056, so that 0.0158 units of mass disappear. The alpha particles produced have large kinetic energies, and it is this energy that represents the vanished mass. From cloud-chamber photographs the speed and therefore the kinetic energy of the alpha particles can be measured; the energy amounts to

about 13.6×10^{-6} erg for each particle. The 0.0185 units of mass, according to Einstein's equation, should be equivalent to $0.0185 \times (3 \times 10^{10})^2$ or 1.67×10^{19} ergs for every 8 g. of helium. Since 8 g. of helium contain 12.12×10^{23} atoms (from the kinetic theory), the energy produced per atom should be 1.67×10^{19} divided by 12.12×10^{23}, or 13.7×10^{-6} erg. This figure is in good agreement with the measured energy, 13.6×10^{-6} erg, showing the correctness of Eq. (29). Sometimes, as in this reaction, the energy formed when mass disappears takes the form of kinetic energy of moving particles; sometimes it appears as gamma radiation, or as both kinetic energy and gamma radiation. In whatever form the energy appears, its total amount is always found to be related to the loss in mass by Einstein's equation.

These relationships mean that the laws of conservation of mass and energy, so long regarded as basic principles of physics and chemistry, are not quite correct. For the two laws we should substitute a single one, that *the sum of mass plus energy remains constant.* Matter may disappear and energy may be created, but always the amount of one which is formed is equivalent to the amount of the other which vanishes. Even in ordinary processes like the burning of a match some energy is produced at the expense of matter, but the amount is so exceedingly small that our instruments cannot detect it. For such processes the old conservation laws are sufficiently accurate, but for reactions of atomic nuclei we must abandon these laws and concern ourselves with energy produced at the expense of matter.

If energy surges out of nuclear reactions in such enormous quantities, why didn't physicists proceed to harness atomic energy as soon as the first bombardment process was discovered? Why was there a lapse of a quarter of a century between Rutherford's pioneer experiment and the atomic bomb? The explanation lies in the *low efficiency* of the bombardment experiments. In all these experiments tiny "bullets" are sent in the general direction of the "target" atoms; a very few bullets score direct hits and set free momentary bursts of energy, but the great majority pass through the target without encountering atomic nuclei. The reactions we have been considering are collisions between *single* atoms and *single* particles, not between large amounts of matter. Although each individual collision releases abundant energy, so many particles pass through the target without collision that the energy which must be supplied to make the particles move is far greater than the total energy produced. It is as if someone had scattered small sticks of dynamite over a hillside, and you were trying to explode them by firing a machine gun at the hill from a long distance away. Once in a while one of your bullets would hit a stick and you would see a small explosion, but most of the bullets would plow harmlessly into the ground. The dynamite has an abundance of stored

energy, but with this technique of releasing it you must supply more energy in your gun than you can obtain from the dynamite.

Atomic Bombs

Uranium Fission. In 1939, just before the outbreak of war, came a startling announcement from two German scientists, Otto Hahn and Lise Meitner: Atoms of the heavy metal uranium, bombarded by neutrons, split into fragments with roughly half the mass of the original atoms. This reaction was different from all other known nuclear processes in at least three important respects: (1) A heavy atom was literally split in two, whereas in other bombardment experiments only small particles were knocked out of the target atoms; (2) the energy set free was enormous even by nuclear standards, some hundreds of times greater than that produced by any other nuclear process; (3) among the products of reaction were neutrons, the same particles that started the reaction. The third characteristic suggested an exciting possibility: Under suitable conditions the neutrons produced from one uranium atom might cause the splitting of others, these in turn producing neutrons which would split still other atoms, the reaction thus spreading through the entire mass of uranium spontaneously—much as a forest fire spreads when the heat from one burning tree ignites others, these in burning ignite still others, and so on (Fig. 162). For a more exact analogy, suppose that the sticks of dynamite which you were trying to explode with a machine gun each contained a number of small pebbles: If you succeeded in hitting one stick, the pebbles would be scattered about and explode others, which in turn would shower their pebbles on still others, until most of the dynamite was consumed. The possibility of setting up such a *chain reaction* made it probable that this splitting, or *fission*, of uranium atoms would supply the long-sought key to the practical use of atomic energy.

The breath-taking possiblities of uranium fission were clear to nuclear physicists the world over when war was declared. In peacetime, scientists from all countries would have cooperated in an effort to turn this possible source of prodigious energy to useful purposes. But with the outbreak of war international cooperation in science as in every other field ceased abruptly. Atomic power loomed suddenly not as a source of industrial energy but as a possible means of destruction more powerful than any ever used before. There began a fantastic race among scientists of the opposing sides, each group trying desperately to solve the problems of a uranium bomb before the other, each realizing that a successful solution might well bring sudden victory to the side which discovered it. As it happened, the end of the war was in sight by the time scientists of the Allied Nations performed their spectacular experiment in New Mexico, but the two bombs used in Japan helped bring victory more

quickly and gave the world an appalling demonstration of what any future war will be like.

Although Hahn and Meitner's discovery seemed to show the possibility of utilizing atomic energy, applying it to actual production of an atomic bomb was an incredibly difficult undertaking. Industrial processes of completely new kinds had to be designed, tried out, and set up for large-scale operation. Even under the terrible urgency of wartime it

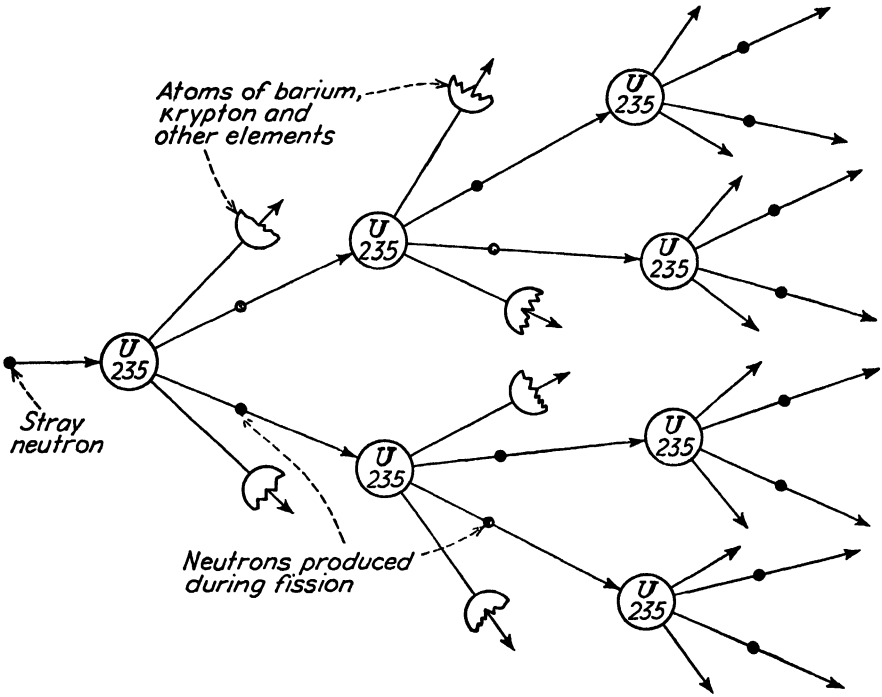


FIG. 162. *Diagram of a chain reaction as it might occur in pure U_{235}. Each collision of a neutron with a U_{235} nucleus produces (1) nuclei of barium, krypton and other elements with roughly half the atomic weight of uranium, (2) one to three neutrons, and (3) energy in the form of gamma radiation. The neutrons continue the chain by colliding with other U_{235} nuclei.*

required more than four years of concerted effort by scientists, engineers, and industrialists, together with an expenditure of many millions of dollars. The same development under normal conditions would very likely have extended over a decade or longer.

Uranium 235. One major difficulty lay in the fact that only a particular kind of uranium undergoes the fission reaction. Ordinary uranium consists of three isotopes: Atoms with a weight 238 are by far the most abundant, about one atom in 140 has a weight 235, and about one in 16,000 a weight 234. Only the atoms of U_{235} are fissionable by neutrons.

Despite the small amount of this isotope, a chain reaction might still take place in natural uranium if a neutron could be depended on to bounce away from collisions with other atoms until it met one of the fissionable variety; unfortunately most neutrons do not bounce but are absorbed into U_{238} nuclei, which means that in natural uranium the chain reaction is quickly interrupted. Evidently one method of helping the chain to take place would be to separate the two isotopes. Separation of isotopes on an industrial scale had never before been attempted (except the hydrogen isotopes, page 294), but methods were devised and plants were set up to produce U_{235} in large quantities—a truly prodigious feat of engineering.

Plutonium. The absorption of neutrons by U_{238}, although it damps out the chain reaction in natural uranium, from another angle was a great help in developing atomic bombs. An atom of U_{238} which has absorbed a neutron becomes another isotope of uranium, U_{239}, which is unstable and quickly breaks down by emission of a beta particle into an atom of a different element, neptunium. This atom is also radioactive, its nucleus presently emitting another beta particle and thereby changing to an atom of plutonium. These changes may be summarized in symbols:

$$n + U_{238} \rightarrow U_{239} \text{ (atomic number 92)}$$
$$U_{239} \rightarrow \text{electron} + Np_{239} \text{ (atomic number 93)}$$
$$Np_{239} \rightarrow \text{electron} + Pu_{239} \text{ (atomic number 94)}$$

Plutonium, like U_{235}, is capable of fission when struck by neutrons. So if plutonium could be produced in quantity, it also could be used in a bomb. Now plutonium is present only in the minutest traces in natural uranium, but it seemed possible the amount might be increased by neutron bombardment. If an increase could be brought about, the problem of separating Pu from the remaining U_{238} should be much simpler than the problem of separating U_{235}, since plutonium is a different element and thus should be separable by ordinary chemical means. The possibility of obtaining plutonium seemed great enough so that much effort was turned in this direction while the plants for separating isotopes were being set up.

Uranium-graphite Piles. The production of plutonium was brought about not by bombardment with neutrons from an external source, but by a clever device for making the U_{235} chain reaction take place slowly in natural uranium. It happens that the neutrons produced in U_{235} fission are high-speed neutrons, which are readily absorbed by U_{238} atoms but which are not as effective as slow neutrons in splitting atoms of U_{235}. Fast neutrons can be slowed down if they are passed through material consisting of light atoms that will not absorb them, like carbon atoms; their energy is gradually lost as they collide with one light nucleus after another. Now if small chunks of uranium could be scattered through a

large block of graphite (pure carbon), some of the fast neutrons originating by fission in one chunk should escape from the chunk, be slowed as they pass through the graphite, and thus be ready to produce fission in the U_{235} atoms of another chunk (Fig. 163). This arrangement permits the chain reaction to take place in spite of the presence of U_{238}. Of course, many neutrons are still absorbed by U_{238} atoms, and these atoms quickly change to plutonium. Such a uranium-graphite *pile*, first set up

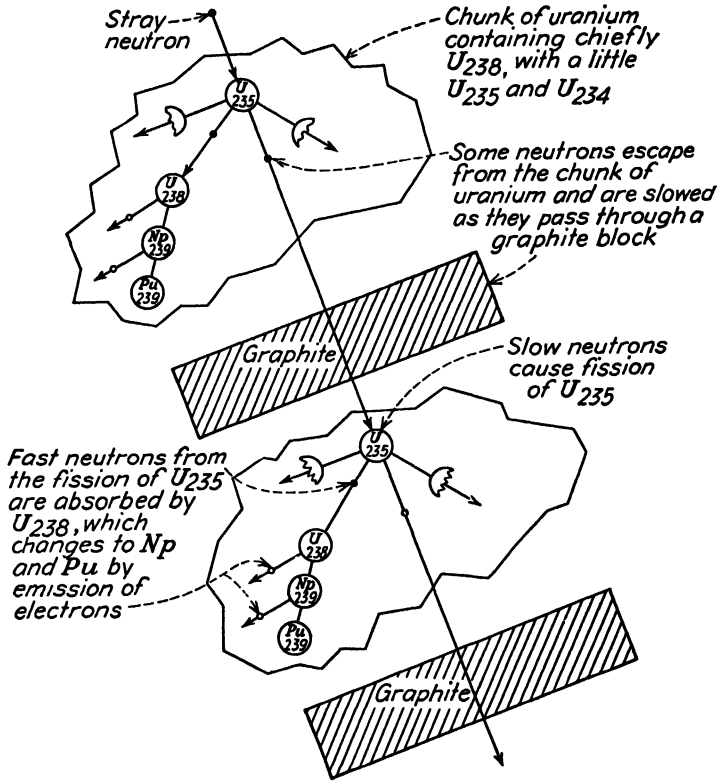


FIG. 163. *Diagram to show the principle of a uranium-graphite pile. Energy is produced at each step of the process.*

at the University of Chicago in 1942, showed by its successful operation that a controlled chain reaction was possible.

The chain reaction in natural uranium diluted with graphite liberates great quantities of energy, chiefly in the form of high-frequency radiation. The reaction is not rapid enough for an effective atomic bomb, but it has much promise as a possible peacetime source of energy. In the piles operated during the war, however, energy production was incidental; the importance of the piles lay in the gradual enrichment of their uranium chunks in plutonium.

Construction of the Bomb. By 1945, enough of both U_{235} and plutonium were available for the construction of bombs, and bombs were made from both materials. But difficulties were by no means ended with successful production of the fissionable elements in pure form. These strange new explosives could not be simply cut up into sticks and exploded by percussion caps, like dynamite. If a large enough mass of one of the ele-

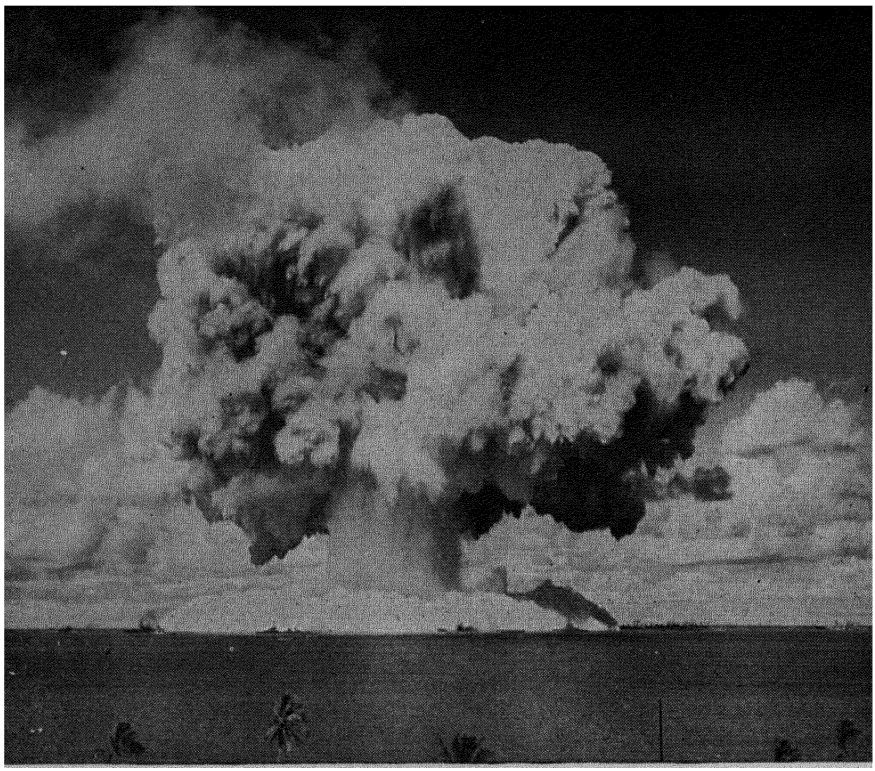


FIG. 164. *Explosion of an atomic bomb at Bikini atoll, July 1946. This bomb was exploded beneath the surface of a quiet coral lagoon. The "stem" of the explosion cloud is a column of water nearly half a mile wide. Ships outlined against the huge wave at the base of the column give some idea of the size of the explosion. (Official U.S. Navy photograph.)*

ments were heaped together, it would explode by itself; there was no way to keep it from exploding, since always in the surrounding air there would be enough stray neutrons to start the chain reaction. On the other hand a small chunk could not be exploded at all, since so many neutrons would escape from it that the chain reaction could not persist. An atomic bomb, then, must consist of small chunks of fissionable material sufficiently separated so that they will not explode, together with a device for bringing them together at the proper moment. The mechanical problem is a

tricky one, for if the separate chunks are not brought together very rapidly the bomb will start to explode prematurely and will fly apart before it can do appreciable damage. It was only when this final problem was solved that the scientists in New Mexico could try their great experiment.

Future Possibilities. Atomic energy so far has brought us only the most frightful instrument of destruction ever devised (Fig. 164). Whether it can be turned to more useful ends is a problem of the future. Certainly uranium-graphite piles are capable of generating energy in great quantity, energy which may perhaps be usable in producing electricity or in turning the wheels of industry. If uranium enriched with plutonium or with U_{235} is used in place of natural uranium, the energy obtainable from a pile is greatly increased. Even the most efficient pile, however, is hardly a suitable power source for automobiles or airplanes, because of its enormous size and weight. The pile itself may be made fairly small if enriched uranium is used, but it must be surrounded by several tons of shielding material to absorb its powerful and deadly radiations.

Prediction of future developments in this field of nuclear physics is of course foolhardy. The atomic bomb has opened our eyes to its practical possibilities, but exploration into possible sources of atomic power has only just begun. Other elements than uranium, other reactions besides fission produced by neutrons, may prove more suitable as energy sources. And theoretical possibilities seem as limitless as practical ones; in no other field are chances so bright for further advances in fundamental knowledge, by men who look on the atom less as a possible source of power than as a fascinating riddle still only partly solved.

Questions

1. What discoveries made necessary a revision in each of the following ideas?
   a. Atoms are indivisible.
   b. All atoms of a single element are identical.
   c. Matter can be neither created nor destroyed.
2. To what conclusions regarding atomic structure did each of the following observations lead?
   a. Alpha-particle tracks through gases are nearly always straight lines.
   b. In some cloud-chamber photographs of alpha-particle collisions, a thin fog track starts from the site of a collision, at an angle to the original alpha-particle track.
3. Radium decomposes spontaneously into radon and helium. Why is radium considered an element rather than a compound of radon and helium?
4. Given only the atomic number of an element, which of the following quantities could you find?
   a. The number of electrons in the electron cloud.
   b. The atomic weight.
   c. The number of isotopes.
   d. The total positive charge on the nucleus.

e. The number of neutrons in the nucleus.
f. The number of protons in the nucleus.

5. The most abundant isotope of radium has an atomic weight of 226, an atomic number of 88. When the nucleus of a radium atom disintegrates with the loss of an alpha particle, what happens to its atomic weight and its atomic number?
6. Why are isotopes so difficult to separate?
7. An atom of radium E disintegrates to form polonium by losing a beta particle from its nucleus. On the assumption that the beta particle is produced by the breakdown of a neutron into a proton and an electron, the electron being then emitted, what effect does emission of the beta particle have on (*a*) the atomic weight and (*b*) the atomic number of the radium E atom? Would polonium be to the left or right of radium E in the periodic table?
8. In what respects is the fission of uranium atoms different from other nuclear processes in atom-bombardment experiments? Of the following substances, which are known to undergo fission when bombarded by neutrons?

Cl, U_{238}, U_{235}, Ra, N, Li, Pu, He

9. What prevents natural uranium from undergoing a chain reaction leading to the production of plutonium? What purpose is served by the graphite in a uranium-graphite pile?
10. The following statements were thought to be true in the nineteenth century. Which of them are now known to be inexact? For each of those known to be inexact, indicate (*a*) what experiments proved the statement wrong, and (*b*) how the statement should be modified to be correct according to modern views.
    a. Atoms are indivisible and indestructible.
    b. Equal volumes of all gases under the same conditions of temperature and pressure contain the same number of molecules.
    c. Energy can be neither created nor destroyed.
    d. Acceleration is directly proportional to force and inversely proportional to mass.
    e. An element is a substance which cannot be decomposed into other substances.
    f. All atoms of any one element are exactly alike.

CHAPTER XXII

Radiation

THE chief problems regarding atomic structure which we have not yet discussed relate to the electron cloud. How are the electrons of a complex atom arranged about the nucleus? Are they in motion or stationary? Why do some atoms lose electrons to become positive ions, while others gain electrons to become negative ions? These and similar questions have been answered by an intensive study of electromagnetic radiation.

The various kinds of electromagnetic radiation which we have touched upon are summarized in Table XVII. The frequencies and wave

TABLE XVII. KINDS OF ELECTROMAGNETIC RADIATION

	Wave length, cm	*Frequency, vibrations per second*
Waves used in wireless telegraphy	2×10^6 to 60,000	15,000 to 500,000
Ordinary radio waves	60,000 to 20,000	500,000 to 1.5×10^6
Waves used in short-wave radio	20,000 to 1,000	1.5×10^6 to 3×10^7
High-frequency waves used in radar, walkie-talkies, etc.	1,000 to 1	3×10^7 to 3×10^{10}
Unnamed waves	1 to 0.02	3×10^{10} to 1.5×10^{12}
Infrared ("radiant heat") waves	0.02 to 0.00007	1.5×10^{12} to 4.2×10^{14}
Visible light	0.00007 to 0.000035	4.2×10^{14} to 8.6×10^{14}
Ultraviolet waves	0.000035 to 4.5×10^{-7}	8.6×10^{14} to 7×10^{16}
X rays	4.5×10^{-7} to 1×10^{-9}	7×10^{16} to 3×10^{19}
Gamma rays	1×10^{-9} to 5.6×10^{-11}	3×10^{19} to 5.4×10^{20}

lengths shown in the table are purely arbitrary boundaries; actually there is a gradual transition from one type of radiation to the next. Note the enormous range of wave lengths, from 2 million centimeters down to 5 hundred-billionths of a centimeter.

The long waves at the top of the table, those used in various kinds of radio transmission, are produced by the rapid back-and-forth motion of electrons along a wire, set up and controlled by the electrical apparatus of a broadcasting station. There is no real upper limit to the length of these waves, but little use is made of wave lengths greater than about 20 kilometers (km) (2×10^6 cm). Waves used in ordinary commercial radio lie between about 600 and 200 m., lengths which correspond to frequencies of 500,000 to 1,500,000 vibrations per second (on the ordinary radio dial these frequencies are indicated as 500 to 1,500 "kilocycles"—each kilocycle being 1,000 vibrations per second). The short-wave radio of amateur transmitters, police departments, and some government agencies uses wave lengths down to about 1 m. Very short waves, with lengths between about 10 m. and 1 cm, proved extraordinarily useful during the war in the development of short-range radio telephones ("walkie-talkies"), radar, proximity fuses, and many other devices; the usefulness of these waves depends on the ease with which they can be "beamed" in particular directions and on their ability to be reflected from opaque objects like ships and airplanes.

For all these waves the connection with the oscillating electric charges that produce them is fairly clear, but for the shorter waves at the bottom of Table XVII the connection with electricity is far less obvious. We perceive these short waves as heat on our skin, as light stimulating our optic nerves, as invisible radiation which affects a photographic plate—properties not ordinarily associated with electrical phenomena. Nevertheless, Maxwell's equations indicate that these high-frequency radiations also should be produced by oscillations of electric charges. The nature of the charges remained a mystery until Thomson, Rutherford, and their colleagues showed that the atom itself contains electrically charged particles.

Investigations into the connection between radiation and subatomic particles have not only clarified the structure of the electron cloud but have led to some perplexing revelations about the nature of radiation and of matter itself.

Spectra

We have mentioned spectra before (page 266)—the beautiful rainbow bands produced when light from a glowing metal passes through a spectroscope, the sharp bright lines from the light of a heated gas. Spectra of either sort are simply light spread out into its different wave lengths. The continuous colored band from red to violet means that all visible wave lengths are present; the discontinuous, "bright-line" spectrum indicates that only a few wave lengths are represented.

The spectrum produced by a light source alone, whether it is a bright-line or a continuous spectrum, is called an *emission spectrum;*

that is, it shows the wave lengths present in light *emitted* from the source (Fig. 145*a*, *c*, *d*, *e*, page 266). Spectra of a different sort, *absorption spectra*, are produced when light from an incandescent solid or liquid passes through a cool gas before entering the spectroscope. The light source alone would give a continuous spectrum, but the gas absorbs certain wave lengths out of the light which passes through it. Hence the continuous spectrum appears to be crossed by dark lines, each line representing one of the wave lengths absorbed by the gas (Fig. 145*b*, page 266). If the bright-line spectrum of an incandescent gas is compared with the absorption spectrum of the same gas, the dark lines in the latter are found to correspond in wave length to bright lines in the emission spectrum. Thus a cool gas absorbs wave lengths of light which it is capable of emitting when heated to incandescence.

The line spectrum of each element (either its bright-line spectrum or its absorption spectrum) contains lines of certain wave lengths which are characteristic of that element (see Plate of Spectra, Frontispiece). These lines can be recognized in complex spectra, so that the element can be identified either in a light source or in an absorbing gas. This fact makes the spectroscope a valuable tool in chemical and metallurgical analysis; even minute traces of most elements are readily identified by the lines in their spectra.

The number, intensity, and position of the lines in the spectrum of an element vary somewhat with temperature, with pressure, with the presence of electric and magnetic fields, and with the method by which the element is made incandescent. Thus an expert can tell by an examination of spectra not only what elements are present in a light source but much about their physical condition. An astronomer, for example, can deduce from the spectrum of a star the composition of its atmosphere, its temperature, whether it is approaching or receding from the earth, and what substances in its atmosphere are ionized.

Spectra are not limited to the narrow band of wave lengths which we call visible light. X rays, ultraviolet radiation (wave lengths between those of X rays and visible light), and infrared radiation (wave lengths longer than red light) can likewise be spread out into bands and sharp lines. For much of this invisible region some other device than a prism must be used for separating the different wave lengths; to record the invisible spectra, specially sensitized photographic film or (in the long infrared) thermocouples are needed. But details of the necessary apparatus are of interest only to the specialist. Our concern here is with the significant facts that electromagnetic radiation from a given source is made up of certain definite wave lengths, whether they be in the infrared, visible, ultraviolet, or X-ray region, and that a study of these wave lengths reveals a great deal about the source.

One curious property of line spectra is the occurrence of lines i series, all lines in each series having frequencies related by a simp formula. A series in the spectrum of sodium is shown in Fig. 165; th regular decrease in the distance between lines is evident. These simp frequency relations were recognized and line series were plotted fc many spectra long before any reasonable explanation was possibl When the foundations of the modern picture of the atom had been lai these carefully studied line series proved to be the needed clue for worl ing out the details of atomic architecture.

Here once more we find the familiar pattern of scientific reasonin from observations to simple mathematical relations between the obse vations, then from these relations to an inclusive theory. As Tych

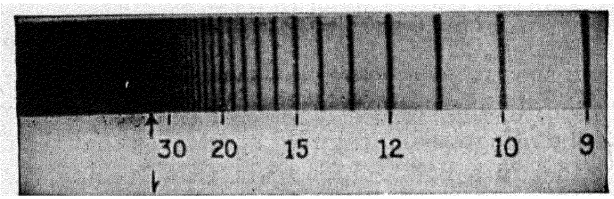


FIG. 165. *A series of lines in the absorption spectrum of sodium.* (*From Elements Physics by Smith.*)

Brahe's observations of the planets were correlated by Kepler's simp formulas, and as these paved the way for Newton's idea of univers gravitation, so in the past half century have observations of spectra le to simple relations among the frequencies of spectral lines and these i turn to modern ideas of atomic structure.

Electron Orbits

Rutherford's demonstration that an atom consists of negative pa ticles at great distances from a positive core posed immediately th question, what keeps the electrons from falling into the nucleus? possible answer was motion in elliptical orbits. Just as the moon i prevented from falling to the earth by the centrifugal force of its motio so perhaps the rapid motion of each electron around the nucleus counte balances the electrical attraction between them.

But this idea presents difficulties. According to Maxwell's theory, a electron so moving should emit electromagnetic radiation continuously Giving out radiation means that it should continuously lose energy, s that its orbit should become steadily smaller, the wave length of it radiation should grow longer, and at length it must collide with th nucleus. Now electrons simply do not act that way, Maxwell's theor or no Maxwell's theory. Atoms under ordinary conditions do not emi radiation; when they are made to emit radiation, wave lengths are ofte

limited to definite values; and the electron quite evidently does not end its career by succumbing to nuclear attraction.

First to see a way out was a Dane, Niels Bohr, who was bold enough to suggest (in 1913) that Maxwell's classical formulas needed modification for systems as small as an atom. Bohr assumed first that electrons can travel around their nucleus without gaining or losing energy, that is, without emitting or absorbing radiation. He assumed next that each electron could use several different orbits, its energy in some orbits being greater than in others. Emission of radiation he imagined to take place when an electron jumps from one orbit to another of smaller energy, the energy it loses escaping as radiation. Thus emission of radiation would be an unusual event in an atom's history, taking place only if an atom is so stimulated that one of its electrons leaps from a high-energy orbit to a low-energy orbit.

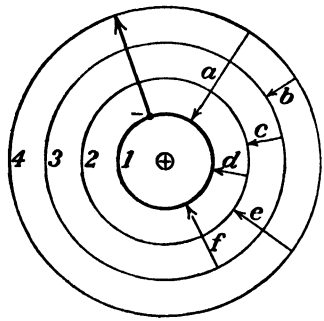


FIG. 166. *Diagram of the hydrogen atom, according to Bohr's hypothesis. Orbit 1 is the electron's normal path, orbits 2, 3, 4 are possible orbits when the atom is excited by heat, electricity, or radiation. If absorption of radiation puts the electron in orbit 4, its possible jumps (with emission of radiation) are shown by the arrows a, b, c, d, e, f.*

For a simple illustration of Bohr's hypothesis, consider the crude diagram of a hydrogen atom in Fig. 166. Here the orbit pursued by the solitary electron under ordinary conditions is represented by the heavy line nearest the nucleus. The lighter lines are other possible orbits, where the electron would possess greater energy than in its normal orbit, since it would be farther from the nucleus (much as a stone at the top of a building has more potential energy than on the ground, since it is farther from the earth's center). Suppose that the electron at first is in the normal orbit. If the atom is supplied with energy—by strong heating, by an electrical discharge, or by powerful radiation—the electron may be induced to jump to a larger orbit. This jump means that the atom has absorbed some of the energy which is being supplied to it. It retains the added energy as long as it remains in the "excited state," that is, as long as its electron stays in the larger orbit. But the excited state is unstable, and in a small fraction of a second the electron spontaneously jumps back to its original orbit (or to another smaller orbit). In this second jump radiation is emitted, the energy of the radiation representing the difference in energy between the excited and normal states of the atom. The amount of energy given out evidently depends on which of the outer orbits the electron followed in the excited state.

To gain a rough picture of the emission of radiation during an electron jump, we may imagine the electron vibrating for a moment as it subsides into the smaller orbit—much as a rubber ball will bounce repeatedly when it loses potential energy by dropping from a table to the floor. The vibrating electron sends out electromagnetic waves, just as the vibrating electrons in a radio antenna send out waves. The electron in the atom vibrates much more rapidly than those in the antenna, so that its waves have a far greater frequency and a smaller wave length than radio waves.

The energy, and likewise the frequency, of the radiation emitted from a hydrogen atom are determined, according to Bohr's hypothesis, by the particular jump which its electron makes. If the electron jumps from orbit 4 to orbit 1 (Fig. 166), the energy (and frequency) of the radiation will be greater than if it jumps from 3 to 1 or 2 to 1. Starting from orbit 4, it may return to 1 not by a single leap, but by stopping at 2 and 3 on the way; corresponding to these jumps will be radiations with energies (and frequencies) determined by the energy differences between 4 and 3, 3 and 2, 2 and 1. Each of these several jumps gives radiation of a single frequency, and will therefore be represented in the hydrogen spectrum by a single bright line. Further, the frequencies of the different lines will be simply related one to the other, since they correspond to different possible jumps in the same set of orbits. And the relations between the lines which Bohr *predicted* by this mechanism matched closely the *observed* relations between lines in the spectrum of hydrogen.

Bohr's model of the hydrogen atom presented several difficulties. It did not explain why the electron's orbital motion failed to produce radiation. It left unexplained such questions as why an electron preferred a small orbit to a large one, why in its jumps it chose one small orbit in preference to another, why it could not move in other orbits than the discrete ones required by the hypothesis. The model implied a radical change in orthodox ideas about the nature of energy, a point to which we shall return in a later section. Nevertheless, Bohr's model found instant favor in scientific circles because of the uncanny accuracy with which it predicted the spectral line series of hydrogen.

Orbits in Complex Atoms

In complex atoms Bohr assigned several possible orbits to each electron, radiation from such an atom being produced by a jump of some one electron from an excited orbit to a smaller orbit. When a complex atom is in its normal state (*i.e.*, when all its electrons are in their normal, stable orbits), its electrons were assumed to be in groups, those of each group having orbits of the same size. In effect, this grouping divides the electron cloud of an atom into layers, or "shells," of electrons, those

in each layer keeping the same average distance from the nucleus. The atom of chlorine, for example, contains seventeen electrons distributed in three layers: an inner layer of two electrons in small orbits, a layer of eight electrons in larger orbits, and a layer of seven electrons in still larger orbits.

In general, bright lines in the visible part of the spectrum given by an element with complex atoms were found to correspond to jumps of the outermost electrons. Lines in the ultraviolet part of the spectrum corresponded to jumps of these electrons or those just below them. X-ray lines represented the more energetic leaps of electrons to orbits deep within the atoms. The continuous spectra of incandescent liquids and solids were at least qualitatively explained by interference of closely packed atoms with each other's electron orbits. Most infrared spectra and some visible spectra were accounted for by vibrations of atoms and groups of atoms within molecules, rather than by electron jumps.

Thus Bohr's ingenious hypothesis provided a simple mechanical explanation for the frequencies of spectral lines. So complete was the explanation that each electron in a complex atom could be assigned to a definite place in the electron cloud, simply from a study of lines in the atom's spectrum.

Further research has shown that Bohr's hypothesis needs extensive modification. For the hydrogen spectrum, line frequencies are accurately predicted on the basis of the assumed electron orbits. But frequencies of lines in more complex spectra agree only roughly with Bohr's predictions, and suggested changes in the assumed orbits have not greatly improved the agreement. Physicists now regard the Bohr atom as a convenient but highly inexact model, misleading because it gives a false impression of the definiteness with which electrons in atoms can be located. To this question of the nature of electrons and electron orbits we shall return later.

The fact that electrons are no longer considered as definite particles racing around precisely located orbits detracts little from the greatness of Bohr's achievement. He was the pioneer in attempts to interpret the atom mechanically, and for many purposes his oversimplified picture of the atom is still useful.

The Quantum Theory

Bohr's hypothesis that light energy originates in electron jumps involved a new idea about the nature of energy.

Imagine a tube of hydrogen excited so that it emits radiation corresponding, say, to an electron jump from orbit 2 to orbit 1 (Fig. 166). The radiation we observe will be made up of contributions from all the myriads of atoms in which the jump is taking place, each atom giving out a certain definite quantity of radiation energy. Let us call this

amount of energy E; it is simply the difference in energy between the excited and normal state of a single atom. Now if conditions could be arranged so that only a single atom were excited, its electron leap would emit exactly E units of energy—no more and no less. It would be impossible for a hydrogen atom to give out light of this frequency associated with any fraction of E units of energy, since the electron must jump all the way to orbit 1. In other words, the wave motion is made up of myriads of separate "packets" of radiation, one from each atom, each possessing an amount of energy E. Usually we regard wave motion as a *continuous* transfer of energy—*i.e.*, a flow of energy without parts, a flow which we can interrupt at any point and from which we can use as large or as small an amount as we choose. But Bohr's hypothesis requires that we consider the flow *discontinuous*—made up of individual, discrete parts, so that we can interrupt the flow only after using an exact whole number of the parts.

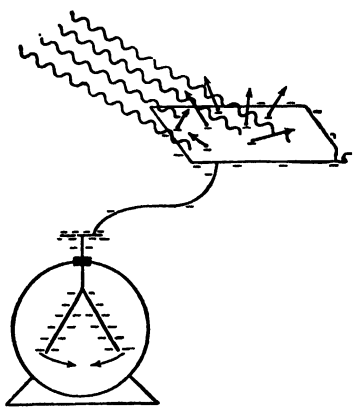


FIG. 167. *The photoelectric effect. Ultraviolet light falling on a zinc plate causes electrons to be emitted.*

This radical suggestion that radiant energy, like matter, consists of tiny "pieces" was not original with Bohr. He borrowed the idea from a German theoretical physicist, Max Planck, who had seen the necessity for discontinuity in radiation shortly after the turn of the century. A detailed discussion of Planck's work would take us too far afield, but we shall consider briefly one simple experiment which demonstrates convincingly that Planck's pieces, or ***photons,*** of radiation are real.

The experiment takes us back to electrostatics. Suppose that a zinc plate is attached to an electroscope (Fig. 167) and that plate and instrument are given a negative charge. Left to themselves, provided the air is dry, they will maintain the charge for some time. But now illuminate the plate with ultraviolet light, say from a mercury-vapor lamp: at once the leaves start to collapse. Our first conclusion might be that the light is simply ionizing adjacent air, after the fashion of X rays or gamma rays (page 273), and that the charge is leaking off due to movement of the ions. We can test the conclusion easily by starting our experiment again, but giving the plate a positive charge. This time the electroscope leaves remain apart. Here we have something different from ionization, for ionized air should be as effective in removing a positive charge as a negative charge.

The zinc loses its negative charge by a direct loss of electrons: the

light waves absorbed by the metal set some of its electrons in such violent motion that they leap clear of the plate and travel rapidly away. This phenomenon, called the *photoelectric effect*, can be demonstrated for a number of metals. It is not necessary that the metal be charged previously, but emission of electrons from an uncharged plate can be detected only with more delicate apparatus. Electrons in very active metals, like potassium and cesium, respond to visible light as well as to ultraviolet.

By suitable arrangements the liberated electrons may be attracted to a positive electrode, setting up a minute electric current. The current is too small to be useful by itself, but it may easily be made to control larger currents. Various ingenious devices are on the market (photoelectric cells, or "electric eyes") which make it possible to use a beam of light to operate switches, open doors, or count people as they pass a ticket office. Television also involves an application of the photoelectric effect.

"Electric eyes" are commonplace today, but to physicists of forty years ago the photoelectric effect was something new and strange. Here was light, an intangible sort of wave motion, evidently able to concentrate its energy on a single electron and kick it free from the positive charge of its parent atom! Still more surprising were measurements of the speed and number of electrons emitted by various kinds of light. With light of any one frequency, the speed of the electrons shooting out from the metal is not changed, however intense or feeble the light is made. Increasing intensity serves only to dislodge more electrons, not to make them move faster. Seemingly each electron takes only a *certain definite amount* of energy from the light, however much light is available. Again we find that light acts as if it is made up of packets, or units, of energy, all the energy in a single packet going into the motion of one electron.

Now if the *frequency* of the light is changed, the speed of the emitted electrons responds at once. Increase the frequency (*i.e.*, use light of shorter wave length) and the electrons jump out more energetically; decrease the frequency, and the electrons move more sluggishly. Decrease the frequency further, and the electrons refuse to leave the metal, however intense the light. Apparently the "packets" of light waves contain more energy if their frequency is high, less energy if their frequency is low; if the frequency is low enough, the energy in each packet is not sufficient to dislodge an electron.

These relations were pointed out by Einstein in 1905. In agreement with Planck, he showed that the energy of a packet, or photon, of light was directly proportional to the frequency of the light. In other words, the energy in a photon may be expressed

$$E = hn \tag{30}$$

where n is the frequency of the light and h is a proportionality constant (often called "Planck's constant"). Like the constant of gravitation, h is a universal constant: its value, so far as we know, does not change anywhere in the universe, or under any known changes of temperature, pressure, or materials. Its value, first determined from measurements of the photoelectric effect, is 6.60×10^{-27} erg-sec. Thus each photon of red light (see Table XVII) has $6.60 \times 10^{-27} \times 4.2 \times 10^{14}$ or 2.77×10^{-12} erg of energy, while the larger photons of violet light contain $6.60 \times 10^{-27} \times 8.6 \times 10^{14}$ or 5.68×10^{-12} erg apiece.

The assumption that energy is made up of photons (also called *quanta*), an assumption generally called the *quantum theory*, has been justified by many investigations since Einstein's work on the photoelectric effect and Bohr's elucidation of the hydrogen spectrum. To physicists of a generation ago the theory was quite as severe a shock as the discovery of the electron, for it meant a rightabout-face in their ideas of radiation. For nearly a century, since interference experiments had discredited Newton's "corpuscular theory," physicists had believed that light was continuous wave motion. Yet here was convincing evidence that light consists of streams of photons, of tiny "hn's" of energy. It seemed almost a reversion to Newton's outmoded corpuscles.

The photons of modern physics, however, bear little resemblance to Newton's corpuscles—or, for that matter, to any other creation of the human mind. They are not easily broken up, they carry energy sufficiently concentrated to move electrons or atoms: in these respects they act like material particles. Yet somehow associated with them must be waves, for an explanation of interference effects without waves is quite as impossible today as it was a century ago. Light seems to have a Jekyll-and-Hyde personality, in one experiment behaving like a hail of tiny bullets, in the next like a series of waves. Imagine a photon any way you choose—a small bunch of waves, a nearly weightless particle which can turn into a wave if suitably provoked, a particle guided by a train of waves; modern physics can give no clearer picture except in the language of mathematics.

Photons of radiations with wave lengths longer than those of visible light represent exceedingly small amounts of energy. Individual effects of such photons are hard to detect, so that the wave aspect of these radiations is the only one commonly observed. At the high-frequency end of the scale the energy associated with each photon is much larger, and the particle aspect becomes more important. Photons of X rays and gamma rays, for instance, will not only eject electrons from atoms, but sometimes bounce off from an electron at an angle as radiation of lower energy (lower frequency)—exactly as a billiard ball might strike another a glancing blow and lose part of its momentum. For these high-fre-

quency radiations the experimental problem of determining whether they actually are waves or particles (in the usual sense of the words) is very difficult, as we have found in earlier discussions.

Cosmic Rays

For the radiation called *cosmic rays* the decision between particles and waves has proved particularly troublesome.

Cosmic rays were discovered as a result of some very simple experiments with one of the oldest of electrical instruments, the electroscope. Insulate a charged electroscope as well as you can—set it in perfectly dry air, place under it carefully dried glass or rubber or sulfur, shield it from light—and still you will find it unable to hold its charge. The discharging is very slow, to be sure, requiring hours or days, but is nevertheless steady and measurable. The steady leaking of charge of either sign means that the air near the electroscope must be ionized, ions of a charge opposite to that of the instrument being attracted to it and so gradually neutralizing its charge. Now the molecules of dry air do not become ionized spontaneously; the only reasonable explanation for the ionization is some kind of radiation, either electromagnetic waves or swiftly moving particles.

For a time the radiation was believed to come from minute traces of radioactive elements known to be present in soil and rocks. But the rate of discharge was found to be unchanged in electroscopes placed over deep bodies of water, which should screen out radiations from the rocks beneath; and strangely, the discharge slackened somewhat when electroscopes were lowered into deep lakes or carried into deep mines. Furthermore, electroscopes carried high above the earth in balloons showed an increased rate of discharge. These observations suggested that the radiation responsible for the ionization comes not from the earth but from some source outside the earth, and that the radiation is partly absorbed in passing through air, water, and rocks. Measurements at widely separated points showed that the radiation had approximately the same strength over all parts of the earth's surface, hence probably did not come from an isolated source like the sun or a planet. Because it seemed to reach the earth from somewhere in space beyond the solar system, the radiation was christened *cosmic rays.*

The strangest property of cosmic rays is their extraordinary penetrating power. Through the entire atmosphere and through great thicknesses of water or solid rock they travel with only a moderate loss in intensity. Such penetrating power implies enormous energy. If the rays are electromagnetic waves, they must have frequencies considerably greater than those of gamma rays. If they are particles, they must be small and moving with velocities near that of light.

The nature of cosmic rays was a subject for heated arguments until

very recently. The decisive experiments were accurate measurements of cosmic-ray intensities at different points on the earth: the intensities are distinctly greater at high latitudes than near the equator. This variation with latitude indicates that the rays are deflected by the earth's magnetic field, which is possible only if they are electrically charged particles. Hence at least in large part cosmic rays are made up of tiny, rapidly moving charged particles—electrons, positrons, and probably others.

Laboratory study of cosmic rays is difficult because they are scanty at the earth's surface and because most of the radiations which do reach the surface are not the original cosmic rays but secondary products formed by collision of the original particles with molecules of the atmos-

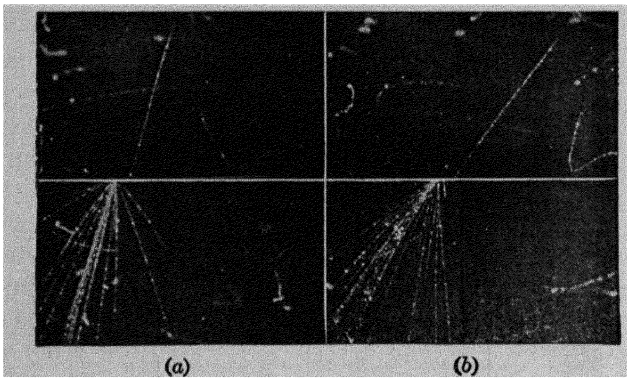


FIG. 168. *Cloud-chamber photograph of the collision of a high-energy cosmic-ray particle with an atomic nucleus. The two pictures are photographs of the same collision, one taken at right angles to the other. (Photograph by J. C. Street and E. C. Stevenson, Harvard University.)*

phere. The collision of a cosmic-ray particle with an ordinary atomic nucleus is exceedingly violent and may produce a number of new particles as well as photons of high-frequency gamma radiation. The secondary products themselves have high energies and may take part in further collisions. Cloud-chamber photographs of such cosmic-ray collisions can be taken by allowing the incoming particle to act as a switch for the cloud chamber (Fig. 168). From such photographs the existence of the positron and the meson was established (page 290).

Still a mystery is the origin of these high-energy particles which bombard our planet from all sides. Perhaps they arise in the interior of the stars, perhaps in the spiral nebulae, perhaps in the dark empty spaces between the stars. Wherever they are formed, it is by a process which has no duplicate on the earth, for no known terrestrial processes give energies so extraordinarily high.

Electrons and Waves

The central idea of the quantum theory, that electromagnetic waves have some of the attributes of particles, is a difficult one to grasp. The difficulty is made worse by experiments which show that this statement can be reversed: particles of matter can be made to show the characteristic properties of waves. If, for example, a stream of electrons falls on a photographic plate after passing through a thin sheet of nickel, the plate shows an unmistakable interference pattern (Fig. 169), of exactly the sort we relied on to establish the wave nature of X rays. The intensity of the electron beam is strong in some directions, weak in other directions, just as we would expect if electrons consist of waves that are in step along some lines and out of step along others.

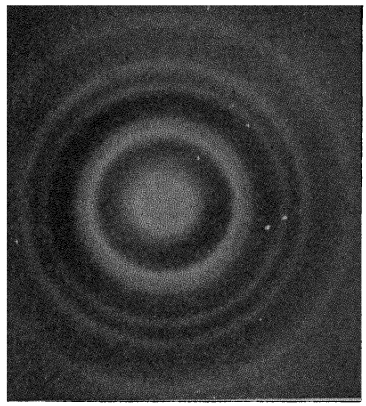


FIG. 169. *Dark and light rings produced by the interference of electrons passing through the crystal lattice of a metal.* (*G. P. Thomson.*)

This brings us back to a question we have argued before: What is an electron? In the earlier discussion we were concerned with the relation of matter to electricity: Is the electron pure electricity, or matter with electricity attached? That query proved meaningless, since matter and electricity are inextricably associated. Now the confusion is worse: Shall we consider the electron a particle because it has mass, momentum, energy, and an electric charge; or shall we regard it as a form of wave motion, because it shows interference? Or, perhaps, will this question too prove meaningless, because we are trying to use concepts of everyday life to describe particles inconceivably small?

Note carefully our line of reasoning. We do certain experiments—Thomson's experiments with cathode rays, for instance—which indicate that the electron is a particle. We form a mental image of a very tiny solid ball, since that is the usual meaning of "particle" in ordinary life. Then we do other experiments, which show that electrons interfere when passed through a crystal. Disregarding Thomson's work, we now form a mental picture of waves, since in ordinary experience only waves show interference. From each experiment we try to make ourselves a picture, a mechanical model in ordinary terms, which will help us to understand how electrons behave. But have we any assurance that a model based on ordinary experience can be applied to particles less than a thousand bil-

lionth of a centimeter in diameter? Isn't it conceivable that the electron is something impossible to picture with everyday concepts, something which in one experiment may act as a particle and in another as a wave?

One of the basic difficulties in trying to understand electrons is that we cannot experiment with them without disturbing them. If I experiment with a rubber ball, I can observe the ball by means of light reflected from it, and the light has no appreciable effect on my experiments. With the electron, because of its small size, this is no longer true: every attempt to observe it introduces a disturbance.

Perhaps a hypothetical experiment will make this difficulty clear. Suppose we could construct a microscope powerful enough to see an electron. We cannot use ordinary light to illuminate the electron on the stage of this microscope, because ordinary light has a wave length on the order of a hundred million times the diameter of the electron, and a microscope will show details only down to a size near the wave length of the light employed. But perhaps we can arrange to use some sort of high-frequency gamma rays, with wave lengths on the order of the electron's diameter. These rays would not be visible to our eyes, but as long as we are imagining things let us suppose that we have some device for actually seeing them.

So we arrange to get an electron on the stage of our microscope, and illuminate it by turning on the gamma-ray light. Expectantly we peer down the microscope tube. We see a momentary flash, and then the field is blank; our electron has disappeared. Then we recall that gamma rays of short wave length have photons of very high energy. At least one of these photons must be reflected from the electron in order for us to see it, but when the electron is struck by the photon it recoils out of the microscope field. Evidently, if the electron is to stay in our field of view, we shall have to be content with gamma rays of longer wave length and lower energy. Making this change to longer waves, we again look down the microscope tube. Now the electron moves more slowly, so that we can actually see it; but we see it only as a hazy blur, because with longer waves our microscope loses its power to show details clearly. No matter how we alter the microscope, we cannot improve matters: It is impossible to obtain a clear look at the electron without disturbing it.

In the first of our experiments we tried to see the electron as a distinct image in a definite spot, and the electron moved rapidly out of our field of view. In the second experiment we kept it in the field long enough to see it, but we did not locate it as a clearly defined image. These disappointing results, together with results of other experiments, both hypothetical and real, are summarized in a law called the *uncertainty principle*, first stated by the German physicist Heisenberg: *It is impossible to obtain accurate values for the position and velocity of an electron simultane-*

ously. We can get rough values for both position and velocity; if we try to get the position more accurately, the measurement of velocity becomes very inaccurate; if we try to determine the velocity precisely, the position becomes uncertain. The indeterminateness is not due to faulty experimental technique, but to the impossibility of observing electrons without disturbing them.

The uncertainty relation makes it impossible to decide experimentally whether the electron is a wave or a particle. Either a particle disturbed by the process of observing it, or a tiny "packet" of waves, would show the necessary indeterminateness of position and velocity. The question we posed at the beginning of this section, "Is the electron a wave or particle?" is meaningless in the sense that no experiment can be devised to give us an answer.

The clearest picture we can get of an electron is to regard it as a particle and to think of the associated waves as "waves of probability." According to the uncertainty principle, if we know anything at all about the velocity of an electron we cannot know its position exactly; but we can express its position as a probability of its being in any particular spot. Not only the present position, but future positions and velocities of the electron may be expressed as probabilities. We cannot say just where an electron will be 2 sec hence, or how fast it will be moving; but we can say that it will more probably be in one place than another, and that its speed will more probably have one value than another.

Thus an essential unpredictability is brought into physics. For larger objects like stones or rifle bullets we describe their behavior in definite laws: Knowing the position and speed of a falling stone, we can predict accurately where it will be at future times, by using the law of falling bodies. We are confident in our predictions, because all falling stones of our experience have behaved similarly. Stones and other objects of everyday life obey the principle of *causality*, that similar events always follow similar causes. But electrons are different; we cannot make accurate predictions about an electron's future motion, and we cannot be sure that two electrons with roughly similar positions and velocities will behave at all alike. The principle of causality, expressing our faith in the uniformity of natural processes, does not hold for electrons.

Other small particles as well as electrons are affected by the uncertainty principle, which means that their motions likewise can be expressed only as probabilities. Since the objects of everyday life are made up of small particles, we might conjecture that even their behavior should be a matter of probability rather than of fixed laws. There is a chance, for instance, that the chair I am sitting on will someday defy the law of gravity and rise straight up in the air. But for objects this large—in fact, even for objects the size of molecules—probabilities are so large as to be prac-

tically certainties. The chance that my chair will continue to obey the law of gravity is so large that I might wait a trillion years without seeing any sign of deviation. Only for electrons and other particles of similar size is there appreciable chance for violations of ordinary physical laws.

All this long discussion gives us no clear picture of what the electron actually is. Something which behaves at times like a particle, at times like a wave, or something which resembles a particle but cannot be definitely located—in vague language like this we have to describe it. In future discussions we shall find it convenient to think of electrons as particles, as miniature billiard balls carrying electric charges, but we must keep in the back of our minds the knowledge that this picture is not entirely correct.

Questions

1. What sort of an experimental setup would you use to observe the absorption spectrum of chlorine?
2. What kind of a spectrum would you expect to observe if you looked through a spectroscope at
   a. The sun?
   b. The tungsten filament of an electric light?
   c. A neon sign?
   d. Light from an electric light passing through sodium vapor?
3. What kind of vibrating particle gives rise to each of the following electromagnetic waves?
   a. Radio waves.
   b. Ultraviolet waves.
   c. Infrared waves.
4. In terms of the Bohr theory, explain why the hydrogen spectrum contains many lines, even though the hydrogen atom has only a single electron.
5. In terms of Bohr's model of the hydrogen atom, can you suggest an explanation for the production of dark absorption lines when light from a heated metal passes through hydrogen? (Remember that light is absorbed when an electron jumps to a larger orbit, emitted when it jumps to a smaller orbit.)
6. Suppose that a hydrogen atom absorbs energy by a jump of its electron from the normal orbit to a larger orbit. What is the relation between the amount of energy absorbed and the amount emitted as radiation when the electron jumps back to the normal orbit? Does this suggest an explanation for the fact that dark lines of the hydrogen absorption spectrum have the same wave lengths as bright lines in its emission spectrum? Explain.
7. What conclusion can be drawn from each of the following observations?
   a. Cosmic rays are deflected in passing through the earth's magnetic field.
   b. When light of a single frequency falls on a metal, the electrons emitted all have the same energy; if the light is made more intense, the number of electrons increases but not their energy.
   c. A charged electroscope, even when well insulated, slowly loses its charge.
   d. Electrons can be made to show interference patterns.
8. Calculate the amount of energy in a photon of ultraviolet light with a frequency of 2×10^{16}. Is this energy greater or less than that in a photon of
   a. Yellow light?
   b. X rays?

CHAPTER XXIII

Subatomic Chemistry

THE nucleus of an atom is its most vital part. The mass of the nucleus is practically the same as the atom's mass, and the charge of the nucleus determines the number of electrons in the electron cloud. Rob the atom of an electron, even several electrons, and its essential nature remains unchanged: if other electrons are available, they will fit into its structure quite as well as the ones it has lost. But let the nucleus be altered, by spontaneous radioactive decay or by collision with a fast-moving particle, and the atom's whole structure is irrevocably changed. Automatically the electron cloud adjusts its numbers and its pattern to fit the new mass and charge of the nucleus.

Yet the nucleus, master though it is of the atom's destiny, is not the part of the atom which affects us directly. In ordinary dealings with atoms we are concerned only with the outer part of the electron cloud. True, when we measure atomic weights we are measuring nuclear weights, and when we observe radioactive changes we are watching a nuclear phenomenon. But aside from these two properties—atomic weight and radioactivity—all the common physical and chemical properties of an element are determined by the arrangement of electrons in the outer layers of its atoms.

So, in order to see how an atom's internal architecture determines its ordinary everyday behavior, we need some further information about the structure of the electron cloud. Fortunately this information is available, principally as a result of detailed study of line spectra. As we found in the last chapter, spectral lines originate in electron jumps from one orbit to another, and correlation of spectral lines makes it possible to determine the approximate location of the different orbits.

Electron Shells

For this discussion we shall imagine the atom to be a simple structure like that pictured by Bohr, with tiny spherical electrons moving rapidly around the nucleus. We shall regard the electronic orbits as simple

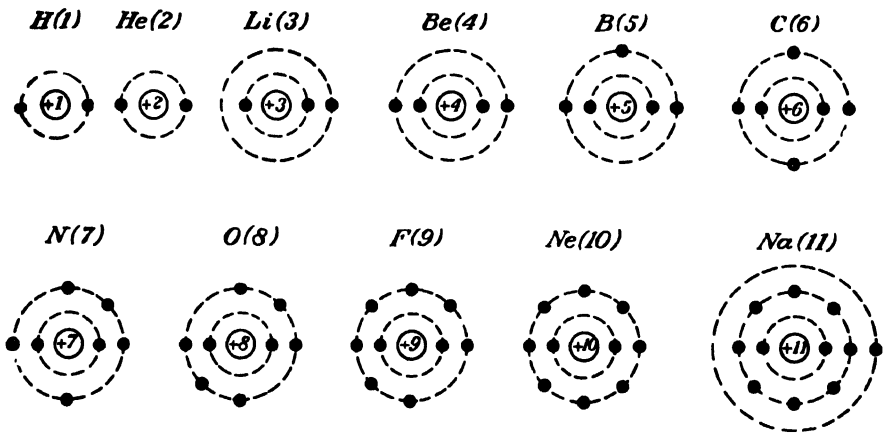
ellipses, like the elliptical orbits of the planets around the sun. Unlike planetary orbits, the electronic paths circle the nucleus on every side, so that the shape of the whole atom is roughly spherical. We shall keep in mind that this model is not a true representation of the atom, but it is a close enough approximation to account for a large part of atomic behavior.

We shall not have to consider electron jumps from one orbit to another, since we are now concerned almost wholly with atoms in their normal states—atoms whose electrons are following their normal orbits of least energy. We shall not even be troubled about the precise shapes of the orbits. Our chief concern with the orbit of an electron is that its size determines the *average distance* of the electron from the nucleus.

In all atoms except the very simplest the electrons are arranged in layers, or *electron shells*, all electrons of a given shell maintaining the same average distance from the nucleus. For simplicity the shells may be pictured as concentric spheres about the nucleus, much like the concentric layers of an onion.

For convenience in diagrams of atomic structure, the electron shells will be pictured as circles rather than spheres, and the electrons themselves will be shown as dots symmetrically placed. These are highly formalized diagrams: remember that the electrons are actually distributed in three dimensions, not two; that they are not fixed points, but rapidly moving; that their distance apart is greater, in proportion to their size, than the diagrams indicate. The nucleus in these diagrams will be represented simply as a circle labeled with its positive charge.

The following diagrams show the atomic structures of the elements with atomic numbers 1 to 11:



The isotopes of any one element all have the same electronic structures, so that the diagrams represent all isotopes of these 11 elements.

Atoms of isotopes differ only in the number of neutrons in their nuclei.

Note that the structure of the He atom, two electrons close to the nucleus, is preserved in the interior of all heavier atoms. That is, the innermost shell of all atoms (except hydrogen) contains only two electrons. The next shell accommodates a maximum of eight electrons; it is partially filled in the elements Li to F, and attains its full quota in Ne. Atoms heavier than Ne have their two inner shells completely filled, with additional electrons partly or completely filling still other shells.

The Periodic Law

As we should certainly expect, the periodic law is closely related to these electron-cloud structures. The relationship becomes apparent if we simply write down the electron patterns for similar elements, as in Table XVIII. To avoid cumbersome diagrams for the larger atoms, the electrons in each shell are indicated simply by a number.

TABLE XVIII. ELECTRON CLOUD STRUCTURES

	H	He	Li	Be	B	C	N	O	F
Electrons in									
1st shell	1	2	2	2	2	2	2	2	2
2d shell			1	2	3	4	5	6	7
	Ne	Na	Mg	Al	Si	P	S	Cl	
Electrons in									
1st shell	2	2	2	2	2	2	2	2	
2d shell	8	8	8	8	8	8	8	8	
3d shell		1	2	3	4	5	6	7	
	A	K	Ca	.....					Br
Electrons in									
1st shell	2	2	2						2
2d shell	8	8	8						8
3d shell	8	8	8						18
4th shell		1	2						7
	Kr	Rb	.....	.....					I
Electrons in									
1st shell	2	2							2
2d shell	8	8							8
3d shell	18	18							18
4th shell	8	8							18
5th shell		1							7

Note, in the first two rows, that each element differs from the one preceding by one electron in its outer shell. In passing from element to

element along these rows, we see first the second shell, then the third shell, being "built up" to its maximum of eight electrons. In the third row begins the "building up" of the fourth shell. The inert-gas atoms (He, Ne, A, Kr) are characterized by complete outer shells—*i.e.*, shells which remain intact in the next succeeding heavier atoms. Helium has a complete shell of two electrons, all others of the group an outer shell of 8. The alkali metals of the next column resemble each other in that each contains a single electron in addition to the completed shells of the preceding inert gas. The elements Be, Mg, and Ca have two electrons outside a complete inner shell. So across the table to the halogens, which are characterized by outer shells of seven electrons.

Evidently chemical similarity between elements is somehow bound up with a similarity in the number of electrons in the outer shells of their atoms. The inactivity of the inert gases is associated with their completed outer shells, the great activity of the alkali metals with their single outer electrons, the activity of the halogens as nonmetals with their groups of seven outer electrons. In terms of these outer electron shells, the periodic law may be restated: *when the elements are arranged in order of increasing atomic number, similar numbers of electrons occur in the outer shells of their atoms at regular intervals.*

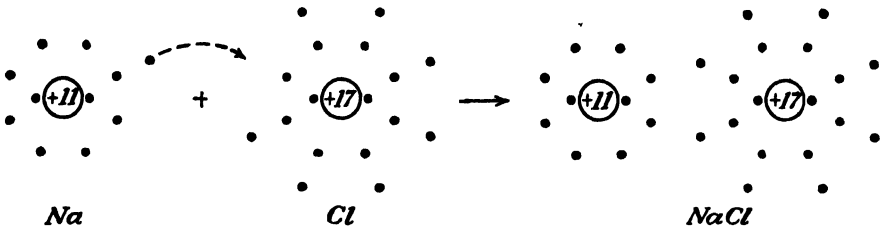
Even the incomplete list of elements in Table XVIII suggests the generalization that metal atoms are characterized by a small number of outer electrons, atoms of active nonmetals by a number just under eight. The most active metals of all are those with a single outer electron in each atom, the most active nonmetals those with seven outer electrons in each atom. Elements whose atoms have three or four outer electrons, like B, C, and Si, are in some respects intermediate in properties between metals and nonmetals.

Chemical Combination

A structure with eight outer electrons seems to be peculiarly stable. Just why is not known. Perhaps this structure is more symmetrical than others, hence more mechanically stable. Whatever the reason, atoms which normally possess eight outer electrons (the inert gases) show no slightest tendency to change their structure by combining with other atoms, while atoms which do not possess this favored number are ready to sacrifice their independence and even their electrical neutrality to gain it. In this tendency of atoms to provide themselves with outer shells of eight electrons lies the explanation of a great many chemical reactions.

Atoms may attain stable outer shells in one of two ways: by the transfer of electrons from one atom to another, or by the sharing of electron pairs between atoms.

Combination by Electron Transfer. The simplest example of this type of reaction is the combination of a metal and a nonmetal. For a specific case, let us consider the burning of sodium in chlorine to give sodium chloride. From Table XVIII, it is evident that an Na atom can attain a stable structure most simply by getting rid of its lone outer electron. A Cl atom, on the other hand, can give itself a stable outer shell most easily by adding one electron to the seven already present. Thus Na and Cl are perfect mates: one has an electron to lose, the other an electron to gain. In the process of combination, an electron is *transferred* from Na to Cl.



So eager are sodium atoms to lose electrons, so anxious are Cl atoms to receive them, that the combination is extremely violent, accompanied by the evolution of much heat and light.

The great stability of the eight-electron arrangement is indicated by the energy given out in this reaction. The resulting salt NaCl is extremely unreactive, since each part of it has a stable electronic structure. To break it apart—*i.e.*, to return the electron to Na, destroying the stable eight-electron structures—requires the expenditure of the same enormous amount of energy that the combination set free.

The Na atom in the compound, shorn of its electron, is of course no longer a normal atom, since it consists of a nucleus with eleven positive charges and only ten electrons. The structure as a whole, therefore, has a positive charge; we call it *sodium ion*, with the symbol Na^+. The Cl atom has one electron in excess of its normal number, and so is charged negatively; we call this structure *chlorine ion* (or *chloride ion*), Cl^-. The solid salt NaCl is made up of a crystal lattice of alternate sodium and chloride ions (Fig. 170).

The fundamental characteristic of all metal atoms is their tendency to lose their outer electrons, as sodium did in the above example. Nonmetal atoms, on the other hand, tend to gain electrons so as to fill in gaps in their outer shells. In most reactions of this sort a metal loses all of its outer electrons, and a nonmetal fills all the gaps in its structure. Thus when sodium combines with sulfur, each S atom has two spaces to fill (Table XVIII), each Na atom only one electron to give; hence two Na

atoms are required for each S atom, and the resulting compound is Na_2S. When calcium combines with oxygen, each Ca atom contributes two electrons to each O atom, and the formula of the compound is CaO.

Compounds formed by electron transfer are called ***ionic compounds.*** In addition to simple compounds like NaCl, $MgBr_2$, K_2S, ionic compounds include substances with more complex formulas like Na_2SO_4, KNO_3, $CaCO_3$, in which electrons from the metal atoms have been transferred to nonmetal atom groups (SO_4, NO_3, CO_3) instead of to single nonmetal atoms. Ionic compounds in general contain a metal and one or more nonmetals, and their crystal lattices are made up of alternate positive and negative ions. Most of them are crystalline solids with high melting points, as might be expected since melting involves a separation of the ions. Another important characteristic of ionic compounds which we shall discuss presently is their ability in the molten state or in solution to conduct electricity.

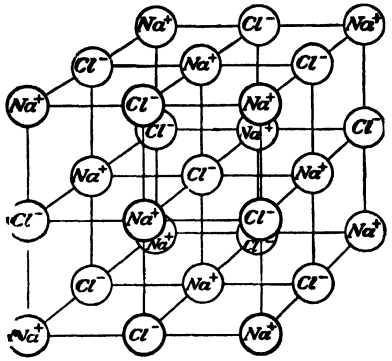
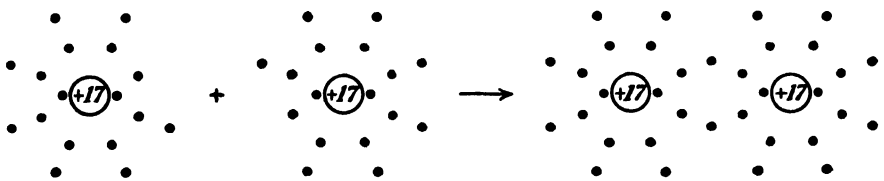
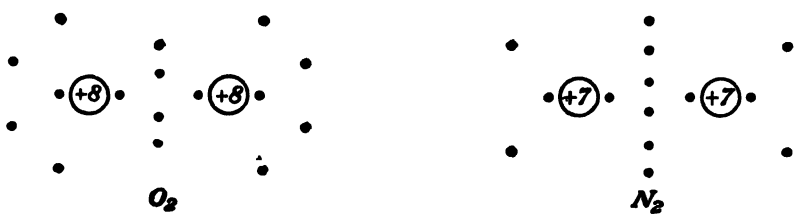


FIG. 170. *The crystal lattice of an ionic compound, NaCl. Actually the ions are crowded more closely together than the diagram shows.*

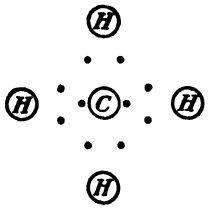
Combination by Sharing of Electron Pairs. Two atoms may combine to give each other stable outer shells even though the one has no more tendency to lose or gain electrons than the other. Thus Cl atoms unite to form a Cl_2 molecule.



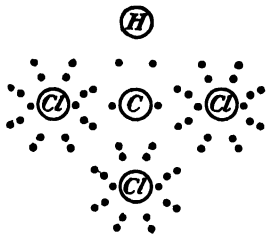
The pair of electrons held in common by the two atoms is said to be *shared.* In some molecules more than one pair is shared; thus the structures of O_2 and N_2 are



Most important of the compounds formed by electron sharing are the so-called "organic compounds," compounds of the element carbon (page 401). The carbon atom has four outer electrons, which can be shared with other atoms to form a stable ring of eight:



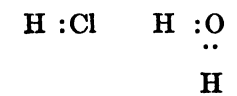
Methane, CH_4



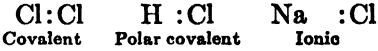
Chloroform, $CHCl_3$

Substances whose atoms are joined by shared electron pairs are called ***covalent substances.*** In general they are nonmetallic elements or compounds of one nonmetal with another, although some compounds of metals belong in this class. Since a pair of electrons shared between two atoms remains at approximately the same distance from each atomic nucleus, neither atom acquires an electric charge; hence the crystal lattices of covalent compounds are made up of atoms or molecules rather than ions. In general, covalent substances are poor conductors of electricity.

The distinction between covalent and ionic compounds is not sharp. In both kinds a pair of electrons is held between two atoms; the distinction depends on whether the pair is held chiefly by one atom (ionic) or equally by the two atoms (covalent). Some covalent compounds have an intermediate character, in that the electron pair is somewhat closer to one atom than the other. Two examples are HCl and H_2O:



These substances are called *polar* covalent compounds, because one part of the molecule is relatively negative and another part positive. All gradations can be found between symmetrical covalent molecules at one extreme, through polar covalent molecules to ionic compounds at the other extreme. For example:



Activities of Metals and Nonmetals

Since metals are characterized by the tendency of their atoms to give up outer electrons to other atoms, we may describe an active metal as one

whose atoms give up their electrons more readily than do the atoms of other metals. Thus magnesium is a more active metal than platinum because the magnesium atom loses its outer electrons more readily than the platinum atom. Similarly we may describe an active nonmetal as one whose atoms readily take electrons from metal atoms to complete their outer shells. Thus chlorine reacts more violently with metals than does iodine because its atoms have a greater attraction for the outer electrons of metal atoms.

Now an earlier discussion of the periodic table (page 205) brought out the fact that elements within each horizontal row and each vertical column show a progressive change in metallic or nonmetallic activity. In a horizontal row properties change from those of an active metal (alkali group) at one side to those of an active nonmetal (halogen group) on the other. The change takes place through a series of progressively less active metals and then a series of progressively more active nonmetals. In the vertical columns is a much less striking change, tending as a rule toward more active metals (and less active nonmetals) at the bottom of the table. Can we find an explanation for these progressive changes in properties in terms of atomic structure?

Consider first the metals in a horizontal row, say the second row in Table XVIII. Skipping the inert gas neon, we find at the left side of the row three metals, sodium, magnesium, and aluminum. Within the outer shell of the sodium atom is a structure consisting of 10 electrons and a nucleus with a charge of $+11$; within the outer shell of the magnesium atom are 10 electrons and a nucleus with a charge of $+12$; within the outer shell of the aluminum atom are 10 electrons and a nucleus with a charge of $+13$. Thus the solitary outer electron of sodium is attracted to the inner part of the atom by a charge of $+1$ $(+11 - 10)$; each outer electron of magnesium is attracted to the inner part of the atom by a charge of $+2$; each outer electron of aluminum is attracted to the inner part by a charge of $+3$. Hence the sodium atom loses its outer electron more easily than magnesium, and magnesium loses its outer electrons more readily than aluminum. So we should expect to find sodium the most active of the three and aluminum the least active, a conclusion in good agreement with experiment. The other elements in this row, with 4 or more electrons in the outer shells of their atoms, are all nonmetals.

Now consider the vertical column containing the alkali metals. Atoms of these elements all have single outer electrons, but differ in the number of electrons between the outer one and the nucleus. Cesium has the largest atom of the group, with fifty-four electrons inside the orbit of the outer one. This means that cesium's outer electron is relatively far from the nucleus and is screened from the positive nuclear charge by

many other electrons. Hence the cesium atom should lose its outer electron more easily than the smaller atoms of the other metals. This conclusion checks the experimental fact that cesium is the most active of the alkali metals.

Similar reasoning with only slight further complications will explain differences in activity among the nonmetals.

Valence

Another property of an element which finds a simple explanation in terms of atomic structure is its valence.

In a previous discussion of chemical combination (page 201), we defined the valence of a metal as the number of Cl atoms combined with each metal atom in the compound of the metal with chlorine. Thus potassium has a valence of 1, since the formula of its chloride is KCl; calcium has a valence of 2 and aluminum a valence of 3, since the formulas of their chlorides are $CaCl_2$ and $AlCl_3$, respectively. Now in terms of electrons, each K atom combines with only 1 Cl atom because it has only a single outer electron to lose; each Ca atom combines with 2 Cl atoms because it has 2 electrons in its outer shell; each Al atom combines with 3 Cl atoms because it has 3 electrons to lose. Hence *the valence of a metal is simply the number of electrons in the outer shell of each atom.*

The valence of a nonmetal was defined as the number of hydrogen atoms combined with each nonmetal atom in the compound of the nonmetal with hydrogen. Thus chlorine has a valence of 1 (HCl), oxygen a valence of 2 (H_2O), nitrogen a valence of 3 (NH_3). In terms of electrons each Cl atom combines with 1 H atom because it needs only 1 electron to complete a stable outer shell of 8; each O atom combines with 2 H atoms and each N atom with 3 H atoms because O has room for 2 more outer electrons and N for 3 more. *The valence of a nonmetal is evidently the number of electrons needed to complete an outer shell of 8.*

In ionic compounds the valence of the metal is considered *positive* and that of the nonmetal *negative*. These terms are justified by the fact that in such compounds electrons have been transferred from metal atoms to nonmetal atoms, leaving the former positive and the latter negative. Thus solid sodium chloride is made up not of sodium atoms and chlorine atoms but of positive sodium ions (valence +1) and negative chloride ions (valence −1).

In covalent compounds the valence of an element is defined as the number of electron pairs which each of its atoms shares with other atoms. Positive and negative valences are generally not distinguished, since electron pairs are shared between adjacent atoms instead of being strongly attracted to one or the other. For example, carbon has a valence of 4 in either CH_4 or $CHCl_3$, H has a valence of 1 and Cl has a valence of 1 (see diagrams,

page 327). In SO_2, sulfur has a valence of 4 and oxygen of 2, since each sulfur atom shares 4 electron pairs and each oxygen atom shares 2.

Since the number of electrons in the outer shell of an atom determines its normal valence, these outer electrons are often called *valence electrons*. Thus H has 1 valence electron, C has 4 valence electrons, Cl has 7, Ca has 2.

Ions

How useful modern ideas of atomic structure are in explaining ordinary chemical properties should by this time be evident. We shall consider here only one further application of these ideas, their application to the electrical conductivity of various materials.

An electric current through a solid is chiefly a movement of electrons (page 231). The movement takes place by a shift of electrons from one atom to the next, and occurs readily only if atoms of the solid have outer electrons which are easily detached. Hence it is only reasonable that metals should be good conductors, since their valence electrons are readily lost, while nonmetals are poor conductors because their valence electrons are firmly held.

Electricity is conducted through gases and nonmetallic liquids by a different mechanism. The current almost always consists not of moving electrons but of moving *ions*. These structures, which we have mentioned several times before, are electrically charged atoms or groups of atoms—structures resembling ordinary atoms and molecules, but possessing either less than or more than enough electrons to neutralize the nuclear charges present. Conduction through a gas or liquid usually involves movement of both positive and negative ions.

We have noted the ionization of air produced by various powerful radiations—X rays, cosmic rays, and the alpha, beta, and gamma radiations from radioactive elements. These agents are capable of stripping off one or two outer electrons from some of the air molecules in their path, producing positive ions; the liberated electrons attach themselves temporarily to adjacent neutral molecules, making them negative ions. Left to themselves, positive and negative ions in the course of a few minutes neutralize each other by an exchange of electrons. If a charged electroscope is present when the ions are produced, ions of opposite sign are attracted to it and the instrument is quickly discharged; this fact we have used again and again as a test for the presence of radiations. In the cloud chamber (page 284) positive and negative ions form the nuclei for fog droplets which make visible the paths of high-speed particles.

The light and heat of an electric spark are due to the violent disturbance created by ionized gas molecules moving rapidly between the electrodes.

Ionic compounds like ordinary salt are good conductors when liquefied. These compounds, as we have seen, at ordinary temperatures are solids with crystal lattices made up of positive and negative ions. The solids are nonconductors, since their ions are held firmly in position; but in the liquid state the ions are free to move about, hence to conduct electricity from one electrode to another.

Molten salt is made up of the two ions Na^+ (sodium ion, a sodium atom with its outer electron missing) and Cl^- (chloride ion, a chlorine atom with one excess electron). When two electrodes are immersed in the liquid, Na^+ ions are attracted to the cathode (— electrode) and Cl^- ions to the anode (Fig. 171). At the anode each Cl^- is *neutralized;* it gives up its extra electron to the electrode and becomes a normal chlorine atom. At the cathode Na^+ is neutralized by adding to itself an electron from the electrode and becomes an atom of ordinary metallic sodium. Thus electrons move from electrode to the liquid at the cathode, from liquid to electrode at the anode. In effect, therefore, a current passes through the liquid, the current in the liquid itself consisting of a movement of ions.

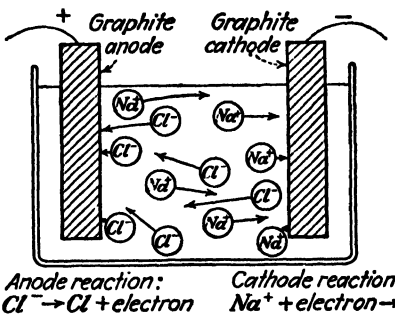


FIG. 171. *Electrolysis of molten sodium chloride.*

Note that the current, forced through the molten salt, effects the breaking up of the compound NaCl into its elements.

$$2NaCl \rightarrow 2Na + Cl_2$$

The sodium, a liquid at the temperature of molten salt, collects around the cathode; chlorine gas bubbles up from the anode. A process of this sort, in which free elements are liberated from liquids by an electric current, is called *electrolysis.* In the preparation of many elements, and in electroplating, electrolysis finds important commercial applications: for instance, the procedure just outlined is the one commonly used to prepare metallic sodium.

Electrolysis is more frequently carried out in water solutions than in molten salts. Pure water contains an exceedingly small number of ions, hence is practically a nonconductor. But ionic compounds dissolve in water as separate ions: thus a solution of ordinary salt contains Na^+ and Cl^-; a solution of copper bromide ($CuBr_2$) contains Cu^{++} (a Cu atom with its 2 valence electrons missing) and Br^- (a Br atom with 1 electron

added to its normal outer shell of 7). The ability of these dissolved ions to move through a solution makes the solution a good conductor.

Details of electrolysis are much the same in solutions and in molten salts. In electrolysis of $CuBr_2$ solution, for example, Cu^{++} ions move to the cathode, Br^- ions to the anode. The cathode gives 2 electrons to each Cu^{++}, changing the ion to a Cu atom; the anode takes 1 electron from each Br^-, changing the ion to neutral Br. Thus copper is plated out at the cathode, while free bromine is liberated and goes into solution at the anode. Every time one Cu^{++} ion and two Br^- ions are neutralized, 2 electrons are in effect transferred from cathode to anode, so that a continuous current is maintained across the solution.

Let us for a moment consider electrolysis quantitatively. Suppose we connect together an electrolytic cell containing molten NaCl and one containing $CuBr_2$ solution, with a battery in the circuit to supply electricity (Fig. 172). The battery, as usual, starts electrons around the circuit from its negative terminal, and attracts electrons from the circuit to its positive terminal. Now imagine that we can observe the progress of a single electron setting out from the negative side of the battery. Moving along the wire to the cathode of the NaCl cell, it here adds itself on to a Na^+ ion from the liquid. At the same time, to keep the liquid neutral, a Cl^- ion gives up an electron to the anode. This electron moves to the cathode of the $CuBr_2$ cell, where it neutralizes half the charge of a Cu^{++} ion. At the same time a Br^- ion gives an electron to the anode, and this electron returns to the battery at its positive side. Movement of a single electron thus liberates a Na atom, a Cl atom, and a Br atom and performs half the task of neutralizing a Cu^{++} ion. Passage of 10 electrons would set free 10 Na atoms, 10 Cl atoms (or 5 Cl_2 molecules), 10 Br atoms, and 5 Cu atoms. No matter how many electrons the battery pushes around the circuit, equal numbers of atoms of Na, Cl, and Br must be liberated, and half as many Cu atoms.

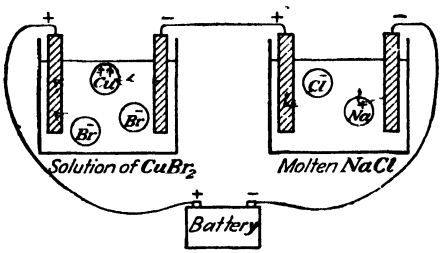


FIG. 172. *Electrolysis of $CuBr_2$ solution and molten NaCl.*

Now the atomic weight of Cl (35.5) is roughly 1½ times that of Na (23.0), and the atomic weight of Br (79.9) roughly 3½ times as great. If we should weigh the amounts of these elements in our electrolytic cells, we ought to find about 1½ times as much Cl as Na, and about 3½ times as much Br, since we should be weighing equal numbers of atoms. The atomic weight of Cu (63.6) is roughly three times the weight of Na; since only half as many atoms of Cu are formed as atoms of Na, the weight of Cu should be about three-halves the weight of Na.

Without pursuing the subject further, we may at least conclude that the amounts of various elements liberated during electrolysis by a given amount of electricity should be simply related to their atomic weights and their valences. We reach this conclusion by starting with the assumptions that electricity travels around a circuit in tiny discrete particles and that liberation of elements at electrodes takes place by means of these particles. A century ago Michael Faraday, first to study electrolysis quantitatively, discovered the same simple weight relationships from careful experiments. By reversing the reasoning we have just used, Faraday might have concluded from his experiments that electricity was not a continuous fluid, but composed of discrete units. Faraday did not make this deduction, but his work on electrolysis truly presaged the discovery of the electron sixty years before Thomson turned his attention to cathode rays.

Questions

1. What general methods of investigation have provided information concerning (*a*) structure of the electron cloud, (*b*) structure of the nucleus?
2. Draw a diagram showing the electron structure of (*a*) sulfur; (*b*) calcium; (*c*) neon; (*d*) chloride ion; (*e*) magnesium ion.
3. The rare element selenium has the following arrangement of electrons: first shell, 2; second shell, 8; third shell, 18; fourth shell, 6. Would you expect selenium to be a metal or a nonmetal? What would be its normal valence? To what group in the periodic table would it belong?
4. Explain in terms of atomic structure why argon forms no compounds with other elements.
5. By means of electronic diagrams represent the reactions between (*a*) a lithium atom and a fluorine atom, (*b*) a magnesium atom and a sulfur atom. Would you expect lithium fluoride and magnesium sulfide to be ionic or covalent compounds?
6. What is the valence of (*a*) carbon in CO_2, (*b*) sulfur in SO_3, (*c*) nitrogen in N_2O, (*d*) phosphorus in P_2O_5, (*e*) carbon in CH_2O? Assume in each case that the oxygen atom shares two electrons with another atom.
7. For each of the following elements, list (*a*) the most common valence; (*b*) the number of electrons in the outer shell of its atoms; (*c*) the number of positive or negative charges on the corresponding ion when the element is part of an ionic compound:

Sodium (*a*) +1, (*b*) 1, (*c*) +1	Sulfur
Potassium	Chlorine
Calcium	Bromine
Aluminum	Iodine
Magnesium	

   What is the general relationship between the number of valence electrons and the number of charges on an ion?
8. Which would you expect to be the more active metal, potassium or sodium? Explain in terms of atomic structure. What experimental facts support your conclusion?
9. What is the difference in atomic structure between the two isotopes of chlorine? How would you account for the great chemical similarity of the two isotopes in terms of atomic structure?

10. Which of the following substances would you expect to have crystal lattices made up of ions?

SO_3 diamond KBr $NaNO_3$ CCl_4 solid CO_2 $CuCO_3$ $Al_2(SO_4)_3$

11. What substances would be produced by the electrolysis of a solution of copper chloride ($CuCl_2$)? Describe how a current is carried through the solution.
12. What part of the atom is principally involved in each of the following processes?
    a. Burning of charcoal.
    b. Radioactive disintegration.
    c. Production of X rays.
    d. Ionization of air.
    e. Production of visible light.
    f. Rusting of iron.
    g. Liberation of sodium by electrolysis of salt.
13. In general, electrons are liberated from metals but not from nonmetals when illuminated with visible or ultraviolet light (photoelectric effect, page 312). Explain. From metals of what group would you expect electrons to be liberated most easily?
14. Into what kind of energy is electrical energy converted during electrolysis? In what common device is this energy change reversed?

Suggestions for Further Reading—Part III

General.

STEWART, O. M.: *Physics*, Ginn and Company, Boston, 1939. See p. 98.

SAUNDERS, F. A.: *A Survey of Physics*, Henry Holt & Company, Inc., New York, 1936. See p. 98.

LEMON, H. B.: *From Galileo to the Nuclear Age*, University of Chicago Press, Chicago, 1946. See p. 98.

On applications of electricity to communication:

HARRISON, G. R.: *Atoms in Action*, William Morrow & Co., New York, 1939.

On atomic energy and fundamental particles:

SMYTH, H. D.: *Atomic Energy for Military Purposes*, Princeton University Press, Princeton, N.J., 1945. The official story of the development of the atomic bomb. Contains an excellent discussion of fundamental particles and nuclear reactions. Not technical, but requires a fair knowledge of elementary physics.

GAMOW, GEORGE: *Atomic Energy in Cosmic and Human Life*, The Macmillan Company, New York, 1946. Elementary and well-written account of research dealing with atomic energy, and of the political and economic consequences of this research.

On applications of modern physics in chemistry:

DEMING, H. G.: *Fundamental Chemistry*, John Wiley & Sons, Inc., New York, 1940. See p. 208.

On the development of ideas regarding electricity and light:

EINSTEIN, A., and L. INFELD: *The Evolution of Physics*, Simon & Schuster, Inc., New York, 1938. A clearly written, nonmathematical account of fundamental physical ideas from Galileo to Einstein.

On the history of radioactivity:

CURIE, E.: *Madam Curie*, translated by Vincent Sheean, Doubleday & Company, Inc., New York, 1937. An excellent biography of Madame Curie and an account of her work on radioactivity, written by her daughter.

PART IV

FUNDAMENTAL PROCESSES

ELECTRIC charges at rest and magnetic poles at rest obey laws of force much like the law which expresses the gravitational attraction between masses. But a moving charge exerts on a magnet a force of altogether different type: a force which is not an attraction or repulsion but a sidewise push, a force which depends not only on distance but on rate of motion. A similar sidewise force which varies with speed is exerted by a moving magnet on an electric charge. These forces destroy all hope of constructing the universe on a simple mechanical basis from particles influenced only by forces of attraction and repulsion.

To express the complex forces brought into being by moving charges and moving magnets, we introduced the idea of a field, a region of space different from surrounding space because in it forces act on magnets or charges or both. Maxwell's equations describing the structure and behavior of fields proved useful not only in correlating electric and magnetic phenomena but in explaining the properties of light. Newton's strictly mechanical picture of light in terms of "corpuscles" had yielded to an interpretation by means of waves, and Maxwell showed that these waves were periodic variations in electric and magnetic fields.

The field concept is fundamentally different from simple mechanical ideas in that it focuses attention on the *structure of space* rather than on the motion of particles. Now space can have a structure in mechanical terms only if it is filled with some kind of material; for structure in emptiness has no mechanical meaning. Hence to extend mechanical ideas to fields, we must imagine, as Maxwell did, that space is filled with all-pervading "ether." But experiments designed to show the existence of ether have given negative results, so that it seems more reasonable to abandon the mechanical concept entirely and ascribe reality to fields in empty space.

Approaching the study of electrical phenomena from another angle, Thomson established the existence of particles smaller than atoms,

thereby paving the way for a wide extension of purely mechanical explanations. Movements of electrons and ions accounted for electric currents. Orbital motions of electrons restrained by the simple electrical attraction of a positive nucleus gave a reasonable picture of atomic structure. Emission of light and related forms of electromagnetic radiation found an explanation in terms of the motion of electrons within atoms and molecules. Even chemical phenomena yielded to an interpretation by means of electrons.

Yet these explanations were not wholly satisfying. Probed more deeply, they led to the strange discovery that matter may be converted into energy and energy into matter, and to the equally strange discovery that light waves sometimes behave as particles and material particles sometimes as waves. Here again mechanical ideas lead to a hopeless impasse. Electrons, protons, and neutrons, despite their usefulness in helping us to picture atomic structure, are nonetheless impossible to visualize completely in mechanical terms.

We have delved as far as we can into the inner structure of matter and energy. We find here a strange confusion of matter and energy, electricity and electromagnetic waves. Yet in this confusion there are some tangible ideas which will help us in understanding more familiar phenomena. We have sought to dissect the world down to its ultimate building blocks; now we shall use these building blocks in reconstructing the world of everyday experience.

CHAPTER XXIV

Ionic Reactions

CHEMICAL combination, we have learned, takes place either by the transfer of electrons from one atom to another, or by the sharing of pairs of electrons between atoms. The distinction is not sharp but serves as the basis for a convenient separation of two kinds of compounds: ionic compounds formed by electron transfer, and covalent compounds formed by electron sharing. An ionic compound nearly always contains a metal and one or more nonmetals; its crystal lattice is made up of ions; it has a high melting point, since melting involves separating its ions; in the liquid state or in solution it conducts an electric current, since its ions are free to move about. A covalent compound most commonly contains two nonmetals; the crystal lattice of its solid state consists of molecules or of atoms; it is a poor conductor either in the liquid state or in solution.

With these ideas as a background, let us inquire into the properties of solutions and the processes by which solutions are formed.

Solubility

A solution is a mixture, a very intimate mixture of the tiny particles of two or more different substances. Solutions may be formed by any of the three states of matter: air is a solution of several gases, sea water is a solution of various solids and gases in a liquid, many alloys are "solid solutions" of two or more metals. Here our chief concern will be solutions in liquids.

In a solution containing two substances, the one present in larger amount is called the *solvent*, the other the *solute*. When solids or gases dissolve in liquids, the liquid is always considered the solvent. Thus when sugar is stirred into water, the sugar is the solute, the water the solvent. Water is by far the commonest and most active of all solvents.

Solutions, like compounds, are homogeneous, but unlike compounds they do not have fixed compositions (page 151). To a solution of 10 g.

of salt in 100 g. of water, for instance, a little more salt or a little more water may be added; the composition of the solution is altered, but it remains homogeneous. Some pairs of liquids form solutions in all proportions: thus with 100 g. of water any amount of alcohol may be mixed to form a homogeneous liquid. More commonly, however, a given liquid will dissolve only a limited amount of another substance. Common salt can be stirred into water at 20°C until every 100 g. of solution contains 26.4 g. of salt; further additions of salt will not dissolve, no matter how prolonged the stirring. This figure, 26.4 g. per 100 g. of solution, is called the ***solubility*** of salt in water at 20°. Formally, *the solubility of a substance*

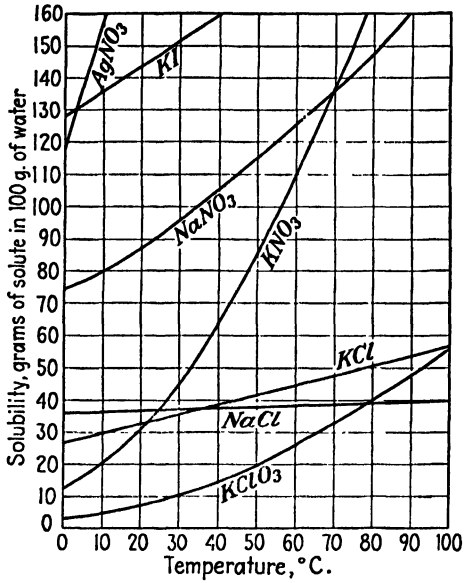


FIG. 173. *Graph showing variations of solubility with temperature.*

is the maximum amount which can be dissolved by stirring in a given quantity of solvent at a given temperature. It is most often expressed in grams of solute per 100 g. of solution but may be given as grams per liter, ounces per quart, etc. A solution which contains the maximum amount of solute is called a ***saturated*** solution.

Solubilities change markedly with temperature, as Fig. 173 shows. In general (this rule has several exceptions), solubilities of gases decrease as the temperature rises, while solubilities of solids increase. If a solution of a solid is saturated at a high temperature and allowed to cool to a temperature at which its solubility is smaller, some of the solid usually crystallizes out. Thus the solubility of KNO_3 is 71 g. per 100 g. at 100°C and 24 g. per 100 g. at 20°C, so that cooling 100 g. of saturated solution through this 80° range would force 47 g. of solid KNO_3 to

crystallize out. Sometimes, if the cooling is allowed to take place slowly and without disturbance, a solute may remain in solution even though its solubility is exceeded, forming a *supersaturated solution.* Supersaturated solutions are in general unstable, the solute crystallizing out suddenly when the solution is jarred.

Solubilities of different materials vary widely. Water, for instance, dissolves readily such diverse substances as salt, sugar, alcohol, ammonia, but it will not dissolve materials like camphor, fat, sulfur, or diamond. Gasoline, on the other hand, dissolves fat but will not affect salt or sugar. Can we account for these relationships in terms of molecular and atomic structure?

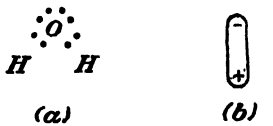


FIG. 174. *Two ways of representing by diagrams the polar molecules of water.*

The explanation in part depends on the electrical structures of different kinds of molecules. Water, for instance, has polar molecules, which behave as if negatively charged at one end and positively charged at the other (page 327). This is possible because water molecules are "bent" and because the electrons shared between O and H are considerably closer to the oxygen (Fig. 174). We call water a *polar liquid,* while a liquid like gasoline, whose molecules have positive and negative charges symmetrically arranged, is *nonpolar.*

Water and other strongly polar liquids consist in large part of molecular aggregates rather than simple molecules: The molecules join together in groups, positive charges against negative charges (Fig. 175). Water molecules can likewise pair up with polar molecules of other substances, such as alcohol and sugar (Fig. 176), so that water dissolves these substances readily. Molecules of fats and oils, on the other hand, are nonpolar, hence not attracted by water molecules. If oil is shaken with water, the strong attraction of the polar molecules of water for each other "squeezes out" the oil molecules from between them, so that the liquids quickly separate into layers. Oil or fat molecules mix readily, however, with the similar nonpolar molecules of gasoline (Fig. 177).

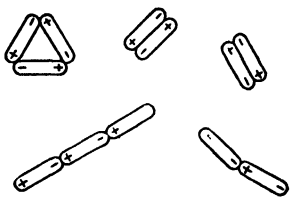


FIG. 175. *Aggregates of polar water molecules.*

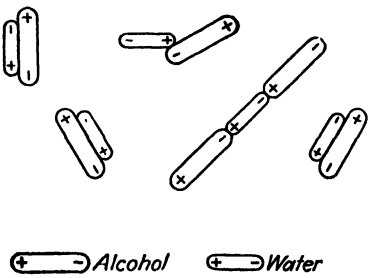


FIG. 176. *Alcohol dissolved in water.*

So the solubility of a *covalent substance which has distinct molecules* depends in large part on the electrical structure of its molecules. It dissolves only in liquids whose molecules have similar electrical structures.

Ionic compounds are highly polar in the sense that negative charges are concentrated on some atoms, positive charges on others, although in the solid state they do not consist of distinct molecules. They dissolve

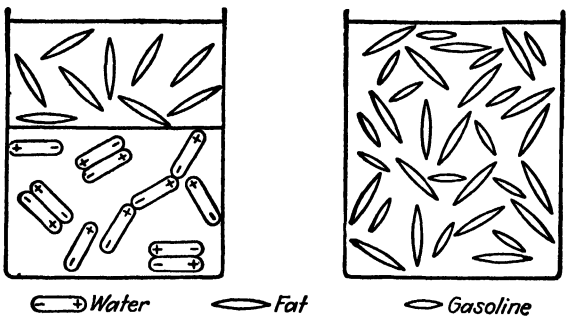


FIG. 177. *Gasoline dissolves fat, water does not.*

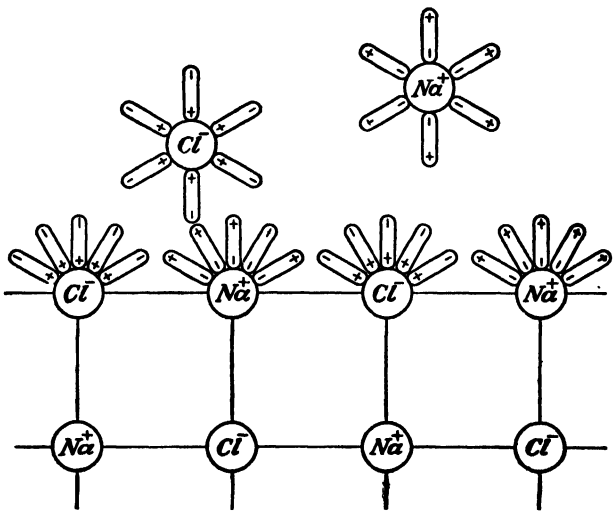


FIG. 178. *Diagram showing how a salt crystal dissolves in water.*

only in highly polar liquids. The process of solution, somewhat different from that for polar covalent compounds, may be visualized with the aid of Fig. 178, which represents the solution of NaCl in water. At the surface of the salt crystal, water molecules cluster around the ions, positive ends toward negative ions and negative ends toward positive ions; the attraction of so many water molecules is sufficient to overcome the electrical forces within the crystal lattice, and each ion moves off into the solution with its retinue of solvent molecules. As each layer is removed

the next is attacked, until the salt is completely dissolved or until the solution becomes saturated. Since this process depends on the overcoming of forces within the crystal by electrical forces exerted by the solvent molecules, it is understandable why ionic compounds will not dissolve in any but the most polar solvents.

These oversimplified explanations and diagrams do not tell the complete story of solubility by any means, but suggest only the principal reasons for the differences in solubility of different substances.

Ions in Solution

Since water is the most highly polar of all ordinary liquids, it has a unique capacity for dissolving ionic compounds, according to the mechanism just described. This ability to dissolve materials *as ions* is a major reason why water is such an immensely important substance on our planet.

Many fluids in our bodies consist largely of dilute solutions of ions. Some of the material that plants manufacture into their food and ours enters their roots as ions dissolved in water. Rain water wears away solid rocks, in part because of its ability to dissolve their material as ions. Ions which make up the salt of the ocean form part of the food of marine plants and furnish material for the shells of marine animals. Reactions between ions in solution are responsible for salt beds, for some ore deposits, for the hardening of sedimentary rocks. Industry takes advantage of ionic reactions in ways too numerous to mention. Because ions in solution play so important a role in our lives and in all manner of natural processes, we have every reason to give them careful attention.

The ions which a given compound forms on dissolving are simply its constituent atoms or atom groups with electric charges determined by their excess or deficiency in electrons. In the example of the last section, every sodium ion is a sodium atom minus its outer electron, and every chloride ion is a chlorine atom with eight electrons instead of the normal seven in its outer shell—precisely the structures, of course, which are present in the original crystal lattice. A potassium nitrate crystal contains positive potassium ions, which are potassium atoms shorn of their outer electrons, and negative nitrate ions, which are nitrate groups with one electron in excess of the total number of protons; when the crystal dissolves, these same two ions are free in the solution. We say commonly that an ionic compound "ionizes" when it dissolves, but "ionization" in this sense means only that ions already present in the crystal are set free to move about independently in the solution.

Formulas for ions are the symbols of the atoms or atom groups with enough + or − signs attached to indicate their charges. The two ions of NaCl are Na^+ (Na minus an electron) and Cl^- (Cl plus an electron),

and those of KNO_3 are K^+ and NO_3^-. $CaCl_2$ ionizes to form Ca^{++} and Cl^- (two of the latter for every one calcium ion), Na_2SO_4 to form Na^+ and SO_4^{--}. Since the charge on an ion is determined by how many electrons it has gained or lost in electron transfer, the charge is the same as the valence of the atom or atom group.

Substances which separate into free ions on solution in water are called *electrolytes.* Electrolytes include all ionic compounds which are soluble in water and some covalent compounds containing hydrogen (for example HCl), which form ions by reaction with water. Other soluble covalent compounds, like sugar and alcohol, which do not ionize in solution, are *nonelectrolytes.*

A property of electrolytes by which they may quickly be recognized, the property which gives them their name, is the ability of their solutions to conduct an electric current. Conduction is possible, of course, because the ions are free to move; positive ions migrate through the solution toward the negative electrode, negative ions toward the positive electrode. The migration of ions, together with the reactions which occur at the electrodes, make up the process of electrolysis (page 331).

Arrhenius

The hypothesis that many substances exist in solution as ions was proposed in 1887 by a young Swedish chemist, Svante Arrhenius. So radical was the idea that older chemists derided it, and Arrhenius all but ruined his career at its beginning by defending his hypothesis too vigorously. Later, as the weight of accumulating evidence supported him more and more strongly, Arrhenius won world-wide acclaim. He was for several years president of the University of Stockholm, and in later life Director of the Nobel Institute.

Today the idea of ions in solution follows naturally from the electrical structure of matter. We learn that some compounds are formed by the loss of electrons from one kind of atom to another, so that some of the atoms gain a positive charge and the others a negative charge; and we find no difficulty in imagining that a liquid like water can separate these electrically charged atoms. But in 1887 this modern picture of the atom was not even dreamed of. Neither the electron nor radioactivity had been discovered. Without this modern knowledge Arrhenius's contemporaries had good reason to consider his idea of neutral substances breaking up into electrically charged fragments a bit fantastic.

Only one sort of direct experimental evidence suggested the existence of charged particles in solution—the evidence from electrolysis. When a current passed through a solution, one kind of material patently migrated to the anode, another kind to the cathode, just as if the two kinds had opposite electrical charges. Faraday had supposed that passage of the

current *caused* the substance in solution to break down into ions, and this explanation of electrolysis was generally accepted.

Arrhenius boldy proposed that *in solution ions are formed, not by the passage of a current, but whenever an electrolyte dissolves.* Positive and negative charges are formed in equal numbers, so that both the solution and the original solute appear to be electrically neutral. Of the detailed evidence which Arrhenius presented in support of his theory, we shall mention only two of the principal items:

1. Reactions between electrolytes which take place instantaneously in solution often are very slow or do not occur at all if the electrolytes are dry. As a single example, consider the reaction between silver nitrate and sodium chloride. If solutions of the two electrolytes are mixed, a white *precipitate* (a solid which forms as a result of a chemical reaction in solution) of the insoluble substance silver chloride appears immediately, at first as a general cloudiness through the solution, later as tiny crystals which settle to the bottom. If the liquid is separated from these crystals by filtration and evaporated to dryness, another solid, sodium nitrate, will be left. The equation for the reaction may be written

$$AgNO_3 + NaCl \rightarrow AgCl \text{ (ppt.)} + NaNO_3 \text{ (in sol'n)}$$

Seemingly the silver and sodium have "exchanged partners." Now if dry salt is mixed with dry silver nitrate, no reaction of this sort occurs. Apparently the exchange of partners is aided by solution, which suggests that the "partners" are already free in the solutions, ready to react as soon as they are mixed.

2. Solutions of electrolytes have abnormally low freezing points. Any substance dissolved in water lowers its freezing point somewhat: we take advantage of this fact, for instance, by dissolving various materials in the water of automobile radiators during cold weather to keep the water from freezing. Careful study shows that the amount by which the freezing point is lowered depends only on the number of molecules of dissolved material present, not on the nature of the material. Equal numbers of molecules of sugar, alcohol, and glycerin dissolved in 1 l. of water lower the freezing point by almost exactly the same amount. But the same number of "molecules" of salt (assuming the molecule to be NaCl) lowers the freezing point nearly twice as much; so also does the same number of "molecules" of $MgSO_4$, KBr, $AgNO_3$. Arrhenius concluded that in solutions of these substances twice as many particles were actually present as would be indicated by the simple formulas, because the substances were broken down into ions. Similarly a substance like $CaCl_2$ lowers the freezing point nearly three times as much as sugar does, because each "molecule" is broken down into three particles—one calcium ion and two chloride ions. Boiling points and vapor pressures of

electrolytic solutions show similar abnormalities when compared with solutions of nonelectrolytes, and these abnormalities also find ready explanation in terms of ions.

Arrhenius's hypothesis, based on careful study of these and other lines of evidence, has been brilliantly confirmed by modern research into the electrical structure of the atom. Not all of Arrhenius's ideas are still accepted; he believed, for instance, that electrolytes consisted of separate molecules in the solid state, a hypothesis that has been disproved by X-ray studies of crystals. But Arrhenius's basic idea that electrolytes dissolve in water to form free ions capable of moving about in solution remains the foundation of all modern work on solutions in water.

Properties of Ions

One of the early objections to Arrhenius's theory was that sodium chloride was assumed to break down into separate particles of sodium and chlorine, yet the solution remains colorless. Why, if chlorine is present as free ions, should we not find the characteristic greenish-yellow color of chlorine in the solution? Arrhenius replied that chloride ion has altogether different properties from chlorine gas—a different color, a different taste, different chemical reactions. The answer is straightforward enough, but its full significance is not easy to grasp.

Henceforth we must regard a solution of sodium chloride not as a solution of NaCl, but as a solution of two substances with the formulas Na^+ and Cl^-. Further, *each of these substances has its own set of properties*, properties which are quite different from those of the active metal and the poisonous gas whose formulas are so similar, Na and Cl_2. We must change our point of view toward all electrolytic solutions, so that we may think of *each ion* as a new and separate material with characteristic properties of its own.

By "the properties of an ion," we mean, of course, the properties of solutions in which the ion occurs. A solution of a single kind of ion, all by itself, cannot be prepared: always positive ions and negative ions must be present together, so that the total number of charges of each sign is the same. But each ion gives its own characteristic properties to all solutions containing it, and these properties can be recognized whenever they are not masked by other ions. For example, a property of copper ion (Cu^{++}) is its blue color; all solutions of this ion are blue, unless some other ion is present which has a stronger color. A characteristic of hydrogen ion (H^+ or H_3O^+) is its sour taste, and all solutions containing this ion (acids) are sour. Silver ion (Ag^+) shows the property of forming a white precipitate, AgCl, when mixed with solutions of chloride ion (Cl^-); any solution of an electrolyte containing silver when mixed with a solution of any chloride will give this precipitate.

To demonstrate this last statement experimentally, portions of a silver nitrate solution may be poured into solutions of HCl, NaCl, $CuCl_2$, and $FeCl_3$, and then portions of a sodium chloride solution into solutions of $AgNO_3$, $AgC_2H_3O_2$ (silver acetate), and Ag_2SO_4. In each case, regardless of the other ions present, Ag^+ is being mixed with Cl^-, and in each case a white precipitate of AgCl appears.

To emphasize the difference between the properties of an ion and the properties of the corresponding neutral substance, we may set down in tabular form a comparison of the properties of the two substances Cl^- and Cl_2:

Cl_2	*Cl^-*
Greenish-yellow color	Colorless
Strong, irritating taste and odor	Mild, pleasant taste
Combines with all metals	Does not react with metals
Combines readily with hydrogen	Does not react with hydrogen
Does not react with Ag^+	Forms AgCl with Ag^+
Very soluble in CCl_4	Insoluble in CCl_4

In general, Cl_2 is much more active—as might be expected, since its atoms have only seven valence electrons while chloride ions have eight.

So for each ion we may write down a list of properties, which means a list of the properties common to all its solutions. In general, *the properties of a solution of an electrolyte are the sum of the properties of the ions which it contains.* The properties of sodium chloride are the properties of Na^+ and the properties of Cl^-; the properties of copper sulfate are the properties of Cu^{++} and SO_4^{--}. This fact is one of the reasons why the ionic hypothesis is so valuable in the study of solutions: instead of learning the individual properties of several hundred different electrolytes, we need only learn the properties of a few ions to be able to predict the properties of any electrolytic solution which contains them.

Ionic Equations

To understand equations for reactions among ions requires no new ideas, but only emphasis on the basic principle that *a chemical equation should be a summary of an actual chemical change.*

We may use the formula of a solid compound, just as in preceding chapters, to mean the relative numbers of atoms present. Thus the formula of solid sodium chloride is NaCl, even though the crystal actually consists of sodium ions and chloride ions. In solution, we write the formula of an electrolyte in terms of its ions, signifying that the ions are now independent substances rather than parts of a crystal lattice. The change from solid salt to salt in solution, for instance, may be summarized

$$NaCl \rightarrow Na^+ + Cl^-$$

The equation means that sodium ions and chloride ions have been set free to move about in the solution. If the solution is cooled or evaporated so that solid sodium chloride crystallizes out, the re-formation of the crystal lattice may be written

$$Na^+ + Cl^- \rightarrow NaCl$$

Once an electrolyte is dissolved, each ion is independent and its reactions need not involve the other ions present. For example, consider again the reaction between silver nitrate and sodium chloride discussed on page 343. Both $AgNO_3$ and NaCl, when dissolved, are completely dissociated into ions. Of the products, $NaNO_3$ is soluble and completely ionized, but AgCl is a solid precipitate not ionized at all. Hence we might write the equation

$$Ag^+ + NO_3^- + Na^+ + Cl^- \rightarrow AgCl + Na^+ + NO_3^- \qquad (31)$$

On each side of this equation appear the formulas Na^+ and NO_3^-; evidently these substances have not changed during the reaction but have merely remained in solution. Actually the only chemical change which has occurred is the disappearance of Ag^+ and Cl^- and the formation of solid AgCl.

$$Ag^+ + Cl^- \rightarrow AgCl \qquad (32)$$

This brief equation is *a complete summary of the reaction.* The other ions are present but take no part. If KCl had been used instead of NaCl, or $AgC_2H_3O_2$ instead of $AgNO_3$, the ions would have been different but the actual chemical change would still have been the union of Ag^+ and Cl^- to form AgCl. Equation (31) is not wrong, but Eq. (32) emphasizes more clearly that the reaction with chloride ion is a property of Ag^+ and not of some particular silver compound.

A second very similar example is furnished by the precipitation of the white solid calcium carbonate ($CaCO_3$) when a solution of calcium ion is added to a solution of carbonate ion. If $CaCl_2$ is used to supply Ca^{++} and Na_2CO_3 to supply CO_3^{--}, the equation may be written

$$Ca^{++} + 2Cl^- + 2Na^+ + CO_3^{--} \rightarrow CaCO_3 + 2Cl^- + 2Na^+$$

Sodium ions and chloride ions take no part in the reaction, so we may simplify the equation to

$$Ca^{++} + CO_3^{--} \rightarrow CaCO_3 \qquad (33)$$

When two electrolytes are mixed, *no reaction occurs unless some unionized substance like AgCl and $CaCO_3$ can form.* If solutions of $Cu(NO_3)_2$ and NaCl are mixed, for instance, a possible reaction might be an exchange of partners to form $NaNO_3$ and $CuCl_2$. But both of these are sol-

uble salts, completely ionized in solution, so no reaction can occur. In symbols

$$Cu^{++} + 2NO_3^- + Na^+ + Cl^- \rightarrow Cu^{++} + Cl^- + Na^+ + 2NO_3^-$$

The same substances are represented on each side of the "equation," so that no chemical change has taken place.

For another example of an ionic process, consider the reaction between active metals and acids which we discussed in an earlier chapter (page 181). Without bothering about ions, we wrote equations like

$$Zn + H_2SO_4 \rightarrow H_2 + ZnSO_4$$

Let us try to revamp this equation in terms of ions. Zinc is a solid metal, hence not ionized; H_2 represents the covalent molecule of a gas, also un-ionized. Sulfuric acid and zinc sulfate, however, are ionized in solution:

$$Zn + 2H^+ + SO_4^{--} \rightarrow H_2 + Zn^{++} + SO_4^{--}$$

This is one way to write the ionic equation, but inspection shows that it can be simplified. SO_4^{--} appears on both sides of the equation, and so may be omitted.

$$Zn + 2H^+ \rightarrow Zn^{++} + H_2 \qquad (34)$$

This shorter statement implies that the only chemical changes are the formation of zinc ion from zinc metal and the formation of hydrogen gas from hydrogen ion.

For important ionic processes which we shall meet in future chapters we shall set down both the "complete" ionic equations and the simplified equations with nonessentials omitted. The former will show the specific substances used, the latter the actual chemical changes. Both will emphasize that solutions of electrolytes contain independent ions, each with its own separate properties.

Questions

1. Why is the solubility of one gas in another unlimited?
2. How could you tell experimentally whether or not a solution of sugar is saturated?
3. Give an example of (*a*) a polar liquid, (*b*) two nonpolar liquids; and give examples of substances which are soluble in each.
4. If a solution of potassium nitrate is just saturated at 50°C, how many grams of KNO_3 are dissolved in every 100 g. of water (Fig. 173)? If the solution is heated, will it remain saturated, become unsaturated, or become supersaturated?
5. How much does the solubility of KNO_3 change between 30° and 40°C? How much does the solubility of NaCl change between these two temperatures?
6. Can you suggest any evidence to show that the solubility of gases decreases as the temperature rises?
7. Write the formulas of the ions which would be present in dilute solutions of each of the following substances and name each ion. KOH, $AgNO_3$, $MgSO_4$, $CuCl_2$, Na_2CO_3.

8. Give the formulas and valences of the following ions: nitrate ion, calcium ion, carbonate ion, aluminum ion, sulfide ion, bromide ion.
9. How could you distinguish experimentally between an electrolyte and a non-electrolyte?
10. Contrast the properties and electron structures of Na and Na^+. Which would you expect to be more chemically active?
11. Name one property by which you could distinguish
    Cl^- from NO_3^-.
    Ag^+ from Na^+.
    Ca^{++} from Na^+.
    Cu^{++} from Ca^{++}.
12. Write an equation showing what happens when (*a*) a solution of magnesium sulfate is evaporated to dryness, (*b*) sodium carbonate is dissolved in water.
13. Write ionic equations for the reactions between the following:
    Silver nitrate and potassium chloride.
    Calcium and hydrochloric acid (giving hydrogen gas).
    Calcium nitrate and sodium carbonate.
14. What is the actual chemical change shown in the following equation?

$$Ba^{++} + NO_3^- + Na^+ + SO_4^{--} \rightarrow BaSO_4 \text{ (precipitate)} + Na^+ + NO_3^-$$

Write a simplified equation to show this change.

CHAPTER XXV

Acids, Bases, and Salts

WE CONTINUE our investigation into the behavior of ions in solution by considering now the three important classes of electrolytes: acids, bases, and salts.

Acids

Acids have been described on an earlier page (page 189) as substances containing hydrogen whose water solutions taste sour and change the color of a dye called litmus from blue to red. Strong acids like hydrochloric acid (HCl) and sulfuric acid (H_2SO_4) are poisonous, cause painful burns if allowed to remain on the skin, and are injurious to clothing. Weak acids like carbonic acid (H_2CO_3) and citric acid, far from being harmful, add a pleasant sour taste to foods and drinks. Acids are widely used in industry for cleaning metal surfaces, as solvents, and to control chemical reactions; the work of acids can be traced in many natural processes; control of acidity in the human body is important in medicine.

What is it that gives an acid its characteristic properties? Following the line of reasoning in the last chapter, we might suppose that these hydrogen compounds ionize on solution in water, and that "acid" properties are simply properties of the hydrogen ion set free:

$$HCl \rightarrow H^+ + Cl^-$$

This equation is roughly correct, and for some purposes it is convenient to speak of an acid solution as one that contains free hydrogen ions. To be accurate, however, the equation should be a little more complicated. Acids differ from substances like NaCl and $AgNO_3$ in that their molecules are *covalent*, meaning that the hydrogen is attached to the rest of the molecule by a shared electron pair (page 327). Acids which are liquid in the pure state, for instance, do not conduct an electric current because they are not ionized—whereas melted NaCl conducts a current readily by movement of its ions. Nor do acids ionize enough to conduct a current when they are dissolved in nonpolar liquids like carbon tetrachloride

and toluene. *Only when an acid is dissolved in certain polar liquids like water, alcohol, or liquid ammonia do its acid properties appear.* This fact suggests that an "acid" solution is formed only when the original acid *reacts with* a polar solvent.

When hydrochloric acid dissolves in water, for instance, HCl molecules do not simply break up into ions. They must be pulled apart by the water; and the pulling apart is principally a reaction of the hydrogen with H_2O molecules:

$$HCl + H_2O \rightarrow H_3O^+ + Cl^- \qquad (35)$$

The ion H_3O^+ is a combination of H^+ with H_2O. We could write it $H^+.H_2O$, and call it *hydrated hydrogen ion*, but more commonly it is abbreviated to H_3O^+ and called *hydronium ion*. It is present in all solutions of acids in water, and it gives these solutions their common properties. When we say that acid solutions taste sour, turn litmus pink, and liberate hydrogen gas by reaction with metals, we mean that the hydronium ion will do these things.

Equation (35) shows, in effect, that a hydrogen ion, H^+, has left an HCl molecule and attached itself to an H_2O molecule. The ion H^+ is a hydrogen atom that has lost an electron; since the normal H atom consists simply of one electron and one proton, the H^+ particle is the proton alone. We could say, then, that HCl has given up a proton to H_2O. Any acid dissolving in water behaves similarly: The essential characteristic of an acid is its possession of protons that can be transferred to another substance in this manner. We may define an ***acid*** concisely as *any substance that can give up protons to another substance.* This is a very broad definition; for most purposes acids include only substances that can give up protons to water molecules to form hydronium ion. An *acid solution* in water is always a solution containing this ion.*

When we speak of the "ionization of an acid," we mean a process like that shown by Eq. (35), in which protons are transferred to the solvent. Some acids, called ***strong acids***, give all their available protons to the solvent, and thus are completely ionized. Others, called ***weak acids***, give up only a small fraction of their available protons; solutions of weak

* Many chemistry texts, particularly older ones, discuss acids in terms of the simple hydrogen ion H^+, rather than the hydronium ion. Expressions like "hydrogen ion activity" and "hydrogen ion concentration" are common in applications of chemistry to agriculture, medicine, and industry. For most purposes it makes little difference whether hydrogen ion or hydronium ion is used.

Hydronium ion seems preferable here, both because experiment has demonstrated its presence in acid solutions and because reactions of acids are simplified if we describe them all as transfers of protons from one substance to another. At first sight there seems little more reason to write H_3O^+ than to write $Na(H_2O)^+$ or $Cl(H_2O)^-$ (page 340); but these latter hydrates are known to be less stable, and their use would make ionic reactions more complicated rather than more simple.

acids contain mostly un-ionized molecules, together with a small concentration of ions. The three most common strong acids are HCl (hydrochloric), H_2SO_4 (sulfuric), and HNO_3 (nitric). Some familiar weak acids are acetic acid ($HC_2H_3O_2$), the acid in vinegar; carbonic acid (H_2CO_3), the acid in soda water; boric acid (H_3BO_3), found in every medicine cabinet; and citric acid, the acid of citrus fruits. The greater ionization of strong acids means that in solutions of similar concentration a strong acid has a much larger amount of hydronium ion than a weak acid. The contrast may be strikingly shown by comparing dilute solutions of hydrochloric and acetic acids: the HCl has a much sourer taste, it is a better conductor of electricity, and if the two acids are poured on zinc the evolution of hydrogen gas is much faster from the HCl.

Note that the *stronger* an acid, the *weaker* is the attachment of protons in its molecules. For a strong acid like HCl the attachment is so weak that a dilute water solution consists entirely of H_3O^+ and Cl^- ions, with no HCl molecules at all. For a weak acid like $HC_2H_3O_2$, on the other hand, the attachment is strong enough so that most of the molecules remain undissociated; a solution consists chiefly of $HC_2H_3O_2$ molecules, with only minor amounts of H_3O^+ and $C_2H_3O_2^-$. In a solution that contains 6 g. of acetic acid in 1 l., for instance, about 1 molecule in every 100 is ionized. Carbonic acid (H_2CO_3) is a much weaker acid than acetic acid, which means that the attachment of CO_3^{--} to H^+ is stronger than the attraction between $C_2H_3O_2^-$ and H^+; in a saturated solution 999 molecules out of every 1,000 remain as undissociated H_2CO_3, only 1 breaking down to form H_3O^+ and CO_3^{--}.

Bases

A substance whose molecules combine with protons set free by an acid is called a base. We may define *bases* formally as *substances capable of taking up protons*. Like the definition of acids above, this is a very broad statement; it includes as bases substances which ordinarily are not considered bases at all. For instance, in Eq. (35) the substance which takes up protons is water, although normally we do not regard water as a base.

A common base is the ion OH^- (hydroxide ion) which reacts with protons to form water.

$$OH^- + H^+ \rightarrow H_2O$$

Other common bases are ions of weak acids, like acetate ion ($C_2H_3O_2^-$) and carbonate ion (CO_3^{--}). These have a sufficient attraction for protons to take protons away from water molecules.

$$CO_3^{--} + 2H_2O \rightarrow H_2CO_3 + 2OH^- \quad (36)$$

$$C_2H_3O_2^- + H_2O \rightarrow HC_2H_3O_2 + OH^- \quad (37)$$

Many other bases react similarly with water to give OH^-. Just as H_3O^+ is the characteristic ion of acid solutions, so OH^- is the characteristic ion in water solutions of bases. We have earlier described bases in terms of their common properties, particularly their bitter taste and their ability to turn red litmus blue (page 190); these are properties of the hydroxide ion in water solution.

Since OH^- is the characteristic ion of basic solutions, substances which dissociate to give this ion directly are often called bases. Such substances are the soluble hydroxides, of which the three commonest are NaOH, KOH, and $Ba(OH)_2$. KOH is ordinary lye or caustic potash, and NaOH is soda lye or caustic soda; these substances are quite as poisonous and quite as destructive to flesh and clothing as are the strong acids. Like the strong acids they are extensively used in chemical industry. Another hydroxide that gives a fairly basic solution is calcium hydroxide [$Ca(OH)_2$, ordinary limewater]; the amount of OH^- obtainable is limited because the compound is not very soluble. All other common hydroxides, like $Al(OH)_3$, $Fe(OH)_3$, $Cu(OH)_2$, are so insoluble in water that the amount of OH^- obtainable from them is negligible.

A substance which takes up protons readily is a ***strong base,*** *one which takes them up less readily a* ***weak base.*** For solutions in water this means that strongly basic solutions have high concentrations of OH^-, weakly basic solutions only small amounts of OH^-. The hydroxides of sodium, potassium, and barium give strongly basic solutions; solutions of carbonate ion are moderately basic, solutions of acetate ion weakly basic. The difference may be illustrated by preparing dilute solutions of NaOH, Na_2CO_3, $NaC_2H_3O_2$, and NaCl, all of similar concentration; the characteristic bitter taste is most pronounced in the hydroxide solution, less in the carbonate, still less in the acetate, and absent in the chloride.

Weakly basic solutions are common household items, useful as cleaning agents because of their ability to dissolve, or "cut," grease. Ordinary soap gives a slightly basic solution, but its cleaning ability is due more to its emulsifying action (page 436) than to its basic nature. For difficult cleaning jobs more basic solutions are commonly used—ammonia water (see below), a solution of washing soda (sodium carbonate, Na_2CO_3), or a solution of borax ($Na_2B_4O_7$, which ionizes to give the weak base $B_4O_7^{--}$). Especially stubborn dirt may require still more basic solutions, obtained with lye (KOH or NaOH).

Another name often used for substances that dissolve to give basic solutions is *alkali.* The alkali that crystallizes in desert basins is Na_2CO_3 ("white alkali") or K_2CO_3 ("black alkali"). Because NaOH and KOH are strong alkalies, sodium and potassium are often called "alkali metals" (page 200). An *alkaline solution* is any solution with appreciable quantities of OH^-; the terms *alkaline* and *basic* are practically synonyms.

Reactions between Strong Acids and Bases

When sodium hydroxide solution is slowly added to hydrochloric acid, there is no visible sign that a reaction is taking place. Both of the original solutions are colorless, and the resulting solution is colorless also. That a reaction does occur may be shown in several ways: (1) The mixture becomes warm, showing that chemical energy is changing to heat energy; (2) the taste of the acid becomes less and less sour as the base is added; (3) if small samples of the acid are taken out while the base is being added, the samples show progressively less and less active evolution of hydrogen gas when poured on zinc. The base evidently destroys, or *neutralizes*, the characteristic acid properties, and the reaction is called *neutralization*. In the same way the characteristic properties of a base may be neutralized by adding a strong acid.

What is the chemical change in the neutralization of HCl by NaOH? We could write for a rough preliminary equation

$$HCl + NaOH \rightarrow H_2O + NaCl$$

To make the equation more precise, consider the ions involved. HCl, a strong acid, ionizes completely in water to give H_3O^+ and Cl^-; NaOH, a strong base, ionizes into Na^+ and OH^-; the product NaCl, likewise a soluble electrolyte, remains ionized in solution. Of the four substances shown, only water is a nonelectrolyte, so it should appear un-ionized in the equation

$$H_3O^+ + Cl^- + Na^+ + OH^- \rightarrow 2H_2O + Na^+ + Cl^-$$

Since Na^+ and Cl^- appear on both sides, they may be omitted.

$$H_3O^+ + OH^- \rightarrow 2H_2O \qquad (38)$$

This is the actual chemical change, stripped of all nonessentials. *The neutralization of a strong acid by a strong base in water solution is essentially a transfer of protons from hydronium ion to hydroxide ion, forming water.*

To carry out a neutralization reaction in the laboratory, some method is necessary to determine when just enough acid has been added to neutralize all the base present, or vice versa. Otherwise the resulting solution will be either acidic or basic, depending on which is in excess. One convenient method for determining the "neutral point" is to add a few drops of litmus solution to the base before the reaction is carried out, giving the basic solution a blue color. As the acid is added, the blue color persists until all the base is used up, then changes to pink. Similarly, if a base is to be added to an acid, the acid may first be made pink with litmus solution; then its color will change to blue as soon as enough base has been added to neutralize the acid.

A substance like litmus, whose color enables a chemist to tell whether a solution is acid or basic and whose sharp change in color shows him when neutralization has occurred, is called an *indicator*. Many different indicators are used in chemical laboratories, some being more useful than litmus because their color changes during neutralization are more abrupt. Two common ones are phenolphthalein, which is pink in basic solution and colorless in acid, and methyl orange, which is yellow in basic solution and salmon pink in dilute acid. All indicators are complex compounds containing carbon, hydrogen, and oxygen; several are commercial dyes. One familiar "indicator" is the red coloring matter of cherry juice: when a red cherry stain is washed with soap (which gives a weakly alkaline solution), its color changes abruptly to blue, showing that the acid of the fruit juice has been neutralized.

Salts

When NaOH solution is neutralized with HCl, the resulting solution should contain nothing but the ions Na^+ and Cl^-. If the solution is evaporated to dryness, the ions combine to form the white solid NaCl. This substance, ordinary ***salt,*** gives its name to an important class of compounds, most of which are crystalline solids at ordinary temperatures and most of which consist of a metal combined with one or more nonmetals. Typical salts are KBr, $MgSO_4$, $Al(NO_3)_3$, and $ZnCO_3$. Crystal lattices of salts consist of alternate positive and negative ions, which means that practically all soluble salts will ionize* completely in water solution. No salt is completely insoluble in water, but some like AgCl and $CaCO_3$ are so very slightly soluble that we often refer to them as "insoluble"; such salts can ionize, of course, only to a very small extent. The following summary should aid in remembering which salts are soluble, hence ionized in solution, and which are practically insoluble:

All nitrates (salts containing the NO_3 group) are soluble.
All acetates (salts containing the $C_2H_3O_2$ group) are soluble.
All chlorides are soluble, except AgCl and a very few others.
All sulfates are soluble, except $BaSO_4$ and a few others.
All carbonates are insoluble, except Na_2CO_3, K_2CO_3, and $(NH_4)_2CO_3$.
All sulfides are insoluble, except Na_2S, K_2S, $(NH_4)_2S$, CaS, and BaS.
All salts of Na, K, and NH_4 are soluble.

Any salt may be formed by mixing the appropriate acid and base and evaporating the solution to dryness. Thus KNO_3 is formed when solutions of KOH and HNO_3 are mixed and evaporated; $CuSO_4$ is formed when H_2SO_4 is poured on the insoluble hydroxide $Cu(OH)_2$ and the resulting solution is evaporated. We could say in general that many neutral-

* Remember that "ionize" is used in the sense of "dissociate into free ions."

ization reactions (in water solution) give water and a solution of a salt. It is important to remember, however, that the salt is not produced directly by the neutralization. Neutralization is essentially a reaction between substances which liberate protons and substances which take up protons; as a result of this process ions may be left in solution which on evaporation will unite to form a salt.

Reactions Involving Weak Acids and Weak Bases

Consider the reaction between NaOH and a typical weak acid, acetic acid ($HC_2H_3O_2$). (Note that in $HC_2H_3O_2$ only one of the four H's has any tendency to form H_3O^+ in solution; the other three remain always a part of the acetate ion $C_2H_3O_2^-$.) As NaOH is added, the vinegarlike odor of $HC_2H_3O_2$ decreases and finally disappears. We might write the reaction roughly

$$NaOH + HC_2H_3O_2 \rightarrow NaC_2H_3O_2 + H_2O$$

The $NaC_2H_3O_2$ (sodium acetate) could be obtained as a white salt by evaporating the solution. But for present purposes it is more important to consider the actual ionic process involved in the neutralization, rather than the salt that could be obtained by evaporation. NaOH, of course, ionizes to give the strong base OH^-; $HC_2H_3O_2$, being a weak acid, is present in solution chiefly as un-ionized molecules. So we write

$$Na^+ + OH^- + HC_2H_3O_2 \rightarrow Na^+ + C_2H_3O_2^- + H_2O$$

Or, eliminating Na^+ which appears on both sides,

$$OH^- + HC_2H_3O_2 \rightarrow C_2H_3O_2^- + H_2O \qquad (39)$$

The acetate ion ($C_2H_3O_2^-$) we have described as a weak base, so we might say that the essential process here is the combination of a strong base and a weak acid to form a weak base and water.

The acetate ion is strong enough as a base to take some protons away from water and liberate OH^- in solution [Eq. (37)]. This means that the neutralization process [Eq. (39)], which is precisely the reverse of Eq. (37), cannot go to completion. If original quantities of solution are used which contain equal numbers of OH^- ions and $HC_2H_3O_2$ molecules, then when the two are mixed the final solution will not consist simply of Na^+ and $C_2H_3O_2^-$, but will contain some unused OH^- and $HC_2H_3O_2$. If we make another solution by simply dissolving $NaC_2H_3O_2$ in water, we would find a similar composition—chiefly Na^+ and $C_2H_3O_2^-$, but a little OH^- and $HC_2H_3O_2$— since the reaction of Eq. (37) will proceed to some extent In effect, the two bases OH^- and $C_2H_3O_2^-$ are competing for protons; the OH^- is by far the stronger base, so it gets most of the protons, but the acetate ion holds on to some.

In general, if a strong base is added to a weak acid, molecule for molecule, the resulting solution is somewhat basic. And in general a salt containing a metal and a basic negative ion gives a basic solution, unless the metal happens to react with water to form an acid. Thus when $HC_2H_3O_2$ reacts molecule for molecule with $Ba(OH)_2$ the solution is slightly basic, and $Ba(C_2H_3O_2)_2$ gives a basic solution when dissolved in water; KOH reacting molecule for molecule with carbonic acid (H_2CO_3) remains slightly basic, and K_2CO_3 gives an alkaline solution.

We might predict that similar rules would hold for reactions between strong acids and weak bases. Consider, for instance, the neutralization of the weak base $C_2H_3O_2^-$ with HCl. If we use sodium acetate as a source of acetate ion, the reaction can be written

$$NaC_2H_3O_2 + HCl \rightarrow HC_2H_3O_2 + NaCl$$

To write the ionic equation, we need only remember that all these substances except acetic acid are completely ionized.

$$\left.\begin{aligned} Na^+ + C_2H_3O_2^- + H_3O^+ + Cl^- \\ \rightarrow HC_2H_3O_2 + Cl^- + Na^+ + H_2O \\ C_2H_3O_2^- + H_3O^+ \rightarrow HC_2H_3O_2 + H_2O \end{aligned}\right\} \quad (40)$$

The fundamental reaction, then, produces acetic acid and water from the weak base $C_2H_3O_2^-$ and the strong acid H_3O^+. In general, a strong acid plus a weak base gives a weak acid and water. This fact provides a general laboratory method for preparing weak acids: thus hydrocyanic acid ("prussic acid") is formed when HCl is added to potassium cyanide (KCN),

$$CN^- + H_3O^+ \rightarrow HCN + H_2O$$

and carbonic acid is formed when sulfuric acid is added to sodium carbonate,

$$CO_3^{--} + 2H_3O^+ \rightarrow H_2CO_3 + 2H_2O$$

In many of the above reactions, molecules of water either take up or lose protons. When a strong acid dissolves in water, protons are added to H_2O [Eq. (35)]; when a base dissolves in water [Eqs. (36) and (37)], protons are removed from H_2O. Thus water behaves in some reactions like a base, in others like an acid.

If we regard water as a very weak acid in some reactions, a very weak base in others, we may restate some of the above rules for reactions of acids and bases in a general form:

A strong acid plus a strong base gives a very weak acid and a very weak base. [Equation (38); here one molecule of water is an "acid," the other a "base."]

A strong acid plus a weak base gives a weak acid and a very weak base. [Equation (40); here water is a base.]

A strong base plus a weak acid gives a weak base and a very weak acid. [Equation (39); here water is an acid.]

These are general rules which apply to all acid-base reactions, even those which do not take place in water solution. They are simply an expression of the fact that in all these reactions the fundamental process is transfer of protons from one substance to another.

To use these rules we need to know the relative strengths of common acids and bases. The following table shows acids in the left-hand column arranged in order from strong to weak, and bases in the right-hand column arranged in order from weak to strong.

TABLE XIX. STRENGTHS OF ACIDS AND BASES

	Acids	*Bases*	
Strong	HCl	Cl^-	Very weak
	H_2SO_4	SO_4^{--}	
	H_3O^+	H_2O	
Weak	$HC_2H_3O_2$	$C_2H_3O_2^-$	Weak
	H_2CO_3	CO_3^{--}	
	H_2S	S^{--}	
	NH_4^+	NH_3	
Very weak	H_2O	OH^-	Strong

Any acid in this table should react completely or fairly completely with any base *below* it, according to the rules in the paragraph above. For instance, we might predict that $HC_2H_3O_2$ would react with S^{--}:

$$2HC_2H_3O_2 + S^{--} \rightarrow H_2S + 2C_2H_3O_2^-$$

We can verify the prediction experimentally by dropping a chunk of sodium sulfide (Na_2S) in vinegar (largely $HC_2H_3O_2$); the rotten-egg odor of the H_2S set free is at once apparent. On the other hand, we could predict that a solution of H_2S would react only very slightly with $C_2H_3O_2^-$, since the base is *above* the acid in the table. This prediction can be tested by adding a hydrogen sulfide solution to sodium acetate; not enough acetic acid is formed to be detected by its vinegarlike odor.

Thus the table and the simple rules for using it make possible predictions about the behavior of any common acid-base combination.

Ammonia and Ammonium Ion

The base NH_3 (ammonia) merits some individual attention. Unlike most common bases ammonia is a neutral molecule rather than a negative ion. The pure substance is a gas with a strong odor, extremely soluble in water; the water solution, called "ammonia water," is a widely used cleaning agent.

Ammonia is a moderately strong base, as is shown by its position near the bottom of Table XIX. It reacts readily with all common acids, either in the pure state or in water solution. With gaseous hydrogen chloride, for instance, ammonia gas reacts to form the white solid NH_4Cl (ammonium chloride):

$$NH_3 + HCl \rightarrow NH_4Cl$$

This reaction is conspicuous when bottles of HCl and ammonia solutions are opened side by side; the two gases escaping from the solutions mix and form a dense cloud consisting of tiny particles of NH_4Cl.

Ammonium chloride (often called "sal ammoniac"), although made up of three nonmetals, is a white crystalline substance which behaves like a salt. Its crystal lattice is made up of the ions NH_4^+ and Cl^-, the NH_4^+ apparently playing the role of a metal ion like Na^+. When the salt dissolves in water, the two ions are set free in the solution. Thus a reaction between two covalent substances, NH_3 and HCl, by a transfer of protons gives an ionic compound. Similar ionic compounds, called in general *ammonium salts*, are formed by the reaction of ammonia with other acids—ammonium sulfate $[(NH_4)_2SO_4]$; ammonium nitrate (NH_4NO_3); ammonium carbonate $[(NH_4)_2CO_3]$; and so on.

The same sort of reaction takes place when ammonia is added to an acid in solution rather than to a pure acid. If the acid is strong, the principal substance in solution is H_3O^+ and the reaction is

$$NH_3 + H_3O^+ \rightarrow NH_4^+ + H_2O \qquad (41)$$

In other words, NH_3 and H_2O compete for the available protons; since NH_3 is a much stronger base (Table XIX), most of the protons combine with it rather than staying with H_2O. Ammonia is not strong enough, however, to acquire all the available protons; if amounts of solution are used which contain equal numbers of NH_3 molecules and H_3O^+ ions, the mixture is still sufficiently acid to turn litmus pink. In other words, the reaction shown in Eq. (41) does not continue until all the H_3O^+ is used up. This means that we could start the reaction in the opposite direction, say by dissolving ammonium chloride in water, and the ammonium ion would react to a slight extent with H_2O:

$$NH_4^+ + H_2O \rightarrow NH_3 + H_3O^+$$

In other words, a solution of ammonium chloride (or of several other ammonium salts) is slightly acidic.

Although NH_3 is above OH^- in Table XIX, it is a strong enough base to react slightly with water:

$$NH_3 + H_2O \rightarrow NH_4^+ + OH^- \qquad (42)$$

This reaction cannot go far, because OH^- has a greater attraction for protons than NH_3 (in other words, OH^- is the stronger base); but enough OH^- is produced by the reaction to give ammonia water the bitter taste, the ability to react with litmus, and the ability to dissolve grease which are characteristic of alkaline solutions.*

Since the reaction shown by Eq. (42) does not go far, we might predict that the reverse process would go nearly to completion. If we add a sodium hydroxide solution to ammonium chloride, for instance, we should expect to have NH_3 liberated in solution:

$$\left.\begin{aligned} Na^+ + OH^- + NH_4^+ + Cl^- &\rightarrow NH_3 + H_2O + Na^+ + Cl^- \\ OH^- + NH_4^+ &\rightarrow NH_3 + H_2O \end{aligned}\right\} \quad (43)$$

This is a typical example of a strong base added to a weak acid (NH_4^+) giving a moderately weak base (NH_3) and a very weak acid (H_2O). We can show that the reaction has taken place by gently heating the mixture: the odor of ammonia gas, vaporized by the heating, is at once apparent. Similar reactions take place between solid hydroxides and ammonium salts:

$$Ca(OH)_2 + (NH_4)_2SO_4 \rightarrow CaSO_4 + 2NH_3 + 2H_2O$$

These reactions with strong bases, either in solution or in the solid state, provide a delicate test for the presence of ammonium salts.

Acidic and Basic Oxides

In an earlier chapter (page 196) we learned that one important distinction between metals and nonmetals is the fact that oxides of the former often react with water to form bases, while oxides of the latter often react with water to form acids. For example,

$$CaO + H_2O \rightarrow Ca(OH)_2 \quad (44)$$

$$SO_3 + H_2O \rightarrow H_2SO_4 \quad (45)$$

$$CO_2 + H_2O \rightarrow H_2CO_3 \quad (46)$$

Using ions, we can now rewrite the first two equations more accurately

$$CaO + H_2O \rightarrow Ca^{++} + 2OH^-$$

$$SO_3 + 3H_2O \rightarrow 2H_3O^+ + SO_4^{--}$$

Equation (46) we do not change, since H_2CO_3 is a weak acid and therefore ionizes only slightly.

Not only do these oxides dissolve in water to give acidic and basic solutions, but in some reactions they may themselves play the roles of

* An ammonia solution is often regarded as a solution of the hypothetical compound NH_4OH (ammonium hydroxide) but there is little positive evidence for the existence of this substance. In most reactions it makes little difference whether the formula of ammonia solution is written NH_4OH or $NH_3 + H_2O$.

acids or bases. For example, Na_2O can "neutralize" H_2SO_4 quite as effectively as can NaOH, in the sense that it destroys the acidic properties:

$$\left.\begin{aligned} Na_2O + 2H_3O^+ + SO_4^{--} &\rightarrow 2Na^+ + 3H_2O + SO_4^{--} \\ Na_2O + 2H_3O^+ &\rightarrow 2Na^+ + 3H_2O \end{aligned}\right\} \quad (47)$$

Some oxides (CuO, Fe_2O_3, SiO_2) will not dissolve in water. Their relationship with acids and bases is shown, however, by the fact that they can be prepared by reactions of the type

$$Cu(OH)_2 \rightarrow CuO + H_2O$$
$$2Fe(OH)_3 \rightarrow Fe_2O_3 + 3H_2O$$
$$H_4SiO_4 \rightarrow SiO_2 + 2H_2O$$

These equations show what happens when copper hydroxide, iron hydroxide, and silicic acid are heated.

Because any oxide may be regarded as derived from an acid or base (sometimes the acid or base is purely hypothetical), oxides of metals in general are referred to as *basic oxides* and oxides of nonmetals as *acidic oxides*. Thus SiO_2 (the chief constituent of ordinary sand) is called an acidic oxide, although it is insoluble in water and so neither tastes sour nor turns litmus red, while Fe_2O_3 (the chief constituent of hematite ore) is called a basic oxide, although it neither tastes bitter nor turns litmus blue.

Carbonic Acid and Carbonates

From these many pages it should be evident how indispensable are the ions of Arrhenius's theory in discussing the intricate chemistry of electrolytes. For reactions between acids and bases we add the idea of proton transfer, usually between ions or between ions and molecules. We can hardly claim that the subject of acids and bases is made simple by this interpretation; the experimental facts themselves are too varied and too complex for any one theory to make them appear really simple. But ions and proton transfers do furnish a rational basis on which a bewildering variety of observational data may be linked together.

Unfortunately, the principles underlying the behavior of acids and bases can be made clear only by using simple compounds which are familiar enough in chemical laboratories, but which are not often encountered in everyday life—substances like $NaC_2H_3O_2$, NH_4Cl, NaOH. But application of the principles is by no means limited to these playthings of the chemist. In a multitude of ways the control of acidity and alkalinity affects our lives and natural processes in the world about us. Every farmer knows the importance of maintaining the proper alkalinity in soil to promote the growth of plants. Our bodies maintain a small concentration of OH^- in our blood, and if the concentration varies

slightly from its normal value we become seriously ill. A small amount of H_3O^+ in our stomachs is necessary for digestion. Such industrial processes as dyeing, bleaching, fermentation, and the manufacture of dyes and drugs depend on a careful regulation of the concentration of H_3O^+ and OH^-. But rather than multiply examples, we shall pick one acid, carbonic acid, which we encounter every day of our lives, and see how the principles outlined on preceding pages apply to its behavior.

The Chief Facts. Carbon dioxide is the oxide of a nonmetal, hence an *acidic oxide*, which dissolves readily in water to form the weak acid H_2CO_3, carbonic acid—not to be confused with the poisonous carbolic acid, an altogether different substance. Since CO_2 is a minor constituent of the atmosphere, any water in contact with air becomes a very dilute solution of carbonic acid. Because it dissolves so very readily to form an acid, CO_2 is often called commercially "carbonic acid gas."

Carbonic acid is *extremely unstable*, decomposing into CO_2 and H_2O.

$$H_2CO_3 \rightarrow H_2O + CO_2$$

It cannot be prepared in pure form and even in water can exist only as a dilute solution; under normal pressure and temperature a saturated solution contains about 0.25 g. in every liter. The acid is decomposed whenever a solution is heated or allowed to evaporate.

Carbonic acid *ionizes very slightly.*

$$H_2CO_3 + 2H_2O \rightarrow 2H_3O^+ + CO_3^{--}*$$

It is neutralized by bases, forming salts called *carbonates.* All carbonates are insoluble in water except Na_2CO_3, K_2CO_3, and $(NH_4)_2CO_3$. Many of the insoluble carbonates occur in nature as minerals.

Now let us see where these facts and the principles of preceding paragraphs can lead us.

Carbonic Acid. Since CO_2 is a gas, we should expect its solubility, and therefore the concentration of H_2CO_3, to increase when pressure is increased and when temperature is decreased. Since H_2CO_3 is a very weak acid, we should expect its solution to have a slightly sour taste, to turn litmus faintly pink, and to liberate hydrogen very slowly by reaction with metals.

These expectations are amply confirmed by ordinary soda water, which is a solution of CO_2 under pressure. When soda water stands in an open glass, the decreased pressure allows H_2CO_3 to decompose and bubbles of CO_2 to escape all through the solution. Heating the solution makes the bubbles form faster. The evolution of gas and the pleasantly sour taste are, of course, the qualities which soda water adds

* For strict accuracy, this equation should be written in two steps:

$H_2CO_3 + H_2O \rightarrow H_3O^+ + HCO_3^-$ and $HCO_3^- + H_2O \rightarrow H_3O^+ + CO_3^{--}$

to drinks. In the laboratory we could easily show its slight effects on indicators and on active metals.

We might further anticipate that a base in a solution exposed to the air would be slowly neutralized by absorbing CO_2. This reaction is most strikingly shown by allowing limewater [$Ca(OH)_2$] to stand in an open dish. Gradually the solution grows cloudy as the insoluble salt $CaCO_3$ is formed:

$$Ca^{++} + 2OH^- + CO_2 \rightarrow CaCO_3 + H_2O$$

This reaction is often used to test for the presence of CO_2 in a gas mixture.

Carbonates. We should expect that a carbonate will form whenever a base is neutralized by H_2CO_3 or CO_2, as in the reaction just mentioned. Insoluble carbonates like $CaCO_3$, $ZnCO_3$, $MgCO_3$ should form also when a solution containing CO_3^{--} is added to a solution containing Ca^{++}, Zn^{++}, or Mg^{++}. Thus Na_2CO_3 solution added to $CaCl_2$ solution gives a white precipitate of $CaCO_3$.

$$2Na^+ + CO_3^{--} + Ca^{++} + 2Cl^- \rightarrow CaCO_3 + 2Na^+ + 2Cl^-$$

[Eq. (33), page 346]

Solutions of Na_2CO_3 and K_2CO_3 should be somewhat alkaline, since CO_3^{--} is a weak base. This conclusion is borne out by washing soda and by the "alkali" of desert basins, both chiefly Na_2CO_3, both giving fairly basic solutions.

Since H_2CO_3 is a weak acid, we could predict that its molecules will form whenever a solution containing H_3O^+ is added to a solution containing CO_3^{--}. Since the acid is unstable, we might predict further that if its molecules form in any considerable amounts, some will decompose into water and CO_2 gas. To test these predictions, we need only add an acid to a solution of washing soda; CO_2 bubbles off rapidly, causing a violent commotion:

$$2H_3O^+ + 2Cl^- + 2Na^+ + CO_3^{--} \rightarrow H_2CO_3 + 2Cl^- + 2Na^+ + 2H_2O$$

followed by

$$H_2CO_3 \rightarrow H_2O + CO_2$$

So weak is H_2CO_3 that H_3O^+ will take CO_3^{--} away from even insoluble carbonates. A drop of HCl on marble (chiefly $CaCO_3$) causes instant effervescence as CO_2 is liberated (Fig. 179):

$$\left.\begin{aligned} 2H_3O^+ + 2Cl^- + CaCO_3 &\rightarrow 3H_2O + CO_2 + Ca^{++} + 2Cl^- \\ 2H_3O^+ + CaCO_3 &\rightarrow 3H_2O + CO_2 + Ca^{++} \end{aligned}\right\} \quad (48)$$

(This equation shows the complete reaction, omitting the intermediate formation of H_2CO_3.) All carbonates behave similarly; this effervescence

with acids is, in fact, the most convenient way of testing a rock or a salt for the presence of carbonates.

So easily are carbonates dissolved by acids that they will react even with H_2CO_3 itself. The reaction does not form CO_2, but *bicarbonate ion* (HCO_3^-):

$$CaCO_3 + H_2CO_3 \rightarrow Ca^{++} + 2HCO_3^- \qquad (49)$$

A few sentences back we learned that carbonates are formed when carbonic acid reacts with bases; now we learn that these same carbonates are dissolved by more carbonic acid. The statement is not as contradictory as it sounds: carbonates are *formed* only when a *base or a basic oxide* is present, *dissolved* only when *free carbonic acid* is present.

Bicarbonate ion is unstable, decomposing readily to give CO_2:

$$2HCO_3^- \rightarrow H_2O + CO_2 + CO_3^{--}$$

So a carbonate dissolved by H_2CO_3 will reappear as a solid if the solution is heated or evaporated.

Limestone Caverns. The ability of H_2CO_3 to dissolve carbonates is responsible for the caverns so commonly found in limestone regions. Limestone, like marble, is composed chiefly of $CaCO_3$, nearly insoluble in water but readily attacked by carbonic acid. Even the very dilute acid solution formed by rain water falling through air dissolves limestone slowly, which accounts for the characteristically pitted and grooved surfaces of limestone which has been long exposed to the weather. Rain water and river water seeping downward through cracks in the rock, carrying CO_2 in solution, dissolve minute quantities of rock material along the cracks. This slow solution, continued through long ages, at length enlarges the cracks into huge caverns.

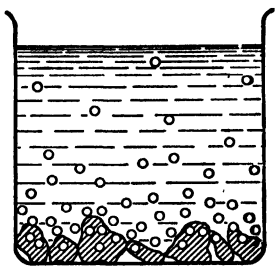


FIG. 179. *Marble and other carbonates dissolve readily in acid, liberating CO_2 gas.*

The material dissolved by the dilute acid is carried in solution underground to some point where the water can evaporate and is there redeposited as $CaCO_3$. If the water is tapped by wells and used for domestic purposes, it is considered very "hard," and deposits its $CaCO_3$ as "scale" in teakettles and bathtubs. If the water comes to the surface in a spring, its evaporation may leave $CaCO_3$ in white terraces around the spring. Or the water may partly evaporate in slowly dripping from the roof of a cave, each hanging drop depositing a bit of $CaCO_3$ before it falls, so that ultimately long "stone icicles," or stalactites, are formed (Fig. 180).

Baking. The "rising" of bread and cake depends on the liberation of

CO_2 in the dough while it is baking, the gas ordinarily being supplied by baking soda, baking powder, or both.

Baking soda or sodium bicarbonate ($NaHCO_3$) is a "half-neutralized" form of carbonic acid. Even the partial neutralization is sufficient to destroy the acid properties; sodium bicarbonate gives, in fact, a slightly

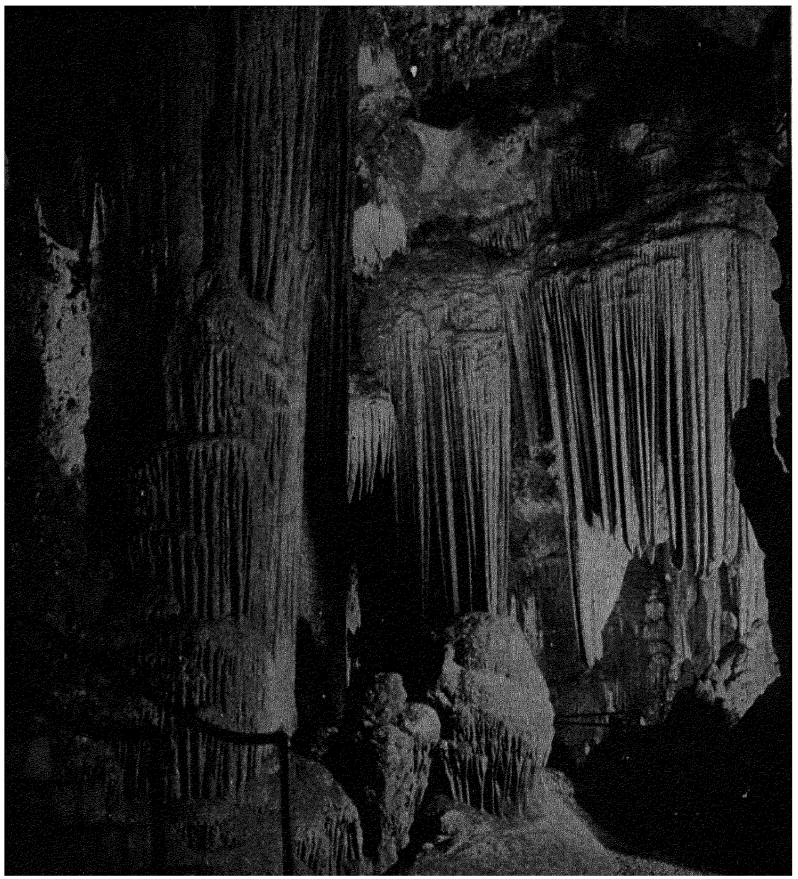


FIG. 180. *Stalactites and pillars of calcium carbonate in the cave[illegible]ay, Virginia. (Courtesy of Luray Caverns Corporation.)*

basic solution because bicarbonate ion takes a few protons away from water:

$$HCO_3^- + H_2O \rightarrow H_2CO_3 + OH^-$$

The solution is so weakly alkaline that it may be taken internally, and for this reason ordinary baking soda is a principal constituent of commercial preparations for "alkalizing the system." With acids $NaHCO_3$

reacts much like normal carbonates, liberating CO_2. It is this property which makes it useful in baking.

The acid used with baking soda should be a fairly weak one, so that evolution of CO_2 will not be too rapid. Sometimes the acid is supplied by the lactic acid of sour milk, but more commonly it is added as a constituent of baking powder. Baking powders contain sodium bicarbonate mixed either with an acid or with a salt which can liberate H_3O^+ in solution; either type gives off CO_2 slowly when water is added and so can furnish the necessary gas for leavening dough.

Questions

1. Which of the following electrolytes are completely ionized in solution? NaCl, HNO_3, NH_4Cl, $CaCl_2$, KBr, H_2CO_3, KOH.
2. Which of the following substances are practically insoluble? $CaCO_3$, $ZnCl_2$, $HC_2H_3O_2$, $Fe(OH)_3$, NaOH, AgCl, HgS, $BaCO_3$, $Al(NO_3)_3$.
3. Which of the following are weak acids and which are weak bases? H_2SO_4, $HC_2H_3O_2$, NH_3, OH^-, HCl, CO_3^{--}, H_3BO_3.
4. Would you expect HBr to be a weak or strong acid? Why?
5. Write the ionic equation for the neutralization of KOH by HNO_3. What actual chemical change does this equation show?
6. How could you tell whether an unknown solution is acidic, basic, or neutral?
7. What reaction takes place when a solution of $Ca(C_2H_3O_2)_2$ is added to a solution of H_2SO_4?
8. How could you tell whether an unknown mixture of salts contains a salt of ammonium ion?
9. What reaction would take place if $FeCl_3$ solution were added to NaOH solution?
10. Boric acid (H_3BO_3) is a weaker acid than carbonic acid. What would happen if solutions of sodium borate (Na_3BO_3) and HCl are mixed? Would you expect a solution of Na_3BO_3 to be acidic, basic, or neutral?
11. If the following salts are dissolved in water, which would give acidic solutions, which alkaline solutions, which neutral solutions? Na_2CO_3, KCl, $KC_2H_3O_2$, $BaCl_2$, $(NH_4)_2SO_4$, $NaNO_3$.
12. Which of the following oxides dissolve in water to give basic solutions and which to give acidic solutions? Which are basic oxides and which acidic oxides?

$$CO_2 \quad SiO_2 \quad SO_3 \quad Fe_2O_3 \quad CaO \quad K_2O$$

13. Write an equation to show what happens when an acid reacts with a carbonate. Explain the importance of this reaction (or very similar reactions) in (*a*) baking and (*b*) the formation of limestone caverns.
14. Use Table XIX to predict what will happen when the following pairs of substances are mixed: (*a*) solutions of sulfuric acid and potassium sulfide; (*b*) solutions of H_2S and $CaCl_2$; (*c*) solutions of ammonium nitrate and sodium acetate; (*d*) ammonia gas and pure acetic acid.

CHAPTER XXVI

Chemical Energy

EVER since our remote ancestors learned the value of fire, mankind has been putting chemical energy to practical use. Today we transform it not only into heat and light, but into mechanical energy and electrical energy as well. Locked up in the atoms of matter, chemical energy has long remained a mystery. The electronic theory gives us some insight into the origin of this energy, but the mystery is yet far from completely solved.

Exothermic and Endothermic Reactions

Any chemical reaction can be made to take place so that all or nearly all the liberated energy appears in the form of heat. The heat can be measured accurately by determining the rise in temperature it produces in a given quantity of water, and so provides a convenient means for comparing the chemical energies of different reactions.

Chemical changes which liberate heat are called ***exothermic reactions.*** Familiar examples are the burning of coal and the explosion of a mixture of hydrogen and oxygen. The heat liberated is often expressed in the equation; thus

$$C + O_2 \rightarrow CO_2 + 94{,}400 \text{ cal} \qquad (50)$$

$$2H_2 + O_2 \rightarrow 2H_2O + 116{,}800 \text{ cal} \qquad (51)$$

These figures represent the heat produced when an amount of each substance is used equal to its molecular weight expressed in grams, multiplied by its coefficient in the equation. When 12 g. of carbon are burned, 94,400 cal are produced; when 4.032 g. of hydrogen are burned, 116,800 cal are produced. These particular amounts are chosen so that heats liberated for similar numbers of molecules may be compared for different reactions.

Chemical changes which take place only when heat or some other kind of energy is supplied are called ***endothermic reactions.*** Thus water

may be decomposed into hydrogen and oxygen only by heating to very high temperatures, or by supplying electrical energy (during electrolysis):

$$2H_2O + 116{,}800 \text{ cal} \rightarrow 2H_2 + O_2 \qquad (52)$$

The formation of nitric oxide (NO) from its elements is an endothermic reaction, which takes place only at high temperatures:

$$N_2 + O_2 + 43{,}200 \text{ cal} \rightarrow 2NO$$

Again each substance is assumed to be present in an amount equal to its molecular weight in grams multiplied by its coefficient.

From the law of conservation of energy we might predict that if a given reaction is exothermic, the reverse reaction will be endothermic, and further that the amount of heat liberated by one must be equal to the amount absorbed by the other. This prediction is borne out by Eqs. (51) and (52) above, and might be checked by any number of other reactions. For a single example, sodium burning in chlorine liberates 196,800 cal for every 46 g. of sodium:

$$2Na + Cl_2 \rightarrow 2NaCl + 196{,}800 \text{ cal}$$

NaCl is decomposed in an endothermic process requiring the absorption of this same amount of heat:

$$2NaCl + 196{,}800 \text{ cal} \rightarrow 2Na + Cl_2$$

Energy changes accompanying ionic reactions are measured and represented in equations in the same manner as energy changes for other reactions. The ionization of most salts is an endothermic process; for instance, when KNO_3 is dissolved in water the container becomes cold, since ionization of the salt absorbs heat from its surroundings:

$$KNO_3 + 8{,}500 \text{ cal} \rightarrow K^+ + NO_3^-$$

Neutralization is a good example of an exothermic ionic process. If concentrated solutions of NaOH and HCl are mixed, for instance, the mixture quickly becomes too hot to touch:

$$H_3O^+ + OH^- \rightarrow 2H_2O + 13{,}700 \text{ cal}$$

The neutralization of any strong acid by any strong base liberates almost precisely this same amount of heat for each 1.008 g. of H^+—as might be expected, since the actual chemical change in all cases is simply the transfer of protons from hydronium to hydroxide ions.

The heat given out or absorbed in a chemical change is an approximate measure of the chemical energy possessed by the substances which react, hence also a measure of their stability. If much energy is required to decompose a substance, that is, if its decomposition is strongly endo-

thermic, the substance is (with rare exceptions) relatively stable; if its decomposition is weakly endothermic or exothermic, the substance is in general unstable. From the reactions above, we can see at a glance that CO_2, H_2O, and NaCl are stable compounds, since the formation of each is strongly exothermic and its decomposition endothermic. NO, on the other hand, is unstable, since its decomposition liberates heat. We can say further that the combinations H_2 and O_2, Na and Cl_2, H_3O^+ and OH^- are relatively unstable, since they react with evolution of much energy, while N_2 and O_2 form a stable mixture.

The general interpretation of chemical-energy changes in terms of electrons follows readily from earlier discussions of chemical combination (page 325). When sodium reacts with chlorine, for instance, an electron from each sodium atom is transferred to the outer shell of a chlorine atom, a position in which it has a smaller amount of potential energy with respect to the atomic nuclei. When carbon reacts with oxygen, the atoms are joined by electron pairs, the formation of pairs involving a decrease in the potential energy of the electrons. Thus *chemical energy is initially potential energy of electrons;* when the electrons move to new positions, some of this potential energy is transformed into other kinds. Just how the transformation occurs is not clear. Apparently the excess energy of the shifted electrons causes a violent disturbance in the new-formed molecules: it may give the molecules themselves motions which we would feel as heat energy, or it may displace outer electrons into new orbits from which they jump back with emission of radiant energy. In endothermic reactions, some other form of energy must be supplied to *increase* the potential energy of electrons.

Activation Energies

Coal burns in air to give great quantities of heat; how, then, can coal be kept indefinitely at ordinary temperatures in contact with air? The decomposition of nitric oxide liberates considerable energy; why does it not therefore break up spontaneously? How can the compound exist at all? A mixture of hydrogen and oxygen will produce a violent explosion; why should heat or an electric spark be necessary to start the explosion? Why, in general, do not all exothermic reactions take place instantaneously of their own accord?

Apparently many exothermic processes occur only if some energy is provided to start them. A mixture of hydrogen and oxygen may be likened to the car of Fig. 181: the car's potential energy may be converted into kinetic energy if it moves down into the large valley *C*, but it can move of its own accord only if it is first given sufficient energy to climb to the top of the hill *B*. Similarly, the chemical energy stored in the electrons of hydrogen and oxygen can be liberated as heat only if the

molecules have sufficient energy, or are sufficiently "activated," to make the reaction start. The energy necessary for activation, corresponding to the energy required to move the car from *A* to *B*, is called the ***activation energy*** of the reaction.

The electronic picture of chemical combination gives a plausible reason why activation should be necessary. The combination of oxygen and hydrogen involves the formation of electron-pair bonds* between O and H atoms, a process which gives out energy; but before these bonds can be formed, the electron pairs which bind hydrogen atoms in H_2 molecules and oxygen atoms in O_2 molecules must be broken. To break these bonds requires energy. The energy is in part supplied by the

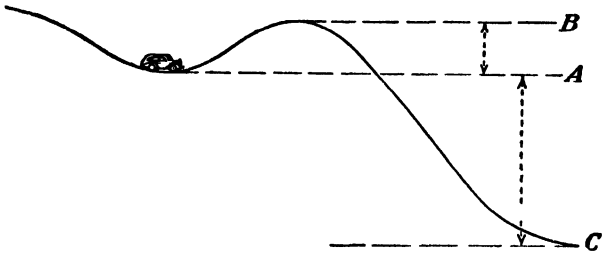


FIG. 181.

thermal energy of the molecules, and after the reaction starts by the energy it gives out; but initially some additional energy must be supplied to loosen the bonds.

A molecule with sufficient energy above the average to enable it to react is called an *activated molecule*. In some gas reactions an activated molecule may be actually broken down into atoms. Other activated molecules may simply have a high kinetic energy, or they may have one or two electrons displaced from their normal orbits. In reactions which take place spontaneously at room temperatures (for instance, the reaction between hydrogen and fluorine), enough of the molecules have sufficient heat energy to break the necessary bonds without further activation. In many ionic reactions no bonds need be broken: ions are, so to speak, already activated and react almost instantaneously. But a great number of exothermic reactions require the preliminary activation of some of their molecules before they can take place at appreciable rates.

Once an exothermic reaction is started, it commonly supplies its own activation energy. In other words, the energy given out when some of the molecules react supplies neighboring molecules with sufficient energy to activate them, so that the reaction spreads quickly. Thus a mixture

* "Chemical bonds" are simply the forces binding atoms together in molecules. In the same sense we might speak of a gravitational "bond" between the earth and the moon, or an electrostatic "bond" between two unlike charges.

of hydrogen and oxygen need only be touched with a flame for the reaction to spread so rapidly that an explosion results. When a bed of coal is set on fire it continues to burn, since the heat liberated in one part is sufficient to ignite the coal in adjacent parts.

Activation energy may be supplied by heat, by light, or by electricity. Thus a coal or wood fire is started by supplying heat. A mixture of hydrogen and oxygen may be exploded by heating or by the passage of an electric spark. A hydrogen-chlorine mixture may be exploded by either of these methods or by strong illumination.

Stability

Activation energies give a new approach to the idea of stable substances and stable mixtures. Ordinarily we think of a stable compound as one which resists attempts to decompose it, while an unstable compound breaks up on the slightest provocation. We try to make these ideas more precise by saying that a stable compound, in general, is one whose decomposition is highly endothermic, while the decomposition of an unstable compound is exothermic. But now we find that the two methods of describing stability are not the same, since a compound which liberates energy in decomposing may have a sufficiently high activation energy to prevent its breaking up readily.

A glance at Fig. 181 will make the difficulty clearer. The car's position at A is "unstable," in the sense that the car can lose energy by rolling down to C, but there is no imminent danger of its position changing until it acquires energy to lift it over B. The energy which the car can give out by falling through the vertical distance AC represents the heat which an exothermic reaction can liberate; the energy required to elevate it to B is the activation energy. A substance is "unstable" as long as AC is large, but it will not decompose readily as long as AB is large also. Thus nitric oxide is unstable in the sense that its decomposition is strongly exothermic (large AC), but ordinarily it shows no disposition to break up because its activation energy is high (large AB). Similarly a hydrogen-oxygen mixture, though even more "unstable," shows no tendency to react at ordinary temperatures. In spite of this ambiguity, we shall find it convenient to keep on using the term "stability" with reference to the amount of heat which a compound or mixture liberates or absorbs during reaction.

We have already discussed (pages 197 and 328) the stability of the compounds of various elements in relation to their electronic structures and their positions in the periodic table.

Stability is at best a relative notion. Ordinary lime (CaO) is unstable, since it reacts with CO_2 and water vapor in the air, but out of contact

with these gases it will keep indefinitely. Milk is unstable, since it sours easily, yet if sterilized and kept sealed at low temperatures it remains sweet a long time. Water is a highly stable material under ordinary conditions, but at 3000°C it is largely decomposed into hydrogen and oxygen. Carbonic acid is stable under high pressures, relatively unstable at ordinary pressures. The stability of a substance depends largely on conditions around it—on its temperature, its pressure, and the other substances with which it comes in contact.

Deep in the earth's interior exist conditions of high temperature and enormous pressure which cannot be duplicated at the surface. Under these conditions compounds and mixtures of compounds are stable which would not be stable at ordinary temperatures and pressures, and many familiar compounds cannot exist. Many rocks now at the earth's surface were originally formed at these depths, and their compounds are of types stable only at extreme temperatures and pressures; on the surface, exposed to air and water, such rocks gradually decay, their materials being converted to compounds that are stable under ordinary conditions.

In the sun and stars are conditions even more extreme—temperatures ranging from thousands to millions of degrees, pressures up to hundreds of millions of atmospheres. In such an environment no ordinary compounds are stable, and only a few compounds of any sort exist. Atoms ions, and electrons are the stable forms of matter.

Energy Transformations

Any endothermic reaction implies a change of some other form of energy into chemical energy. Priestley's preparation of oxygen by heating mercuric oxide, for instance, involves a change of heat energy into the chemical energy of free mercury and oxygen. The preparation of lime by heating limestone,

$$CaCO_3 \rightarrow CaO + CO_2$$

is another change of the same sort; that CaO and CO_2 have considerable chemical energy is evident from the heat evolved when they react at ordinary temperatures. When water is decomposed by electrolysis, electrical energy is converted into the chemical energy of free hydrogen and oxygen.

One of the most important of all chemical reactions from the human standpoint, the formation of carbohydrates in the leaves of green plants, involves a direct conversion of radiant energy from sunlight into chemical energy. Carbohydrates are complex compounds of carbon, including sugar, starch, and cellulose—the first two vitally important foods, the

last a major constituent of wood, paper, and cloth. Plants are able to manufacture carbohydrates out of water which enters through their roots and CO_2 taken from the air:

$$6CO_2 + 5H_2O + 671{,}000 \text{ cal} \rightarrow C_6H_{10}O_5 + 6O_2$$

The reaction is highly endothermic, the necessary energy coming from sunlight. The energy is not absorbed by the CO_2 and H_2O directly but by a substance called *chlorophyll*, which is part of the green coloring matter of leaves; the chlorophyll is not changed by the reaction, but serves to pass on the sun's energy to the reacting molecules. This important reaction is often called *photosynthesis*, since light is necessary for its occurrence. It has not been duplicated in the laboratory, and its details are far from understood. But mankind depends on it, not only for the chemical energy in carbohydrates, but also for the constant replenishment of oxygen in the atmosphere.

Processes in which chemical energy is transformed into other forms are more familiar than the reverse changes. The heat of burning fuels comes from their chemical energy. Burning always produces light as well as heat; some other chemical processes, such as the reaction which occurs in the abdomens of fireflies, produce light energy without appreciable heat. Batteries use chemical reactions to produce electric currents. Explosives are substances whose chemical energy is readily converted into mechanical energy. We shall discuss two examples of these transformations in more detail.

Fuels

The first requirement of a good fuel is naturally that its combination with oxygen should be a strongly exothermic reaction. Other requirements are that it should be cheap, that it should be easy to store, that the products of combustion should be easily disposed of. Many substances satisfy the first requirement, but only a few fulfill the other three as well. Sodium, for instance, would be an excellent fuel as far as its heat-producing qualities go, but it is expensive, it must be stored under oil, and getting rid of its oxide would be a knotty problem.

The substances which best fulfill the requirements for fuels are carbon and some of its compounds. These occur in nature as wood, coal, petroleum, and natural gas, materials which may themselves be used as fuels or which may be converted into artificial fuels for special purposes. Fuels containing carbon are abundant and cheap, and their inertness at ordinary temperatures makes them easy to store. The principal products of combustion are two gases, CO_2 and water vapor, which escape into the air; the ashes from wood and coal are inactive materials which can be easily removed.

Coal is the single most important source of the energy on which modern industry depends. Its heating value is often expressed as the number of calories produced by the combustion of 1 g.; thus the heating value of good bituminous (soft) coal is about 7800 calories per gram (cal/g.), that of good anthracite coal about 7400 cal/g. These are close to the heating value of pure carbon (7900 cal/g.), but the similarity is partly accidental: much of the heat produced by burning coal comes from the combustion of carbon compounds, and coal always contains a fairly large percentage of unburnable ash. *Coke* is a fuel derived from coal by heating it in the absence of air; volatile constituents are driven off, leaving free carbon and ash in the coke. Coke is used in place of coal where a hotter and less smoky fire is desirable. *Wood* is chiefly cellulose, one of the carbohydrates which plants produce by photosynthesis. Its heating value is less than that of coal, ranging from 2500 to 4500 cal/g. The exact formula for cellulose is unknown, but its burning may be represented roughly by the equation

$$C_6H_{10}O_5 + 6O_2 \rightarrow 6CO_2 + 5H_2O + 671{,}000 \text{ cal}$$

This reaction is the reverse of photosynthesis: radiant energy from sunlight, stored in wood as chemical energy, is released by burning as heat energy. The energy of burning coal comes ultimately from the same source, for coal consists of plant material buried beneath layers of sediment and altered by long ages of slow chemical change.

Most liquid fuels are obtained from the black, oily liquid called *petroleum*, a complex mixture of compounds of carbon and hydrogen. Distillation separates petroleum into simpler mixtures: some of the more volatile constituents form the mixture called gasoline, those with slightly higher boiling points kerosene, those with still higher boiling points gas oil, lubricating oil, vaseline, and paraffin. By far the most important liquid fuel is *gasoline*, used not only for heat energy but more commonly to produce mechanical energy in internal combustion engines. The ability of gasoline to produce mechanical energy, however, depends on the large amount of heat (11,000 cal/g.) given out when it burns: the gaseous products of burning expand very suddenly because they are intensely heated by the reaction. The equation for the burning of heptane, one of the constituents of gasoline,

$$C_7H_{16} + 11O_2 \rightarrow 7CO_2 + 8H_2O$$

shows that the volume of the products is scarcely greater than that of the original substances at ordinary temperatures, so that their expansion is due almost entirely to sudden heating.

Gas fuels, like liquid fuels, leave no solid ash on burning but are converted entirely to CO_2 and H_2O. *Natural gas* is often found with petro-

leum, and like petroleum consists of carbon-hydrogen compounds; its chief constituent is methane (CH_4). Artificial gas fuels are produced from coal or coke. The most widely used one is *water gas*, a mixture of hydrogen and carbon monoxide produced by passing steam over hot coke:

$$H_2O + C \rightarrow CO + H_2$$

Hydrogen itself would be an excellent gas fuel, since its heating value is very high (34,500 cal/g). It is too expensive for ordinary purposes but is widely used for the oxyhydrogen blowtorch.

One purpose of food is to act as a fuel for our bodies, supplying us with heat and muscular energy. We derive this energy, as it is needed, by slow oxidation. For instance, the simple carbohydrate glucose is oxidized according to the equation

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6H_2O + 6CO_2 + \text{heat}$$

The necessary oxygen enters through our lungs as we inhale, and the waste product CO_2 leaves our bodies when we exhale. *In effect, we and other animals reverse the process of photosynthesis: plants use radiant energy to produce carbohydrates, and animals oxidize carbohydrates to recover this energy as heat.*

The slow oxidation of foods in our bodies is different from burning them only in the rate of the reaction. The final products and the amount of energy obtained are the same, whether foods are digested or are heated in oxygen. Hence the fuel value of a food may be determined by actually burning it and measuring the number of calories produced. Those who anxiously "count calories" are adding up fuel values determined in this way.

The calories which dieticians watch so carefully are "large calories," each equal to a thousand ordinary calories. Of these large calories, an adult needs 2000 to 3500 per day, depending on his weight and on how active a life he leads. Approximate values for a few foods are:

Starchy foods (cereals, flour, beans, etc.)	1650 Cal/lb.
Sugar	1860 Cal/lb.
Fats and oils	3650 Cal/lb.
Meat	1600 Cal/lb.

Of course, counting calories alone will not ensure an adequate diet, since our bodies use food for many purposes besides energy production.

Explosives

Explosives are unstable substances which react to liberate large quantities of gas suddenly, the expansion of the gas producing the desired mechanical energy. We refer to mixtures of hydrogen and oxygen, or of gasoline vapor and oxygen, as "explosive mixtures," since the gases

produced by their reactions are heated enough to expand suddenly; but the term "explosive" usually means a solid or liquid material which reacts to form gas on slight disturbance.

Most explosives contain compounds of nitrogen, and the chemistry of explosives is in large part the chemistry of this element. Nitrogen itself is an extremely inactive gas, uniting with other elements only when strongly heated. The chief reason for this inactivity is the strength of the electron-pair bonds between nitrogen atoms in N_2 molecules (page 326). If these bonds are broken momentarily by passing an electric discharge through nitrogen, the resulting N atoms are much more active than the ordinary molecules. Because the N_2 molecule is so stable, nitrogen atoms unite to form it very readily, even when they are already joined with other atoms in compounds. In other words, compounds containing nitrogen are unstable, decomposing readily to form nitrogen gas. This property makes nitrogen compounds valuable in explosives

The earliest widely used explosive was *gunpowder*, made by mixing potassium nitrate (KNO_3) with charcoal and sulfur. Gunpowder explodes because the charcoal takes the oxygen from KNO_3, forming CO and CO_2, while nitrogen gas is set free; most of the sulfur unites with potassium to form solid particles of smoke. The reaction is highly exothermic and takes place very suddenly, its expanding gaseous products exerting enormous pressures in every direction.

Modern explosives give gaseous products entirely, and contain carbon, oxygen, and nitrogen united in the same molecule so that the explosion can be even more rapid. They are made by the action of nitric and sulfuric acids on various carbon compounds, such as glycerin, cellulose, and toluene; from these three, respectively, come the explosives *nitroglycerin* [$C_3H_5(NO_3)_3$], *nitrocellulose* [$C_6H_7O_2(NO_3)_3$], *trinitrotoluene* or *TNT*[$C_7H_5(NO_2)_3$]. These are alike in that they contain the same four elements, and that their products are nitrogen, the oxides of carbon, and water vapor. For example, when nitroglycerin explodes,

$$4C_3H_5(NO_3)_3 \rightarrow 12CO_2 + 10H_2O + 6N_2 + O_2$$

Nitroglycerin is dangerous to handle, for it explodes at the slightest shock. Nitrocellulose and TNT require a fairly strong shock, often provided by a small amount of another explosive in "percussion caps." *Dynamite* is a mixture of nitrocellulose and nitroglycerin, with various amounts of sawdust, flour, and certain salts added to adjust the rate of explosion.

Chemical Energy and Atomic Energy

Thousands of times more powerful than the best chemical explosives are the materials used in atomic bombs, uranium 235 and plutonium (page 300). The explosion of these elements is an altogether different

process from the explosion of nitroglycerin or TNT: no chemical reaction is involved, but rather the splitting of atomic nuclei when struck by neutrons. The nuclear disintegration itself does not produce appreciable amounts of expanding gas, but the enormous quantity of energy suddenly set free in a small space causes the heating and rapid expansion of surrounding air.

The possible use of uranium as an industrial "fuel" in the uranium-graphite pile (page 300) likewise involves a process entirely different from the chemical reactions on which we normally depend for energy. Again the fundamental process is nuclear disintegration, rather than the rapid oxidation of combustible material.

Both the energy of chemical reactions and the energy of nuclear processes are in a sense "atomic" energy, in that both involve the tiny particles which make up atoms. But the parts of the atoms concerned are different. The chemical energy of burning and of ordinary explosions depends on rearrangements in the outer part of the electron clouds of atoms; the energy of atomic bombs and uranium piles, popularly called "atomic energy," is produced by particles within atomic nuclei. So much more energy is stored in the nuclei of atoms than in their electron clouds that atomic energy in the future may very possibly supplant chemical energy for many purposes.

Questions

1. Which of the following are exothermic reactions and which endothermic? (*a*) The explosion of dynamite. (*b*) The burning of methane. (*c*) The decomposition of water into its elements. (*d*) The decomposition of water into ions. (*e*) The burning of iron in chlorine. (*f*) The combination of zinc and sulfur to form zinc sulfide (page 146).
2. From the observation that the slaking of lime [addition of water to CaO to form $Ca(OH)_2$] gives out heat, would you conclude that the following reaction is endothermic or exothermic?

$$Ca(OH)_2 \rightarrow CaO + H_2O$$

3. Most salts absorb heat on going into solution. Would you expect the crystallization of a salt from solution to be an exothermic or an endothermic process?
4. Methane is an important constituent of the atmosphere of Jupiter. Why should it be stable in Jupiter's atmosphere and not in the earth's?
5. Which of the following are good fuels? Why are the others not good fuels? Potassium, coke, methane, sulfur, ether.
6. Why is nitrocellulose a better explosive than gunpowder? What are the chief products resulting from its explosion?
7. In what ways is the burning of wood a similar chemical process to the oxidation of starch in the human body?
8. In what ways is the explosion of an atomic bomb different from the explosion of dynamite?

CHAPTER XXVII

Reaction Rates and Equilibrium

THE extraordinary attraction of nitrogen atoms for one another is important to mankind in other ways, besides making some nitrogen compounds unstable enough to act as explosives (page 375). The solid flesh of our bodies consists largely of nitrogen compounds, complex compounds with carbon, hydrogen, and oxygen called *proteins*. These proteins, without which we cannot live, our bodies manufacture from other proteins in food. The ultimate source of all our protein material is plants, although much of it comes to us secondhand as animal proteins (meat, eggs, milk). Plants in turn manufacture their proteins from simpler nitrogen compounds. How plants accomplish this feat of molecule-building we do not know, but we do know that the necessary nitrogen enters their roots in simple compounds, chiefly nitrates. Green plants are unable to break up the stable molecules of free nitrogen in the air around them; all their nitrogen, and therefore all the nitrogen that goes into animal bodies as well, comes from nitrogen compounds in the soil.

Human welfare depends directly on the combined nitrogen, often called *fixed nitrogen*, in the soil. All the nitrogen molecules which we breathe, all the nitrogen molecules which beat against our skins can do us no good, for the atoms in these molecules of the free element are united by a tie which our bodies have no power to break. Like a shipwrecked mariner surrounded by water but dying of thirst, so mankind is surrounded by an ocean of nitrogen but would perish except for the combined nitrogen which plants can absorb through their roots.

The formation of plant and animal proteins continually removes nitrogen compounds from the soil. Just as continually fixed nitrogen is returned to the soil by the decay of animal excrement and of dead plants and animals, the nitrogen of proteins being converted by decay into ammonia and ammonium salts which are then oxidized to nitrates by soil

bacteria. But the replenishment is never complete: either in decay or during combustion, some of the nitrogen atoms of the proteins manage to join together into molecules and escape into the air as nitrogen gas. Some nitrogen is also lost permanently from the soil by solution of nitrates and ammonium salts in streams and rainwash, and by bacteria which decompose nitrates into free nitrogen. Nature makes good these losses in two ways: another kind of soil bacteria, the "nitrogen-fixing" bacteria, have the unique ability to break down the stable nitrogen molecule and to manufacture compounds from the atoms; and electric discharges during thunderstorms cause some combination of atmospheric nitrogen and oxygen into nitrogen oxides, which are carried to the soil in solution in rain water. So in nature nitrogen goes through a continuous, rather complicated cycle (Fig. 182), which keeps the amount of fixed nitrogen in the soil approximately constant.

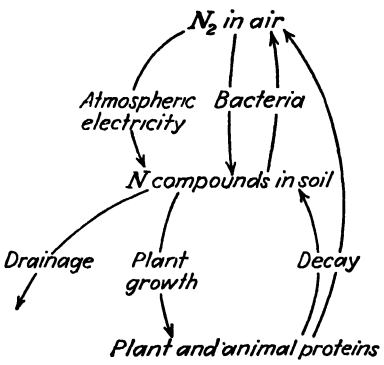


FIG. 182. *The nitrogen cycle.*

Civilized man rudely disturbs this natural cycle. Much of the protein material that goes into his body is never returned to the soil, and his habit of using plant material for fires greatly accelerates the conversion of combined nitrogen into unavailable free nitrogen. One of the problems of modern nations is to keep enough nitrogen supplied to agricultural lands to make up for this steady depletion. The problem is partially solved by the use of manure and by planting crops on which nitrogen-fixing bacteria can grow, but these expedients cannot supply all the needed combined nitrogen. The widespread use of explosives, especially in wartime, demands further immense quantities of combined nitrogen which cannot possibly be obtained from the soil or from organic wastes.

One place to look for additional supplies of combined nitrogen should seemingly be in rocks: compounds of almost all other elements are found as mineral deposits, and there is no immediately obvious reason why nitrogen should be an exception. The difficulty is that nearly all inorganic nitrogen compounds are soluble in water. They cannot accumulate under ordinary conditions, because rain water, streams, and underground seepage carry them away too rapidly. Only in the world's most arid regions are nitrogen compounds found in appreciable quantities; only in one place, the desert of northern Chile, where rain falls less than once a decade, have deposits formed which are large enough for commercial use. The chief compound here is sodium nitrate ($NaNO_3$, Chile saltpeter), a material useful directly as a fertilizer and easily convertible into nitric

acid for the manufacture of explosives. Exploitation of the Chilean deposits began in the early nineteenth century, and for more than 100 years much of the world's fixed nitrogen came from this little patch of desert.

But importation of Chilean nitrate is slow, expensive, and in wartime dangerous. By the opening of the twentieth century the need for another source of fixed nitrogen was growing acute, and scientists in several countries set to work to find one. Some nitrogen compounds could be obtained from coal when it was heated to form coke, but not nearly enough to satisfy industrial requirements. The only possible source seemed to be the atmosphere itself; the great difficulty was to make nitrogen atoms leave their partners and combine with atoms of another element. Several processes were tried, the most successful one being the union of nitrogen with hydrogen to form ammonia.

$$N_2 + 3H_2 \rightarrow 2NH_3 \qquad (53)$$

The ammonia gas produced may be converted to ammonium salts for fertilizers by combination with sulfuric or nitric acid, or it may be oxidized by atmospheric oxygen to nitric acid for use in making explosives.

The major difficulties in the development of the ammonia process were (1) the reaction is extremely slow at ordinary temperatures, and (2) at higher temperatures, where the speed is greater, the yield of ammonia is small. To make the process useful, some way had to be found either to speed up the reaction at moderate temperatures or to increase the yield at high temperatures. The problem was solved in 1913 by the German chemist Haber, by making use of the basic principles governing the speeds of chemical reactions. Let us inquire into these principles and see how Haber used them to make his country independent of Chilean nitrate during the first World War.

Reaction Rates

Some chemical changes are practically instantaneous. In neutralization, for instance, acid and base react as soon as they are stirred together; silver chloride is precipitated immediately when solutions of silver ion and chloride ion are mixed; the reaction involved in a dynamite explosion is immeasurably rapid. Other chemical changes, like the formation of ammonia, the souring of milk, the rusting of iron, take place slowly. For many of these slow reactions we can set up experiments to measure exactly how fast they are going, that is, what fractions of the original substances have disappeared at various times after the reactions start.

To study reaction rates scientifically, we should proceed along the lines which we have followed so many times before: choose a simple case—that is, a reaction whose rates are easily measurable and show no obvious

complications; perform experiments under various conditions—that is, measure rates for the reaction at different temperatures, different pressures, different amounts of material; see if the experimental results follow simple rules—that is, see if there is any mathematical connection between reaction rates and temperature, pressure, etc.; and finally, study other reactions to see if these rules are general laws. We shall state here the results of such a study, without actually going through the long hours of work or experiencing the disappointments and the triumphs which it would entail in the laboratory.

Reaction rates depend first of all on the nature of the reacting substances. Obviously some materials undergo chemical change more rapidly than others: meat decays more quickly than wood; iron rusts more rapidly than copper. For any particular reaction, the rate is influenced by four principal factors: (1) temperature, (2) concentrations of the reacting substances, (3) amount of surface exposed (in reactions involving solids), and (4) catalysts.

Temperature. Reaction rates are always increased by a rise in temperature. We use this simple fact, of course, when we put food in the refrigerator to retard its decay, and when we use hot water rather than cold for washing. As a rule, reaction rates are approximately doubled for every 10° (centigrade) rise in temperature.

The kinetic theory suggests one obvious reason for the increase of rate with temperature: most reactions depend on collisions between particles, and the number of collisions increases with rising temperature because molecular speeds are increased. But a 10° rise at ordinary temperatures produces only a slight acceleration of molecules, not nearly sufficient to make a reaction double its pace. To find an adequate explanation, we must go back to the idea of activation energy which we encountered in the last chapter.

If molecules must be activated before they can react, reaction rates should depend, not on how many ordinary collisions occur each second, but on *the number of collisions between activated molecules.* Now activated molecules in a fluid may be produced by ordinary molecular motion as a result of exceptionally energetic collisions; they remain activated for only a small fraction of a second, losing their excess energy by further collisions unless they react in the meantime. So in any fluid, provided the temperature is not too low, a fraction of the particles should be activated at any one instant. The fraction may be very small at ordinary temperatures, but it increases rapidly as the temperature rises and molecular motion speeds up. Reaction rates increase with temperature, therefore, chiefly because the number of activated molecules grows larger.

A mixture of hydrogen and oxygen, for instance, at room temperature contains very few molecules with sufficient energy to react, and the

reaction is so exceedingly slow that the gases may remain mixed for years without appreciable change. Even at 400°C the rate is negligibly small; but at 600° enough of the molecules are activated to make the reaction fast, and at 700° so many are activated that the mixture explodes. This sort of behavior is typical of many reactions, especially those involving molecules whose bonds must be broken: at low temperatures the chemical change is so slow that for all practical purposes it does not occur, in a range of intermediate temperatures reaction is moderately rapid, and at high temperatures it becomes instantaneous. Reactions between ions, on the other hand, are instantaneous even at room temperatures, the ionic state itself being a form of activation.

Concentration. The general effect of concentration on reaction speed is well shown by rates of burning in air and in pure oxygen: the pure gas has almost five times as many oxygen molecules per cubic centimeter as air has, and rates of burning are very much greater. The concentration effect appears even more spectacularly in the burning of iron wire in liquid air; despite the low temperature, oxygen molecules are so abundant and so close together in the liquid that the metal burns brightly. As a general rule, *the rate of a chemical reaction is directly proportional to the concentration of each reacting substance.* This is an experimental result, for which the kinetic theory gives a simple and reasonable explanation: the number of collisions between activated molecules, which determines the reaction speed, should depend on the total number of collisions, and this in turn on how many molecules each cubic centimeter contains.

Surface. When a reaction takes place between two solids, or between a fluid and a solid, its rate depends markedly on the amount of solid surface exposed. A finely powdered solid presents vastly more surface than a few large chunks, and reactions of powders are accordingly much faster. Granulated sugar dissolves more rapidly in water than lump sugar; finely divided zinc is attacked by acid quickly, larger pieces only slowly; ordinary iron rusts slowly, but if the metal is very finely powdered its oxidation is fast enough to produce a flame. A kinetic explanation is obvious enough: the greater the surface, the more quickly molecules can get together to react. For a similar reason, efficient stirring speeds up reactions between fluids.

Catalysts. Harder to explain is the action of catalysts. *These are substances with the singular property of changing the rate of a chemical reaction without being themselves altered or used up.* A catalyst may either speed up or retard a reaction. For a simple example, hydrogen peroxide solutions are unstable at ordinary temperatures, slowly decomposing into water and oxygen. If a little of the black powder called manganese dioxide is added to hydrogen peroxide, the decomposition becomes much more rapid, oxygen bubbling from the solution in large quantities. At the end of

the reaction the manganese dioxide catalyst can be recovered unchanged. Ordinary solutions of hydrogen peroxide contain a little of a carbon compound called acetanilid, a catalyst of the opposite sort which slows down the decomposition. Water plays the role of a catalyst in promoting the decay of wood. Another example is the action of chlorophyll in photosynthesis (page 372): the chlorophyll transfers energy from sunlight to molecules of CO_2 and H_2O, but is not itself changed.

Catalysts remain almost as mysterious to experts as to laymen. Some catalysts are known to form unstable intermediate compounds with one of the reacting substances, which decompose again as the reaction proceeds. Others, notably certain metals, can affect reaction rates by producing activated molecules at their surfaces. For the action of many others no adequate explanation has been suggested. In general, a given reaction is influenced only by a few catalysts, and these may or may not affect other reactions. Catalysts are highly important in many industrial processes, but in searching for new ones a chemist must usually rely more on experience and trial-and-error methods than on any definite knowledge as to how catalysts work.

Chemical Equilibrium

Most chemical reactions are *reversible*. That is, the products of a chemical change, under suitable conditions, can usually be made to react to give the original substances. We have discussed many examples in other connections. Hydrogen and oxygen combine to form water when ignited, and water may be decomposed into its elements by extreme heat or by electricity. Mercury and oxygen combine when heated moderately, and mercuric oxide decomposes when heated more strongly. Carbon dioxide dissolves in water to form carbonic acid, and carbonic acid decomposes into carbon dioxide and water.

There is no reason, of course, why the forward and reverse processes of a chemical change cannot take place simultaneously, provided their rates are approximately equal. If CO_2 is kept over water in a stoppered bottle, for instance, the gas forms H_2CO_3 by reaction with water. But H_2CO_3 is unstable at room temperature and begins to dissociate into CO_2 and H_2O as soon as an appreciable concentration is built up. The rate of its dissociation increases as its concentration grows larger, until finally as much dissociates each second as is being formed by the dissolving of CO_2. At this point the rates of the forward and backward reactions are the same, and no further change occurs in the amounts of CO_2 and H_2CO_3. We may represent this situation by a single equation

$$H_2O + CO_2 \rightleftharpoons H_2CO_3 \qquad (54)$$

the double arrow indicating that reactions in both directions occur to-

gether. Another example of simultaneous forward-and-backward reactions from our study of acids is the ionization of acetic acid:

$$HC_2H_3O_2 + H_2O \rightleftharpoons H_3O^+ + C_2H_3O_2^- \qquad (55)$$

The un-ionized acid reacts continually with water to form ions, and the ions recombine at the same rate, so that at any given instant the same small fraction of the acid is ionized.

A situation of this kind is called a ***chemical equilibrium.*** It is a state of balance determined by two opposing processes, much like the balance between vaporization and condensation at a liquid surface in a closed container (page 138). The two processes do not reach equilibrium and stop, but continue indefinitely, maintaining a balance because one constantly undoes what the other accomplishes. For a crude analogy, imagine a man walking down an escalator while the escalator is moving upward: if he walks as fast in one direction as the escalator carries him in the other, the two motions will be in equilibrium and he will remain at the same place indefinitely.

Very many chemical changes reach a state of equilibrium, instead of going to completion in one direction or the other. Equilibrium may be established when a reaction is very nearly complete, or when it is only just starting, or when both products and reacting substances are present in considerable amounts. At what point equilibrium occurs depends entirely on the rates of the opposing reactions; one reaction takes place until a sufficient concentration of products is built up for the reverse reaction to go at the same rate. For example, consider the ionizations of different acids: (1) HCl dissociates completely into H_3O^+ and Cl^-; here there is no reverse reaction, no equilibrium, except in very concentrated solutions. (2) $HC_2H_3O_2$ ionizes to a slight extent; when a small concentration of ions is built up, the recombination goes at the same rate as the dissociation. (3) H_2CO_3 ionizes very slightly; the dissociation is so slow that a very low concentration of ions permits the acid molecules to be formed as fast as they break up. The extent to which an acid is ionized depends on *how fast* its molecules break down into ions, compared with *how fast* the ions recombine.

Conditions Affecting Equilibrium

Frequently a chemist encounters this problem: he wishes to prepare a compound but finds that the reaction which produces it reaches equilibrium before much of the compound has been formed. Once equilibrium is established, waiting for more of the product to form is futile, for its amount thereafter does not change How can the equilibrium conditions be altered, so that the yield of the product will be larger?

Since equilibrium depends on a balance between two rates, a solution

to this problem depends on finding a way to change the speed of one reaction or the other. If the man on the escalator walks faster, he will move slowly downward against the escalator's motion; if the escalator is made to run faster, he will move upward in spite of his walking. So if the forward reaction of an equilibrium is speeded up, more of the products will form; if the backward reaction is made faster, more of the original substances will appear. Similarly, of course, the yield of a product may be increased if the backward reaction is slowed down, or decreased if the forward reaction is slowed down.

Speeding up or retarding one of the reactions in an equilibrium is not quite so simple as changing the rate of a single reaction, but we might expect that the same general factors which affect reaction rates would influence equilibrium also. Careful study shows that a chemist has three chief ways at his disposal for shifting an equilibrium in one direction or another: (1) changing the concentration of one or more substances, (2) changing the temperature, or (3) changing the pressure (especially in gas reactions). Using a catalyst and changing the amount of exposed surface do not influence an equilibrium, since these factors always affect both forward and reverse reactions alike.

Concentration. Suppose that equilibrium has been established between H_2O, CO_2, and H_2CO_3 in a closed bottle [Eq. (54)] and that the stopper is then removed. CO_2 can now diffuse out into the air, so that the concentration of CO_2 above the water surface is diminished. This means that fewer collisions between H_2O and CO_2 molecules will take place, so that the reaction between these two is slowed down. The decomposition of H_2CO_3 is not affected but continues at its usual rate. So the forward reaction of Eq. (54) becomes slower than the backward reaction, and more H_2CO_3 decomposes than is being formed. Hence the concentration of acid in the solution decreases. If a vacuum pump is connected to the bottle and CO_2 is continuously removed, the forward reaction is still further slowed down and presently all the H_2CO_3 is decomposed. On the other hand, if a tank of CO_2 is connected to the bottle so that the gas can be introduced under pressure, its concentration is increased and the forward reaction is speeded up, giving a larger amount of H_2CO_3 in solution.

For a more complicated example, consider the reaction between solutions of ammonia and iron chloride ($FeCl_3$). As the ammonia is added, a gelatinous brown precipitate forms in the solution and slowly settles to the bottom. The precipitate is iron hydroxide [$Fe(OH)_3$], an extremely insoluble substance which forms even when only a trace of iron is present; the precipitate forms so readily and is so easily recognized that this reaction is often used in analytical work as a test for the presence of iron compounds. The reaction involves a displacement of the equilibrium in the ammonia solution,

$$NH_3 + H_2O \rightleftharpoons NH_4^+ + OH^- \qquad (56)$$

The forward reaction, discussed on page 358, occurs when ammonia gas dissolves in water; the reverse reaction takes place when NaOH is added to a solution of an ammonium salt. Both reactions are occurring simultaneously in any solution of ammonia. Since OH^- is a stronger base than NH_3, the reverse reaction goes more nearly to completion than the forward reaction—which means that an ammonia solution consists chiefly of NH_3 and H_2O, with only a little NH_4^+ and OH^-. If this solution is added to iron chloride solution, which contains the ions Fe^{+++} and Cl^-, the ions of iron react immediately with the OH^- present to form $Fe(OH)_3$. This lowers the concentration of OH^- and thus slows down the backward reaction of Eq. (56). The forward reaction is unaffected, NH_3 and H_2O reacting to form more OH^- and NH_4^+. The OH^- is removed as fast as it forms, so that more and more NH_3 changes to NH_4^+. The process may be represented

$$NH_3 + H_2O \rightleftharpoons NH_4^+ + \overbrace{OH^- \quad Fe^{+++}}$$

and the final result is summarized in the equation

$$3NH_3 + 3H_2O + Fe^{+++} \rightarrow Fe(OH)_3 + 3NH_4^+$$

In general, changing the concentration of one of the products in an equilibrium affects the backward reaction without affecting the forward reaction, whereas a change in concentration of one of the original substances influences the forward reaction only. Thus we may change the rate of either reaction separately and hence shift the equilibrium in one direction or the other.

Temperature. If one reaction in an equilibrium is exothermic, the opposite reaction must be endothermic (page 367). An increase in temperature, of course, speeds up both reactions, but the endothermic reaction, the one which *absorbs heat*, is speeded up more than the other. We could not have predicted this result from our brief study of reaction rates, but it seems at least reasonable that a reaction which uses up heat should be especially favored by high temperatures.

For an example, consider the formation of NO from N_2 and O_2, a moderately endothermic reaction:

$$N_2 + O_2 + 44{,}000 \text{ cal} \rightarrow 2NO$$

At fairly high temperatures NO decomposes into its elements (by an exothermic reaction), so that an equilibrium is established:

$$N_2 + O_2 \rightleftharpoons 2NO$$

The forward reaction absorbs heat, the backward reaction evolves

heat. Hence a rise in temperature favors the formation of NO, and for good yields of this gas the reaction must be carried out at very high temperatures.

Thus a knowledge of the energy change in an equilibrium reaction gives us another method of causing the equilibrium to shift in one direction or the other. If the desired product is formed by an endothermic process, the reaction should be carried out at high temperatures; if it is formed by an exothermic reaction, low temperatures are more favorable.

Pressure. If a gas reaction involves a change in volume, its equilibrium position may be shifted by changing the total pressure. An increase in pressure causes the reaction to take place in a direction which gives the smaller number of molecules, while a decrease in pressure favors the opposite reaction. High pressure, so to speak, squeezes the reaction mixture into as small a volume as possible. For example, at fairly high temperatures equilibrium is established between the three gases SO_2, O_2, and SO_3 (sulfur trioxide):

$$2SO_2 + O_2 \rightleftharpoons 2SO_3$$

If pressure is constant, the SO_3 occupies a volume only two-thirds as great as the volume of the gases from which it formed. Hence an increase in pressure favors the forward reaction: to obtain large yields of SO_3, the reaction should be carried out at as high a pressure as possible.

Synthetic Ammonia

We have discussed at some length the various means which a chemist can use to influence the rates of chemical reactions and to increase the yields of substances involved in equilibriums. Let us see how use of these principles led Haber to a successful commercial method for the preparation of ammonia from nitrogen and hydrogen.

The reaction is reversible, forming an equilibrium represented by the equation

$$N_2 + 3H_2 \rightleftharpoons 2NH_3$$

The forward reaction is exothermic, so that the yield of ammonia is greatest at low temperatures. But the reaction has a high activation energy and is exceedingly slow at ordinary temperatures; only above 700°C does it become rapid. At 700°, however, the yield of ammonia is far too small for practical use, as the second column of Table XX shows. So Haber's problem was to find a method either to speed up the reaction at moderate temperatures, or to increase the yield at high temperatures.

A possible way to increase the rate of reaction was to discover a catalyst. Extensive search brought to light several fairly good ones: the rare metals osmium and uranium, and a mixture of the more common

metals iron and molybdenum. But even the best catalyst could not make the reaction go at a reasonable speed below 500°, and at this temperature the yield of ammonia is still negligible. One other possibility remained: forcing the equilibrium to shift in the direction of the forward reaction by a change in pressure. The equation shows that ammonia occupies only half the volume of the gases from which it is made (when all are measured at the same pressure), so that its yield should be improved by increasing the pressure. How the yield actually changes with pressure is indicated by the experimental results in Table XX. For temperatures near 500° the yields at high pressures proved large enough for commercial

TABLE XX

Temperature, °C	*Yield of ammonia, per cent*		
	At 1 atm	*At 100 atm*	*At 200 atm*
500	0.13	10.2	17.6
600	0.05	4.5	8.2
700	0.02	2.0	4.0
800	0.01	1.1	2.2

use. A mixture of nitrogen and hydrogen could be allowed to reach equilibrium, the 10 to 20 per cent of ammonia which had formed could be removed by freezing or by solution in water, then more nitrogen and hydrogen could be added to the remaining gases and the process repeated.

This brief account tells nothing of the technical difficulties that Haber faced in putting the ammonia process into actual operation; it merely outlines his solution for the purely chemical problems of speeding up a sluggish reaction and increasing the yield from an equilibrium. Commercial production of "synthetic" ammonia by Haber's process began in Germany in 1913. During the First World War, when imports of Chilean nitrate were cut off by the English blockade, synthetic ammonia became Germany's chief source of fixed nitrogen. Without it Germany could hardly have survived more than a single year of war.

Perfected in the war years, Haber's process since 1918 has been used by many nations as a source of combined nitrogen for fertilizers and explosives. It competes successfully with Chilean nitrate and gives assurance that even when Chile's deposits are exhausted the world need never fear a nitrogen shortage.

So brief a summary of rates and equilibriums as we have sketched in this chapter can suggest only vaguely the importance of these subjects. We have mentioned a single industrial application, chosen from hundreds of similar examples. In pure research, in industry, and in applying their science to natural processes, chemists are guided continually by their

knowledge of the factors which determine the speeds of chemical reactions and the shifting of chemical equilibriums.

Questions

1. For what two chief purposes do civilized nations need nitrogen compounds?
2. By what natural processes is fixed nitrogen put into the soil? What additional sources of fixed nitrogen are available?
3. Give two examples of reactions which are (*a*) practically instantaneous at room temperatures, (*b*) fairly slow at room temperatures.
4. Suggest three ways to increase the rate at which zinc dissolves in sulfuric acid.
5. Under ordinary circumstances coal burns slowly, but fine coal dust in mines sometimes burns so rapidly as to cause an explosion. Explain the difference in rates.
6. Explain why a reaction with high activation energy is slow at room temperature.
7. Ammonia gas dissolves in water and reacts according to the equation

$$NH_3 + H_2O \rightleftharpoons NH_4^+ + OH^-$$

   How would the amount of ammonium ion in solution be affected by (*a*) increasing the pressure of NH_3? (*b*) Pumping off the gas above the solution? (*c*) Raising the temperature? (*d*) Adding a solution of HCl? (*e*) Adding a solution of iron sulfate [$Fe_2(SO_4)_3$]?
8. Hydrogen sulfide gas dissolves in water and ionizes very slightly,

$$H_2S + 2H_2O \rightleftharpoons 2H_3O^+ + S^{--}$$

   How would the acidity of the solution (concentration of H_3O^+) be affected by (*a*) increasing the pressure of H_2S? (*b*) Raising the temperature? (*c*) Adding a solution of KOH? (*d*) Adding a solution of silver nitrate? [Silver sulfide (Ag_2S) is insoluble.]
9. Limestone ($CaCO_3$) dissolves in carbonic acid to form calcium bicarbonate. The latter decomposes readily, so that an equilibrium is set up

$$CaCO_3 + H_2CO_3 \rightleftharpoons Ca^{++} + 2HCO_3^-$$

   How would this equilibrium be affected by (*a*) raising the temperature? (*b*) Allowing the solution to evaporate? (*c*) Increasing the pressure of CO_2, thereby increasing the concentration of H_2CO_3? Under what natural conditions, then, is limestone most soluble? Under what conditions will it be precipitated from solution?
10. The reaction $2SO_2 + O_2 \rightarrow 2SO_3$ is exothermic. How will a rise in temperature affect the yield of SO_3 in an equilibrium mixture of the three gases? Will an increase in pressure raise or lower this yield? In what possible way can the speed of the reaction be increased at moderate temperatures?
11. The three gases H_2, O_2, and H_2O are in equilibrium at temperatures near 2000°C. Write the equation for the equilibrium. Would the yield of H_2O be increased or decreased by raising the temperature? By raising the pressure?

CHAPTER XXVIII

Oxidation and Reduction

We have been using the term "oxidation" to mean the chemical combination of a substance with oxygen, the term "reduction" to mean the removal of oxygen from a compound. In the language of electrons we may extend these ideas to describe an important general class of chemical reactions.

Valence Changes

The valence of an element in a compound is the number of electrons which each atom has gained, lost, or shared. In ionic compounds, the valence of the metal is positive, that of the nonmetal (or nonmetals) negative. In covalent compounds valences are properly neither positive nor negative, but the more active nonmetal is often arbitrarily assigned a negative valence. The valence of an ion is the amount of its charge. Elements in the free state are regarded as having a valence of zero. This is a hasty summary, to call to mind earlier discussions of valence (page 329).

*The **oxidation** of an element is defined in general as a chemical change in which its positive valence is increased or its negative valence decreased. The **reduction** of an element is a chemical change in which its positive valence is decreased or its negative valence increased.*

When zinc is oxidized by burning in oxygen to form ZnO, its valence increases from 0 to +2. Similarly, when zinc burns in chlorine

$$Zn + Cl_2 \rightarrow ZnCl_2 \qquad (57)$$

its valence again increases, and this reaction is considered an oxidation, although no oxygen is involved. The solution of zinc in hydrochloric acid is a further example of oxidation, for the valence of the metal jumps from 0 to +2:

$$Zn + 2H_3O^+ + 2Cl^- \rightarrow Zn^{++} + H_2 + 2Cl^- + 2H_2O$$

The decomposition of mercuric oxide on heating means an oxidation of

oxygen itself, for the negative valence of oxygen decreases from −2 to 0.

$$2HgO \rightarrow 2Hg + O_2$$

The reduction of iron by heating its oxide in a stream of hydrogen

$$Fe_2O_3 + 3H_2 \rightarrow 2Fe + 3H_2O$$

gives a valence change in the opposite direction, for the original valence of +3 changes to zero. In similar fashion, iron in ferric chloride may be reduced by hydrogen.

$$2FeCl_3 + 3H_2 \rightarrow 2Fe + 6HCl \qquad (58)$$

When chlorine reacts with sodium, its negative valence increases from 0 to −1, so this too is an example of reduction.

$$2Na + Cl_2 \rightarrow 2NaCl$$

An increase in the positive valence of a metal means that its atoms have lost electrons, while a decrease means that its atoms have gained electrons. Thus iron atoms combine with either oxygen or chlorine by losing electrons to the nonmetal atoms; to reduce iron from these compounds to the free metal requires that electrons be given to the iron atoms. When a nonmetal increases in negative valence, on the other hand, its atoms gain electrons; when it decreases in negative valence, its atoms lose electrons. So we may simplify our definitions of oxidation and reduction to read: *an element is oxidized when its atoms lose electrons, reduced when its atoms gain electrons.*

We run into difficulties, of course, in trying to apply this definition to covalent compounds. Carbon, we say, is "oxidized" by burning in oxygen,

$$C + O_2 \rightarrow CO_2 \qquad (59)$$

since its valence increases from 0 to +4. The + sign indicates merely that carbon is a less active nonmetal than oxygen, not that electrons are transferred from carbon atoms to oxygen atoms. Again, when phosphorus reacts with chlorine

$$2P + 3Cl_2 \rightarrow 2PCl_3$$

the phosphorus gains in positive valence, although electrons are shared between its atoms and chlorine atoms. For these compounds we shall find it best to keep the definition of oxidation and reduction in terms of valence rather than loss or gain of electrons.

Note that oxidation of one element is always accompanied by reduction of another. For instance, the burning of zinc in chlorine [Eq. (57)] means the reduction of chlorine as the zinc is oxidized; in other words,

electrons *lost* by the zinc atoms are *gained* by chlorine atoms. When carbon burns in oxygen [Eq. (59)], the oxidation of carbon is accompanied by the reduction of oxygen from a valence of 0 to a valence of −2; for each electron shared by a carbon atom, one electron must, of course, be shared by an oxygen atom. Since in chemical reactions atoms cannot lose, gain, or share electrons all by themselves, oxidation and reduction must take place together.

Reactions involving changes of valence are often called *oxidation-reduction reactions*. They make up a large and important group of chemical changes, some of which we have already studied. Here we shall examine in more detail a few typical oxidation-reduction processes.

Displacement Reactions

If a piece of copper wire is covered with a solution of silver nitrate and allowed to stand for a few hours, the wire becomes coated with gray crystals and the solution turns pale blue. The crystals are metallic silver, and the telltale blue color shows the presence of copper ions in the solution. Evidently some copper has become ionized, and at the same time silver has been set free. This reaction is summarized by the equation

$$Cu + 2Ag^+ + 2NO_3^- \rightarrow Cu^{++} + 2Ag + 2NO_3^-$$

Or, since nitrate ion is not affected,

$$Cu + 2Ag^+ \rightarrow Cu^{++} + 2Ag \qquad (60)$$

This is an oxidation-reduction reaction in its simplest form. Each copper atom has lost two electrons, to become a copper ion; each silver ion has gained one electron, to become a neutral silver atom. Copper is oxidized, silver is reduced. We may describe this reaction by saying that copper has *displaced* silver from solution.

A similar reaction takes place when a steel knife blade is held in a solution of copper sulfate. After a few moments the blade is coated with a reddish film of copper, and chemical tests would show the presence of iron ions in solution. The equation is therefore

$$Fe + Cu^{++} \rightarrow Fe^{++} + Cu$$

Iron reduces copper ion to free copper and is itself oxidized to a positive ion. Or we may say that iron has displaced copper from solution.

Iron gives up electrons to copper ion, copper gives up electrons to silver ion. Thus we might arrange the three metals in the order Fe, Cu, Ag, showing their relative abilities to give up electrons. By studying other displacement reactions, we should find that other metals may be added to this series, each metal being capable of giving electrons to the ions of

metals which follow it:

Na Ca Mg Al Zn Fe Pb Cu Hg Ag Au

This sequence is in the order of decreasing ability to lose electrons, or of decreasing ability to reduce the ions of other metals. Magnesium placed in a solution of copper chloride gives its electrons to the copper ions; lead placed in a silver nitrate solution reduces the silver ions. This is likewise the order which we described on an earlier page as the *order of activity* of the metals (pages 197 and 327). Oxidation-reduction reactions, in fact, furnish a precise measure for the activities of different metals.

Hydrogen may be placed in the above sequence, for the solution of a metal in acids is a typical displacement reaction:

$$Zn + 2H_3O^+ \rightarrow Zn^{++} + H_2 + 2H_2O$$

or

$$Zn + 2H^+ \rightarrow Zn^{++} + H_2$$

Zinc gives up electrons to hydronium ion, going into solution as zinc ion and setting hydrogen free. To find where hydrogen belongs in the series, we need only see which metals will dissolve in acids and which will not. We should find that all the metals from Na to Pb will reduce H_3O^+, while those from Cu to Au are unaffected by ordinary acids. Hence hydrogen belongs between Pb and Cu.

Displacement reactions among nonmetals are shown especially well by the halogens. If chlorine is added to a solution of potassium bromide, for instance, the solution turns brownish because bromine is set free:

$$Cl_2 + 2Br^- \rightarrow Br_2 + 2Cl^- \qquad (61)$$

Chlorine oxidizes bromide ion; in other words, each chlorine atom takes an electron from Br^-, setting free a Br atom and itself becoming ionized. Similarly bromine displaces iodine, and fluorine displaces any one of the other halogens. By means of these reactions, nonmetals also may be arranged in an activity series (page 198):

F Cl Br O I S

This sequence is in the order of decreasing ability to gain electrons, or of decreasing ability to oxidize the ions of other nonmetals. Note that the activity of nonmetals is measured by their oxidizing ability, that of metals by their reducing ability—which is just another way of stating the fundamental fact that metals enter chemical combination by losing electrons to other elements, whereas nonmetals enter chemical combination by gaining electrons.

Note that oxidation-reduction reactions mean a transfer of *electrons*, while acid-base reactions mean a transfer of *protons*. In many respects the two kinds of reactions are analogous: Thus an acid is strong if it gives

up protons readily, and a reducing agent is strong if it gives up electrons readily; an acid, no matter how strong, cannot give up protons unless a base is present to receive them, and a reducing agent cannot give up electrons unless an oxidizing agent is present to receive them.

Compounds of Iron

Many elements show two or more different valences in their compounds. Mercury, for example, forms compounds in which it has a valence of +1, like Hg_2O, Hg_2Cl_2, $Hg_2(NO_3)_2$, Hg_2SO_4 (mercur*ous* oxide, mercur*ous* chloride, etc.), and other compounds in which it has a valence of +2, like HgO, $HgCl_2$, $Hg(NO_3)_2$, $HgSO_4$ (mercur*ic* oxide, etc.). Tin in some compounds ($SnCl_2$, SnO) has a valence of +2, in others a valence of +4 ($SnCl_4$, SnO_2). Carbon forms the two oxides CO and CO_2, sulfur the two oxides SO_2 and SO_3. Changes from one valence state to another are, of course, oxidation-reduction reactions. For instance, carbon is oxidized when carbon monoxide is burned,

$$2CO + O_2 \rightarrow 2CO_2$$

and sulfur is reduced when sulfur trioxide is strongly heated:

$$2SO_3 \rightarrow 2SO_2 + O_2$$

Especially important are valence changes among the compounds of iron. This element shows two principal valences, +2 and +3; thus it forms the two oxides FeO (ferr*ous* oxide) and Fe_2O_3 (ferr*ic* oxide), the two chlorides $FeCl_2$ and $FeCl_3$, the two sulfates $FeSO_4$ and $Fe_2(SO_4)_3$. Many compounds of each type are soluble in water, ionizing to give, respectively, the ions Fe^{++} and Fe^{+++}. Most ferrous compounds in solution are pale green in color, most ferric compounds pale yellow. (Solutions of ferric compounds often show stronger tints of yellowish to reddish brown, because ferric ion has a tendency to form complex ions by attaching itself to water molecules and to other ions.) As a general rule, iron compounds are strongly colored, both in solution and in the solid state: ferric compounds have shades ranging from yellow to brown and red, while ferrous compounds are gray, green, or black.

Changes from one valence state of iron to the other are easily brought about. If chlorine, for example, is bubbled into a solution of ferrous sulfate, the greenish color of the solution quickly changes to yellow, showing that Fe^{++} has been converted to Fe^{+++}:

$$Cl_2 + 2Fe^{++} \rightarrow 2Cl^- + 2Fe^{+++}$$

Each chlorine atom oxidizes a ferrous ion by taking one electron away from it. One way of effecting the opposite change is to add an iodide,

say KI, to a solution of a ferric compound; the yellow solution changes to deep brown as iodine is liberated:

$$2I^- + 2Fe^{+++} \rightarrow I_2 + 2Fe^{++}$$

Each iodide ion gives an electron to a ferric ion, thereby reducing it to ferrous ion. Many other oxidizing and reducing substances besides Cl_2 and I^- may be used in these reactions.

In the presence of air, most ferrous compounds are slowly oxidized to ferric compounds by atmospheric oxygen. A solution of ferrous sulfate allowed to stand in an open dish for several days loses its clear green color, becoming yellowish and developing on its surface a yellow-brown scum of ferric hydroxide [$Fe(OH)_3$]. Another way to demonstrate this oxidation is simply to leave a moist knife blade or razor blade exposed until it begins to rust. The formation of rust is a complex process, involving first the formation of Fe^{++} by carbonic acid in the water (formed, of course, by solution of CO_2 from the air), then the further oxidation of Fe^{++} by oxygen. The final product is ferric oxide (Fe_2O_3) combined with a considerable amount of water. The extent to which the oxide is combined with water, or "hydrated," determines the color of the rust: Fe_2O_3 itself is brick red, but in hydrated form its color changes to yellow and brown.

Deep in the earth, at high temperatures and out of reach of the atmosphere, ferrous compounds are more stable than ferric compounds. Most of the iron in rocks formed at these depths, therefore, goes into ferrous compounds, chiefly ferrous silicates. When such rocks appear at the surface, either brought up by volcanic activity or exposed by long erosion, their iron compounds are no longer stable. Slowly oxygen attacks the iron, aided again by the slight solvent action of carbonic acid in streams and rain water. The complex silicates are broken down, and the iron is oxidized as in ordinary rusting to various hydrates of Fe_2O_3. The rusty stains on the surfaces and in cracks of so many common rocks are due to this iron oxide, formed by slow oxidation of ferrous compounds present in the rocks.

Many of the bright and somber hues in nature we owe to the colorful compounds of this one element. Not only the rusty surface stains, but yellow-brown and reddish-brown colors in rocks themselves are nearly always due to ferric compounds. Black rocks and green rocks usually get their colors from ferrous compounds. Sand, clay, and soil likewise owe their brown tints to ferric compounds, their gray and black shades often to ferrous compounds. The colors of ferric compounds are so strong that only a little iron is needed to color a rock or a soil conspicuously.

Reduction of ferric compounds at the earth's surface is accomplished chiefly by carbon compounds produced by plants and animals. Soils

containing much organic matter are commonly black, since their iron is kept in the form of ferrous compounds whose relatively weak colors cannot obscure the dark carbonaceous materials. Often we find a thin layer of black soil resting on brown soil, the dark color showing the depth to which decaying organic matter keeps the iron of the soil reduced.

Thus in nature we find continuous transformations of ferric compounds to ferrous compounds and back again.

Oxygen in the Atmosphere

We have discussed at some length the effect of atmospheric oxygen on iron and its compounds. We extend the discussion now to a few other oxidation processes in which free oxygen plays a part.

Oxygen is a powerful oxidizing agent, but most of its reactions under ordinary conditions are so slow that it seems relatively inactive. Moderately high temperatures speed its reactions greatly, and the pure gas is more active than the dilute mixture with nitrogen which makes up ordinary air. The presence of water containing hydronium ion also increases the rate at which oxygen reacts with other substances; even the minute concentration of hydronium ion formed by solution of carbon dioxide in water is enough to speed up oxidations enormously. How great the effect of carbonic acid can be is readily shown by exposing to the air two polished iron surfaces, one damp and the other dry: while the dry surface retains its polish for weeks, the damp one begins to rust in a few hours.

Oxidation of Metals. All active metals react slowly with atmospheric oxygen, just as iron does. Some form simple oxides, others react further with water or carbon dioxide to form hydroxides or carbonates. Sodium and potassium "rust" so quickly on exposure to air that they must be kept under oil. Some active metals, like zinc and aluminum, show no visible rusting because their oxides, formed quickly on fresh surfaces, cling to the metal as an impervious film which prevents further oxidation.

Spontaneous Combustion. Carbon and most of its compounds oxidize so very slowly under ordinary conditions that we commonly think of them as stable substances. Some carbon compounds, however, do react with oxygen at perceptible rates. Usually the heat given out by these reactions is dissipated into the surrounding air, but if free circulation of air is prevented, the heat may gradually accumulate. Thus when coal is piled too deeply, heat formed by slow oxidation of coal within the pile cannot escape and may ultimately raise the temperature sufficiently to ignite the pile. Piles of oily rags are another source of danger, for many oils oxidize slowly in air. Burning started in this manner, by accumulation of heat from slow oxidation, is often called "spontaneous combustion."

Decay. The decay of organic matter is principally a slow oxidation by atmospheric oxygen. The processes of decay may be extremely com-

plex, but the chief ultimate products are the two which nearly always result from the complete oxidation of carbon compounds—CO_2 and H_2O. Nitrogen in the original compounds is changed principally to ammonia (page 377). Decay is greatly accelerated by moisture and also by the presence of bacteria which can use the energy of the slow oxidation for their life processes.

Oxidation of Food. The complex carbon compounds that form the principal part of our food are broken down by the processes of digestion into simpler compounds that can be transported by the blood stream and stored in various parts of the body. To obtain energy from food, our bodies must oxidize these stored compounds. The necessary oxygen is taken in through the lungs and is carried to the tissues by a substance in the blood called *hemoglobin*. Probably the oxidation is aided by complex organic catalysts. The products of oxidation, besides energy, are CO_2 and H_2O, the former being carried by the blood to the lungs and thence exhaled.

To summarize briefly: oxygen in the air is a good oxidizing agent but under ordinary conditions reacts slowly. Its rates of reaction in different circumstances may be speeded up by heat, by the presence of dilute carbonic acid, by bacteria, and by catalysts. Active metals, organic substances, and other oxidizable materials are always subject to slow attack by oxygen when exposed to the air.

Metallurgy

Oxidation and reduction reactions are highly important in the various processes by which metals are obtained from their ores. These processes are included in the term *metallurgy*.

An *ore* of a metal is any naturally occurring material from which the metal can be extracted profitably. In an ore, a metal may occur as the free element or in a compound, mixed usually with large amounts of valueless material called *gangue*. For example, a common gold ore consists of particles of free metal in a gangue of sand and gravel. The first problem of metallurgy is to concentrate the ore, that is, to remove as much of the gangue as possible. This is usually a physical problem rather than a chemical one. A common method of concentration is a large-scale application of the gold-miner's pan: the crushed ore is shaken, usually under water, and the heavier metallic particles settle to the bottom. More complicated methods are necessary when the metallic part of the ore has nearly the same density as the gangue.

The metallurgist must then separate the metal from the remainder of the gangue and from any elements which are combined with it. This is simple for the few metals which occur as free elements—platinum, gold, silver, more rarely copper. These may be separated from the gangue

by melting the metal, or by adding a solvent which will remove the metal in solution. Because gold, silver, and copper are so easily obtained from their ores, these were the first metals used by primitive man.

Other metals occur in their ores as compounds, of which the most important types are oxides (Fe_2O_3, SnO_2), sulfides (MoS, ZnS), and carbonates ($MnCO_3$, $PbCO_3$). From these the free metals are obtained by various chemical processes involving reduction. How difficult the reduction is for a given metal is suggested with fair accuracy by the date of its discovery: the metals most easily reduced from their ores were known to primitive man, a few were discovered in the ancient Mediterranean civilizations, a few more were added by the alchemists, whereas those reduced with most difficulty were discovered only in recent times.

Copper, available to primitive man only in limited quantities as the free element, was the first metal which he learned to reduce from its compounds. Probably he discovered the process accidentally, by dropping a chunk of rock in the hot charcoal of his campfire and observing that a few shining drops of liquid metal were formed. In modern language we should say that a copper compound in the rock had been changed by heating to copper oxide (CuO) and that this oxide had been reduced by hot carbon in the charcoal:

$$2CuO + C \rightarrow 2Cu + CO_2$$

Much later, about 1500 B.C., primitive man learned that another sort of rock heated with charcoal produced the metal tin, which could be mixed with copper to give the valuable alloy *bronze*. We would write the equation for the reducing of tin ore as

$$SnO_2 + C \rightarrow Sn + CO_2$$

Lead and mercury, produced by similar crude processes, were known to the ancient Egyptians and Babylonians.

Iron presented a more difficult metallurgical problem. In its principal ores, iron occurs as ferric oxide—the simple oxide in the ore mineral *hematite*, the hydrated oxide in the mineral *limonite*. Like the oxides of copper and tin, ferric oxide can be reduced with hot charcoal, but the reaction requires a higher temperature. Separation of the molten metal from gangue required a further improvement in technique: some substance had to be added as a *flux*, to combine with the gangue and make it liquid; the liquid, a light, glassy substance which we call *slag*, collected on top of the molten iron and could be easily removed. Although a little iron found in meteorites was used even in primitive times, the trick of reducing it from its ores was not learned until about 600 B.C. Large-scale production of iron began only about A.D. 1600, when coal was substituted for charcoal in the reduction process. In modern *blast*

furnaces, huge steel towers (40 to 100 ft high) lined with firebrick, iron is reduced by a continuous process (Fig. 183): a mixture of ore with coke and flux is fed in at the top, heated air is forced in near the bottom to burn the coke and so provide heat for the reaction, and molten iron and slag are drawn off at intervals from the bottom. The flux used depends on the nature of the ore; in most iron ores silicon dioxide is the chief impurity, and for these limestone is the best flux. The reactions which take place in the blast furnace may be summarized in the following equations, although the reduction of the oxide and the formation of slag actually take place in steps:

$$2Fe_2O_3 + 3C \rightarrow 4Fe + 3CO_2$$
$$CaCO_3 + SiO_2 \rightarrow CO_2 + CaSiO_3 \text{ (slag)}$$

Blast furnace slag, as we shall find presently, has a composition similar to that of ordinary glass.

The alchemists added five metals to those known in ancient times: antimony, bismuth, zinc, arsenic, cobalt. Several others were discovered before the end of the eighteenth century, including nickel, manganese, tungsten, molybdenum, chromium, rare metals which have become familiar today because of their use in steel alloys.* Some of these metals, like the earlier metals, were reduced from their ores by reduction with carbon, but others required more elaborate treatment.

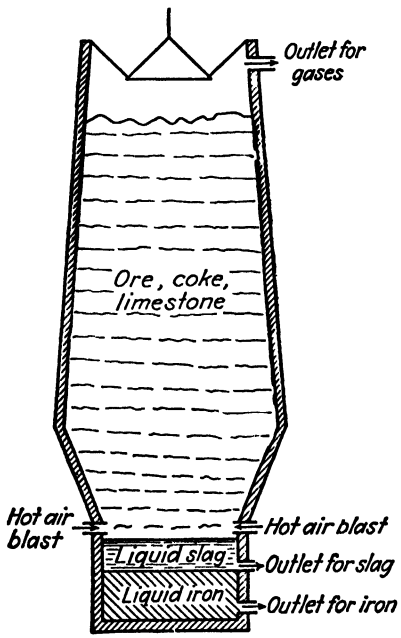


FIG. 183. *Diagram of a blast furnace.*

In the nineteenth century metallurgy gained a powerful new tool, electrolysis. The processes of electrolysis are oxidation and reduction reactions reduced to bare essentials: at the cathode, electrons are given directly to one substance, and at the anode electrons are removed from another substance (page 331). To reduce a metal from a compound by electrolysis, the compound must be liquefied or dissolved so that free metal ions may be liberated; then the passage of a current attracts the positive metal ions to the cathode, where electrons are supplied to reduce

* An *alloy* is a mixture of two or more metals. *Steel* is an alloy consisting of iron mixed with small amounts of other metals and carbon.

them. By means of electrolysis we can today reduce even the most active metals, like sodium, potassium, and calcium. Electrolysis also furnishes a convenient method for the final purification of less active metals, such as copper, tin, zinc, and silver.

One of the great triumphs of electrolysis in metallurgy was the cheap production of the light metal aluminum. Aluminum was first prepared early in the nineteenth century, by reducing its chloride with potassium, but the process was so expensive that the metal's only commercial use for some years was in jewelry. Aluminum is an extremely active element, and its compounds resisted all attempts to decompose them by any reducing agents less powerful and less expensive than the alkali metals. Attempts to produce the metal commercially by electrolysis were balked at first by the difficulty of liquefying or dissolving aluminum ores. The problem was finally solved in 1886 by a young American, Charles Martin Hall, at that time just out of college. His great discovery was that cryolite, a mineral obtained in large quantities from Greenland, when melted would dissolve the chief aluminum ore (Al_2O_3). When a current is passed through the solution between graphite electrodes, aluminum is liberated at the cathode and oxygen at the anode. This process makes possible the manufacture of aluminum in quantities sufficient to satisfy the demands of modern industry.

Metallurgy today includes a great variety of chemical processes, but the two principal ones remain the ancient process of reducing ores with hot carbon, and the recent process of reducing metallic ions by electrolysis.

Questions

1. Which of the following equations represent oxidation-reduction reactions?

$$Zn + S \rightarrow ZnS$$
$$H_3O^+ + OH^- \rightarrow 2H_2O$$
$$MnO_2 + 2Cl^- + 4H_3O^+ \rightarrow Cl_2 + Mn^{++} + 6H_2O$$
$$CaO + H_2O \rightarrow Ca^{++} + 2OH^-$$

2. In each of the following reactions, pick out (*a*) the element which is oxidized (*b*) the element which is reduced, (*c*) the element whose atoms gain electrons, (*d*) the element whose atoms lose electrons:

$$Mg + 2H_3O^+ \rightarrow H_2 + Mg^{++} + 2H_2O$$
$$Ca + S \rightarrow CaS$$
$$2Na + 2H_2O \rightarrow 2Na^+ + 2OH^- + H_2$$
$$F_2 + 2Br^- \rightarrow Br_2 + 2F^-$$
$$2Fe^{+++} + 3H_2S + 6H_2O \rightarrow 2FeS + S + 6H_3O^+$$

3. Which loses electrons more easily, Na or Fe? Al or Ag? I^- or Cl^-? Which gains electrons more easily, Cl or Br? Hg^{++} or Mg^{++}?
4. In what part of the periodic table are the elements which are most easily reduced? In what part are those which are most easily oxidized?
5. What would you expect to happen when a knife blade is held in a solution of silver nitrate? Write an equation to show the reaction. (The iron forms Fe^{++}.)

6. Which of the halogens will displace bromine from solution? Which will bromine displace from solution? Write an equation for one of these reactions.
7. How could you demonstrate that magnesium is a better reducing agent (*i.e.*, more easily oxidized) than hydrogen?
8. A common gangue material in ore deposits is the brassy yellow mineral *pyrite*, or "fool's gold" (FeS_2). Account for the yellow-brown and red-brown colors often found in rocks and soil near such deposits.
9. Account for the fact that the color of soil in marshes is nearly always black.
10. Would you expect the danger from spontaneous combustion to be greater in a coal pile containing principally large chunks or in one containing finely pulverized coal? Why?
11. One of the compounds whose oxidation provides our bodies with energy is a simple sugar with the formula $C_2H_{12}O_6$. Write an equation showing the oxidation of this compound. Where does the necessary oxygen come from? What becomes of the products?
12. Mercury is recovered from its principal ore (cinnabar, HgS) by simply heating the ore in air without any reducing agent. The following reaction takes place:

$$HgS + O_2 \rightarrow Hg + SO_2$$

Which elements change in valence during this process? Which ones are reduced and which oxidized?

13. Explain how the reduction of sodium is accomplished by the electrolysis of molten sodium chloride.

CHAPTER XXIX

Carbon Compounds

CARBON has figured prominently in many discussions of other chapters. We have studied the two forms of elementary carbon, diamond and graphite; the two oxides, carbon monoxide and carbon dioxide; the compounds with oxygen and metals called carbonates; and more briefly some of the hydrocarbons and carbohydrates. We undertake now a more detailed study of this remarkable element.

Because carbon compounds are the chief constituents of living things, the chemistry of carbon is often called *organic chemistry*, while the chemistry of all the other elements is included under *inorganic chemistry*. At one time it was thought that carbon compounds, with the exception of the oxides, the carbonates, and a few others, could be produced only by plants and animals, or from other compounds produced by plants and animals. Carbon was supposed to unite with other elements only under the influence of a mysterious "vital force" possessed by living things. This ancient idea was exploded in 1828 by the German chemist Wöhler, who prepared the "organic" compound urea by heating the "inorganic" compound ammonium cyanate. Since Wöhler's time a great number of organic compounds have been made in the laboratory from inorganic materials, but the general distinction between the chemistry of carbon compounds and inorganic chemistry remains a useful one.

The most astonishing single fact about the carbon compounds is their number. Over 300,000 have been identified, more than ten times the number of all other known compounds. What peculiarity of this element carbon enables it to enter into so many different combinations?

For answer we look first to the periodic table. Carbon stands at the head of the middle group, which means that it has a small atom with four valence electrons. Unable either to lose these electrons completely or to capture enough additional ones to fill a shell of eight, carbon forms compounds exclusively by sharing electron pairs. Its bonds with other atoms are covalent and are especially strong because of the small size of the

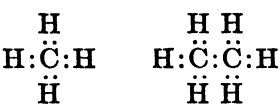
carbon atom. A carbon atom can attach itself firmly not only to the atoms of many metallic and nonmetallic elements *but to other carbon atoms as well.* How strong the attraction between carbon atoms can be is suggested by the hardness of diamond, in which each carbon atom is linked to four others by electron-pair bonds. In the molecules of carbon compounds are long chains, branching chains, and closed rings of carbon atoms linked together. *It is this unique capacity of carbon atoms to join together that makes possible the prodigious number of carbon compounds.* A few elements near carbon in the periodic table—boron, silicon, nitrogen—show this same ability to a slight extent, but their chains of atoms are short and unstable.

Because the bonds formed by carbon atoms are covalent, carbon compounds are mostly nonelectrolytes, and their reactions are usually slow. The great attraction of carbon atoms and hydrogen atoms for oxygen makes many organic compounds subject to slow oxidation in air and rapid oxidation if heated. Even in the absence of air, organic compounds are in general stable only at ordinary temperatures, few resisting decomposition at temperatures over a few hundred degrees centigrade.

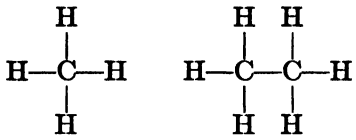
Structural Formulas

The formula of an organic compound is determined in the same manner as other formulas: the molecular weight is found by measuring the density of its vapor (or by other methods), and the proportions by weight of the different elements are found by analysis. Like inorganic formulas, the formula of a carbon compound tells us how many atoms of each kind are present in a molecule. But this information may not be sufficient to describe an organic substance accurately, for often *two or more entirely different compounds have the same formula.* For instance, analysis shows that three different liquids have the formula $C_2H_4O_2$: one of these is the familiar acid of vinegar, acetic acid, from which the other two are immediately distinguishable by their odors.

Evidently some device besides the simple formula is needed to describe the properties of an organic compound. One such device is a diagrammatic formula showing not only the number but the arrangement of atoms in each molecule. For example, the simple formulas CH_4 (methane) and C_2H_6 (ethane) may be written

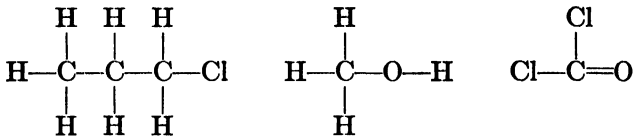


the dots representing the paired valence electrons of C and H. Dashes are often used instead of electron pairs to represent the bonds:



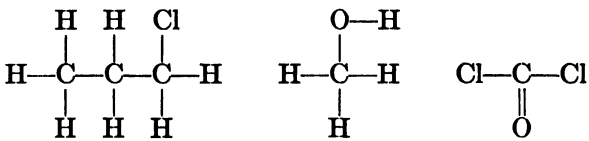
Diagrammatic formulas of this sort are called *structural formulas*, while the simpler CH_4 and C_2H_6 are called *molecular formulas*. Besides the information which the molecular formulas give, these structural formulas show that in both methane and ethane molecules each hydrogen atom is attached to a carbon atom, and that in ethane the two carbon atoms are linked together.

Structural formulas are written according to the ordinary rules of valence. Each carbon atom, with its valence of 4, must be connected with other atoms by four bonds or dashes; each hydrogen or chlorine atom can have only one dash; each oxygen atom has two. Since the bonds are all covalent, the positive or negative character of the valence is immaterial. The following formulas illustrate the valence rules:

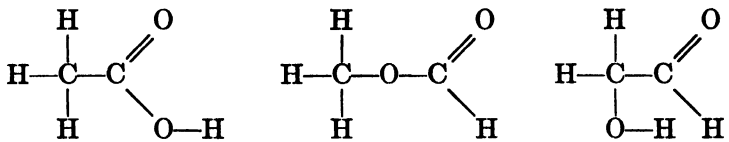


Propyl chloride (C_3H_7Cl) Methyl alcohol (CH_4O) Phosgene ($COCl_2$)

Since the formulas show only which atoms are attached together, the arrangement of atoms around each carbon atom does not matter. Thus the above formulas might equally well be written



Now if we try to write a structural formula corresponding to $C_2H_4O_2$, we find at once a reason for the existence of three different substances with this same molecular formula. Three possible arrangements of atoms satisfy the rules of valence:

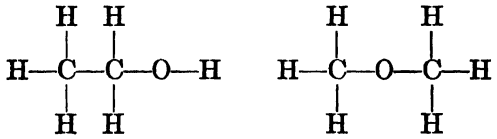


Acetic acid Methyl formate Glycolic aldehyde

The differing properties of the three compounds are determined by these different atomic arrangements. *Compounds of this sort, with the same molecular formulas but different structural formulas, are called* ***isomers.*** **Examples of isomerism are very numerous in organic chemistry and**

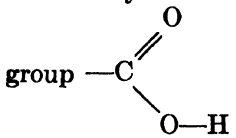
supply one excellent reason for the use of structural formulas rather than molecular formulas.

Figuring out the possible structural arrangements corresponding to a given molecular formula is a pencil-and-paper pastime, akin to solving crossword puzzles. But to decide which of several possible structures best describes the properties of a given substance is more difficult, requiring extensive laboratory tests. For instance, the valence rules suggest two possible structures for ethyl alcohol (C_2H_6O).



Laboratory tests like the following show that the first of these is preferable: (1) Sodium reacts with alcohol, liberating hydrogen and forming the compound C_2H_5ONa. No further reaction with the other five H's of the alcohol molecule takes place, so that one H must be attached in a different manner from the others. (2) Alcohol reacts with HCl to give water and the gas ethyl chloride (C_2H_5Cl). An O and an H have been replaced by a Cl atom, which suggests that the O and H were together in the original molecule.

Simple examples of this sort give the organic chemist a background of experience which enables him to solve more difficult problems. For instance, suppose that he is asked to determine which of the three formulas on page 403 best fits the properties of an unknown liquid with the molecular formula $C_2H_4O_2$. Experience tells him that simple organic compounds containing the group OH are nearly always soluble in water, so if the liquid is not miscible with water he would suspect at once that it is methyl formate. He knows further that compounds with the larger



group (often written COOH) are acids, so that a red color with litmus would suggest that the unknown compound was acetic acid. He knows that moderately strong oxidizing agents will oxidize the CH_2OH group (as in glycolic aldehyde) to COOH, whereas structures like acetic acid and methyl formate would be affected only by very strong oxidizing agents. Thus he looks in the structural formulas for various *groups of atoms*, which past experience tells him will give the molecules certain properties, and then compares the observed properties with those which each formula suggests.

Just as atom groups like SO_4 and NO_3 appear in formula after formula of inorganic chemistry, so do various groups appear in organic

structural formulas, each group giving a compound certain recognizable characteristics. The properties of an organic compound are the sum of the properties of its various atom groups, each one modified somewhat by the presence of the others. By investigating the properties of a substance, an organic chemist can usually deduce what groups are present, and by piecing them together can make a good guess at the formula. Organic chemistry is a study in molecular architecture, dealing with arrangements of atoms in larger units and the fitting together of these units to make molecules.

Structural formulas are not strictly accurate pictures of molecules. While they probably do show something about the actual connections between atoms, they obviously cannot portray the true positions of the atoms in space, for molecules are 3-dimensional structures rather than 2-dimensional. *Structural formulas are primarily a means for summarizing concisely the experimentally determined properties of organic compounds.* Their amazing ability to accomplish this is largely responsible for the development of modern organic chemistry.

Hydrocarbons

The simplest organic substances are the ***hydrocarbons***, compounds which contain only the two elements carbon and hydrogen. These compounds will merit detailed study, not only because many of them are important industrially, but also because they illustrate in simple fashion how properties vary with differences in molecular structure.

TABLE XXI. THE METHANE SERIES OF HYDROCARBONS

(Constituents of natural gas and petroleum)

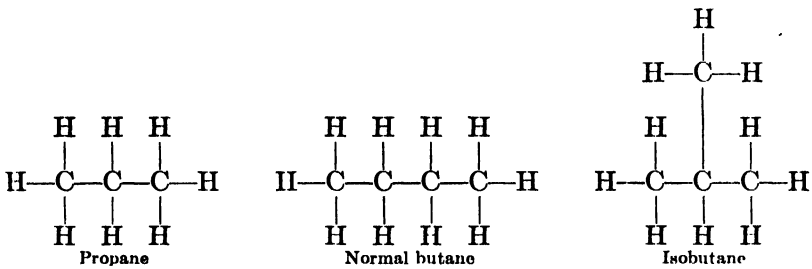
Formula	*Name*	*Freezing point, °C*	*Boiling point, °C*	*Density*	
CH_4	Methane	−184	−161		Fuel gases
C_2H_6	Ethane	−172	− 88		
C_3H_8	Propane	−190	− 45		
C_4H_{10}	Butane	−135	1		
C_5H_{12}	Pentane	−132	36	0.631	Petroleum ether
C_6H_{14}	Hexane	− 94	69	0.660	(naphtha)
C_7H_{16}	Heptane	− 90	98	0.684	
C_8H_{18}	Octane	− 57	125	0.707	Gasoline
C_9H_{20}	Nonane	− 51	151	0.718	
$C_{10}H_{22}$	Decane	− 32	174	0.747	
$C_{11}H_{24}$	Undecane	− 27	197	0.741	Kerosene
. . .	. . .	. . .	. . .	. . .	
$C_{16}H_{34}$	Hexadecane	+ 20	288	0.775	

$C_{17}H_{36}$ to $C_{22}H_{46}$, semisolids, constituents of vaseline and lubricating oil

$C_{23}H_{48}$ to $C_{29}H_{60}$, constituents of paraffin

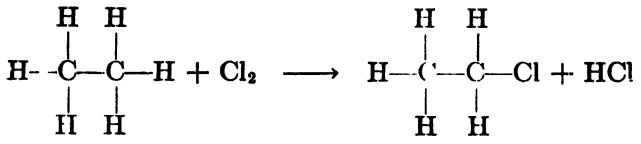
Methane Series. Hydrocarbons of this series are the chief constituents of petroleum and natural gas (page 373). The series includes a long list of compounds with the general formula C_nH_{2n+2}—*i.e.*, the number of H atoms per molecule is always 2 more than twice the number of C atoms. The names, formulas, and a few physical properties of several members of the series are shown in Table XXI. Note that density, boiling point, and freezing point increase regularly with increasing molecular weight. The first four members of the series are gases at ordinary temperatures, those from pentane to hexadecane are liquids, those with molecules heavier than hexadecane are solids.

For methane, ethane, and propane only one structural formula is possible. In the butane molecule, however, the atoms may be arranged in two different ways; corresponding to the two possible arrangements, two isomers of butane with slightly different properties are known experimentally. The formulas for methane and ethane have already been given (page 403); those for propane and the two butanes are



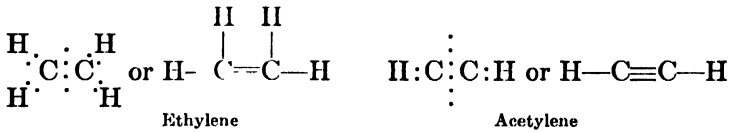
The number of possible isomers corresponding to a given molecular formula increases enormously as the molecular weight increases. For pentane three arrangements are possible; $C_{13}H_{28}$ has 813 theoretically possible isomers, and $C_{20}H_{42}$ has 366,319. Only a few of the possible isomers have actually been prepared.

Molecules of the methane hydrocarbons have either "straight chains" of carbon atoms, as in propane and normal butane, or "branched chains," as in isobutane. Such symmetrical structures we might expect to give nonpolar molecules, that is, molecules with neither end appreciably more positive or negative than the other. In accordance with this nonpolar character, the methane hydrocarbons are insoluble in water (page 339). Chemically they are unreactive: neither concentrated acids and bases nor most oxidizing agents will affect them at ordinary temperatures. When ignited, they burn readily in air or oxygen. Chlorine reacts with them slowly at room temperatures in the presence of light, chlorine atoms being substituted for one or more hydrogen atoms in the hydrocarbon molecules:



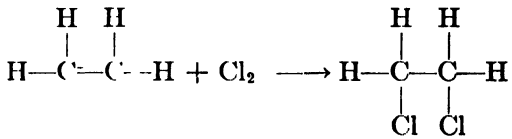
(62)

Unsaturated Hydrocarbons. The gases ethylene and acetylene have the formulas C_2H_4 and C_2H_2, respectively. Structural formulas for these gases are impossible to write with simple electron-pair bonds: there simply are not enough H atoms to go around. We can stick to the rule of 4 valences for each carbon atom only by supposing that two carbon atoms can share more than one electron pair between them. In ethylene the atoms apparently share two pairs, in acetylene three pairs. Such linkages are called *double bonds* and *triple bonds*, and are represented by two dashes and three dashes, respectively:

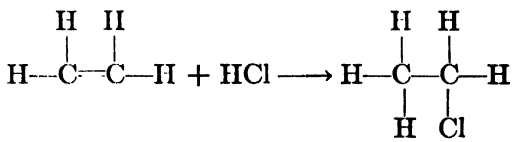


Ethylene Acetylene

Compounds of this sort are much more reactive than the methane hydrocarbons. Both Cl_2 and HCl combine readily with ethylene, for instance,



(63)

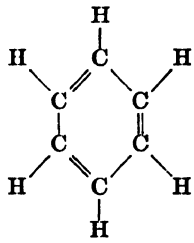


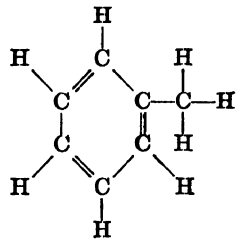
(64)

The other halogens and many other acids behave similarly. Since compounds with double and triple bonds can thus react by *adding* other atoms to their molecules, they are called *unsaturated compounds*, in distinction to *saturated compounds* like the methane hydrocarbons.

Turpentine and rubber are examples of unsaturated hydrocarbons with far more complicated molecules.

Benzene Hydrocarbons. Unsaturated hydrocarbons of a special type, with closed rings of carbon atoms in their structural formulas, are obtained as by-products when coal is heated to form coke. The simplest of these are benzene (C_6H_6) and toluene (C_7H_8).

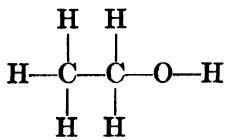

Benzene

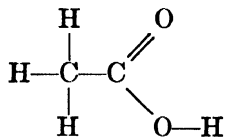

Toluene

Naphthalene, the white solid used in moth balls, is a hydrocarbon with a more complicated ring structure in its molecule. In general, these compounds do not react by addition as readily as do other unsaturated hydrocarbons, the ring structure apparently making the double bonds more stable.

Simple Hydrocarbon Derivatives

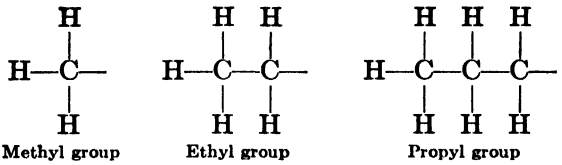
The variety and complexity of organic compounds containing only two elements is amply demonstrated by the hydrocarbons. With one or two more elements added, the number of possible compounds grows bewilderingly large. To simplify the problem of classifying these compounds, they are often considered as *derivatives* of hydrocarbons—*i.e.*, as compounds obtained by substituting other atoms or atom groups for some of the H atoms in hydrocarbon molecules. Ordinarily carbon compounds are not prepared in this manner, but their structural formulas suggest that they might be. For instance, ethyl alcohol and acetic acid may be considered as derivatives of ethane and methane, respectively.


Ethyl alcohol


Acetic acid

The formula of alcohol is derived from that of ethane by substituting an OH group for an H atom, and that of acetic acid is derived from CH_4 by substituting a COOH group for an H atom. No one in his right mind would try actually to prepare alcohol or acetic acid from these hydrocarbons, except as a laboratory stunt, but their formulas indicate that the preparation would be possible.

The carbon-hydrogen atom groups which appear in hydrocarbon derivatives are named from the hydrocarbons. Groups corresponding to the hydrocarbons methane, ethane, and propane are



These groups are often abbreviated CH_3, C_2H_5, C_3H_7. Thus the compound CH_3Cl is methyl chloride, C_3H_7I is propyl iodide, C_2H_5OH is ethyl alcohol, $CH_3C_2H_5SO_4$ is methyl ethyl sulfate.

Halogen Derivatives. One or more of the H atoms in a hydrocarbon molecule may be replaced by halogen atoms, giving compounds like CH_3Br, CH_2I_2, C_2H_5Cl, C_7H_7Cl. The simpler compounds of this sort are gases and volatile liquids, but, as in the methane series, their boiling points and melting points rise with increasing molecular weight. They may be prepared directly by the illumination of mixtures of hydrocarbons and halogens, or by the addition of halogens and halogen acids to unsaturated hydrocarbons [Eqs. (63) and (64)], but are more conveniently made indirectly from alcohols. The halogen derivatives are particularly important to the organic chemist because the halogen atoms are easily replaced by other groups in building up complex molecules. A few of the simpler ones are useful for other purposes: $CHCl_3$ is the anesthetic chloroform; CCl_4 is carbon tetrachloride, an important cleaning fluid; CCl_2F_2 (dichlorodifluoromethane) is a gas used in electric refrigerators.

Alcohols. These are a group of hydrocarbon derivatives in which one or more H atoms in the molecule have been replaced by OH groups. The two commonest members of the group are ethyl alcohol or grain alcohol (C_2H_5OH) and methyl alcohol or wood alcohol (CH_3OH). The OH group makes alcohol molecules somewhat polar, so that the simpler alcohols are soluble in water. The polarity is not great enough, however, to prevent alcohols from mixing also with a great variety of less polar organic substances. These properties make alcohols, especially ethyl and methyl alcohol, valuable in industry as solvents.

Alcohols are, so to speak, organic hydroxides, but unlike inorganic hydroxides they do not ionize appreciably in water. They react slowly with acids, forming compounds called *esters* (see below):

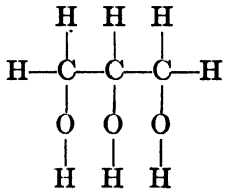
$$CH_3OH + HCl \rightarrow CH_3Cl + H_2O \qquad (65)$$

Like water, alcohols react with alkali metals to liberate hydrogen:

$$2C_2H_5OH + 2Na \rightarrow 2C_2H_5ONa + H_2$$

Because either OH or H may thus be replaced by other atoms or atom groups, the alcohols like the halogen derivatives are important as intermediates in the preparation of complex molecules.

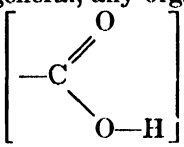
A familiar alcohol with more than one OH group in its molecule is the sweetish, viscous liquid *glycerin:*



Substitution of an OH group in the molecule of a benzene hydrocarbon produces a compound with properties somewhat different from ordinary alcohols. The simplest is the antiseptic *phenol*, or carbolic acid (C_6H_5OH), a very weak but highly poisonous acid.

Organic Acids. Partial oxidation of ethyl alcohol gives CH_3COOH (acetic acid). In previous chapters we have written this formula $HC_2H_3O_2$. Either formula indicates that only one of the four H atoms is capable of ionizing, but the former shows a little more about the molecular structure. Acetic acid itself is a colorless liquid, freezing at 17°C and boiling at 118°C, miscible with water in all proportions. It is used industrially as a solvent and in the manufacture of dyes, drugs, flavors, and plastics.

In general, any organic compound whose molecule contains the group



is an acid, the H atom of this group being capable of ionizing. Most of these organic acids are very weak. Acids corresponding to the methane hydrocarbons form a series, as do the halogen derivatives and the alcohols, in which the boiling points and freezing points increase steadily with increasing molecular weights. The simpler members of the series, like acetic acid, are soluble in water, the more complex ones insoluble. Three members of the series are butyric acid (C_3H_7COOH), the acid of rancid butter; capric acid ($C_9H_{19}COOH$), which has an odor like that of goats; stearic acid ($C_{17}H_{35}COOH$), whose sodium salt is a constituent of soap. More complex organic acids are the citric acid of citrus fruits, the tartaric acid of grapes, the lactic acid of sour milk.

Organic acids are often produced as intermediate steps in the decay of organic matter. Black soils with abundant decaying plant material, for instance, are often too acidic for the successful growing of crops, unless lime or some other basic substance is added to neutralize the acidity. The acids of decay often aid in the slow disintegration of rocks.

Esters. Alcohols react with acids to produce water and compounds called esters [Eq. (65)]. These reactions are superficially similar to the neutralization of an acid by a base, but in contrast to neutralization are

slow and incomplete. Esters are analogous to the salts of inorganic chemistry but are nonelectrolytes. They may be formed from either organic acids or inorganic acids:

$$2CH_3OH + H_2SO_4 \rightarrow 2H_2O + (CH_3)_2SO_4 \text{ (methyl sulfate)}$$
$$C_2H_5OH + HOOCCH_3 \rightarrow H_2O + C_2H_5OOCCH_3 \text{ (ethyl acetate)}$$

Many esters have pleasant flowerlike or fruitlike odors, and find extensive use in perfumes and flavors. The explosive *nitroglycerin* is an ester formed by the reaction of nitric acid with glycerin:

$$C_3H_5(OH)_3 + 3HNO_3 \rightarrow C_3H_5(NO_3)_3 + 3H_2O$$

This brief survey of the simple hydrocarbon derivatives is very far from complete but gives some idea of the important structural types and the variety of properties which these structures represent.

Foods

The materials we have been discussing are the ordinary compounds of the organic chemistry laboratory, substances whose structures and reactions are as familiar to the organic chemist as arithmetic is to a mathematician. In turning to the more complex carbon compounds which make up foods, we enter a branch of the science about which far less is known. Here we must deal with gigantic molecules whose structures are known only in part, whose complicated reactions can be followed only in a general way. This is on the frontiers of organic chemistry, a field in which active research is under way in many laboratories.

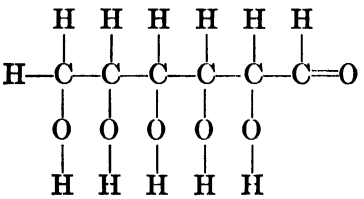
The German chemist Liebig, toward the middle of the last century, classified the chief organic constituents of food as ***carbohydrates, fats,*** and ***proteins.*** To these we now add another group, discovered in the present century, the ***vitamins.*** Compounds of these four classes, together with small amounts of ***mineral salts,*** make up the food of man.

Foods of all kinds are *digested* in the stomach and intestines. Digestion consists essentially of the breaking down of complex molecules into simpler ones, small enough to pass through the walls of the intestine into the blood, which distributes them to various parts of the body. This breakdown of food molecules takes place in a series of chemical reactions, many of them aided by organic catalysts called *enzymes.* Enzymes are highly complex organic compounds secreted by the pancreas and by various parts of the digestive tract. Their structures and the details of their reactions are unknown, but small amounts, merely by their presence, bring about chemical changes which otherwise would be very slow. Each enzyme aids one specific reaction; for instance, the enzyme ptyalin in saliva aids the conversion of starch into sugar, and the enzyme pepsin in the gastric juice breaks down complex proteins

into simple proteins. Enzymes are produced not only in the digestive systems of animals, but by plants as well; enzymes associated with yeast plants, for instance, bring about alcoholic fermentation.

Carbohydrates. These are compounds of carbon, hydrogen, and oxygen, whose molecules contain two atoms of hydrogen for every one of oxygen. They are manufactured in the leaves of green plants from CO_2 and water by the process called *photosynthesis*, energy for the reaction being absorbed from sunlight by the catalyst chlorophyll (page 372). Only the very simplest carbohydrates have been prepared artificially, and these only at great expense. Carbohydrates are of three principal kinds: sugars, starches, and cellulose.

1. *Sugars.* These are carbohydrates of fairly simple structure, characterized by their sweet taste. The important sugars belong to two groups of isomers, one group having the formula $C_6H_{12}O_6$, the other, the formula $C_{12}H_{22}O_{11}$. As an example of a structural formula for a sugar, the molecule of *glucose* (also called *dextrose*) is represented by



All sugars have similar chains of carbon atoms in their molecules, with several OH groups attached. They are, then, complex alcohols somewhat similar to glycerin and show many of the characteristic alcohol reactions.

Ordinary sugar *sucrose* ($C_{12}H_{22}O_{11}$) cannot be absorbed directly into the blood stream. During digestion an enzyme breaks it down into the simpler sugars glucose and fructose, which can be absorbed.

$$C_{12}H_{22}O_{11} + H_2O \rightarrow C_6H_{12}O_6 + C_6H_{12}O_6$$

This same reaction may be brought about in the laboratory by heating a sugar solution with dilute acid, the H_3O^+ acting as a catalyst. The simple sugars absorbed into the blood may be oxidized to supply the body with energy or, if present in excess, may be converted to glycogen (a form of starch) or fat and stored for future use.

Enzymes in yeast catalyze not only the decomposition of sucrose into glucose and fructose, but the further disintegration of the simple sugars into alcohol.

$$C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2$$

This is the commercial method for the preparation of ethyl alcohol.

2. *Starches.* Probably by linking together simple sugar molecules, plants build up the more complex carbohydrates called starches. These are compounds of high molecular weight with unknown formulas; an empirical formula is often written $(C_6H_{10}O_5)_x$, the subscripts showing the relative numbers of atoms and the x indicating the unknown number of units in a molecule. Starch is stored by plants in tubers, seeds, and fruits, and a similar substance called *glycogen* is stored by animals in the liver. During digestion, starch is broken down by enzymes into simple sugars which can pass through the intestinal wall.

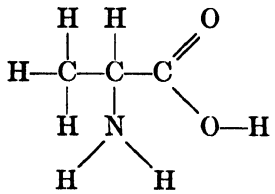
3. *Cellulose.* This most complicated of the carbohydrates has the same empirical formula as starch, but the x is a larger number. Cellulose is the chief constituent of the woody parts of all plants; cotton and linen are practically pure cellulose. Man secretes no enzymes which can digest cellulose, but some animals, like cows and horses, can use the simpler celluloses of grass. Some insects apparently can digest even the tough cellulose of wood.

That cellulose, like sugar, has OH groups in its molecule is suggested by its ability to form esters with various acids. *Nitrocellulose* or guncotton, for instance, is a cellulose ester of nitric acid. The plastic *pyroxylin* is a similar ester with fewer nitrate groups per molecule.

Fats. These are esters of glycerin with long-chain organic acids; for example, a typical fat is the ester of stearic acid $[C_3H_5(OOCC_{17}H_{35})_3]$. Like carbohydrates, fats contain only the elements C, H, and O, but the atoms are combined in different ratios. Liquid fats, or oils, are esters of unsaturated acids, solid fats esters of saturated acids. In the digestion of a fat, an enzyme catalyzes the addition of water to the fat molecule, so that it breaks down into glycerin and acid. These are absorbed into the blood and may be oxidized directly to produce energy, as are the simple sugars. If not used immediately, the fragments of a fat may be recombined and stored for later use.

Proteins. The proteins are compounds of carbon, hydrogen, oxygen, and nitrogen. Some contain sulfur and phosphorus as well, and a few contain still other elements. Their molecules are enormous, the molecular weights of the simpler proteins being at least 15,000. Their structures and formulas are unknown. Proteins are the chief constituents of living cells; other familiar materials made up largely of proteins are fingernails, hair, egg albumen, hemoglobin, the casein of milk. During digestion, proteins are disintegrated by enzymes into simpler compounds called *amino acids,* which can be absorbed into the blood and used by the body to build up its own proteins. Any present in excess are converted into carbohydrates and urea $[CO(NH_2)_2]$, the carbohydrates being stored in the body and the urea excreted by the kidneys.

Amino acids are compounds containing both COOH groups and NH_2 groups in their molecules. A simple example is alanine



Probably amino acids are the units of which protein molecules are constructed.

Vitamins. In the early years of this century experiments on animal nutrition showed that a diet of pure carbohydrate, fat, and protein, together with the essential mineral salts, would not sustain life. Some additional substances, present in a normal diet in minute amounts, were apparently necessary. Different members of this group of substances, called vitamins, were identified by the "nutritional" diseases caused by their absence—scurvy, beriberi, rickets, pellagra. At present over a dozen different vitamins are recognized; many have been isolated in pure form, and a few have been prepared artificially. They have nothing in common, except that they are all complex organic compounds and that they are all necessary in small amounts to maintain normal life.

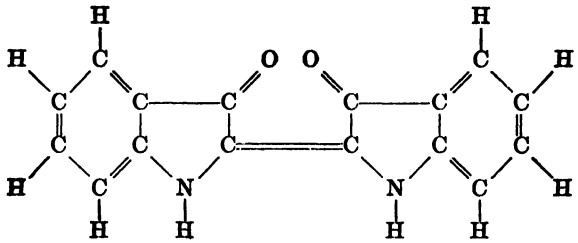
Industrial Organic Chemistry

The organic chemist can claim only modest success in deciphering the formulas of the compounds in foods and almost no success in preparing them artificially, but he has been spectacularly successful in the preparation of other substances essential to human welfare. The long list of modern industries based on his work includes the manufacture of dyes, drugs, perfumes, plastics, explosives, lacquers, artificial textiles, and other materials.

Many an industry grew from the successful attempt to duplicate a complex compound, like a dye or a perfume, which is produced in nature. An attempt of this sort involves two problems: (1) an *analysis* of the natural compound, to determine its structural formula; and (2) the *synthesis*, or building up, of the compound from simple and inexpensive raw materials. To solve the first problem, a chemist investigates the properties and reactions of the compound, to see what information they can give about the structural units in the molecule, and then tries to fit the various units into a consistent structural formula. The second problem, usually the harder one, requires a long study of conditions under which simple compounds containing the desired structural units

will react to put these units together in proper fashion. The organic chemist is an architect on a minute scale, designing complex molecular structures from building blocks contained in other molecules.

A classical example is the synthesis of indigo, a valuable blue dye produced for centuries from a plant grown in India. The formidable structure of indigo was worked out by a German, von Baeyer, in 1882.



A few years later the molecule was prepared in the laboratory, and after another decade of experiment a way was found to make the synthesis on a commercial basis. So successful is the process that the once-flourishing industry based on preparing indigo from plants has been completely wiped out. Many other tales could be told of one industry succeeding another as the result of some new feat of molecule building by an organic chemist.

Often the chemical architect not only can reproduce a naturally occurring molecule, but can even improve on nature's handiwork. For example, when the structure of indigo became known, other dyes could be produced by constructing molecules of similar pattern with slightly different structural groups in odd corners. Again, the drug cocaine is a valuable anesthetic, but dangerous because it is habit forming; by working out the structural formula of cocaine and then building similar molecules, chemists have prepared artificial compounds like novocaine and procaine which have the same anesthetic properties as cocaine but are not habit forming.

Some complex natural products defy the best efforts of organic chemists to reproduce them artificially. Rubber, for instance, has been studied intensively for years, but the work of the rubber plant still cannot be duplicated in the laboratory. By taking a hint from the structure of natural rubber, however, organic chemists have been able to manufacture several "synthetic rubbers"—rubberlike compounds which have many of the properties of real rubber and which for some purposes are better than the natural product. When rubber is analyzed, it proves to be a hydrocarbon, containing carbon and hydrogen in the ratio of 5C atoms to 8H atoms. The simplest compound with this atomic ratio is an unsaturated liquid hydrocarbon called isoprene (C_5H_8); molecules of

rubber contain isoprene units linked together in long chains. Most attempts to synthesize rubber have started with isoprene (easily prepared from hydrocarbons in petroleum), but so far have failed to make the small isoprene molecules join together, or *polymerize*, into rubber molecules. The most successful of the synthetic rubbers, "buna-S," is built not from isoprene but from a closely related hydrocarbon called butadiene (C_4H_6). Butadiene, mixed with a small proportion of a benzene derivative called styrene and heated, polymerizes into long chain molecules with structures and properties similar to those of natural rubber. During the Second World War, the production of buna-S became a matter of life-and-death importance to this country, when Japan blocked shipments of natural rubber from the East Indies. So successfully were the problems of large-scale production solved that synthetic rubber will henceforth be a formidable competitor of natural rubber even in peacetime.

Still other synthetic organic materials have been produced without any attempt to duplicate or improve on nature's models. Most of the *plastics*, for instance, have little resemblance to natural substances. These are materials which can be molded when hot, materials which are being manufactured in ever-increasing quantity and variety for brush handles, doorknobs, fountain pens, insulators, instrument boards, costume jewelry, and which may even prove useful as structural materials for houses, automobiles, and airplanes. Most plastics, like synthetic rubber, are made by polymerization—heating substances of low molecular weight, usually with a catalyst, until the small molecules join together in long, complex chains.

The raw materials of organic chemical industry are very numerous, but three are particularly important: *coal tar*, a black, sticky, unpromising liquid obtained when coal is heated to make coke; *cellulose*, obtained chiefly from wood and cotton; and *petroleum*.

Distillation of coal tar gives a variety of organic compounds, most of them derivatives of the ring hydrocarbon benzene; among others are benzene itself, toluene, xylene, naphthalene, anthracene, phenol. These relatively simple compounds are the building blocks for an incredible array of synthetic dyes, drugs, perfumes, and other products. Indigo, alizarin, turkey red, mercurochrome are a few of the dyes. Oil of wintergreen, oil of bitter almonds, synthetic vanilla, the compounds responsible for the fragrance of roses, violets, heliotrope, and narcissus are among the flavors and perfumes. Representative drugs are cocaine, novocaine, aspirin, epinephrine, and sulfanilamide. The explosive trinitrotoluene, TNT, is prepared by the action of nitric and sulfuric acids on toluene. The earliest plastic, bakelite, was made by polymerization of a mixture of phenol and formaldehyde.

Petroleum and natural gas furnish simple hydrocarbons—methane, ethane, ethylene, propane—which can be used as starting points in building up complex molecules. Some kinds of petroleum contain also ring hydrocarbons like benzene, and these ring hydrocarbons can be built from the simpler ones if not initially present. Thus petroleum and coal tar can both be used as sources of many kinds of hydrocarbons; which one is used in a given industry depends on availability, cost, and ease of handling. The two ingredients of synthetic rubber, for instance, butadiene and styrene, can each be produced from either coal tar or petroleum (or from other raw materials); at present it is most convenient to use petroleum for butadiene and coal tar for styrene.

Cellulose treated with nitric acid gives cellulose nitrate, or pyroxylin; further addition of nitric acid gives the explosive guncotton. Pyroxylin is an important plastic and is also an intermediate step in the manufacture of celluloid, photographic film, and lacquers. Cellulose and its esters may be dissolved in various solvents and squirted through small openings into other liquids, which precipitate the cellulose; in this manner are produced threads of rayon and sheets of cellophane.

Whether an organic chemist is engaged in fundamental research or in the preparation of new commercial materials, he is guided continually by structural formulas. These molecular diagrams not only summarize the properties of the compounds with which he deals, but serve as plans or blueprints to aid in the building of new structures. Without structural formulas, or some similar means of visualizing complex atomic arrangements, the chemistry of the carbon compounds would be a hopeless jumble of isolated facts.

Structural formulas depend on a few simple generalizations about the valences of different elements and the way in which carbon atoms join together. On these generalizations, applied to observational facts, are based all structural formulas from the simplest to the most complex. Proposed long before the electronic theory, these generalizations find today a clear interpretation and a new significance in terms of the tiny charged particles in atoms. Here once again we have the familiar story of progress in science: simple generalizations designed to summarize and correlate observational facts; establishment of the generalizations by their success in predicting new facts; and finally, their explanation in terms of a wider theory.

In few fields of science have simple generalizations proved more successful—more capable, that is, of correlating a wide variety of facts and of accurately predicting new ones. In few fields also has this success been more apparent to the man in the street; for one can scarcely

contemplate, even casually, the wide use of carbon compounds in our civilization without realizing the extraordinary power of the organic chemist's methods.

Questions

1. Write structural formulas for carbon tetrachloride, propane, propyl alcohol, and butyl chloride. Which of these compounds would you expect to have isomers?
2. Show by means of structural formulas the reaction between methyl alcohol and acetic acid to produce methyl acetate.
3. Compare the properties of a simple ester, like methyl chloride, with those of a salt like sodium chloride.
4. In each of the following pairs, which substance would you expect to have (*a*) the higher melting point (*b*) the lower density?

C_6H_{14}	and	$C_{11}H_{24}$
$C_6H_{13}Br$	and	$C_{11}H_{23}Br$
$C_5H_{11}COOH$	and	$C_{10}H_{21}COOH$

5. Which of the following would (*a*) dissolve in water, (*b*) turn a litmus solution red, (*c*) be gases at ordinary temperatures, (*d*) react with Na to liberate hydrogen, (*e*) react with ethyl alcohol to give esters, (*f*) react with acetic acid to give esters?

C_2H_5COOH C_3H_8 C_2H_5OH C_2H_4 HCl $C_3H_5(OH)_3$ (glycerin)

6. Name one property by which you could distinguish (*a*) C_2H_4 from CH_4, (*b*) CH_3COOH from CH_3OH, (*c*) C_3H_7OH from $C_5H_{11}OH$, (*d*) C_2H_5OH from H_2O, (*e*) CH_4 from O_2.
7. To which group of organic substances does each of the following belong: benzene, glycerin, nitroglycerin, sucrose, pyroxylin, chloroform, acetylene, glyceryl stearate?
8. Into what products is sucrose changed by (*a*) enzymes during digestion, (*b*) burning in oxygen, (*c*) yeast enzymes during fermentation?
9. Write a reaction showing the photosynthesis of carbohydrates by plants. Why is this reaction important for animal life besides providing carbohydrates for food?
10. What products of digestion are used directly to supply the body with energy? What happens to these substances if they are not used immediately for energy?
11. What organic raw material is used in the preparation of each of the following: cellophane, TNT, ethyl alcohol, gasoline, synthetic rubber, indigo, benzene?
12. Name two unsaturated hydrocarbons, two esters, two organic acids, two alcohols, two sugars, two derivatives of methane, and two dyes made from coal-tar products.

CHAPTER XXX

Silicon Compounds

NEXT to oxygen, the most abundant element in the earth's crust is silicon. Silicon never occurs free in nature, and the pure element is a rarity even in chemical laboratories, but its compounds make up some 87 per cent of the rocks and soil which compose the earth's solid outer portion. In the chemistry of naturally occurring inorganic materials, silicon has the same sort of central role which carbon plays in the chemistry of living things.

The Element

Silicon is just below carbon in the fourth group of the periodic table. Its atoms, therefore, should be somewhat larger than carbon atoms and should have similar outer shells of four electrons. Silicon should form compounds in which it has a valence of 4, and in these compounds its behavior should be less actively nonmetallic than that of carbon. These general predictions are fulfilled, but in detail the chemistry of silicon does not resemble that of carbon as closely as we might expect.

Free silicon finds some commercial use in the manufacture of hard, resistant steels. It is made by reducing the dioxide (SiO_2) at high temperatures with active metals or with carbon. Two forms of the element may be prepared: an "amorphous" brown powder probably made up of minute crystals, and the more common gray, crystalline solid. The latter has a structure like that of diamond, with each Si atom linked to four others. It is hard enough to scratch glass and has a high melting point and boiling point. If an excess of carbon is used in the reduction of SiO_2, the liberated silicon combines with carbon to form the important abrasive *carborundum* (SiC, silicon carbide); this compound has a diamondlike structure, with Si and C atoms alternating, and has a hardness only slightly less than that of diamond itself.

Like carbon, silicon is inactive at room temperatures but combines directly with active metals and nonmetals at moderately high tempera-

tures. Again like carbon, silicon enters chemical combination by sharing the four outer electrons of its atoms, and nearly always shows a valence of 4. It forms two oxides (SiO and SiO_2), analogous to CO and CO_2; a tetrachloride ($SiCl_4$), analogous to CCl_4; and compounds with metals and oxygen called ***silicates,*** some of which (*e.g.*, Na_2SiO_3) resemble in their

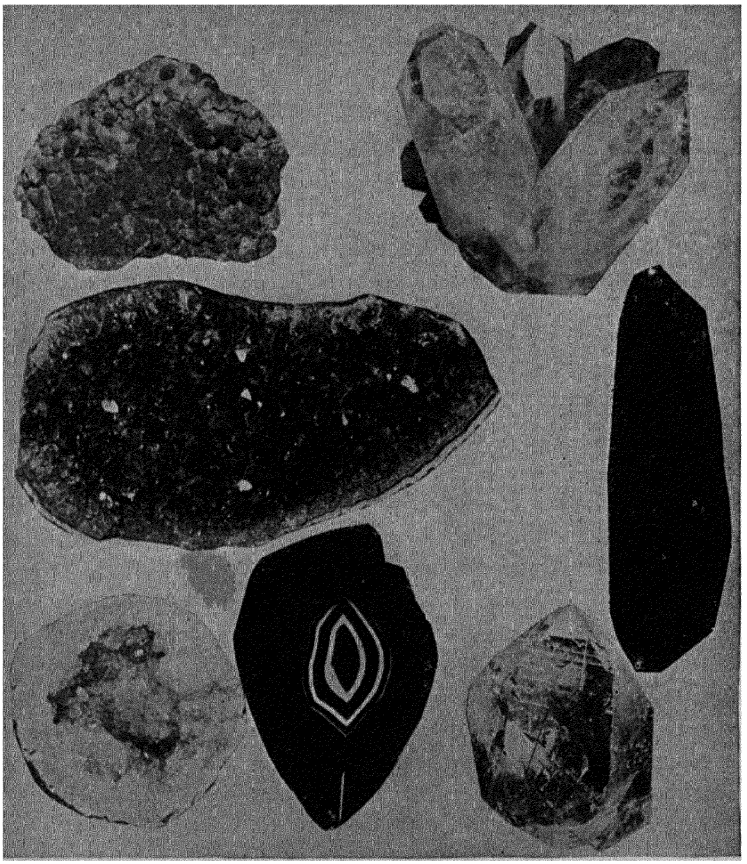


FIG. 184. *Silicon dioxide, showing various forms which occur in nature.* (*Courtesy of Ward's Natural Science Establishment.*)

formulas the carbonates (Na_2CO_3). That silicon atoms can share electrons with each other is shown by their ability to fit into the diamond structure; the bond is not as strong as the bond between carbon atoms, however, and silicon atoms show only in slight degree the ability of carbon atoms to form chains. The series SiH_4, Si_2H_6, Si_3H_8, . . . is analogous to the methane series CH_4, C_2H_6, C_3H_8, . . . , but the silicon compounds are far less stable than the hydrocarbons.

Silicon dioxide, commonly called ***silica,*** resembles CO_2 in its formula but in few other properties. CO_2 is a gas, moderately soluble in water; SiO_2 is a hard, insoluble solid with a high melting point. Silica is extremely common in nature, occurring when pure as clear crystals of *quartz* or rock crystal, and with minor impurities as amethyst, agate, onyx, flint, jasper, opal, etc. (Fig. 184). It is the chief component of most sands and sandstones. Clear quartz is transparent not only to visible light but to much of the ultraviolet, a property which makes it valuable for optical instruments. Heated above 1700°C silica becomes a viscous liquid, which on rapid cooling solidifies to a hard, amorphous solid resembling ordinary glass; since this "fused quartz" has much of the ultraviolet transparency of crystalline quartz and is far cheaper, it has found wide use in instruments for the production and study of ultraviolet radiation. Silica glass is also highly prized in the laboratory for its ability to withstand enormous temperature changes without cracking.

Silicates

Nearly all the earth's silicon is either combined with oxygen in silica or combined with oxygen and one or more metals in the various silicates. In number and variety the silicates will hardly stand comparison with the carbon compounds, but their complexities of structure are sufficient to make the chemistry of silicon an intricate and difficult subject.

Sodium silicate (Na_2SiO_3, ordinary water glass), dissolves in water to form the ions Na^+ and SiO_3^{--}. If a calcium salt is added to this solution, insoluble calcium silicate is precipitated:

$$Ca^{++} + SiO_3^{--} \rightarrow CaSiO_3$$

If a strong acid is added to a sodium silicate solution, the very weak silicic acid is formed:

$$2H_3O^+ + SiO_3^{--} \rightarrow H_4SiO_4 + H_2O$$

These reactions seem very similar to the reactions of carbonate ion: thus a sodium carbonate solution gives a white precipitate of $CaCO_3$ when a calcium salt is added, and it gives the weak acid H_2CO_3 when treated with a strong acid.

$$Ca^{++} + CO_3^{--} \rightarrow CaCO_3$$
$$2H_3O^+ + CO_3^{--} \rightarrow H_2CO_3 + 2H_2O$$

Seemingly we might proceed to study the chemistry of the silicates in terms of the atom group SiO_3, much as we have studied the chemistry of the carbonates in terms of the atom group CO_3.

It happens, however, that very few of the important silicates are appreciably soluble, and the ion SiO_3^{--} is seldom encountered. Further-

more, the silicates exhibit a peculiar complexity of structure which makes a classification of their reactions in terms of simple atom groups like SiO_3 all but impossible. Thus we find not one sodium silicate, but several:

Na_2SiO_3 $Na_4Si_3O_8$
Na_4SiO_4 $Na_2Si_2O_5$
$Na_4Si_2O_6$ $Na_2Si_3O_7$, etc.

To study these in the usual manner would mean gathering a separate set of facts about each silicate group—and then we should find that these groups often do not remain intact during reactions.

The above formulas become more intelligible if we write

Instead of Na_2SiO_3, $Na_2O.SiO_2$
Instead of Na_4SiO_4, $2Na_2O.SiO_2$
Instead of $Na_4Si_2O_6$, $2Na_2O.2SiO_2$
Instead of $Na_4Si_3O_8$, $2Na_2O.3SiO_2$, etc.

That is, the differences between the various sodium silicates may be considered simply as differences in the ratios of the two oxides Na_2O and SiO_2. Other complex silicate formulas may be broken down similarly into combinations of SiO_2 and metallic oxides: thus $MgSiO_3$ and Mg_2SiO_4, two magnesium silicates which occur widely in nature, may be written $MgO.SiO_2$ and $2MgO.SiO_2$, and the more complex silicate $CaAl_2Si_2O_8$ may be written $CaO.Al_2O_3.2SiO_2$. Some justification for this procedure besides mere convenience is the fact that many silicates may be prepared artificially by heating oxides together; the two magnesium silicates just mentioned can be made by heating mixtures of MgO and silica in the right proportions, and calcium silicate ($CaSiO_3$) can be made by heating CaO with silica. Reactions between silicates, moreover, are often best interpreted as interchanges of different oxides.

It is the singular ability of its oxide to form innumerable different compounds by combining with various amounts of metallic oxides that distinguishes silicon from other elements and makes possible the great variety and complexity of the silicates.

We might, of course, have studied other compounds in terms of their component oxides: Na_2SO_4 might be written $Na_2O.SO_3$, $CaCO_3$ might be written $CaO.CO_2$, etc. The reason that we do not use such formulas is simply that in the reactions of these compounds the atom groups which remain together, which recur in compound after compound, which appear as ions in solution, are not the oxides but the groups SO_4 and CO_3. In silicate reactions, on the other hand, the groups of elements which seem to cling together are the oxides, so we shall frequently use them rather than larger groups as chemical units.

As a class, the silicates are crystalline solids, melting at high temperatures to give viscous liquids. Their variations in composition and struc-

ture are reflected in a variety of colors, hardnesses, and crystal forms. The softness of talc, the hardness of zircon and beryl, the transparency of topaz and the deep color of garnet, the platy crystals of mica and the fibrous crystals of asbestos give some idea of the range of silicate properties.

Silicate Structures

Like so many other solids, crystalline silicates are composed not of separate structures which we might call "molecules," but of continuous crystal lattices. The simple formula of a silicate, even when written out in terms of oxides, tells us only the ratio of the numbers of different atoms present, nothing about how these atoms are put together. Actual structures of the simpler silicates have been worked out only in recent years, largely through the use of X rays. No convenient method like the structural formulas of organic chemistry has been proposed for diagramming the complex silicate structures, but the general patterns on which silicate crystals are built have been deciphered.

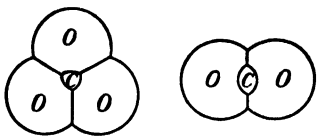


FIG. 185. *Structures of CO_3^{--} and CO_2 (diagrammatic).*

The chemical differences between silicon and carbon hinge on the difference in the sizes of their atoms. Around the tiny carbon atom only two (in CO_2) or three (in carbonates) oxygen atoms can find room; so strong is the covalent bond between C and O that the oxygen atoms are warped out of their usual approximately spherical shape to fit snugly about each carbon atom (Fig. 185). In CO_2 each oxygen atom is tightly bound to carbon by two shared electron pairs; since valences are satisfied within each CO_2 molecule, the molecules have little attraction for one another, and they fit into the crystal lattice of solid CO_2 only at low temperatures. The larger size of the silicon atom and its consequent smaller attraction for its shared electrons enable four oxygen atoms to cluster around it without warping. Between each Si and O is a single electron-pair bond, so that every oxygen atom has one valence free for attachment to another atom. In SiO_2 this valence is satisfied by another Si atom; Si and O atoms are linked in a continuous network, each Si joined to four O's and each O joined to two Si's (Fig. 186). Thus there are no separate SiO_2 molecules: the atoms are tightly bound in a continuous lattice, and SiO_2 is accordingly a solid with a high melting point.

In silicates as well as in silica, each silicon atom is linked to four oxygen atoms. The O's are symmetrically placed around Si, at the corners of a tetrahedron (Fig. 187). In the simplest silicates (*e.g.*, Mg_2SiO_4) the SiO_4 tetrahedra form separate negative ions, linked together by positive metallic ions—in a general way like the ionic lattice of NaCl, with SiO_4 groups taking the place of Cl ions. In other silicates two, three, or more

SiO_4 tetrahedra may join together into larger ions, like $Si_2O_7^{-6}$ and $Si_3O_9^{-6}$ (Fig. 187), some oxygen atoms lying between two silicon atoms.

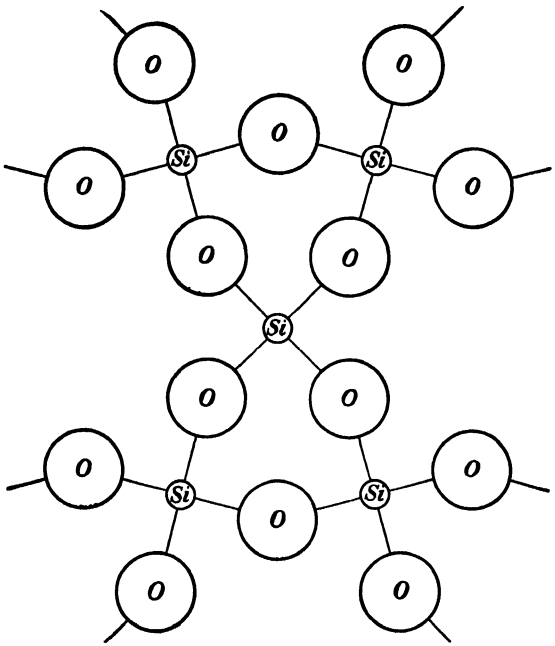


FIG. 186. *Structure of SiO_2 (quartz). Actually this is a three-dimensional structure, each SiO_4 group being a tetrahedron (Fig. 187).*

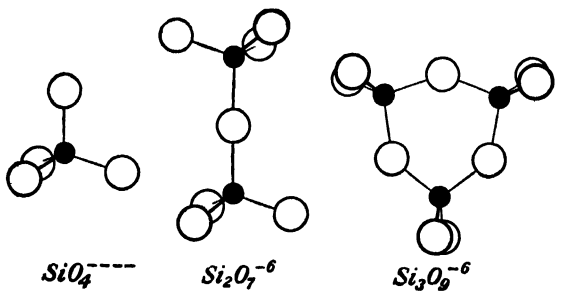


FIG. 187. *Silicon-oxygen structures in simple silicates. These structures alternate with positive metallic ions in the crystal lattices. (Black circles = Si atoms, open circles = O atoms.)*

In more complex structures the tetrahedra are linked together in continuous chains (Fig. 188) or sheets, with metal ions lying between. Still more intricate structures involve three-dimensional arrangements of tetrahedra, with some of the Si atoms replaced by Al atoms.

Thus some of the oxygen atoms in a silicate structure link silicon atoms together, while others are held jointly by silicon atoms and metal ions. Just as each silicon atom surrounds itself with four oxygen atoms, so each metal ion tries to surround itself with a certain number determined by its size and valence. The silicon-oxygen structure in any particular case adapts itself to the needs of the metal ions present, leaving enough oxygen atoms with free valences to satisfy the metal ions.

So the analysis of silicate structures with X rays suggests a new explanation for the number and variety of these compounds: silicon forms all manner of stable structures with oxygen, from simple negative ions to a continuous lattice, the structure which appears in a particular compound depending on the metal ions present. If silicon were more

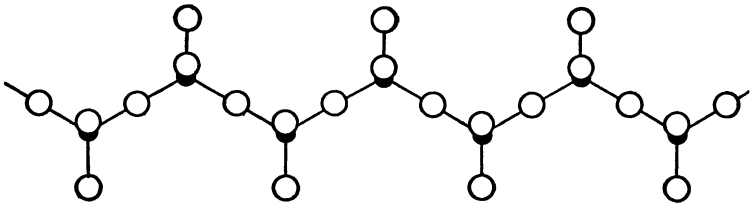


FIG. 188. *Chain of SiO_4 tetrahedra. These chains are linked sidewise by metal ions.*

strongly nonmetallic, like carbon, its atoms would be more closely united to oxygen and it would form simple ions (SiO_3^{--} or SiO_4^{----}) exclusively. If silicon were more metallic, it would form positive ions like other metals and silicates would have a simple ionic type of lattice. But silicon is neither metallic nor strongly nonmetallic, and its intermediate character makes possible a wide variety of silicate structures. This is a far more penetrating explanation, of course, than our earlier statement that silica combines with metallic oxides in many different ratios.

Artificial Rocks

For illustrations of the behavior of silica and silicates, we turn to the preparation of three familiar materials—glass, pottery, and cement. So similar are these substances in composition and general properties to the silicate rocks that for purposes of this discussion we may christen them "artificial rocks."

Glass. Molten silicates as a rule are highly viscous liquids. If cooled very slowly they solidify, like other liquids, into crystalline solids. But if the cooling is fairly rapid, some silicates, instead of crystallizing, simply grow more and more viscous until they become amorphous solids or *glasses.*

The formation of a glass depends on the slowness with which particles move about in a viscous liquid. Cooling does not allow the sluggish atom

groups sufficient time to find their places in the normal crystal lattice but freezes them in the random arrangement of the liquid state. Because of the random arrangement of its particles a glass has no definite melting point, but softens gradually to a viscous liquid when heated. Some glasses slowly crystallize or devitrify on standing at room temperature, their particles apparently moving sufficiently even in the solid to arrange themselves in lattice structures. All glasses can be made to devitrify if maintained for a long time at a temperature just below the softening point, where molecular motion is more rapid.

Glasses are usually prepared not by melting silicates but by heating together certain mixtures of silica with other oxides. Silica alone forms an excellent glass (page 421), but it is too difficult to make and too expensive for ordinary use. Since a glass in general is a complex mixture of silicates made by melting together various oxides, its composition cannot be expressed by a simple chemical formula. Its properties, such as transparency, color, hardness, tendency to devitrify, depend on the amounts and kinds of the oxides used, and it is no small part of the glassmaker's art to know just how different oxides and different proportions will affect the final product.

Ordinary glass is a mixture of complex silicates formed from the oxides Na_2O, CaO, and SiO_2. It is made by heating together clean quartz sand (SiO_2), limestone ($CaCO_3$), and sodium carbonate (Na_2CO_3). Heat drives off CO_2 from the carbonates, leaving Na_2O and CaO.

$$Na_2CO_3 \rightarrow Na_2O + CO_2$$
$$CaCO_3 \rightarrow CaO + CO_2$$

The two metallic oxides then combine with silica to form the silicates of the glass. Amounts of the three ingredients are used which will give a final ratio of the three oxides of about $Na_2O{:}CaO{:}5SiO_2$. Any considerable deviation from this ratio gives a glass which is partially opaque or which devitrifies easily. Many glass articles made carelessly several decades or several centuries ago have become cracked and opaque through slow devitrification.

Small amounts of other substances added to the glass often change markedly its color and other properties. The green glass used in making cheap bottles has a small amount of iron oxide, originally present as an impurity in the sand or limestone. A little cobalt gives the glass a blue color, a little manganese oxide a violet color. Use of potassium oxide instead of sodium oxide gives a glass which softens at a higher temperature than ordinary glass. Pyrex glass, more resistant to temperature changes than ordinary glass, contains a relatively high percentage of silica, considerable B_2O_3, a little K_2O, and only small amounts of Na_2O and CaO. Lead glass, a soft, brilliant glass extensively used in

optical instruments, is made up of the three oxides PbO, K_2O, and SiO_2.

The art of glassmaking is almost as old as written history, dating back to the predynastic period in Egypt. The early Egyptian glass was colored and opaque, suitable only for ornaments and glass vessels. Not until the time of the Roman emperor Augustus was transparent glass made in sufficient quantities to be used in houses. In Europe during and after the Renaissance the glassmaker's craft was widely and elaborately developed. The quality of ordinary glass was greatly improved, vessels and apparatus of intricate design were constructed, and new kinds of glass were discovered. Up until very recent times, however, glassmaking remained strictly a craft, improvements arising from trial and error rather than from scientific experiment. Even today scientists have only begun the difficult study of the reactions involved in glass manufacture.

Pottery and Porcelain. Older than glassmaking is the art of making pottery, practiced often with amazing skill by primitive tribes of the past and present. Pottery manufacture was carried to a high stage of perfection in Europe during and after the Renaissance. The trick of making fine white porcelain was discovered in China in the seventh or ninth century A.D.; when importations of Chinese porcelain reached Europe about the time of Columbus, European pottery makers sought in vain to duplicate it. The secret was finally discovered more than two centuries later in Meissen, Germany, whence the manufacture of porcelain spread to all parts of Europe and America. As with glass manufacture, improvements in pottery and porcelain until recent years have been the result of accident or of trial-and-error experiment by expert craftsmen rather than the result of scientific knowledge.

Pottery and porcelain, including earthenware, stoneware, chinaware, terra cotta, etc., together with bricks, tiles, and firebricks, are called *ceramic products.* In the manufacture of all of them the chief material used is clay. Ordinary clay is a variable mixture of many substances of which the principal one is the hydrous aluminum silicate *kaolin* (from the Chinese term *kao-ling*) ($H_4Al_2Si_2O_9$ or $2H_2O.Al_2O_3.2SiO_2$). Other constituents commonly present are oxides of iron, carbonates of magnesium and calcium, and decayed vegetable matter. Kaolin has the property of absorbing large amounts of water between its particles, producing a plastic mass which can be molded into any desired form. On heating or "firing," the clay loses both the water it has absorbed and water from its own molecules, giving a hard, rocklike product. The characteristics of the product depend on (1) the purity of the clay, (2) the temperature of firing, and (3) the kind of glaze.

Ordinary *bricks* are made from clays containing large amounts of iron oxide and sand. For *pottery* somewhat purer clays are used and the firing temperatures are higher. The hardening of bricks and pottery in

the firing process is due at least partly to the formation of silicates which melt and bind together the clay particles. The red and brown colors of the products are due to varying amounts of iron oxide. *Porcelain* is made from a mixture of pure white kaolin with finely ground silica and feldspar (a silicate of potassium and aluminum) fired at high temperatures (1300 to 1500°). The mass partially melts and gives a dense, somewhat translucent product.

Pottery and porcelain are usually *glazed* during the firing process. For the cheaper varieties of earthenware this is often accomplished by simply throwing common salt into the kiln; at the high temperature the salt is vaporized and a thin film of sodium silicate glass is formed on the surface of the articles. Tableware is usually given a lead glaze by dipping the articles into a paste of lead oxide, clay, and silica; the resulting glaze is a surface layer of lead silicate glass. Other types of glazes are produced by the use of various metallic compounds which form glasses with silica either added or present in the material being fired.

Cement. Cements are substances which form pastes with water and harden or set on standing. The commonest of these, the one to which the term cement usually refers, is *Portland cement*, a material whose use in structural work has been widespread since about 1900.

Portland cement is made by heating a mixture of limestone ($CaCO_3$) and clay ($H_4Al_2Si_2O_9$) to a temperature of 1400 to 1600°, where the mixture partially melts. When cooled, the cement is ground to a fine powder which slowly hardens when mixed with water. Ordinarily cement is mixed with sand or gravel to form *concrete.*

The chemical reactions involved in the formation and setting of Portland cement are not well understood. The three principal products formed by heating clay and limestone are calcium aluminate ($Ca_3Al_2O_6$ or $3CaO.Al_2O_3$) and two calcium silicates (Ca_2SiO_4 or $2CaO.SiO_2$ and Ca_3SiO_5 or $3CaO.SiO_2$). The formation of these substances becomes understandable if the formulas for limestone and clay are written $CaO.CO_2$ and $2H_2O.Al_2O_3.2SiO_2$, respectively. Heat drives off CO_2 and H_2O, and the other oxides rearrange themselves into the three compounds just mentioned. When water is added to cement, hydrated silicates are formed from these compounds, together with gelatinous substances like aluminum hydroxide [$Al(OH)_3$ or $Al_2O_3.3H_2O$] and silicic acid (H_4SiO_4 or $SiO_2.2H_2O$). Setting of the cement apparently involves the slow hardening of these gels, which bind together the other materials.

This brief discussion of glass, ceramics, and cement shows clearly the advantage of dissecting silicate formulas into their constituent oxides. It tells us further something about the nature of silicate reactions: for the most part they are slow, high-temperature processes involving

rearrangements of chemical partners in viscous liquids. We shall find later that reactions of this sort are important in the formation of silicate rocks.

Because silicates commonly are formed only at high temperatures, because they cannot easily be dissolved or separated from one another, the chemistry of silicate reactions is difficult to study and remains imperfectly understood. Here we have no great body of experimental facts organized by a few simple principles, as we found in the chemistry of electrolytes and in the chemistry of carbon compounds. The electronic theory and X-ray analysis of crystal structure have suggested general reasons for the peculiarities of silicon and the silicates, but have so far made little headway in predicting the course of specific reactions.

Questions

1. Write the formulas of the following silicates in terms of oxides:

   Zn_2SiO_4 Al_2SiO_5 $H_4Mg_3Si_2O_9$
   $CaMg_3Si_4O_{12}$ $Ca_3Al_2Si_3O_{12}$ $KAlSi_3O_8$

2. Below silicon in the fourth group of the periodic table is the rare element germanium. Would you expect germanium to be more metallic or less metallic than silicon? Would you expect it to form compounds with hydrogen analogous to the hydrocarbons? Write the formula you would expect for (*a*) the most common oxide of germanium, (*b*) the most common chloride, (*c*) sodium germanate.
3. Contrast the chemical behavior of silicon with that of carbon. To what differences in their atomic structure is the contrast due?
4. In terms of molecular structure, why is SiO_2 a solid while CO_2 is a gas?
5. Water and CO_2 from the air slowly convert many silicates in rocks into clay. A typical reaction is

$$2KAlSi_3O_8 + 2H_2O + CO_2 \rightarrow H_4Al_2Si_2O_9 + K_2CO_3 + 4SiO_2$$

   Rewrite this equation by expressing all formulas as combinations of oxides.
6. Which of the following substances are used in the manufacture of (*a*) ordinary glass, (*b*) Portland cement, (*c*) porcelain, (*d*) bricks?

   kaolin, sodium carbonate, feldspar, calcium carbonate, silica.

7. What is the difference in composition between glass and blast-furnace slag (page 398?)

CHAPTER XXXI

The Colloidal State

CONSIDER two simple experiments: (1) A little dilute ferrous sulfate solution ($FeSO_4$) is added to a dilute solution of gold chloride ($AuCl_3$). Gold is one of the most inactive metals, so should be easily reduced from its compounds, and Fe^{++} is a good reducing agent. By all rules, then, mixing the two solutions should give at once a precipitate of metallic gold. But no precipitate appears; instead the mixture turns a clear blue color. Somehow or other the liberated gold has remained in solution. (2) A bit of gelatin is dissolved in hot water and allowed to cool. For a time the solution remains fluid, but presently it "sets" to a clear, transparent jelly.

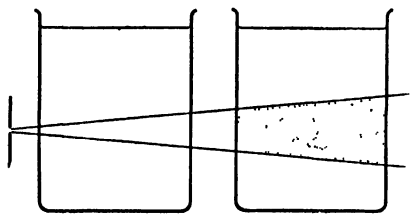


FIG. 189. *Passage of a beam of light through a true solution (left) and through a colloidal solution (right).*

Nothing in our previous study suggests an explanation for these peculiar solutions, one which contains an insoluble metal, the other which becomes a solid on standing. Ordinary tests show only that they seem to be normal solutions: they are transparent and to all appearances perfectly homogeneous; nothing settles out on standing; analysis of small portions and examination with a powerful microscope give further evidence that the liquids are homogeneous.

Tests of another sort, however, reveal differences between these liquids and ordinary solutions. Suppose that beakers of the two liquids and a third beaker of sodium chloride solution are illuminated with a strong beam of light (Fig. 189). Through the first two beakers the path of the light is outlined by a soft glow, but through the sodium chloride solution it is invisible. Something in the gold solution and the gelatin solution *scatters* light to the side.

A liquid can scatter light in this manner only if it contains tiny suspended particles, particles whose diameters are comparable to the wave

length of the light. We have encountered similar scattering of X rays by the far tinier particles in crystal structures, and we are all familiar with the scattering of ordinary light by tiny dust particles which makes a light beam visible in a darkened room. A still more convincing demonstration that suspended particles are present in the gold solution is possible by examining it under a microscope while strongly illuminated *from the side* (instead of from underneath or above, as is customary in microscopic work). Observed in this manner the liquid appears to contain myriads of bright points of light, moving erratically on a black background. Each point of light marks the position of one particle; the particles themselves are not visible, since they are smaller than the wave length of light, but their positions are shown by the radiation which they scatter. The motion of the particles is the Brownian movement (page 129), produced by their continual buffeting by water molecules.

So the gold and gelatin solutions are actually suspensions of tiny solid fragments, larger than ordinary molecules but smaller than particles whose outlines can be seen under a microscope. Liquids of this sort are called ***colloidal solutions*** or ***colloidal suspensions,*** and the study of their peculiar properties is called *colloid chemistry.* Colloidal solutions are particularly important in the chemistry of biological processes, for much of the material in living tissue consists of such solutions.

Like the chemistry of silicates, colloid chemistry is a relatively new and unexplored branch of physical science. We can expect here no single theory to unify and correlate colloidal phenomena, because the basic phenomena themselves are only now being studied. We shall find, as we should find in the early stages of development in any science, an array of diverse observational facts with a few simple generalizations going only a little way beyond the facts.

Kinds of Colloidal Systems

Colloidal solutions are intermediate between suspensions and true solutions. Fine sand shaken with water and allowed to stand settles out almost at once. Mud particles remain suspended much longer, but they are easily visible with a microscope and eventually settle out. Particles in colloidal solutions are submicroscopic in size and remain suspended indefinitely. Still smaller are the particles of true solutions. Evidently there is a continuous gradation from one extreme to the other: dividing lines between solutions, colloidal solutions, and suspensions must be arbitrarily chosen. Usually a suspension is considered colloidal if its particles have an average diameter between 0.000001 mm (about the size of the largest ordinary molecules) and 0.0005 mm (the smallest particles which can be seen in a microscope).

Any substance which can be obtained in the form of particles within this size range is called a *colloid.* Research in recent years, however, has shown that almost any substance can exist in this form, so that the term "colloid" is practically useless. We can speak more profitably of *colloidal solutions* or *colloidal systems,* referring to suspensions of colloidal particles of one substance in another. The peculiar properties of a colloidal system are determined to a larger extent by the *size* of the colloidal particles than by their chemical composition.

Colloidal particles are commonly said to be *dispersed* in another substance. In the colloidal gold solution of a previous paragraph the gold is dispersed in water. The water is the *dispersing medium* and gold is the *dispersed substance.*

The dispersing medium of a colloidal system is most commonly a liquid, but it may be a gas or solid. The dispersed substance likewise may be either gas, liquid, or solid. Dust and smoke suspended in air are examples of solid colloidal particles suspended in a gas; fog is a colloidal suspension of liquid droplets in a gas. In whipped cream gas bubbles are dispersed in a liquid; in pumice and gray hair gas bubbles are dispersed in solids. A colloidal system formed by one liquid dispersed in another is called an *emulsion;* familiar emulsions are mayonnaise (olive oil in vinegar) and whole milk (butterfat in water). Ruby glass is a suspension of solid gold particles in solid glass.

We shall be concerned chiefly with the commonest type of colloidal systems, dispersions of solids in liquids. Soap solutions, glue, many paints and lubricants, many of the materials in living cells are colloidal systems of this type.

Properties of Colloidal Systems

We can best discuss the idiosyncrasies of colloidal systems by contrasting them with true solutions on the one hand and suspensions on the other. Unless otherwise indicated, the following remarks apply only to systems in which solids are dispersed in liquids.

From true solutions, colloidal systems differ in:

1. *Their ability to scatter light.* The path of light through a colloidal solution shows itself by a hazy glow, as we have mentioned before. This property, one of the most convenient for distinguishing colloidal solutions from true solutions, is often called the *Tyndall effect* after an English physicist who studied the phenomenon in detail.

2. *Their rate of diffusion.* If pure water is poured carefully on top of a colloidal gold solution in one beaker and a sodium chloride solution in another, gold particles in one case and ions of sodium and chlorine in the other will slowly diffuse upward into the pure water. But the big, slow-moving colloidal particles diffuse many times more slowly.

3. *Their ability to pass through membranes.* Many kinds of membranes, like parchment paper, egg skin, bladder, collodion films, have pores which are large enough for ions and ordinary molecules to pass through, but which are too small for colloidal particles. This inability of colloidal particles to diffuse through membranes furnishes an easy way to remove electrolytes from colloidal systems (Fig. 190).

4. *The changes in certain properties of the dispersing liquid produced by the dispersed substance.* These properties include vapor pressure, boiling point, freezing point, viscosity, surface tension, and others. Ordinary solutes affect such properties in a regular, predictable fashion. We have not discussed these changes in properties for solutions, but their predictability is suggested by the fact that deviations from the usual rules in the case of electrolytes led Arrhenius to the ionic theory (page 343). Some colloidal particles do not alter the properties of their dispersing media perceptibly, while others produce striking and largely unpredictable changes.

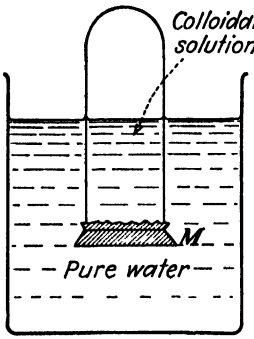


FIG. 190. *Separation of colloidal particles from electrolytes. A colloidal solution is placed on one side of the membrane M, pure water on the other side. Ions diffuse through the membrane into the water, but colloidal particles remain behind.*

5. *Their ability to form gels.* Many colloidal solutions, like a gelatin solution, have the ability to form solids of low rigidity called jellies or *gels.* True solutions never exhibit this property.

From suspensions, colloidal systems differ in:

1. *The failure of the dispersed substance to settle.* Some colloidal solutions are unstable: their particles adhere to one another during collisions, eventually growing into large aggregates which settle out. When $AgNO_3$ solution is added to NaCl solution, for instance, the precipated AgCl (page 343) may first appear only as a faint opalescence in the solution. On standing, the solution grows less and less transparent as the particles increase in size, and presently white curds of the precipitate settle to the bottom. But most colloidal solutions are stable, showing no tendency to separate into their components as do ordinary suspensions.

2. *Their ability to pass through filter paper.* The solid particles of an ordinary suspension may be separated from the liquid by pouring the mixture through a filter paper, since the pores of the paper are small enough to prevent the particles from passing through. Colloidal particles, tinier than the pores of the paper, pass through with the liquid. This property of colloidal solutions is often a nuisance in chemical analysis,

since a precipitate which appears in colloidal form cannot easily be separated from the liquid.

3. *The electric charges of the dispersed particles.* Because colloidal particles have a large surface area in relation to their weight, they readily adsorb on their surfaces considerable numbers of molecules and ions. Colloidal particles of a given substance usually adsorb preferentially ions of only one sign, so that they acquire a positive or negative charge. The existence of a charge on colloidal particles may be demonstrated experimentally by placing positive and negative electrodes in a colloidal solution: the dispersed substance migrates to one electrode or the other, just as ions do in electrolysis. Metals and metallic sulfides dispersed in water have negative charges, while the particles of metallic hydroxides and many dyes have positive charges. Since all the particles of a given kind have charges of the same sign, they repel each other and cannot coalesce unless discharged. This circumstance, together with the molecular bombardment which produces the Brownian movement, keeps colloidal particles from settling out of the dispersing liquid.

Such a review of the properties of colloidal solutions shows clearly that we cannot consider these systems as simple heterogeneous mixtures of solids and liquids, nor yet as homogeneous solutions. Matter in this peculiar state of fine division acquires certain characteristics which it does not possess either in large aggregates or when completely broken down into its ultimate particles.

Preparation and Precipitation of Colloidal Particles

Some materials go into solution of their own accord as colloidal particles. These include glue, soap, starch, and proteins like gelatin and albumen. For the most part they are complex organic compounds with structures containing carbon chains of unknown length. For such substances it seems reasonable to assume that the particles which go into solution as colloidal materials are the actual *molecules* of the compounds. At least there are no experiments to show that the particles ever break down into smaller fragments without undergoing chemical reaction, so that the colloidal particles satisfy the most elementary definition of molecules. On this assumption, the molecular weights of these complex compounds may be found from the actual weights of the colloidal particles, which can be measured with fair accuracy by several methods. The molecular weights so found seem fantastically large: those for the proteins, for example, range from about 15,000 to well over 1,000,000, corresponding to carbon chains hundreds and thousands of atoms long.

Other materials, whose colloidal particles are fragments containing many molecules apiece, are obtained in colloidal form by special methods.

One direct and obvious method is to grind a substance into a fine powder under a liquid. Ordinary grinding equipment does not produce a fine enough powder, but colloidal particles can be obtained by forcing a suspension to flow between two closely spaced metal disks rotating rapidly in opposite directions. Many colloidal solutions can be prepared, like the colloidal gold solution of an earlier paragraph, by chemical reactions in which the dispersed substances are among the products. Sulfur, $Fe(OH)_3$, and As_2S_3 are substances which often appear in colloidal form when precipitated by mixing two solutions. Other precipitates may be obtained as colloidal particles by careful regulation of the temperature and concentration of ions during reaction.

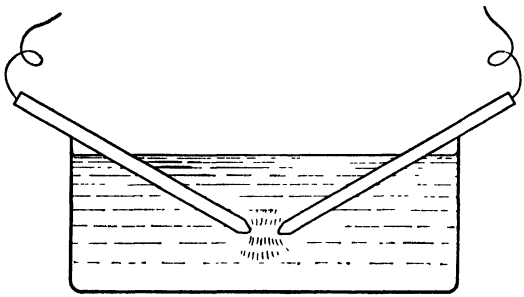


FIG. 191. *Arc process for preparing colloidal solutions of metals in water.*

A convenient method for preparing colloidal solutions of inactive metals (gold, platinum, silver, and many others) is to pass an electric arc between pieces of the metal under water (Fig. 191). Two pieces of the metal are connected to an electric circuit, and their ends are allowed to touch under water; then they are separated slightly and an arc is formed. Colloidal particles are formed both by the ripping loose of fragments of the metal by the discharge and by condensation of metal vapor produced at the high temperature of the arc.

Sometimes the preparation of colloidal particles is aided by the presence of another colloidal substance in solution. For example, if a little gelatin is present in an $AgNO_3$ solution to which NaCl is added, the liberated AgCl remains in colloidal form instead of precipitating. The gelatin is called a *protective colloid*, since it "protects" the AgCl particles from joining together into larger fragments. The silver bromide particles on photographic plates are protected in this way by gelatin. In the manufacture of ice cream, gelatin is often added as a protective colloid to prevent the formation of ice and sugar crystals. The action of protective colloids has not been fully explained, but it is most probably due to the formation of films of one colloid around particles of the other.

An *emulsion*, a colloidal system in which droplets of one liquid are dispersed in another, may be formed simply by shaking two immiscible

liquids together. When kerosene is shaken vigorously with water, the resulting milky fluid contains millions of kerosene droplets suspended in water. This emulsion is not stable, however; as it stands, the kerosene droplets coalesce into larger and larger drops until finally the two liquids separate into distinct layers. To make the emulsion stable, a small amount of a third substance, called an *emulsifying agent*, must be added. Soap, gelatin, various gums, and albumen are good emulsifying agents. The cleansing action of soap is due in large part to its ability to emulsify grease in water. The protein called casein acts as an emulsifying agent for butterfat in milk. In mayonnaise olive oil is dispersed in vinegar with egg yolk acting as the emulsifying agent. The ability of emulsifying agents to stabilize emulsions has been variously explained by the lowering of surface tension between the liquids, by the formation of hydrates, by the formation of protective films around the dispersed droplets; but no entirely satisfactory explanation has been given.

The *precipitation* or *coagulation* of colloidal particles requires that they join together into larger fragments. Sometimes coagulation takes place spontaneously, the particles apparently adhering in collisions. But the particles of a stable colloidal system repel each other because of their electric charge and will not coagulate unless this charge is neutralized. One way of neutralizing the charge is to place electrodes in the solution: the dispersed substance migrates to one or the other, depending on the sign of its charge, and appears as a precipitate or as a coating on the electrode. Another way is to add a suitable electrolyte, the colloidal charges being neutralized by ions of opposite sign. Some ions are much more effective in producing coagulation than others: H_3O^+ and OH^- are particularly effective, so that negatively charged colloidal particles are readily precipitated by addition of acids, positively charged particles by addition of alkalies. In general, colloidal particles are stable only in solutions containing low concentrations of ions. An important natural instance of the coagulation of colloids by electrolytes is the deposition of abundant fine sediments at river mouths: colloidal mud particles, stable in fresh river water, are precipitated on contact with the ions of sea water.

Gels

One of the strangest properties of many colloidal solutions is their ability to form *gels* on standing. Gels are formed most readily by those organic substances which dissolve as colloidal particles directly—gelatin, albumen, soaps, glue, etc. Often surprisingly dilute solutions of these materials form gels; a 2 per cent solution of gelatin in hot water, for example, forms a stiff gel on cooling. Usually gels of this sort may be converted into liquids again by heating.

Some inorganic colloids also form gels, a conspicuous example being silicic acid. This weak, slightly soluble acid does not precipitate when HCl is added to Na_2SiO_3, but remains dispersed in colloidal form. If the clear liquid is allowed to stand, it presently solidifies into a stiff, transparent gel, the time required depending on the temperature and the concentration of the solutions.

On standing, gels often shrink somewhat and give up part of their water. In a dry atmosphere the shrinking is much greater, and most of the water is lost. When placed in water, many dried gels take up some of the liquid and swell to their former volume. The swelling and bursting of seeds is due to absorption of water by protein gels; the swelling of bodily tissues after injury is a similar expansion of protein material. Muscular activity is in part explained by the swelling and contracting of protein gels. Some gels absorb other liquids besides water: rubber, for instance, swells prodigiously when placed in benzene.

In recent years studies by means of X rays have given considerable information regarding the structure of gels. Probably particles of the dispersed substance in a gel have coalesced to form long filaments or flat plates, lying in loose array like piles of brushwood, with water or some other liquid held in the open spaces. The filaments and plates of gel structures are colloidal materials only in the sense that their thicknesses lie within the range 0.000001 to 0.0005 mm; their lengths may be far greater.

A considerable number of familiar materials are gels: the protein material of the body, fruit jellies and gelatin desserts, soap, rubber, resins, bakelite, etc. Agate, flint, and onyx are forms of dehydrated silicic acid gels. The swelling of clay on addition of water is due to the presence of gel-like materials consisting of tiny flat plates, which take up the liquid between them.

Questions

1. Would you consider a colloidal solution heterogeneous or homogeneous? What is the difficulty with these two terms?
2. Can a colloidal system be formed with one gas dispersed in another?
3. Suggest two possible methods for demonstrating that a colloidal solution of platinum is not a true solution, and two methods for showing that it is not an ordinary suspension.
4. How could you show the presence of colloidal dust particles in ordinary air?
5. Why couldn't a colloidal solution of sodium in water be prepared by the arc process?
6. Suggest a method for preparing a colloidal system containing a gas dispersed in a liquid.
7. Colloidal particles of As_2S_3 have a negative charge; those of $Fe(OH)_3$, a positive charge. What would you expect to happen if colloidal solutions of these two substances were mixed?

8. How would you prepare typical examples of each of the following: (*a*) an emulsion, (*b*) a colloidal solution in which the dispersed particles have a negative charge, (*c*) a gel?

Suggestions for Further Reading—Part IV

General:

SNEED, M. C., and J. L. MAYNARD: *General College Chemistry*, D. Van Nostrand Company, Inc., New York, 1944. A standard text, with an especially good discussion of the chemistry of electrolytes.

DEMING, H. G., and B. C. HENDRICKS: *Introductory College Chemistry*, John Wiley & Sons, Inc., New York, 1942. A standard elementary text which emphasizes the industrial applications of chemistry.

HATCHER, W. H.: *An Introduction to Chemical Science*, John Wiley & Sons, Inc., New York, 1940. See p. 208.

On the chemistry of carbon compounds:

DEMING, H. G.: *In the Realm of Carbon*, John Wiley & Sons, Inc., New York, 1930. An easy introduction to organic chemistry, based on its historical development.

On industrial chemistry:

ARRHENIUS, S. A.: *Chemistry in Modern Life*, translated by C. S. Leonard, D. Van Nostrand Company, Inc., New York, 1925. A detailed but readable book by the author of the ionic theory of solution.

SLOSSON, E. E.: *Creative Chemistry*, rev. ed., D. Appleton-Century Company, Inc., New York, 1930. A simple and well-written account of the outstanding accomplishments of chemistry in modern industry.

HAYNES, W.: *This Chemical Age*, Alfred A. Knopf, New York, 1942. Popularly written but fairly detailed discussion of synthetic fabrics, dyes, drugs, perfumes, rubber, and plastics.

RIEGEL, E. R.: *Industrial Chemistry*, Reinhold Publishing Corporation, New York, 1937. A standard text.

PART V

THE BIOGRAPHY OF THE EARTH

We have surveyed a wide variety of chemical processes. In many reactions which take place between substances dissolved in water, we have found ions playing a central role. We have seen the importance both in natural processes and in industry of control over the energy liberated by chemical reactions and over the speed with which reactions take place. We have interpreted the processes of decay, the production of muscular energy, the recovery of metals from their ores in terms of reactions involving valence changes. Carbon compounds and their reactions have given us an insight into life processes, silicon compounds and their reactions an insight into the nature of rocks and soil. We have had a glimpse at the behavior of colloids, substances in the border zone between suspensions and true solutions.

This is the panorama of modern chemistry. Through the diversity of its materials and its processes we have clung to one unifying idea, the electronic structure of atoms. In spite of the mystery which surrounds its ultimate nature, the electron is amazingly useful in simplifying and correlating the facts of chemistry.

Thus we have begun our reconstruction of the world from its tiny, electrically charged building blocks by seeing how these particles behave in fundamental chemical processes. The next stage of the reconstruction is to find what part is played by these chemical processes and by fundamental physical processes in the natural world. We turn our attention now from the laboratory to a study of the earth itself.

In shifting the emphasis of our study from experimental science to natural science we shall find at least two important changes in point of view. For one thing, in physics and chemistry we try continually to check our conclusions by laboratory experiment; in geology and astronomy we must deal with processes like a volcano in action or a star emitting radiation for which direct laboratory checks are impossible. In the second place, geology and astronomy are concerned not only with natural processes taking place at present, but also with the record of natural processes that operated in the past. In order to work with phenomena that occurred in the remote past, we shall have to devise new techniques of observation and reasoning.

CHAPTER XXXII

Earth Materials

TO BEGIN our intensive study of the earth, let us review briefly a few pertinent facts from earlier discussions.

Our planet is a nearly perfect sphere about 8,000 mi in diameter. This shape is a consequence of gravitation: no part of the surface can project much farther from the center than another part, since its greater weight would cause material beneath it to flow out under regions of lower pressure. So great are pressures at depths of a few miles that even solid rock would flow in response to the weight of any considerable projection of the surface.

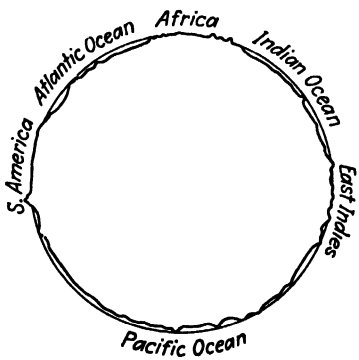


FIG. 192. *Cross section of the earth at the equator, with the vertical scale much exaggerated. Note the large extent of ocean basins compared to continents.*

Slight distortion from a true spherical shape is caused by the centrifugal force of the earth's rotation, the distance from pole to pole being about 27 mi shorter than the diameter at the equator. Lesser deviations from the spherical form give us continents and ocean basins, maintained probably because rock material under the continents is lighter than that under the oceans, hence can project farther from the earth's center. The vertical distance between the highest elevation on land (about 29,000 ft, at the top of Mt. Everest) and the deepest part of the ocean (about 35,000 ft, near the Philippine Islands) is little more than 12 mi, only a small fraction of the earth's radius. The earth's lofty mountains seem impressively high from our human point of view, but on the broad surface of the planet they are the merest wrinkles—smaller in comparison with its size than the tiny wrinkles on the surface of an orange (Figs. 192, 193).

The earth rotates on its axis once a day and revolves around the sun once a year. The inclination of axis to orbit gives us our seasons, for the sun's direct rays strike different parts of the earth at different times during the revolution.

In general, the earth's materials fall into three broad divisions: an outer envelope of gas, called the ***atmosphere;*** a thin, discontinuous envelope of water, called the ***hydrosphere;*** and the solid body of the planet, called the ***lithosphere.*** In the following paragraphs we shall examine these divisions briefly, paying particular heed to the amount and kind of information obtainable about each.

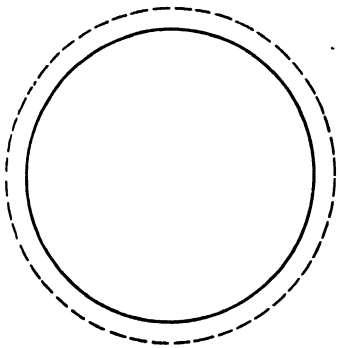


FIG. 193. *Cross section of the earth, to true scale. The width of the solid line includes all surface irregularities and the entire thickness of the hydrosphere. The dotted line is the "top" of the atmosphere (600 miles) as indicated by the maximum height of the aurora borealis.*

The Atmosphere

We live surrounded by the atmosphere, and some of our sense organs are well adapted to keep us informed about its temperature, its moisture content, its motion, and impurities which may be present in it. For more precise information we depend on scientific instruments rather than sense impressions. Some of these instruments are built for use at the earth's surface, while others, carried in airplanes or balloons, enable us to explore higher levels of the atmosphere. Stratosphere balloonists have taken instruments to a maximum height of 14 mi above sea level; small balloons equipped with automatic devices for recording temperature and pressure and for collecting samples of air have climbed some 8 mi farther. Rockets which have recently been projected over 100 mi into the atmosphere give promise of providing measurements at higher levels, but so far our information about air more than 22 mi from the surface is entirely indirect.

An aviator climbing upward finds that the atmosphere grows steadily thinner. About 3½ mi up his barometer reads 38 cm, half the normal atmospheric pressure at sea level; at about 7 mi the reading falls below 19 cm, showing that over three-fourths of the weight of the atmosphere is beneath him. His thermometer also shows steadily lower readings: −20°C at an elevation of 3½ mi and −55° at 7 mi, even though he may have started his climb on a warm summer day. If he ventures higher than 7 mi above sea level, he finds that the air continues to thin

rapidly, so that presently it can no longer support his machine; but the temperature falls no further than −55°C. If he continues his explorations upward by exchanging his airplane for a stratosphere balloon, he still finds no lower temperatures. Even in the rarefied atmosphere penetrated by balloons which carry instruments only, where barometers show a pressure of a fraction of a centimeter, the temperature remains about −55°. Although this constant temperature of the upper atmosphere is uncomfortably cold, it is not as low as surface temperatures which have been recorded in northeastern Siberia.

Besides the constant temperature, an aviator would notice other peculiarities of the atmosphere above a height of 7 mi: here there are practically no clouds and no storms, and dust is almost completely absent. Since the character of the atmosphere changes rather abruptly at the 7-mi level, this is taken as the boundary between two parts of the atmosphere*—the clear, cold upper part or *stratosphere* and the denser lower part or *troposphere*. Such common atmospheric features as clouds and storms, fog and haze, are confined to the troposphere.

That the stratosphere extends upward for many miles beyond the 22-mi limit reached by balloons is shown by several kinds of evidence. (1) The twilight glow of the evening sky and the light of dawn before sunrise are produced by the scattering of sunlight in the upper atmosphere. The duration of dawn and twilight shows that sufficient air to scatter light persists up to nearly 50 mi above the earth's surface. (2) If two observers see the same meteor from different positions, they can use the ordinary triangulation method of surveyors to calculate its height at the beginning of its path. These measurements show that enough air is present at an elevation of 100 mi to raise the temperature of a swiftly moving meteor to incandescence. (3) Streamers of the aurora borealis, or northern lights, sometimes extend upward to a height of 750 mi. This mysterious light of northern skies is almost certainly the effect of electrified particles from the sun striking the extremely rarefied uppermost part of the atmosphere. So far as tangible evidence goes, the 750-mi upper limit of the aurora is the "top" of the atmosphere, although probably stray molecules climb to even greater heights.

Air is a mixture of many gases, whose proportions remain nearly constant to an altitude of at least 22 mi. Pure, dry air contains about 78 per cent nitrogen, 21 per cent oxygen, 1 per cent argon, with minor amounts of carbon dioxide and inert gases of the argon family. Ordinary air contains also a variable amount of water vapor, up to 5 per cent on humid days, and considerable solid material in the form of dust particles.

* The boundary between stratosphere and troposphere is higher near the equator (about 10 mi above sea level) and lower near the poles (about 4 mi).

The changes of temperature and humidity in the lower atmosphere which produce differences in weather and climate will be the subject of the next chapter.

The Hydrosphere

All the water of the earth's surface, in lakes, rivers, and oceans, is included in the hydrosphere. By far the greater part of the hydrosphere, of course, is concentrated in the ocean basins. Inspection of a map shows that ocean basins occupy much more of the surface than do the continents; water bodies of various kinds, including oceans, lakes, and shallow seas, cover about three-fourths of the total area of our planet.

Firsthand observations of the hydrosphere are possible down to a depth of about ½ mi. Below this depth pressures become too great for the best diving equipment to withstand. But sampling devices can bring up specimens of ocean water from far greater depths, so that we have available considerable direct information about all parts of the hydrosphere.

The part of the hydrosphere in the ocean basins has an average thickness of about 2 mi and an extreme thickness in the deeps of the western Pacific of over 6 mi. Pressure increases steadily and rapidly below the surface of the hydrosphere, and temperature decreases. At great depths even in tropical oceans the temperature is only a few degrees above the freezing point. Unlike the atmosphere, the hydrosphere changes in density only slightly in response to the extreme pressures and low temperatures at depth.

Fresh water usually contains a little dissolved material, and the much larger amount in sea water is apparent from its taste. The kind of dissolved material in fresh water is highly variable, since it depends on the nature of the soil and rock with which the water has come in contact. The composition of the dissolved matter in sea water, on the other hand, is surprisingly constant in different parts of the ocean and in samples taken at different depths. The ions of ordinary salt (Na^+ and Cl^-) are the most abundant substances; the ions Mg^{++}, SO_4^{--}, K^+, Ca^{++}, CO_3^{--}, Br^- are present in considerable quantity; and smaller amounts of many other elements, even gold, silver, and radium, can be detected by careful analysis.

In addition to dissolved substances, most parts of the hydrosphere contain considerable finely divided material in suspension.

The Lithosphere

What little we can see of the lithosphere is composed chiefly of solid rock. Soil, vegetation, and fragmental material in places form a thin surface layer, but bedrock is always found when wells or excavations

penetrate beneath this cover. The deepest mines (about 1½ mi) and the deepest wells (about 3 mi) encounter nothing but rock material like that at the surface. We find good evidence that some rock layers now exposed at the surface were once buried to depths of nearly 20 mi; examination of these shows that even 20 mi down the lithosphere contains no materials very different from the substances in ordinary rocks.

Of the earth's 4,000-mi radius 20 mi is about half of 1 per cent. Yet a thin shell 20 mi in thickness is all the lithosphere with which we can claim any direct acquaintance—and even this claim is somewhat boastful, for we do not know that rock 20 mi beneath the ocean basins is at all like rock 20 mi beneath the continental surfaces. The thickness of the solid line in Fig. 193 represents, on a true scale, not only the thickness of the hydrosphere and the maximum relief of the surface, but considerably more of the earth's solid material than we know from firsthand observation.

The part of the lithosphere which we can study directly is often called *the earth's crust*, an old term inherited from days when the interior was believed to be molten. We shall find the expression convenient to refer loosely to an outer layer about 20 mi thick.

The average composition of the crust, as determined by numerous chemical analyses of rocks, is given in Table XXII. The numbers are percentages of the various elements by weight. If the compositions of the oceans and atmosphere are included in such a table, the percentages are changed only slightly.

TABLE XXII. AVERAGE COMPOSITION OF EARTH'S CRUST
(Weight per cent)

O	47.33	Ca	3.47	Ti	0.46
Si	27.74	Na	2.46	H	0.22
Al	7.85	K	2.46	C	0.19
Fe	4.50	Mg	2.24	All others	1.08

The table shows one striking fact about the composition of the crust: a few elements are abundant, but the majority are exceedingly scarce. Oxygen alone makes up nearly half the weight of the crust; some of it is free in the air, some is combined with hydrogen in water, but the greater part is combined with silicon in silica and the silicates. Silicon and the two metals iron and aluminum account for three-fourths of the remainder of the crust's weight. Such familiar metals as copper, tin, lead, silver are too scarce to appear on the list. Nitrogen and the halogens are likewise lumped in the 1.08 per cent which includes "all others." Carbon and hydrogen, present in all living things, together make up less than half of 1 per cent of the total.

As might be guessed from the table, the chief compounds in the rocks of the crust are silicon dioxide and various silicates of the six metals Fe, Al, Ca, Mg, Na, and K. The only common rock in which these compounds are not dominant is limestone, which is chiefly calcium carbonate.

The Deep Interior

We shall leave further discussion of the crust to later chapters and turn our attention briefly to the earth's deep interior. What sort of indirect evidence can we find about conditions below the level of direct observation?

1. Regarding *temperatures* in the interior, we can make some guesses from measurements in deep mines and wells. These measurements indicate that temperature increases steadily with depth, at an average rate of about 1°C for every 110 ft. Probably the increase is not so rapid at lower levels, but at least we may be certain that the earth's interior is much hotter than its surface. The existence of volcanoes, geysers, and hot springs is further evidence of high temperatures within the earth.

2. Enormous *pressures* in the interior are produced simply by the weight of overlying material. With a few assumptions about the distribution of light and heavy material within the earth, the pressure at any depth may be calculated. Such calculations give pressures of several thousand atmospheres for depths of a few miles and about 3,000,000 atm near the earth's center. Laboratory experiments with high pressures show that the toughest rocks cannot maintain open fractures even at so shallow a depth as 20 mi; for the pressure at this level is sufficient to make solid rock "flow" and fill any cracks or cavities which may form. Caverns and underground streams can exist only near the surface. Deep in the earth not even tiny crevices can be maintained against the crushing weight of overlying rocks.

3. The old idea that our planet is a molten globe surrounded by a thin solid crust is proved false by measurements of the earth's *rigidity*. One method of determining rigidity is by measuring the amount of deformation produced in the earth by the moon's gravitational attraction. If the interior were largely molten, the same tidal forces which raise the oceans against their shores would cause great bulges in the body of the earth. Some distortion of the lithosphere by tidal forces can be detected with delicate instruments, but its small amount shows that the earth as a whole behaves like a solid body with a rigidity about twice that of structural steel.

4. Evidence from *density* measurements proves that the material of the interior must be considerably heavier than the compounds which make up ordinary rocks. The average density of the entire earth, found

very simply by dividing its volume into its mass, is about 5.5 g./cc, while the average density of rocks in the crust is only 2.7 g/cc. Evidently the interior must contain some substance with a density greater than 5.5 to give this figure for an over-all average.

5. A hint as to what this heavy material may be comes from analyses of meteorites. On the assumption that planets and meteors were formed in the same cosmic catastrophe, their compositions should be roughly similar. Some meteorites resemble ordinary rocks in composition, but others consist mainly of iron alloyed with a little nickel. The average composition of meteorites shows a much greater abundance of these two metals than is found in rocks; so it is a plausible hypothesis that iron and nickel make up a large part of the earth's heavy core.

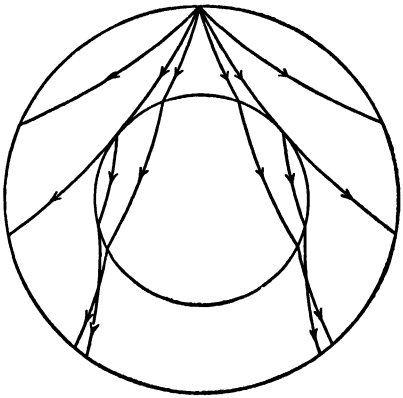


FIG. 194. *Diagram showing the paths of earthquake waves through the earth's interior. Since speeds increase with increasing depth, the vibrations travel most rapidly along curved paths. At major divisions in the interior the waves are bent, or refracted, just as light is refracted on passing from air into water. (For simplicity, only one of the major divisions is shown.)*

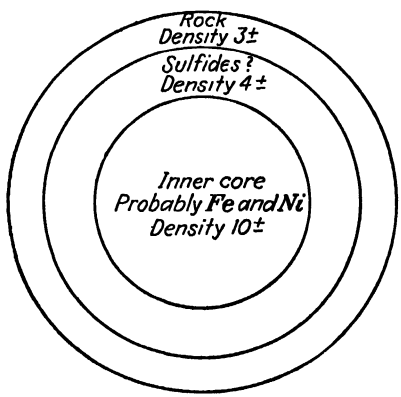


FIG. 195. *Diagrammatic cross section through the earth, showing the major divisions suggested by the speeds of earthquake waves. On this scale the 20-mile "crust" of the earth is included in the thickness of the outer line.*

6. Our most detailed information about the earth's interior comes from earthquake shocks. An earthquake (page 539) is the rapid vibratory movement resulting from a sudden fracture in the solid material of the earth. Like a dynamite explosion, such a fracture sends out vibrations in all directions, vibrations which often are highly destructive near the point of fracture. A big quake may shake the whole earth; its vibrations at distant points are not strong enough to be felt, but may be detected with delicate instruments. Since these vibrations travel through

which they encounter, measurements of the times required for their journey through the earth give us information about materials at different depths (Fig. 194). At a depth of about 300 mi, and again at a depth of about 1,800 mi, earthquake waves show abrupt changes in speed, suggesting that these levels mark important changes in the composition of the interior (Fig. 195). Just how the composition changes at the transition points is not entirely clear: one widely favored hypothesis suggests an inner core composed largely of iron and nickel, an intermediate shell of iron and nickel sulfides, and an outer shell of material like ordinary rocks.

The earthquake vibrations which travel through the earth consist of both longitudinal and transverse wave motions (page 260). Both kinds of waves traverse easily the two outer shells, down to the 1,800-mi level. Since transverse waves can travel only through a rigid medium, this confirms the evidence from tidal deformation that at least the greater part of the earth is a rigid solid. Evidence regarding the inner core is not so conclusive: transverse waves which pass through it are greatly weakened, but not blotted out completely. The weakening suggests that material in the inner core may have some of the properties of a liquid; but it is doubtful whether our ordinary distinction between solid and liquid matter can be correctly applied to material at so high a temperature and under such enormous pressures.

To summarize our knowledge of the earth's interior: Materials within the earth are fairly hot and under high pressures. Despite the high temperatures, the earth is for the most part solid. Its materials are arranged in three well-defined layers, an inner core surrounded by two thick shells. The density of the material in the core is greatest, that of the material in the outer shell the least. A mixture of iron with a little nickel seems the most plausible guess as to the composition of the heavy core.

Perhaps the most important single fact which emerges from this discussion is the extreme limitation of precise knowledge about the materials of our planet. The hydrosphere, the 22 mi of the atmosphere which we can examine directly, the outer 20 mi of the lithosphere are but three films surrounding the vast unknown interior. In succeeding chapters, when we ambitiously examine the processes which have shaped the earth's surface, we must remember our almost complete ignorance of what goes on beneath the surface. If our conclusions are sometimes vague, if our discussions end in a welter of controversial hypotheses, it is because we must often deal with processes which have their origin in the planet's unknown depths.

Questions

1. Indicate which of the following statements are based on direct observational evidence. For each of the others indicate the kind of indirect evidence on which it is based.
   a. The composition of the earth's interior changes abruptly at the 300-mi level and at the 1,800-mi level.
   b. The lower part of the stratosphere has a constant temperature of about −55°C.
   c. The earth's inner core consists largely of iron and nickel.
   d. The atmosphere extends upward for at least 750 mi.
   e. The composition of sea water is nearly constant in different parts of the ocean.
   f. The earth's interior is for the most part solid.
   g. Material in the earth's interior is more dense than that of the crust.
2. Name (*a*) the four most abundant elements in the earth's crust, (*b*) the four most abundant elements in the hydrosphere, (*c*) the three most abundant elements in the atmosphere. What is probably the most abundant element in the earth as a whole?
3. Suppose that you were traveling upward in an airplane not equipped with an altitude meter. How could you tell when you were approaching the top of the troposphere?

CHAPTER XXXIII

Weather and Climate

WE HAVE discussed in various connections the composition of the atmosphere, the pressure of the atmosphere, the extent of the atmosphere. We turn now to those general characteristics of the earth's gaseous envelope which determine the weather and climate of different regions. This branch of earth science is called *meteorology*—a word which, despite appearances, has nothing to do with meteors.

Meteorology is the study of a vast, automatic air-conditioning system. Our spinning planet is heated strongly at the equator, feebly at the poles, and its moisture is concentrated in the great ocean basins. It is the task of the atmosphere, from our point of view, to redistribute this heat and moisture so that large areas of the land surface will be habitable. Air-conditioning by the atmosphere is far from perfect: it fails miserably in desert regions, on mountain summits, in far northern and southern latitudes. On sultry nights in midsummer or on bitter January mornings we may question its efficiency even in our favored part of the world. But the atmosphere does succeed in making a surprisingly large amount of the earth's surface fit for human habitation.

The two chief functions of an air-conditioning outfit are the regulation of air temperature and humidity. In addition to these, we expect the atmosphere to perform a third function: it must provide us at intervals with rain or snow. The weather or climate of a given locality is a description of how effectively these functions are performed. ***Weather*** refers to the temperature, humidity, and rainfall at a certain time, while ***climate*** refers to the average conditions over a period of years. Important in a description of climate is also the variability of temperature and rainfall with the seasons: an outstanding feature of the climate of North Dakota is its extreme warmth in summer and cold in winter, while southern California's well-advertised climate is characterized by equable year-round temperatures and by a concentration of rainfall in the winter

months. Locally barometric pressures and the intensity and direction of wind may be important in descriptions of weather and climate.

In discussing the earth's air-conditioning system, we face the same problems which an artificial system must solve: the control of temperature, the control of moisture, and the maintenance of adequate circulation. We shall study each of these in turn, but the discussions must somewhat overlap.

Atmospheric Temperatures

The earth's only important source of heat is the sun. Energy from the sun reaches us in the form of electromagnetic radiation of short wave length—chiefly visible light, together with some ultraviolet and some infrared radiation. Radiation of this sort is absorbed only slightly by the atmosphere, most of it coming through to the earth's surface. Here it is absorbed and converted into heat. The warm earth then reradiates its excess energy back into the atmosphere, but the energy now is in the form of long-wave-length heat radiation. These long wave lengths are readily absorbed by two of the minor constituents of air, carbon dioxide and water vapor. Their molecules, speeded up by the absorption of heat energy, give some of this energy to other air molecules during collisions. Thus the atmosphere is heated *directly* by heat radiation from the earth, only indirectly by the energy of sunlight.

If the earth had no atmosphere, its heated surface would quickly radiate back into space all the energy which reaches it from the sun. Like the moon, the earth would grow intensely hot during the day, unbearably cold at night. An atmosphere effectively prevents these extremes of temperature: its continual movement makes impossible undue heating of any one region by day, and its ability to absorb and hold the earth's radiation prevents the rapid escape of heat by night. Our atmosphere acts as an efficient trap, admitting the energy of sunlight freely, but hindering its escape.

How hot the atmosphere becomes over any particular region depends on a number of factors. Air near the equator is on the average much warmer than air near the poles, because the sun's vertical rays are more effective in heating the surface than the slanting rays of polar regions (Fig. 196). Air over a mountain summit may become warm at midday but cools quickly because it is thinner and contains less carbon dioxide and water vapor than air at lower elevations. A region covered with clouds usually has lower air temperatures than an adjacent region in bright sunlight. Because the temperature of water is raised more slowly than that of rocks and soil by absorption of radiation, the atmosphere near large bodies of water is usually cooler by day and warmer by night than over regions far from water. Desert regions commonly show

abrupt changes in air temperature between day and night, because so little water vapor is present to absorb heat radiation. In later sections of this chapter, we shall see how the atmospheric temperatures of some regions are profoundly influenced by winds and by ocean currents.

Atmospheric Moisture

The moisture content, or *humidity*, of air refers to the amount of water vapor which it contains. Most of the atmosphere's water vapor comes from evaporation of sea water; a little comes from evaporation of water in lakes, rivers, moist soil, and vegetation. Since water vapor is continually being added to air by evaporation and since it is periodically removed by condensation as rain, snow, and fog, the humidity of the atmosphere is extremely variable from day to day and from one region to another.

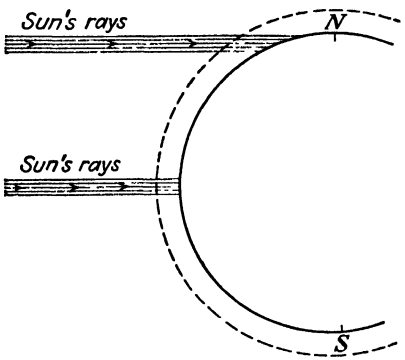


FIG. 196. *Heating of the earth's surface is less effective in the polar regions than near the equator for two reasons: (a) the sun's rays must penetrate a greater thickness of air, and (b) the energy of the rays is distributed over a greater area.*

The amount of water vapor which can be present in air at a given temperature is limited by the vapor pressure of water (page 138). At 20°C the vapor pressure of water is 17.4 mm;* at this temperature water will evaporate until the pressure of its vapor is 17.4 mm, but thereafter no further evaporation will take place. If the temperature increases, the vapor pressure becomes higher and more water can evaporate.

Changes in humidity are most accurately expressed in terms of vapor pressures, but meteorologists commonly describe these changes in somewhat different language. Air is said to be *saturated* with water vapor when it contains the maximum amount which will evaporate at a given temperature: thus at 20° air is saturated when it contains sufficient water vapor to exert a pressure of 17.4 mm. Air is *unsaturated* when its amount of water vapor is less than this limiting value, since it is capable of "holding" more water vapor. In effect, air is regarded as a sort of sponge, filled more or less completely with water vapor. Actually, of course, the air has nothing to do with evaporation; if no air existed, vapor would still escape from bodies of water. But, since air is the agent which transports water vapor from one region to another and since air is the medium

* Pressures in this chapter are expressed as millimeters or centimeters of mercury (p. 120).

in which water vapor condenses as rain or snow, we shall find it convenient to think of the air as "taking up" and "holding" different amounts of vapor.

Commonly we describe air as "humid" if it is saturated or nearly saturated, as "dry" if it is highly unsaturated. Humid weather is oppressive because little moisture can evaporate from the skin into saturated air, so that perspiration does not produce its usual cooling effect. Very dry air is harmful to the skin because its moisture evaporates too rapidly. Meteorologists express the distinction between dry and humid air more accurately in terms of *relative humidity*, a number indicating the degree to which air is saturated with water vapor. Usually relative humidity is expressed in per cent: at 20° the relative humidity is 100 per cent when water vapor has a pressure of 17.4 mm, 50 per cent when the pressure is 8.7 mm, 10 per cent when the pressure is 1.74 mm, and so on.

The amount of moisture which air can hold changes markedly with temperature. At 0°C (32°F) air is saturated when the pressure of water vapor is only 4.6 mm; at 40°C (104°F) water vapor in saturated air has a pressure of 55.3 mm. If air saturated at 20°C (17.4 mm water vapor) is heated to 40°, it can obviously take up more water vapor and so is no longer saturated. (In other words, its relative humidity decreases, although the amount of water vapor does not change.) If, on the other hand, air saturated at 20° is cooled to 0°, some of its water vapor must condense out as liquid water, since at the lower temperature the air can hold only about one-fourth as much vapor as it contained originally. Further, if air at 40°C containing enough water vapor to give a pressure of 17.4 mm is cooled to 0°, it grows steadily more saturated until a temperature of 20° is reached, after which it remains saturated down to 0° and some of its vapor condenses out. *Thus any sample of ordinary air on heating grows less saturated, on cooling grows more saturated; if the cooling is continued past the saturation point, some liquid water (or ice) must condense out.*

The condensation of water from cooled air is an everyday occurrence. Dew forms because the ground surface is cooled by rapid radiation at night, so that air near the ground has its temperature lowered below the saturation point. Fogs are produced when larger masses of air are cooled by contact with cold earth or water bodies. Clouds form when air cools by expansion on rising. Whether the vapor condenses as droplets of liquid water or as tiny particles of ice depends on the temperature.

The growth of tiny water droplets into the larger drops of rain or of minute ice crystals into the lacework patterns of snowflakes can take place only when large masses of air are cooled rapidly. Ordinarily nature provides only one method by which cooling on so grand a scale can take place: the chilling of air by expansion on rising. In the next sections we

shall inquire into the various circumstances which may cause air to rise in large masses.

The Circulation of the Atmosphere

We come now to the third major problem of the earth's air-conditioning system—the movements of the atmosphere which distribute heat and moisture over the earth's surface. These movements are in large part governed by two disarmingly simple facts: (1) heated air rises, because it is lighter than the surrounding cooler air, and (2) air moves horizontally from regions of relatively high pressure to regions of low pressure. To explain atmospheric circulation, then, we must look for causes of differences in temperature and pressure at various points on the earth's surface.

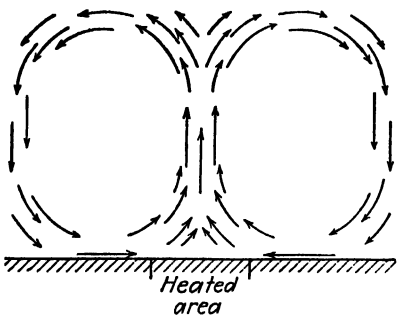


FIG. 197. *Simple convectional circulation.*

A simple idealized example will give us an insight into one important kind of atmospheric movement. Suppose that a small area of the land surface is heated more strongly than adjacent regions (Fig. 197). Air above the heated surface becomes warm, expands, and begins to rise. As it rises, a column of air above it is pushed upward, and air from the upper part of the column spills over into adjacent regions. Because the hot air is light and because air above it moves to each side, pressure over the heated spot is relatively low. Hence cold air from the surroundings moves toward the heated area, is in turn heated, and moves upward. Air currents produced in this manner, as a direct response to unequal heating of the land surface, are called *convection currents.*

The earth as a whole is heated strongly in the equatorial belt, less strongly on either side; so we might expect to find convection currents as a part of the general atmospheric circulation. Imagine for a moment a somewhat idealized earth: suppose that our planet does not rotate, that it is heated strongly near the equator, and that its surface is made up entirely of either land or water. On such an earth air circulation would depend exclusively on the difference in temperature between equator and poles. Air would rise along the heated equator, overflow at high altitudes toward the poles, and at low altitudes move continually from the poles back toward the equator (Fig. 198). We in the northern hemisphere would experience a steady north* wind. Around the equator

* Wind direction is conventionally described as the direction *from which* the wind is blowing.

would be a belt of relatively low pressure, near each pole a region of high pressure.

A huge convectional circulation of this sort exists, but it is profoundly altered by the earth's rotation. To see the nature of the alteration, imagine another sort of idealized earth, spinning on its axis this time but uniformly heated over its entire surface. Just as a bead placed on a moving phonograph record spirals quickly outward, so air molecules on a rotating planet experience a centrifugal force impelling them to move as far as possible from the axis of rotation—*i.e.*, toward the equatorial belt. On a uniformly heated earth, therefore, we should expect the atmosphere to be thickest at the equator, thinnest at the poles.

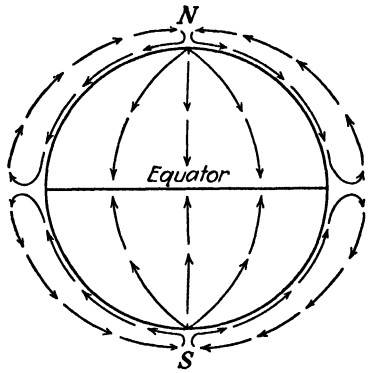


FIG. 198. *The convectional circulation which would exist if the earth were a nonrotating sphere heated uniformly at the equator. (Arrows indicate surface winds except at edges of diagram.)*

On the real earth, we find that a sort of compromise between the effects of rotation and unequal heating distributes the atmosphere in broad belts of relatively high and low pressure. Near the equator centrifugal force

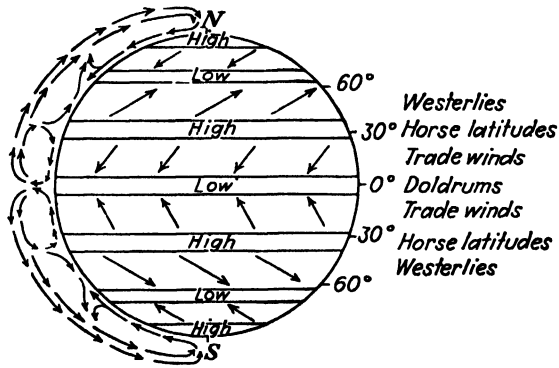


FIG. 199. *Wind belts and belts of permanent high and low pressure. Arrows indicate directions of principal winds. (Highly diagrammatic; actually the belts shift north and south with the seasons, and their positions are modified by the continents. Compare Fig. 200).*

tends to pile up the atmosphere, while high temperature tends to produce a region of low pressure. At the equator itself the temperature effect predominates and pressures are relatively low, but belts of high pressure parallel the equator on either side. Near the poles the tempera-

ture effect is again predominant, giving high pressures, but at the edges of the polar regions are belts of low pressure (Fig. 199). Although these belts of high and low pressure are in places greatly modified by large continental masses, they are in general the chief determiners of the earth's wind systems. Along the earth's surface the dominant wind directions are from the belts of high pressure toward the belts of low pressure.

One further effect on wind directions is exerted by the earth's rotation: everywhere except along the equator, winds are deflected from straight lines into curved paths, the deflection being toward the right in the northern hemisphere and toward the left in the southern hemisphere. Air moving southward from the high-pressure zone in the northern hemisphere instead of moving straight south is deflected toward the southwest; air moving northward from this belt is turned toward the northeast and east.

With these facts in mind, we may summarize the general circulation of air over the earth's surface (Fig. 200). In the low-pressure belt along the equator, strongly heated air rises and overflows at high altitudes to the north and south. Cooling as it leaves the equator and deflected from its poleward course by the earth's rotation, much of this air descends to the surface in latitudes about 30° north and 30° south of the equator, the descending currents forming belts of high pressure. From the high-pressure belts the air in part returns along the surface to the equator, in part moves northward or southward toward the polar belts of low pressure. At the equator, where the principal movement of air is upward, and in the high-pressure belts on either side, where the principal movement is downward, surface winds are light and erratic. These three zones have long been known to sailing vessels as regions of protracted calms—the *doldrums* along the equator, the *horse latitudes* about 30° on either side. Air currents returning to the equator from the horse latitudes form the *trade winds*, winds whose steadiness in direction and intensity were a great aid to navigation before the days of steamships. Deflected by the earth's rotation, the trades of the northern hemisphere blow from northeast to southwest, those of the southern hemisphere from southeast to northwest. Winds which start poleward in both hemispheres from the horse latitudes are so curved by the earth's rotation that their dominant direction is from west to east; these are the *prevailing westerlies* or the *stormy westerlies* of the temperate zones. Concerning air circulation in the polar regions exact information is meager, but at least in part the motion consists of downward currents near the poles themselves and surface currents moving outward and westward toward the adjacent low-pressure belts.

Effects of the general atmospheric circulation are masked in many regions by winds of local origin. Along seacoasts the land is heated more

rapidly than the ocean by day and cools more rapidly at night; hence during the day a breeze often blows from the sea toward the warmer land, and during the night from the land toward the warmer ocean.

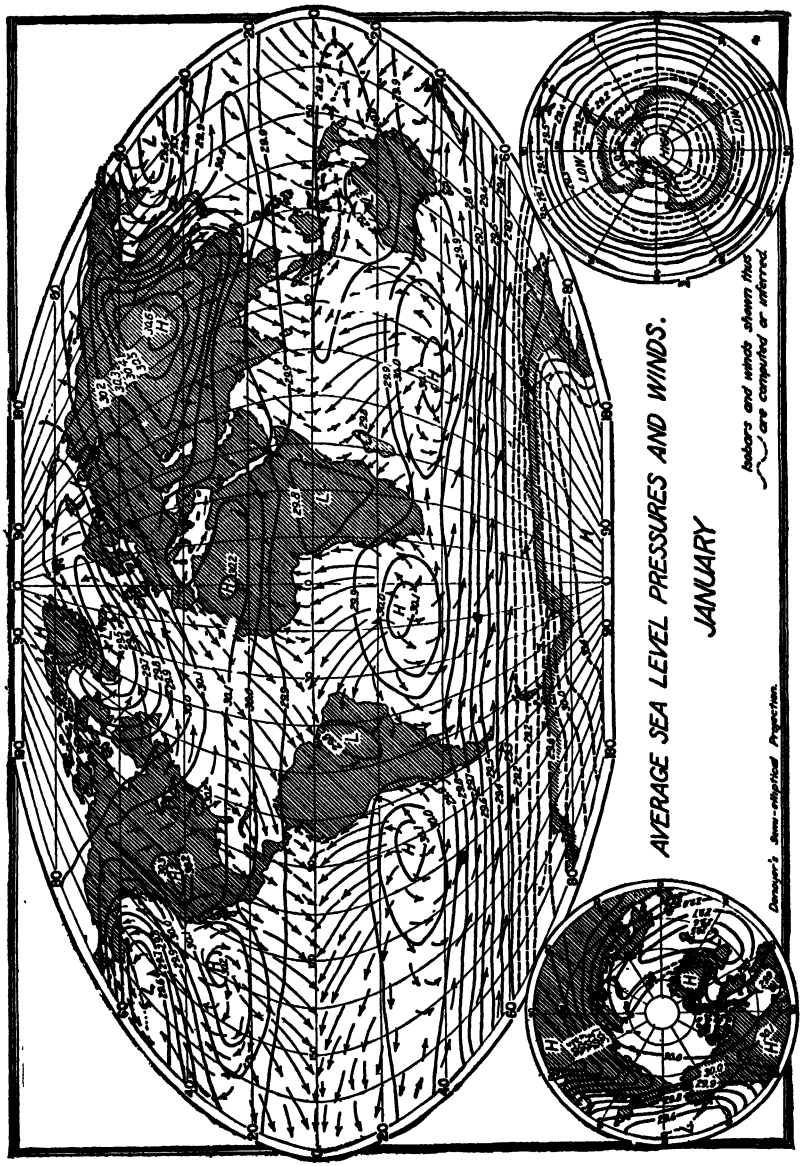


FIG. 200. (*From Finch and Trewartha's Elements of Geography.*)

The *monsoons* of India are similar winds on a grand scale: in summer the huge continent of Asia becomes much warmer than the adjacent Indian Ocean, so that strong south winds blow from sea to land; in winter the

interior of Asia becomes intensely cold, so that the winter monsoon blows southward toward the sea. In mountain valleys winds often blow upward toward the mountain summits by day, downward by night. The influence of storms on local wind directions, especially in the zones of the westerlies, we shall consider in a later section.

Climates

Having gained a fair acquaintance with the mechanism of the earth's elaborate air-conditioning system, we proceed now to find out how successfully it controls the temperature and moisture of various regions. Before starting this inquiry, we need a little more information about the causes of rainfall.

The production of rain or snow requires that large masses of humid air should be cooled to a temperature below the point at which the air is saturated. The necessary water vapor, obtained mostly from evaporation of sea water, may be carried long distances inland by the atmospheric circulation. Sufficient cooling to produce rain is accomplished only by the expansion of air on rising. Three causes for upward movement of large masses of air may be distinguished: (1) convection currents, either as part of the general circulation or as a result of local heating, (2) wind forced upward by mountain barriers, and (3) movements of air masses and associated cyclonic storms (page 461). We shall expect abundant rain or snow, then, wherever air well supplied with moisture moves over a strongly heated area, where mountain ranges lie across the path of prevailing winds, and where cyclonic storms are frequent.

These few facts about rain, together with those we have just learned about atmospheric temperatures and the earth's wind system, make possible a shrewd guess regarding the climate of almost any point on the globe.

The equatorial belt of calms, with its rapid evaporation and strong rising air currents, provides an ideal situation for superabundant rain. Throughout the year the weather is hot, sultry, with almost daily rains and light, changeable winds. Where this belt crosses land occur the steaming jungles of Africa, South America, and the East Indies.

The horse latitudes, roughly 30° north and south of the equator, are also belts of calm, but their climate is anything but humid. Air in these belts moves chiefly downward, becoming compressed and heated, hence less and less saturated with water vapor. Thus the climate is perennially dry, with clouds and rain only at long intervals. In a few regions, however, like the Gulf coast of the United States, the prevailing aridity is modified by moisture-laden winds of local origin.

Air returning to the warm equatorial belt from the horse latitudes on either side has little reason to lose the small amount of water vapor it possesses, except locally where a mountain range or strong convection cur-

rents force it sharply upward. Hence the trade-wind belts, like the horse latitudes, are in general regions of excessive dryness. Seasonal movement of the wind belts to the north and south gives rainfall during part of the year to the equatorial margins of the trade-wind belts. The outer portions of these belts, together with the adjacent horse latitudes, are the regions of the world's great deserts—the Sahara, the deserts of South Africa, the arid districts of Mexico and northern Chile, the great desert of Australia.

The belts of prevailing westerlies in general have moderate average temperatures. Continental interiors show great seasonal variations in temperature, while oceanic islands and the west coasts of continents have equable temperatures throughout the year. Further generalization about these zones is difficult, because their weather and climate are so largely determined by storms of local origin. Winds vary greatly in strength and direction, and conditions of moisture and temperature vary with them. In the northern hemisphere the huge land masses of North America and Eurasia introduce further complications. The complexities of the northern belt of westerlies are well illustrated by a brief survey of climates in the United States: Winds from the Pacific Ocean are forced abruptly upward by a succession of mountain ranges along the West coast, the western sides of the mountains therefore receiving abundant rainfall. Once across the mountain barriers the westerlies have little remaining moisture, so that the region east to the Great Plains is largely arid. If the westerlies maintained their direction as steadily as do the trade winds, arid conditions would continue across the continent to the East coast; but wind directions are continually changed by the cyclonic storms characteristic of this belt, so that moisture-laden air is frequently brought from the Gulf of Mexico and the Atlantic Ocean into the Mississippi Valley and the Eastern states. Rainfall increases eastward across the country, becoming very large along the Gulf of Mexico. Temperatures on the West coast, conditioned by the prevailing wind from the ocean, change relatively little from season to season, but in most other parts of the country the difference between summer and winter is very marked. The fabulous climates of Florida and southern California owe their mildness to nearby warm oceans and to a position approximately at the junction of the belt of westerlies and the horse latitudes.

Poleward from the belts of the westerlies are the bleak arctic and antarctic regions. Summers are short, winters long and cold. Moderate winds are the rule, although violent gales occur at times. The total amount of snow during the year is small simply because the low temperatures prevent the accumulation of much water vapor in the air.

This résumé of world climates is necessarily incomplete. We have tried simply to describe the climates of different latitudes which prevail

over broad areas of the earth's surface, without dallying over the local influence of mountains, large land masses, and ocean currents. Local effects of this sort make the climates of some regions quite different from those we have pictured.

Air Masses

Day-to-day weather is more variable in the temperate zones than anywhere else on earth. If you visit central Mexico or Hawaii, in the belt of the northeast trades, you find that one day follows another with hardly a perceptible change in temperature, moisture, or wind direction; but in nearly all parts of the United States abrupt changes in weather are commonplace. The reason for this variability lies in the movement of warm and cold air masses and of storms derived from them through the belts of the westerlies.

The northern edge of the westerly belt is a fluctuating line separating air moving generally northward from the horse latitudes and air moving southward from the polar regions. Great tongues of cold air at times sweep down over North America, and at other times warm air from the tropics extends itself far northward. The cold air is ultimately warmed and the warm air cooled, but a large body of air can maintain nearly its original temperature and humidity for days or weeks. These huge tongues of air, or isolated bodies of air detached from them, are the *air masses* of meteorologists. The kind of air in an air mass depends on its source: A mass formed over northern Canada is cold and dry, one from the North Atlantic or North Pacific is cold and humid, one from the Gulf of Mexico warm and humid, and so on. Weather prediction in this country depends largely on following the movements of air masses from these various source areas.

The contact between a warm and cold air mass is inevitably a zone of disturbance. In general the lighter air of the warm mass moves up over the heavy air of the cold mass, so that the contact surface is inclined. The surface is called a *frontal surface*, and the line where the surface meets the ground is called a *front*. As the air masses move, the frontal surfaces at their margins move also, generally eastward with the drift of the westerlies. A front with a cold mass to the west and a warm mass to the east is a *cold front;* a front separating warm air on the west from cold air on the east is a *warm front* (Fig. 201). The movement of a cold front brings cold air in place of warm, and the movement of a warm front brings warm air in place of cold.

As warm air rises along an inclined frontal surface it is cooled and part of its moisture condenses out. Clouds and rain, therefore, are commonly associated with both kinds of fronts. A cold frontal surface is generally

steeper, since cold air is actively burrowing under warm air, and the temperature difference is greater, so rainfall on a cold front is heavier and of shorter duration than on a warm front. A cold front with a large temperature difference is often marked by violent thundersqualls and tornadoes.

Storms

Storms, most of them produced along the fronts between shifting air masses, are responsible for much of the changeable weather in the temperate zones. In everyday speech a "storm" is a brief period of violent wind and heavy rain or snow; meteorologists use the word also for larger

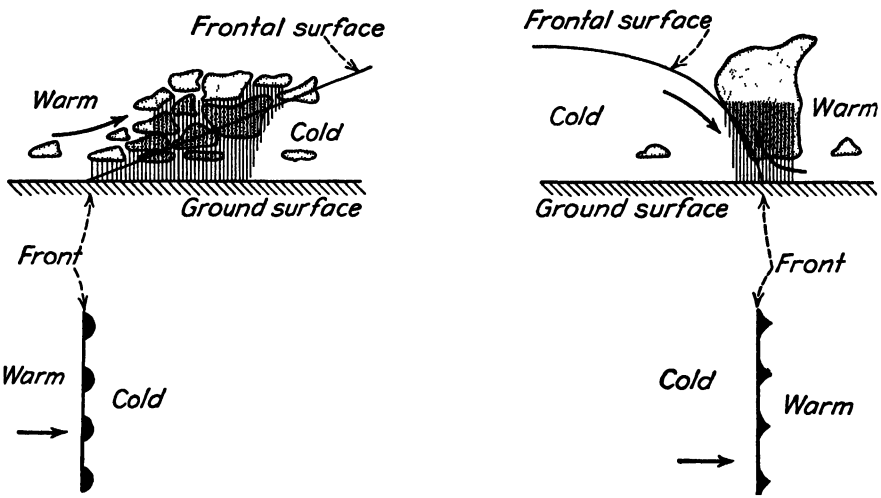


FIG. 201. *Diagrams of a warm front (left) and a cold front (right). The upper diagrams are vertical cross sections, and the lower diagrams show the usual symbols for fronts on surface maps. In each upper diagram the horizontal distance shown is about 200 miles and the vertical height is about 4 miles. Vertical lines indicate rainfall. Dots show areas of cloudiness.*

and less extreme disturbances of the atmosphere. Of the larger storms the commonest are called *cyclones*—a word used popularly as the equivalent of *tornado* (see below), but more properly restricted to huge storms of moderate intensity, several hundreds of miles in diameter, which often can be followed in their movement across country for days or weeks. Cyclones are so common in the United States that over a large part of the country they are the principal determiners of day-to-day weather.

A cyclone is an area of low atmospheric pressure, usually formed in the zone of disturbance along a front. Toward the low-pressure area winds blow in from all sides. Deflected by the earth's rotation, the incoming winds curve to the right (in the Northern Hemisphere), thus moving in

a counterclockwise direction around the cyclone's center. Near the middle of the storm the chief movement of air is upward. (Fig. 202).

Areas of high pressure called *anticyclones* commonly alternate with cyclones in the cross-country drift of the westerlies. In an anticyclone the circulation is opposite to that in a cyclone: air moves downward and spirally outward toward surrounding areas of lower pressure. Anticyclones are not generally regarded as storms, since they are seldom accompanied by rain or strong winds.

Both cyclones and anticyclones form as waves or "kinks" in fronts. After a cyclone is well developed it has the form shown in Fig. 203: a tongue of warm air surrounded by cold air, with a cold front to the west and a warm front to the east. Two stages in the development of an actual cyclone are shown in Figs. 205*a* and 205*b*. The cold front behind a cyclone

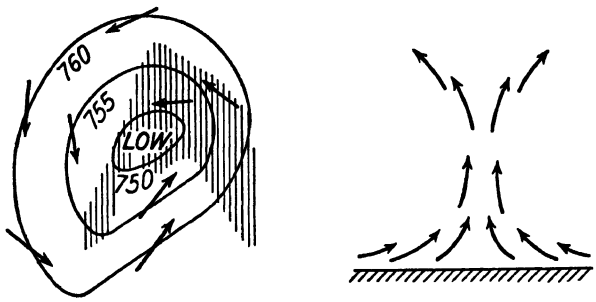


FIG. 202. *Air movement in a cyclone. A. map; B, side view. Lines in A show barometric pressures, given in millimeters of mercury. Vertical ruling shows area where rain or snow is most probable.*

moves slightly faster than the warm front ahead of it, so that ultimately the two fronts come together and raise the warm tongue bodily off the ground. Before it disintegrates a cyclone may travel completely across the United States, and a big one may even continue on over the north Atlantic into Europe. The exact path and speed of a given cyclone or anticyclone are difficult to predict, but the general motion is eastward at about 30 mi/hr or 700 mi/day. So numerous are these storms that the general flow of the westerlies across the United States resembles the movement of a river in which eddy follows eddy so rapidly that the general motion of the stream is apparent only in the bodily drift of the eddies.

Of the smaller and more violent storms the simplest to understand are ordinary *thunderstorms*, common in the equatorial belt and during the summer in middle latitudes. Most thunderstorms are caused by strong convection currents rising from areas which the sun has heated intensely, but some originate in violent air movements along cold fronts. Rainfall

in a thunderstorm is due to abrupt chilling of humid air by expansion in the upward moving currents (Fig. 204). Its spectacular electrical displays

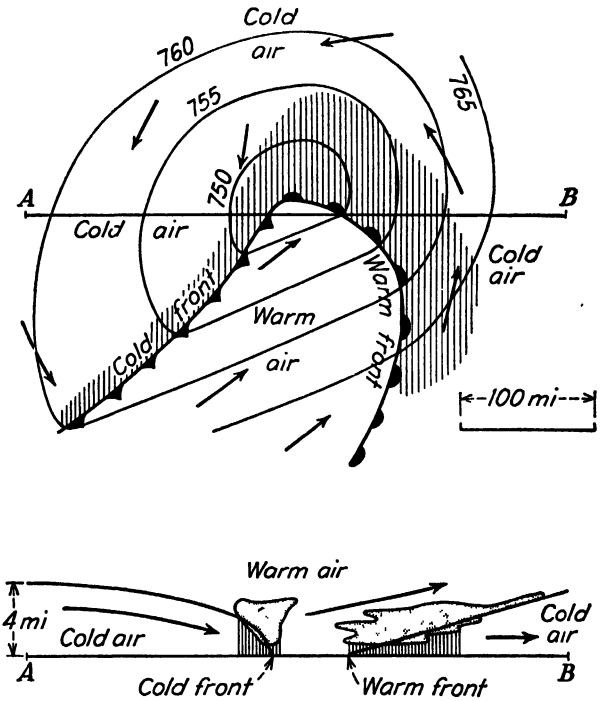


FIG. 203. *Structure of a cyclone. The upper diagram is a map, the lower one a vertical section along the line AB. Vertical lines indicate areas of rain or snow. Dots show areas of cloudiness.*

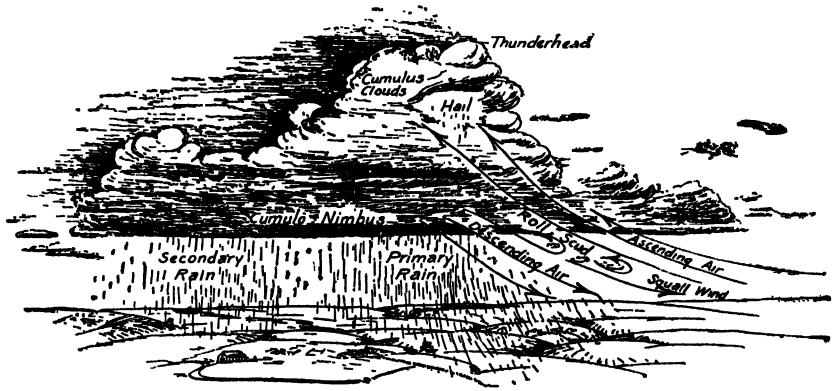


FIG. 204. *Structure of a local thunderstorm.* (*Courtesy of A. K. Lobeck.*)

are the result of accumulations of static charges separated when water droplets and ice crystals are torn apart by violent air movements (page

234). *Tornadoes*, most destructive of all storms, are an exaggerated form of the small dust whirls seen on a summer day—violent eddies with diameters ranging from 300 to 1,500 ft, produced on or near cold fronts with abrupt temperature drops. Tornadoes are often confused with cyclones, but to meteorologists the two are very different: Cyclones are large, with moderate winds, persisting for days or weeks; tornadoes are small, violent, and seldom last more than an hour or two. They are similar only in their inverted-whirlpool motion of air. *Hurricanes* and *typhoons* (Occidental and Oriental names for the same kind of storm) are small, violent cyclones (100 to 200 mi in diameter) of tropical regions, originating over warm oceans near the boundary between the trade-wind belts and the doldrums; they are accompanied by torrential rains and by winds which reach velocities over 100 mi/hr.

Weather Forecasting

Farmers, hunters, aviators, and others whose life exposes them constantly to the weather learn through long experience that certain kinds of clouds, a certain "feel" in the atmosphere, or certain wind directions often foretell a storm or a change in weather. A south or southeast wind, abnormally high temperatures, and a hazy sky generally mean the approach of a cyclone and rainy weather; when the wind veers to the west during a rain and the temperature starts to fall, the center of the storm has passed and the skies will probably clear. Scientific weather forecasting is merely a refinement of such observations and rough guesses.

The principal determiners of weather are air masses, fronts between air masses, and cyclonic storms. An air mass brings to the part of the country over which it lies its own particular humidity and temperature. At its boundaries, where warm air moves upward over cold air, are areas of cloudiness and precipitation—gentle rain of long duration on warm fronts, short and heavy downpours on cold fronts. A cyclone generally brings abundant rain or snow, since moisture-laden winds from the south and southeast are cooled as they move northward and upward into the center of the storm; the west side of a cyclone has clear weather, since winds from the north and west are becoming warmer and drier as they move south. In anticyclones the dominant downward movement of air usually ensures fair, cool weather. Since weather depends so greatly on air masses, fronts, and storms, it is obviously the forecaster's job to find out as much as he can about their positions and movements.

He can get this information only by collecting data on air pressure, temperature, humidity, wind direction, and rainfall from stations scattered over a large area, preferably over the entire country. The government weather service is set up to collect such data several times a day and to assemble the data on maps at a few central stations. After the raw

data are plotted, fronts can be drawn in along lines which seem to separate areas of markedly different temperatures and wind directions. Lines drawn through points of equal pressure (isobars) outline cyclones and anticyclones (designated "lows" and "highs" on weather maps).

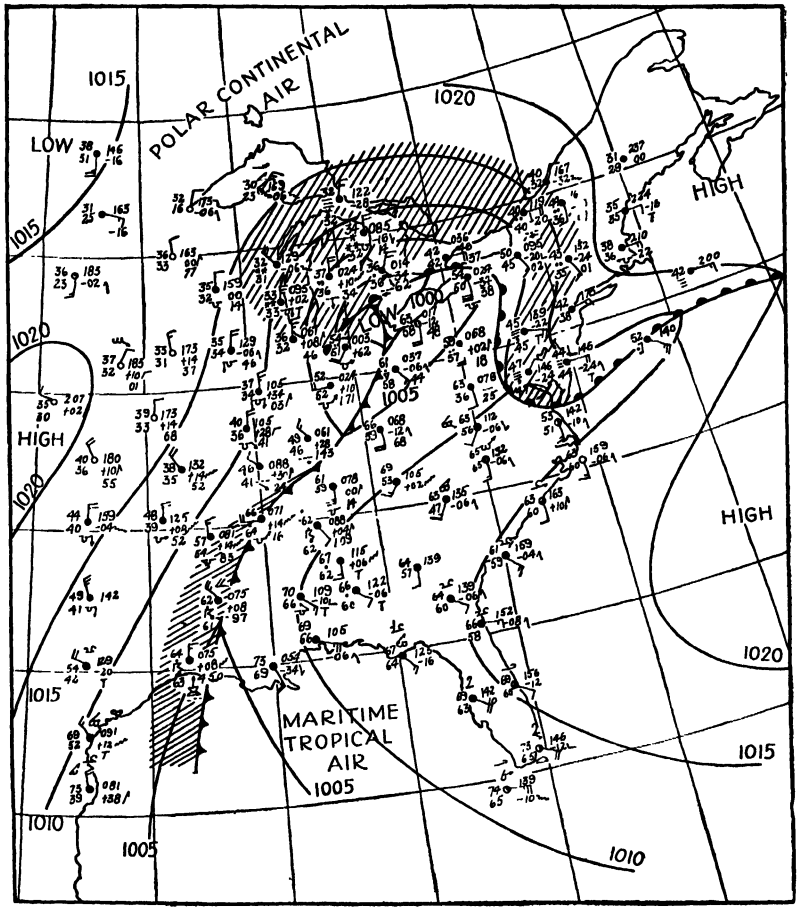


FIG. 205a. *Weather map of the eastern United States for April 18, 1940, at 1.30* A.M., *EST. A cold air mass on the west and north ("Polar Continental Air") is separated from a warm air mass ("Maritime Tropical Air") by a cold front extending from Louisiana to Michigan and by a warm front from Michigan to Virginia. Where the north end of the warm mass lies between the two fronts a cyclone has formed, bringing rain to the Great Lakes region. Compare Fig. 205b. (From Pettersen's Introduction to Meteorology.)*

When the fronts, cyclones, and anticyclones have been located, the map is compared with maps for preceding days. The comparison shows how the air masses and storms are moving and how conditions within them are changing. The forecaster can then project these movements

and changes into the future; knowing the weather types associated with the different fronts and storms, he can predict from their movements the probable weather for the next day or two at any point in the country. The method of forecasting is illustrated by Figs. 205*a* and 205*b*, which

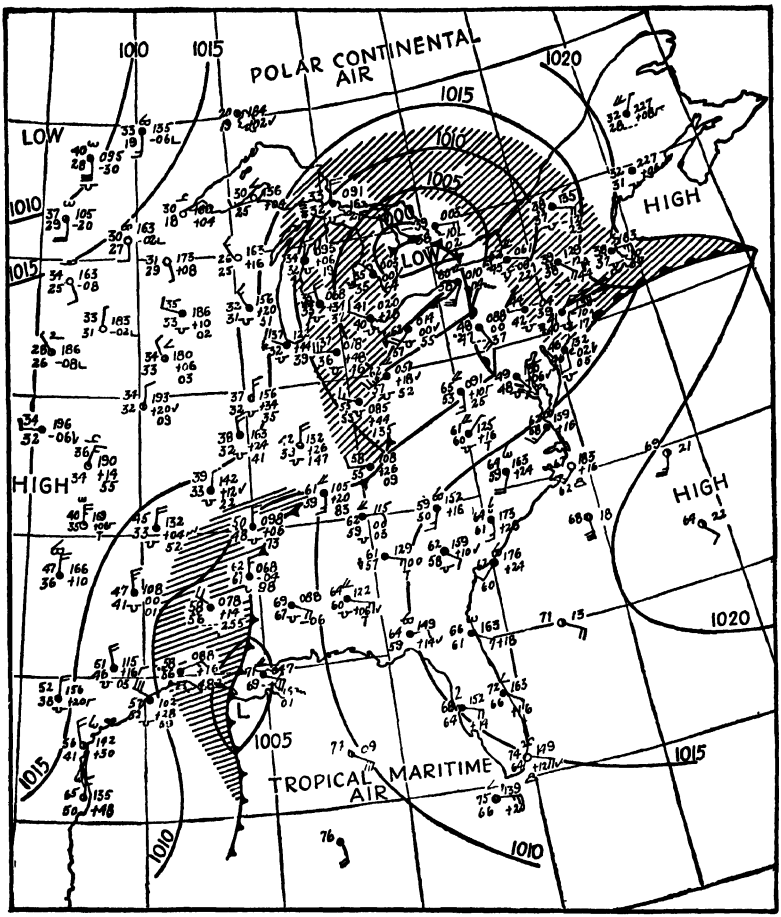


FIG. 205*b*. *Weather map six hours later than the map of F.g. 205a (April 18, 1940,* 7.30 A.M.). *Note that the cyclone has moved northward and eastward and that a new low pressure area is developing as a wave or "kink" in the cold front farther south. (Pressures on these maps are indicated in millibars, one millibar being a pressure of 1,000 dynes per sq cm, or about 0.75 mm of mercury; normal atmospheric pressure = 1,013 millibars.) (From Pettersen's Introduction to Meteorology.)*

show a cold, dry air mass to the west ("polar continental air") and a warm, humid one to the east ("tropical maritime air"), with a cyclone formed where a tongue of the warm mass projects into the cold one. Rain is falling in the cyclone and along the cold front farther south. In the 6-hr

period between the two maps the cyclone, the cold front, and the associated areas of rain have all moved eastward. It is reasonable to suppose that the same general movement will continue for the next 6 hr or longer, shifting the different types of weather to localities farther and farther east.

Cyclones and anticyclones have been recognized as important elements of weather for a long time, but the more fundamental role of air masses has been made clear only within the last twenty-five years, chiefly through the work of the Norwegian meteorologist Bjerknes. Air masses, fronts, cyclones, and anticyclones can be known, of course, only by putting together data on air temperatures, pressures, humidities, and wind directions. Low- and high-pressure areas can be recognized even from fairly rough data, but to outline air masses the data must be more precise. The detailed study of air masses and fronts has been made possible only by recent improvements in instruments for collecting the necessary data, especially by instruments carried in airplanes or in small balloons for automatically recording conditions above the earth's surface.

Despite the advances of recent years, weather forecasts are still all too often inaccurate. The inaccuracies can usually be blamed on insufficient data; if enough information is available, forecasts can be made highly reliable for periods of a few hours and fairly reliable for as long as two or three days. Weather predictions for longer periods are still impossible.

Ocean Currents

One further important influence on climates remains to be considered: currents in the oceans. We have already mentioned the effect of large water bodies in preventing abrupt daily and seasonal temperature changes over adjacent lands; now we turn to the large-scale movements of ocean water which maintain abnormally high temperatures in some parts of the earth, abnormally low temperatures in other parts.

One major type of ocean current, produced chiefly by strong heating of surface water near the equator, is a deep convectional circulation which brings about a slow interchange of water between the tropics and the polar regions. More important in influencing climate are surface drifts produced by the friction of wind on water. Both types of movement are much slower than movements in the atmosphere, surface currents attaining a maximum speed of about 7 mi/hr.

The wind-impelled surface currents parallel to a large extent the major wind systems. The northeast and southeast trade winds drive water before them westward along the equator, forming the "equatorial current." In the Atlantic Ocean this current runs head-on into South America, in the Pacific into the East Indies. At each of these points the

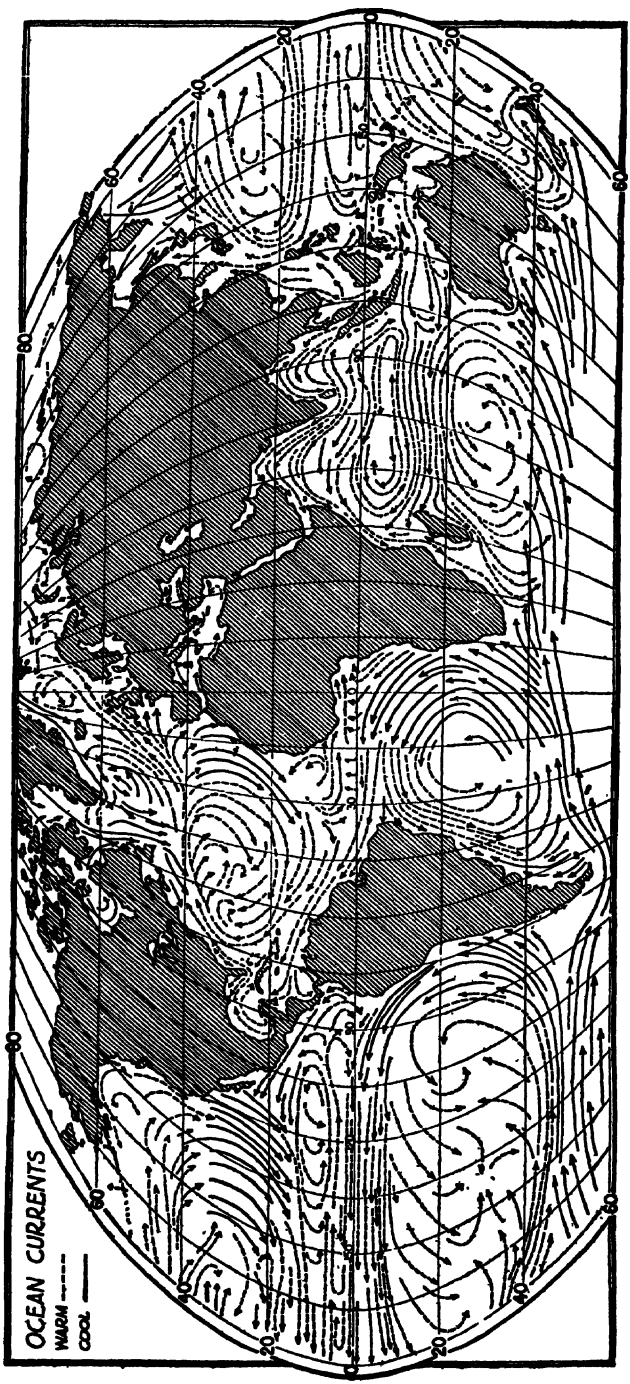


FIG. 206. (*From Elements of Geography by Finch and Trewartha.*)

current divides into two parts, one flowing south and the other north. Moving away from the equator along the continental margins, these currents at length come under the influence of the westerlies, which drive them eastward across the oceans. Thus in both Atlantic and Pacific Oceans are set up two gigantic whirlpools, one on either side of the equator (Fig. 206). Many minor complexities are produced in the four great whirls by islands and continental projections.

The western side of the north Atlantic whirl, a warm current moving partly into the Gulf of Mexico, partly straight north along our southeastern coast, is the familiar "Gulf Stream." Forced away from the coast in the latitude of New Jersey by the westerlies, this current moves northeastward across the Atlantic, splitting on the European side into one part which moves south to complete the whirl, another part which continues northeastward past Great Britain and Norway into the Arctic Ocean. To compensate for the addition of water into the polar sea, a cold current moves southward along the east coast of North America as far as New York; this is the "Labrador Current." Down the west coast of North America moves the "Japan Current," the southward-flowing eastern part of the north Pacific whirl.

Since ocean currents retain for a long time the temperatures of the latitudes from which they come, they exert a direct influence on the temperatures of neighboring lands. The influence is greatest, of course, where the prevailing winds blow shoreward from the sea. Thus the warm Gulf Stream has a much greater effect in tempering the climate of northwest Europe than that of eastern United States, since prevailing winds in these latitudes are from the west. Cyclonic storms bring east winds to the Atlantic seaboard frequently enough, however, so that the Gulf Stream helps to raise temperatures in the south Atlantic states, while the Labrador Current is in part responsible for the rigorous climate of New England and eastern Canada.

Thus the oceans, besides acting as the great water reservoirs of the earth's air-conditioning system, play a direct part in temperature control —both by preventing abrupt temperature changes in lands along their borders and by aiding the winds, through the motion of ocean currents, in their distribution of heat and cold over the surface of the earth.

Questions

1. Account for the abrupt changes in temperature between day and night in desert regions.
2. Would air containing water vapor at a pressure of 16 mm feel dry or humid at 20°C? At 40°C? What would happen if the air temperature dropped to −10°C?
3. Why does the skin feel cooler in a breeze than in quiet air at the same temperature?
4. A temperature of 90°F in Chicago is oppressively warm, while the same temperature in Arizona would be comfortable. Account for the difference.

5. The Japan Current along the California coast is cooler than the ocean to the west. How does this fact explain the numerous fogs on this coast?
6. In what regions of the earth do prevailing surface winds blow in a direction approximately opposite to that which they would have in a simple convectional circulation?
7. The island of Oahu (in the Hawaiian islands) is in latitude 21°N. and is crossed by a mountain range trending roughly northwest-southeast. Account for the fact that the northeastern side of the island has much more abundant rainfall than the southwestern side.
8. Describe the climate which you would expect to find (*a*) on the west side of a continent in latitude 25°S., (*b*) in the interior of a continent in latitude 55°N., (*c*) at an elevation of 17,000 feet in equatorial Africa, (*d*) on the west side of a continent in latitude 45°N., (*e*) on an oceanic island in latitude 30°N.
9. Why is the climate of Ireland so much milder than that of Labrador, although both are in approximately the same latitude?
10. Why are thunderstorms in middle latitudes more common in summer than in winter?
11. At a point 100 mi east of the center of a large cyclone in the Mississippi Valley, would you expect the temperature to be higher or lower than the temperature 100 mi west of the center? Would you expect the humidity to be greater or less?
12. Describe the probable changes in wind direction, temperature, and humidity when the center of a cyclonic storm passes (*a*) 100 mi south of you, (*b*) 100 mi north of you.

CHAPTER XXXIV

Rocks and Minerals

A STUDY of the earth requires first of all an acquaintance with the solid materials of the crust. We have gained considerable knowledge about the ultimate constituents of these materials: we know that compounds in the crust contain ninety-odd different elements and that the atoms of each element are made up of electrons, protons, and neutrons. We need now some information of a more practical sort, information about the actual substances into which the ultimate constituents are combined in nature. In this chapter we seek an acquaintance with the raw rock materials which we find beneath our feet, which we can pick up and examine in the field.

Minerals

Most rocks are heterogeneous solids. The different kinds of material in a coarse-grained rock like granite are apparent to the eye; in a fine-grained rock the separate constituents may be made visible with a microscope (Fig. 207). The separate, homogeneous substances of which rocks are composed are called ***minerals.***

The word "mineral" is used in a variety of ways. Obviously, a doctor who prescribes certain "minerals" for your diet, or an advertizer who tries to sell you water from a "mineral spring," is not thinking of the constituents of rocks. Even geologists do not always agree as to the exact application of the term. Most geological discussions restrict the word to solid, homogeneous, inorganic substances, found in nature either as parts of rocks or in separate deposits. As a rule, minerals are crystalline substances with fairly definite chemical compositions.

Previous discussions of the stabilities of chemical compounds make possible predictions as to the kinds of materials which can occur as minerals. We should expect to find the more inactive elements in the free state, active elements in compounds. Soluble compounds should be common minerals only in arid regions; easily oxidized compounds should

occur only well beneath the surface. So we find free gold, platinum, sulfur, carbon (both graphite and diamond) as minerals, while elements like sodium, chlorine, calcium always occur in compounds. Sodium chloride, sodium carbonate, potassium nitrate form deposits in deserts but are seldom found elsewhere. Such reactive compounds as calcium oxide or phosphorus pentoxide never occur in nature.

Silicates are by far the most abundant minerals; mica, feldspar, garnet, topaz are familiar examples. *Carbonates* are another important class, its most abundant representative being the carbonate of calcium (calcite).

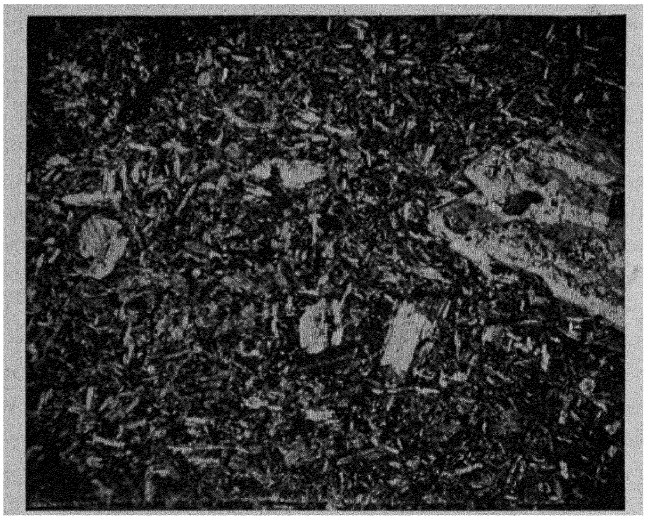


FIG. 207. *Photomicrograph of basalt. To the naked eye this is a dense, dark-gray rock with only a few tiny grains visible, but the microscope shows clearly its crystalline structure. The rock consists chiefly of the two minerals feldspar (the conspicuous, light-colored crystals) and pyroxene (small, dark crystals between the feldspars).*

Oxides and *hydrated oxides* include such common minerals as quartz (oxide of silicon); hematite (ferric oxide), the chief ore of iron; and bauxite (hydrated aluminum oxide), the chief ore of aluminum. Several important metals are obtained from deposits of *sulfide* minerals, such as galena (lead sulfide) and sphalerite (zinc sulfide). *Elements* which occur free, or *native*, were mentioned in the last paragraph. Less common as minerals are sulfates, phosphates, chlorides, etc.

Unfortunately the study of minerals requires the learning of a new list of formidable names, some of them apparently needless duplicates of other names. As an example, the common mineral with the formula $CaCO_3$ is given the name "calcite" instead of the chemical name "calcium carbonate." For this seeming duplication there are two good reasons: (1) The formula $CaCO_3$ expresses not only the composition of calcite, but also that of aragonite, a less common mineral with different crystal

form, hardness, density, etc.; between calcite and aragonite the chemical name alone could not distinguish. (2) Calcite often contains small quantities of $MgCO_3$ and $FeCO_3$, so that its composition is not precisely represented by the formula $CaCO_3$; the iron and magnesium carbonates form an integral part of the calcite crystals, Fe and Mg atoms simply replacing a few of the Ca atoms in the lattice structure. Many other mineral formulas besides that of calcite apply to two or more distinct substances, and most minerals show a similar slight variability in composition. Hence orthodox chemical names are not strictly applicable, and the student of minerals finds necessary a new nomenclature.

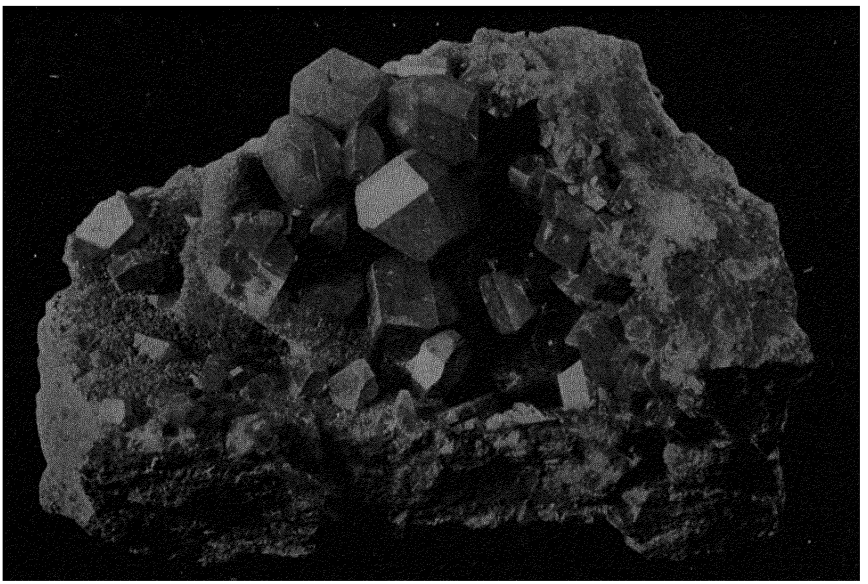


FIG. 208. *Crystals of sulfur.* (*Courtesy American Museum of Natural History.*)

Luckily, for present purposes we need only a few additions to our vocabulary. Some 1,500 different minerals are known, but most of these are rare. Even among the commoner minerals, the greater number occur abundantly only in occasional veins, pockets, and layers. The number of minerals which are important constituents of ordinary rocks is surprisingly small, so small that acquaintance with less than a dozen will be ample for our future geological discussions.

Nature is kind not only in limiting the number of common minerals, but in making them easily recognizable. To distinguish the rarer minerals a specialist must have recourse to elaborate microscopic and chemical tests, but for the minerals which compose ordinary rocks such simple

physical properties as density, color, hardness, and crystal form make identification possible.

In the following descriptions of the important rock-forming minerals, two terms need special attention—*crystal form* and *cleavage*. Most minerals are crystalline solids, which means that their tiny particles (atoms, ions, or atom groups) are arranged in lattice structures with definite geometric patterns. When a mineral grain develops in a position where its growth is not hindered by neighboring crystals, as in an open cavity, its inner structure expresses itself by the formation of perfect crystals, with smooth faces meeting each other at sharp angles. Each mineral has crystals of a distinctive shape (Fig. 208), so that well-formed crystals make recognition of a mineral easy; but unfortunately good crystals are rare, since mineral grains usually interfere with each other's growth. Even when well-developed crystals are not present, however, the characteristic lattice structure of a mineral may reveal itself in the property called *cleavage*. This is the ability of a substance to split along certain planes, determined by the arrangement of particles in its lattice. When a mineral grain is struck with a hammer, its cleavage planes are revealed as the preferred directions of breaking; even without actual breaking, the existence of cleavage in a mineral is usually shown by flat, parallel faces and minute parallel cracks (Figs. 209, 210). The flat surfaces of mica flakes, for instance, and the ability of mica

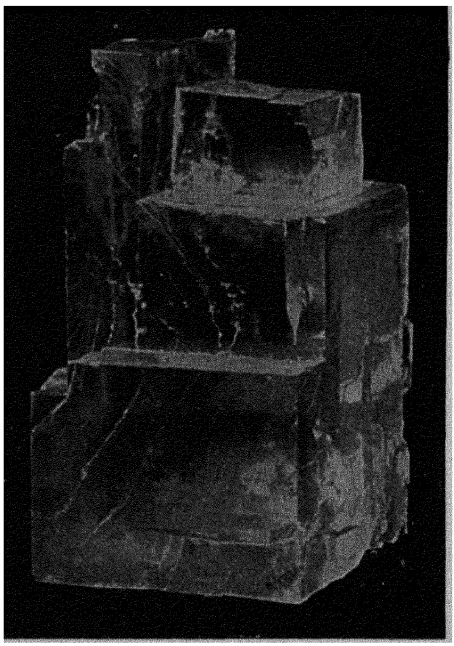


FIG. 209. *A crystal of rock salt showing perfect cleavage in three directions. (Courtesy of Ward's Natural Science Establishment.)*

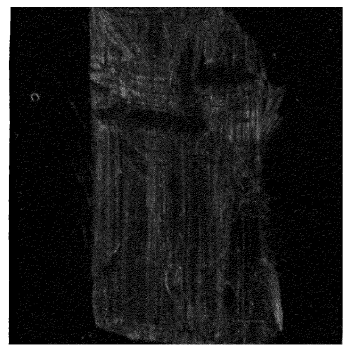


FIG. 210. *Asbestos, a mineral that breaks in minute fibers.*

to peel off in thin sheets, show that this mineral has very perfect cleavage. Some minerals (for instance, quartz) have practically no cleavage; when struck they shatter, like glass, along random, curving lines. The ability to recognize different kinds and degrees of cleavage is an important aid in distinguishing minerals.

Quartz. SiO_2. Crystals when well formed are six-sided prisms and pyramids. No cleavage. Colorless or milky, often gray, pink, or violet because of impurities. Glassy luster. Hard enough to scratch glass and feldspar. Common in many kinds of rocks; abundant in veins; often in aggregates of well-formed crystals on the sides of cavities. Clear quartz (rock crystal) is used in jewelry and in optical instruments (page 421); smoky quartz, rose quartz, and amethyst are colored varieties used in jewelry (Fig. 184, page 420).

Feldspar. This is the name of a group of silicate minerals with very similar properties. The two commonest varieties are a silicate of K and Al (orthoclase) and a silicate of Na, Ca, and Al (plagioclase). Crystals rectangular, with blunt-pointed ends. Good cleavage in two directions approximately at right angles. Color white or light shades of gray and pink, sometimes colorless. Slightly harder than glass, not as hard as quartz. Feldspar is the most abundant single constituent of rocks, making up about 60 per cent of the total weight of the earth's crust. Pure feldspar is used in the making of porcelain and as a mild abrasive.

Mica. The two chief varieties of this familiar mineral are white mica, a silicate of H, K, and Al, and black mica, a silicate of H, K, Al, Mg, and Fe. Easily recognized by its perfect and conspicuous cleavage in one plane. A very soft mineral, only a trifle harder than the fingernail. Large sheets of white mica free from impurities are used as insulators in electrical equipment.

Ferromagnesian Minerals. This name refers to a large group of minerals with diverse properties, all of them silicates of iron and magnesium, nearly all having a dark green to black color. Most of them contain other elements besides iron and magnesium, one of the commonest additional elements being calcium. Black mica belongs to this group, its composition including H, K, and Al in addition to Mg and Fe. No general properties of the group besides color and composition can be set down, since the various minerals differ greatly from one to another. For our purposes it will be sufficient to remember that the most abundant dark-colored constituents of common rocks belong to this group.

Clay Minerals. A group of closely related minerals that are the chief constituents of clay; silicates of H and Al, some with a little Mg, Fe, and K. Aggregates of microscopic crystals, white or light colored when pure, often discolored with iron compounds. Dull luster. Very soft, forming a

smooth powder when rubbed between the fingers. Density low. Distinguished from chalk by softness and lack of effervescence in acids. Kaolin, one of the clay minerals, is the principal ingredient used in the manufacture of pottery and porcelain (page 427).

Calcite. $CaCO_3$. Crystals hexagonal, in general appearance resembling those of quartz. Cleavage perfect in three directions at angles of about 75°, so that fragments of calcite have a characteristic rhombic shape. Colorless or any light shade. Glassy luster. Hard enough to scratch mica or the fingernail, but can be scratched by glass or by a knife blade. Dissolves readily in dilute acid with effervescence (page 362). Like quartz, calcite is a common mineral of veins and crystal aggregates in cavities. It is the chief constituent of the common rocks limestone and marble. It is important commercially for many purposes, especially as a source of lime for glass, mortar, and cement.

Of these six kinds of rock-forming minerals, note that five are compounds of silicon, and that four are silicates. The light-colored silicates all contain aluminum in addition to silicon and oxygen; two of them (feldspar and mica) contain an alkali metal, two of them (mica and clay) contain hydrogen. The dark-colored silicates contain iron and magnesium.

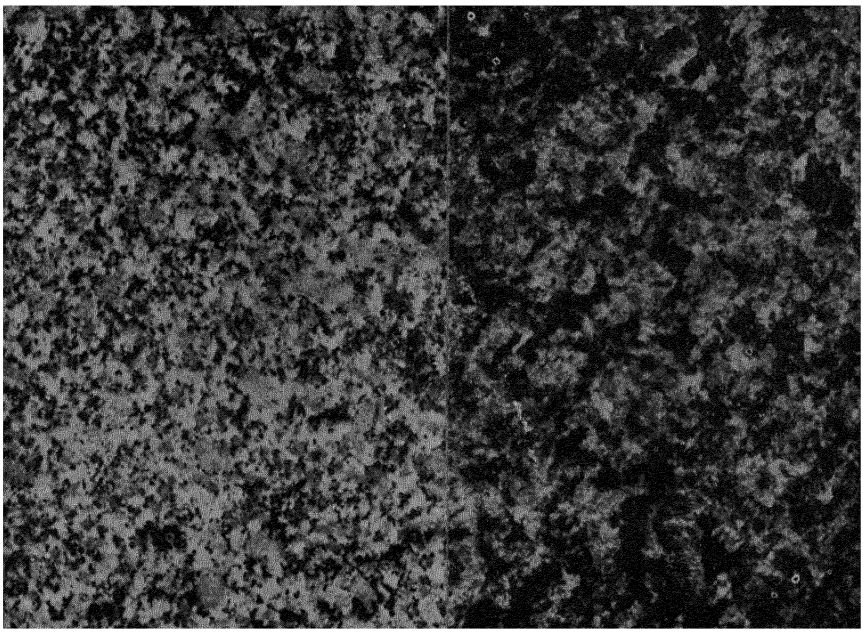
The Classification of Rocks

The study of rocks, like any other branch of science, requires first of all that its materials be classified. We can find coarse-grained rocks and fine-grained rocks, light rocks and heavy rocks, soft rocks and hard rocks, rocks of all different colors; but the origin and behavior of each one remains a problem in itself unless we can find a way to divide rocks into groups with similar characteristics.

The problem of organizing and classifying rocks is more difficult than similar problems in physics and chemistry. When Mendelyeev sought to classify the elements, he had sixty-odd definite substances to work with; each was a distinct material, its properties sharply different from those of the others. In the study of rocks we find no such sharp boundaries between different kinds. We may decide that a light-colored, coarse-grained rock like granite should belong in a different class from a dark, fine-grained volcanic rock; but we can find a whole series of rocks with properties transitional between the two, so that we cannot say with any certainty just where one class ends and the other begins. As in all branches of science dealing with materials that occur in nature, we face the problem of making divisions where natural divisions do not exist. We must fit rocks into pigeonholes as best we can, but we must expect to run across some types with properties transitional between those of different classes. We should encounter a similar problem in trying to classify human beings:

we might begin by sorting them out into farmers, laborers, businessmen, professional men, artists, but we should find many individuals whose occupations would seemingly put them in two or more different groups.

Recognizing the difficulty of the task, how shall we proceed to set up a usable classification of rocks? Since rocks are composed of minerals, we might guess first that they could be classified on the basis of the kinds and amounts of minerals they contain. But we should soon find that rocks of widely different structures and origins have nearly the same mineral composition, so that our classification would group together rocks of



a *b*

FIG. 211. *Coarse-grained igneous rocks. a, granite; b, gabbro. Natural size. (From Grout's Petrography and Petrology.)*

obviously different types. A classification based on chemical composition encounters the same difficulty, since it places in the same pigeonhole rocks which are patently unlike; it has the further disadvantage that chemical compositions are not evident from examination of rocks in the field but require costly analyses. We might disregard composition and try to classify rocks according to their origin. This would be an excellent method if it could be applied to all rocks, but the sad fact is that we simply do not know how some rocks were formed; the origin of many others can be deduced only by detailed study in the laboratory.

From this discussion emerges one of the big obstacles to a satisfactory classification: we expect the classification to fulfill several different pur-

poses. We should like it to summarize something about the origin of different rocks, something about their compositions and structures, and at the same time we should like to be able to apply it to rocks as we find them in the field. The briefest acquaintance with rocks shows immediately that all of these objects cannot be satisfied at the same time. Our only recourse is to adopt some sort of a compromise, a classification which will accomplish each purpose as well as possible without slighting the others. The particular compromise we shall use is not the only possible one, but it is justified by its simplicity and convenience.

FIG. 212. *Conglomerate, a sedimentary rock containing pebbles and boulders of other rocks. About one-half natural size. (From Grout's Petrography and Petrology.)*

A fundamental division of rocks into three main groups according to differences of origin is agreed on by nearly all geologists. ***Igneous rocks*** are those which have cooled from a molten state. Some of these can be observed in process of formation, when molten lava cools on the side of a volcano. For others an igneous origin is inferred from their composition and structure (Fig. 211). ***Sedimentary rocks*** consist of materials derived from other rocks, deposited by water, wind, or glacial ice. Some consist of separate rock fragments cemented together (Fig. 212), while others contain material precipitated from solution in water. ***Metamorphic rocks*** are rocks which have been changed, or metamorphosed, by heat and pressure deep under the earth's surface. The changes produced may

involve the formation of new minerals, or simply the recrystallization of minerals already present.

In the following sections we shall describe the important rock types in each major group. Emphasis here will be on the characteristics by which different rocks may be recognized; in future chapters we shall deal at greater length with processes of rock formation.

Igneous Rocks

The structure of igneous rocks is characterized by random arrangement of grains, by ragged crystal borders, by intertwinings and embay-

FIG. 213. *Photomicrograph of granite. The dark grains clustered in the center of the picture are black mica, and the light-colored ones are quartz and feldspar. Note that most of the crystals have irregular shapes and interlocking borders.* × *45.* (*From Grout's Petrography and Petrology.*)

ments such as one might expect in a mass of crystals growing together and interfering with each other's development. In coarse-grained rocks like granite, this structure is visible to the naked eye; in fine-grained rocks it is revealed by the microscope (Figs. 207, 211, 213). The principal constituents of these rocks are always minerals containing silicon—quartz, feldspar, mica, and the ferromagnesian group.

The siliceous liquids from which igneous rocks form are thick, viscous materials resembling melted glass both in properties and in composition. Sometimes, in fact, molten lava has the right composition, and cools rapidly enough, to form a natural glass—the black, shiny rock called *obsidian*. Usually, however, cooling is slow enough to allow crystalline minerals to form. If cooling is fairly rapid and if the molten material is

highly viscous, the resulting rock may consist of minute crystals, or partly of crystals and partly of glass. If cooling is extremely slow, mineral grains have an opportunity to grow large and a coarse-grained rock is formed. The grain size of an igneous rock, therefore, reveals something about its history and gives us one logical basis for classification.

Mineral composition provides a convenient means for further classification. Nearly all igneous rocks contain feldspar and one or more of the ferromagnesian minerals; many contain quartz as well. A division of igneous rocks based on relative amounts of these three mineral types is shown in Table XXIII. Thus a coarse-grained rock containing quartz,

TABLE XXIII. IGNEOUS ROCKS

(Abbreviations: qz, quartz; fs, feldspar; fmg, ferromagnesian minerals)

	Mineral composition		
	qz, fs, fmg	*fs, fmg; no qz; fs predominant*	*fs, fmg; no qz; fmg predominant*
Coarse-grained rocks	granite	diorite	gabbro
Fine-grained rocks	rhyolite	andesite	basalt

feldspar, and black mica is granite; a fine-grained rock with no quartz and with feldspar in excess of the dark constituents is andesite, and so on. Not all igneous rocks by any means are shown in the table, but these six are the most important.

This classification is convenient for several reasons. (1) Grain size and usually mineral composition can be determined from inspection in the field. Except for a few fine-grained types, a rock can be named without detailed laboratory study. (2) Even if a rock is too fine for its mineral content to be easily determined, its color often shows its place in the table. Granite and rhyolite, containing only a little ferromagnesian material, are nearly always light colored; gabbro and basalt, with abundant ferromagnesian minerals, are characteristically dark; diorite and andesite usually have intermediate shades. (3) Grain size gives an indication, not only of the rate of cooling, but usually of the environment in which a rock cooled. Sufficiently rapid cooling to give fine-grained rocks occurs most commonly when molten lava reaches the earth's surface from a volcano, and spreads out in a thin flow exposed to the atmosphere. Since a fine grain size usually betrays a volcanic origin, rhyolite, andesite, and basalt are often called *volcanic* or *extrusive* rocks. Coarse-grained rocks, on the other hand, have cooled more slowly, well beneath the earth's surface, and are now exposed to view only because erosion has carried away the material which once covered them. Since these rocks do

not reach the surface as liquids but are "intruded" into spaces occupied by other rocks, they are often called *intrusive* rocks.

(4) The change in mineral composition from left to right across the table roughly parallels a steady change in chemical composition. The chemistry of rocks is most simply considered in terms of the oxides which make up their minerals (page 422), and rock analyses are usually given as percentages of these oxides. Thus granite and rhyolite, rocks which contain abundant quartz and little ferromagnesian material, are spoken of as "rocks with a high content of silica," the silica referring not only to SiO_2 which is free as quartz, but also to the combined SiO_2 in the other minerals. Gabbro and basalt, with no quartz and abundant ferromagnesian constituents, show analyses low in silica and high in the oxides FeO and MgO. The decrease in SiO_2 and the increase in metallic oxides across the table are shown by the average analyses of granites, andesites, and basalts in Table XXIV. Since SiO_2 is the oxide of a nonmetal, it is considered an "acidic" oxide (page 360), and rocks like granite and rhyolite, whose analyses show much of this oxide, are commonly called *acidic* rocks. A better term, also commonly used, is *siliceous* rocks—better since "acidic" has so different a meaning in chemistry. Gabbro and basalt, which contain an abundance of the basic oxides FeO, MgO, CaO, are often called *basic* rocks.

TABLE XXIV. CHEMICAL ANALYSES OF IGNEOUS ROCKS

	Granite	*Andesite*	*Basalt*
SiO_2	70.18	59.59	49.06
Al_2O_3	14.47	17.31	15.70
Fe_2O_3	1.57	3.33	5.38
FeO	1.78	3.13	6.37
MgO	0.88	2.75	6.17
CaO	1.99	5.80	8.95
Na_2O	3.48	3.58	3.11
K_2O	4.11	2.04	1.52
Others	1.54	2.47	3.74

From several angles, then, this rather sketchy classification is satisfactory. It groups rocks at least roughly according to their origins; it summarizes information about their mineral and chemical compositions; and finally it makes possible their identification from such obvious characteristics as grain size, color, and mineral content.

Sedimentary Rocks

Sediments laid down by water, wind, or ice are consolidated into rock by the weight of overlying deposits and by the gradual cementing

of their grains with material deposited from underground water. As a class, the resulting rocks are characterized by the usual presence of distinct, somewhat rounded grains, not intergrown as are the crystals of igneous rocks. A few sedimentary rocks, however, do consist entirely of intergrowing mineral grains, formed by precipitation from solution in water. Since sediments are normally deposited in layers, the majority of sedimentary rocks have a banded appearance owing to slight differences in color or grain size from one layer to the next. Sedimentary rocks may often be recognized at a glance by the presence of *fossils*—remains of plants or animals interred with the sediments as they were laid down.

Sedimentary rocks may be divided into two groups according to the nature of their original sediments: (1) *fragmental rocks*, made up of the fragments and decomposition products of other rocks; (2) *precipitates*, formed from material once dissolved in water and deposited either as a chemical precipitate or as the shells and bone fragments of dead organisms. The more abundant rock varieties in each group are listed in Table XXV.

TABLE XXV. SEDIMENTARY ROCKS

(Chief constituents in parenthesis)

Fragmental Rocks	*Chemical and Biochemical Precipitates*
Conglomerate (rock fragments)	Limestone (calcite)
Sandstone (quartz usually most abundant)	Chert (chalcedony)
Shale (clay minerals)	

The three fragmental rocks are distinguished by their grain size. *Conglomerate* is cemented gravel (Fig. 212); its fragments may have any composition, and any size from that of small pebbles to large boulders. With decreasing size of fragment, conglomerate grades into *sandstone* (Fig. 214). Sand grains may consist of many different minerals, but quartz is generally the most abundant. The hardness of sandstone and conglomerate depends in large measure on how well their grains are cemented together; some varieties crumble easily, while others, especially those with silica as the cementing material, are among the toughest of rocks. *Shale* is consolidated mud or silt, a soft rock usually in thin layers. Its chief constituents are usually one or more of the clay minerals.

Limestone, a fine-grained rock composed chiefly of calcite, may be formed either as a chemical precipitate or by the consolidation of shell fragments. Like calcite in larger crystals, limestone is only moderately hard and effervesces readily in acid. Small amounts of impurities may give the rock almost any color. *Chalk* is a loosely consolidated variety of limestone, often made up largely of the shells of tiny one-celled animals.

Most Indian arrowheads are made either of the igneous rock obsidian or the sedimentary rock *chert*, both much prized among primitive peoples

for their hardness and the sharpness of their edges when broken. The chief mineral constituent of chert is a microcrystalline variety of quartz called *chalcedony*, probably formed by the gradual hardening of a silicic acid gel (page 437). Two familiar varieties of the rock are *flint* and *jasper*. Fragments of chert show the same sharp edges and smooth, concave surfaces as broken quartz or obsidian, but the surfaces have a characteristic duller luster resembling that of wax. Impurities may give the rock almost any color; often a single specimen shows bands and pockets of several different colors. Not nearly as abundant as the other sedimentary

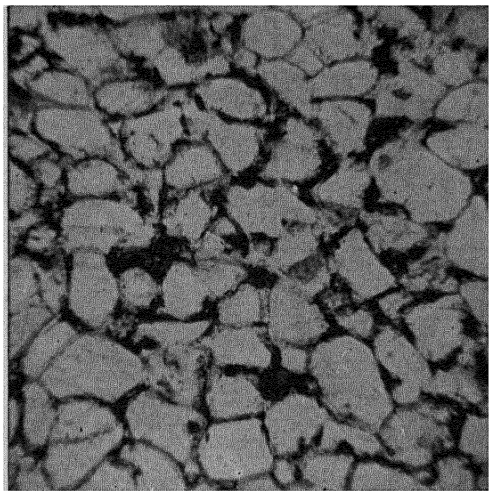


FIG. 214. *Photomicrograph of sandstone, showing angular sand grains cemented by iron oxide.* × *50.* (*From Grout's Petrography and Petrology.*)

rocks just described, chert is nevertheless a common rock in pebble beds and gravel deposits, because its great hardness and resistance to chemical decay enable it to survive much rough treatment by streams, waves, and glaciers.

Metamorphic Rocks

The terrific pressures and high temperatures a few miles below the earth's surface effect profound changes in sedimentary and igneous rocks which become deeply buried. Minerals stable at the surface are often not stable under the new conditions and may react to form different substances. Other minerals remain stable, but their crystals increase in size. Hot liquids permeating the rocks may add some new materials and dissolve out others. So many kinds of change are possible that no satisfactory general rules can be set down for readily distinguishing metamorphic rocks from others.

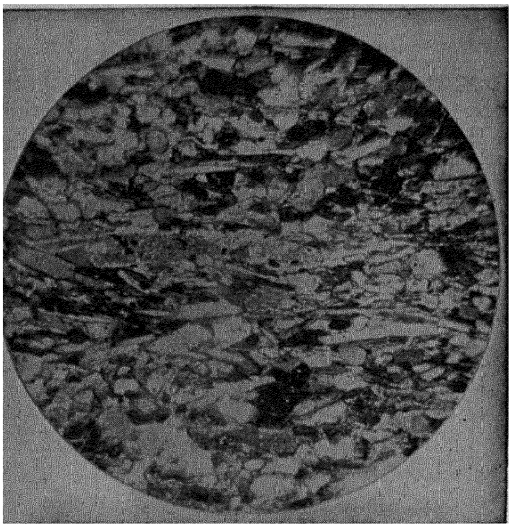


FIG. 215. *Photomicrograph of schist, a foliated metamorphic rock. Note the alignment of needlelike crystals.* × *45.* ***(From Grout's Petrography and Petrology.)***

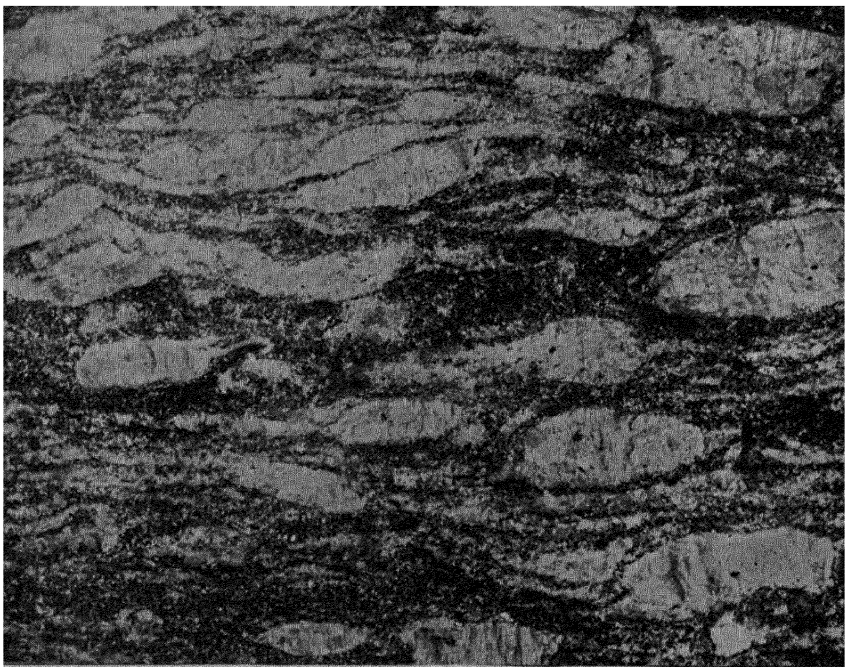


FIG. 216. ***Gneiss, a foliated metamorphic rock. The large white crystals are feldspars; in the bands between are quartz, feldspar, and black mica, the latter in flakes parallel to the bands. About one-half natural size. (From Grout's Petrography and Petrology.)***

Many metamorphic rocks are characterized by a property called *foliation*, which means the arrangement of flat or elongated mineral grains in parallel layers. This arrangement is caused by extreme pressure in one direction, the mineral grains growing out sideward as the rock is squeezed. Foliated rocks always contain a mineral (like mica) which occurs in thin flakes, or a mineral (like some of the ferromagnesian group) which occurs in long needles (Fig. 215). Rocks consisting only of minerals like quartz, feldspar, or calcite cannot show foliation, since these minerals have little tendency to grow longer in one direction than another even under pressure. Foliation gives a rock a banded or layered appearance (Fig. 216), and when broken the rock tends to split along the bands. Layering is also characteristic of sedimentary rocks, but in them the layering is caused by slight variations in color or grain size; layering in metamorphic rocks is due to the lining up of mineral grains.

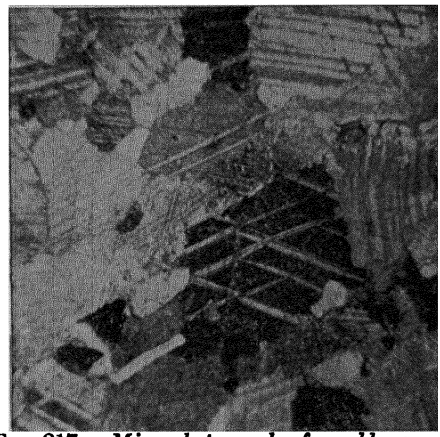


FIG. 217. *Microphotograph of marble, an unfoliated metamorphic rock. The calcite grains have irregular shapes and interlocking edges. (From Grout's Petrography and Petrology.)*

An unfoliated metamorphic rock is produced when the chemical composition prevents the formation of minerals with flat or elongated crystals, or when a rock is metamorphosed simply by heat without excess pressure in one direction. The random growth of individual mineral grains during metamorphism produces in such a rock a structure much like that typical of igneous rocks, each grain embaying and interlocking with its neighbors (Fig. 217). A metamorphic origin for an unfoliated rock can often be recognized only from the fact that its mineral composition is unlike that of typical igneous rocks, or from the presence of other metamorphic rocks adjacent to it in the field.

A thoroughgoing classification of metamorphic rocks would involve a more detailed discussion of metamorphic processes, which we had best

TABLE XXVI. METAMORPHIC ROCKS

(Chief constituents in parenthesis)

Foliated Rocks	*Unfoliated Rocks*
Slate (mica and usually quartz, both in microscopic grains)	Marble (chiefly calcite)
Schist (mica and/or a ferromagnesian mineral; usually quartz)	Quartzite (chiefly quartz)
Gneiss (quartz, feldspar, mica)	

keep for a later chapter. Here we shall simply group the commoner varieties according to the presence or absence of foliation (Table XXVI).

Slate is produced by the low-temperature metamorphism of shale, the unstable clay minerals forming tiny flakes of mica. Although the individual flakes are too small to be seen, parallel sheets of mica crystals are responsible for the shiny surfaces produced whenever slate is split along its foliation. The rock is harder than shale, finely foliated, usually black or dark gray but sometimes lighter colored. *Schist* is produced from shale by higher temperatures, or from fine-grained igneous rocks. In it the mineral grains responsible for the foliation are large enough to be visible, giving the foliation surfaces a characteristic spangled appearance. Schist does not split as easily with the foliation as slate, and its surfaces are rougher. *Gneiss* is a coarse-grained rock, produced under conditions of high temperature and pressure from almost any other rock except pure limestone and pure quartz sandstone. Its composition naturally depends on the nature of the original rock, but quartz, feldspar, and mica are the commonest minerals. In appearance gneiss resembles granite, except for its banding.

The metamorphism of pure limestone and pure quartz sandstone is a relatively simple process. Since each consists of a single mineral of simple composition, heat and pressure can produce no new substances, but simply cause the growth and interlocking of new crystals of calcite and quartz. Thus limestone becomes *marble*, a rock composed of calcite in crystals large enough to be easily visible, whereas sandstone becomes the hard rock *quartzite*. The recognition of marble is a simple matter of applying the usual tests for calcite. Quartzite may resemble sandstone in appearance, but its grains are so firmly intergrown that it splits across separate grains when broken, giving smooth fracture surfaces in contrast to the rough surfaces of sandstone.

In passing, we should remind ourselves once more that this classification of rocks is an artificial one, that in nature we can find no sharp boundaries between the different rock types. A medium-grained igneous rock with no quartz and an abundance of feldspar would be hard to classify either as a diorite or an andesite; some limestones contain so much clay that they may be called either shaly limestones or limy shales; the exact point in the metamorphic process when a fine-grained rock ceases to be a shale and becomes a slate is an excellent topic for academicians to quarrel over. In spite of these obvious defects, the classification is valuable because it *organizes* our knowledge about various rocks: henceforth when we label a rock "schist," we know at once something about its origin and its general properties. The classification gives us a

summary of elementary facts on which we can build the generalizations of geology.

Questions

1. Which of the following naturally occurring substances are minerals: Diamond, diorite, petroleum, ice, soil, wood, salt, coal?
2. Which of the following substances would you expect to occur as minerals: Ag, NaOH, Ca, KNO_3, Fe_2O_3 ZnS, $BaCO_3$, P, Fe?
3. Name (*a*) three minerals which contain carbon, (*b*) three minerals harder than glass, (*c*) three minerals containing both aluminum and silicon, (*d*) one mineral containing sodium and calcium, (*e*) a mineral which dissolves readily in dilute acid, (*f*) two minerals which can be scratched with a knife, (*g*) a mineral with cleavage in two directions.
4. By what characteristics can you distinguish
   a. Granite from gabbro.
   b. Basalt from limestone.
   c. Schist from diorite.
   d. Chert from obsidian.
   e. Conglomerate from gneiss.
   f. Quartz from calcite.
5. Name the following rocks:
   a. A fine-grained, unfoliated rock with intergrowing crystals of quartz, feldspar, and black mica.
   b. A finely foliated rock with microscopic crystals of quartz and white mica.
   c. A fine-grained rock consisting principally of kaolin.
   d. A rock consisting of intergrown crystals of quartz.
   e. The rock resulting from metamorphism of limestone.
   f. An intrusive igneous rock with the same composition as andesite.
6. Which of the following rocks would you expect to be chemically attacked by oxygen and carbonic acid when exposed to the atmosphere? Limestone, basalt, quartzite, granite, quartz sandstone, chert.
7. Arrange the following rocks in three general groups as (*a*) hard, (*b*) moderately hard, (*c*) soft. Indicate which are igneous, which sedimentary, and which metamorphic: gneiss, limestone, quartzite, obsidian, shale, chert, andesite, chalk, marble.

CHAPTER XXXV

Erosion and Sedimentation

STAND on the ocean shore, where granite rises in bold cliffs from the water's edge, and watch the waves at work. Each wave rolls in, breaks noisily against the rock, retreats—and the rock stands apparently unharmed, unchanged in the smallest detail. Years of ceaseless battering by the sea have left the cliffs much as they were in your grandfather's day, and they will remain when your grandchildren have grown old. Or stand at the brink of Niagara, and watch the swift torrent plunge over the precipice. For all the rush of water across its top, all the pounding of spray at its foot, the cliff of solid rock stands firm. Return tomorrow, next month, next year, and you will see no sign of change in the shape of the falls. Wherever hard rock is exposed, preyed upon by waves or rivers, unprotected from the heat of summer or the cold of winter, lashed by wind-driven rain and snow, we find it enduring unharmed year after year, generation after generation. Small wonder that rock has come to be a symbol of strength and permanence in a world of change.

Yet can we be so certain that rocks do not change—slowly perhaps, but none the less steadily? Watch the granite shore again in time of storm, when huge waves lift pebbles and boulders from the beach to grind against the cliff: surely such powerful rasping must dig scratches and grooves into the rock. Note the widened cracks in the granite along the shore, the caves and arches, the rounded granite boulders on the beach: how else could these form than by the slow wearing of waves and wave-driven pebbles? At Niagara, records show that the rocky cliff is yielding slowly to the river's constant gnawing; in places the brink of the falls has retreated as much as 200 ft in fifty years. Even where waves and rivers are not active, careful examination of rock surfaces shows evidence of their slow disintegration and wearing away. Rocks, hills, and mountains are permanent only by comparison with the rapid changes of human life. In a decade, in a century, the changes are seldom great enough to be noticed; but a man who could prolong his existence a thousand years would find many alterations in familiar landscapes.

The long history of the earth extends back not thousands but millions of years. In these immense stretches of time the slow processes by which solid rock is decayed and worn down have wrought great changes in the earth's appearance. Shore lines have been pushed back, waterfalls have disappeared, even mountain ranges have been leveled and broad seas filled with their debris. We can read such events of the past in the rocks of the present day; and we find at least a partial explanation for them in the processes by which these same rocks are now being destroyed.

As we study the various agents at work disintegrating and grinding down the solid materials of the earth's surface, we must keep in mind the long ages through which they have operated. These processes are so very gradual, so commonplace, so unspectacular in action, that they may seem scarcely worth a detailed examination. But we are entering a science now where horizons of time are immensely broadened, as horizons of distance are broadened in astronomy. Against a background of a thousand million years, the importance of slow processes becomes enormously magnified.

Weathering

The obliteration of ancient inscriptions and the rough, pitted surfaces of old stone buildings are good evidence for the decay of rocks long exposed to the atmosphere. This sort of disintegration, brought about simply by rain water and the gases in air, is called ***weathering.*** Weathering is in part a chemical process, in part a mechanical process.

Some of the minerals in igneous and metamorphic rocks are especially susceptible to *chemical weathering*, since they were formed under conditions very different from those at the earth's surface. Ferromagnesian minerals are readily attacked by atmospheric oxygen, aided by carbonic acid (formed by solution of carbon dioxide in water) and by organic acids from decaying vegetation. Their ferrous iron is oxidized to a hydrate of ferric oxide, whose red and brown colors commonly appear as stains on the surface of rocks containing these minerals (page 394); their calcium and magnesium are in large part dissolved as the ions Ca^{++} and Mg^{++}; their silicon goes into solution as colloidal silica. Feldspar usually does not weather as rapidly as the ferromagnesian minerals, but it also gradually succumbs to carbonic acid. The chief product formed by chemical decay of feldspar is one or more of the clay minerals; the alkali metals and the calcium of feldspar are dissolved as the ions Na^{+}, K^{+}, and Ca^{++}.

Among common sedimentary rocks limestone is most readily attacked by chemical weathering because of the solubility of calcite in carbonic acid (page 363). Exposures of this rock can often be identified simply from the pitted surfaces and enlarged cracks which solution produces.

Quartz and white mica are extremely resistant to chemical attack, hence commonly remain as loose grains when the rest of a rock is thor-

FIG. 218. *The weathering of granite. Chemical decay causes an increase in volume of the outer part of the rock, hence a splitting off of successive layers. (Wisconsin Geological Survey photograph, reprinted by permission from Finch and Trewartha's Elements of Geography.)*

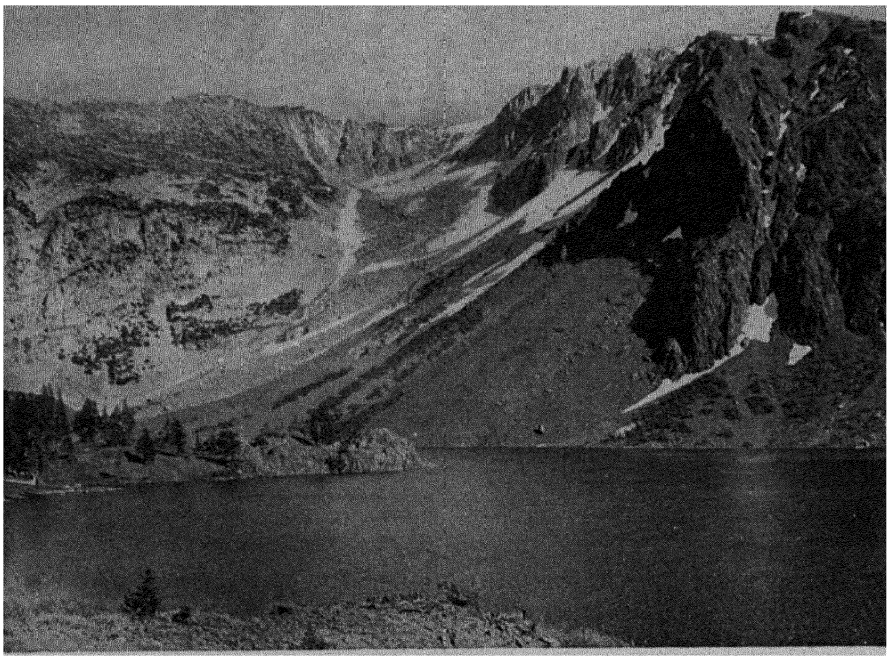


FIG. 219. *Steep slopes of loose rock (talus), consisting largely of fragments wedged loose from the cliff above by the expansion of water freezing in cracks. Sierra Nevada, California. (Photograph by Eliot Blackwelder.)*

oughly decayed. Rocks consisting wholly of silica, like chert and most quartzites, are practically immune to chemical weathering.

Mechanical weathering is often aided by chemical decay: not only is the structure of a rock weakened by the decomposition of its minerals, but fragments are actively wedged apart because the decay of a mineral grain usually results in an increased volume (Fig. 218). The most effective process of mechanical disintegration which does not require chemical action is the freezing of water in crevices. Just as water freezing in an automobile radiator on a cold morning may burst the radiator, so water freezing in tiny cracks is an effective wedge for disrupting rocks

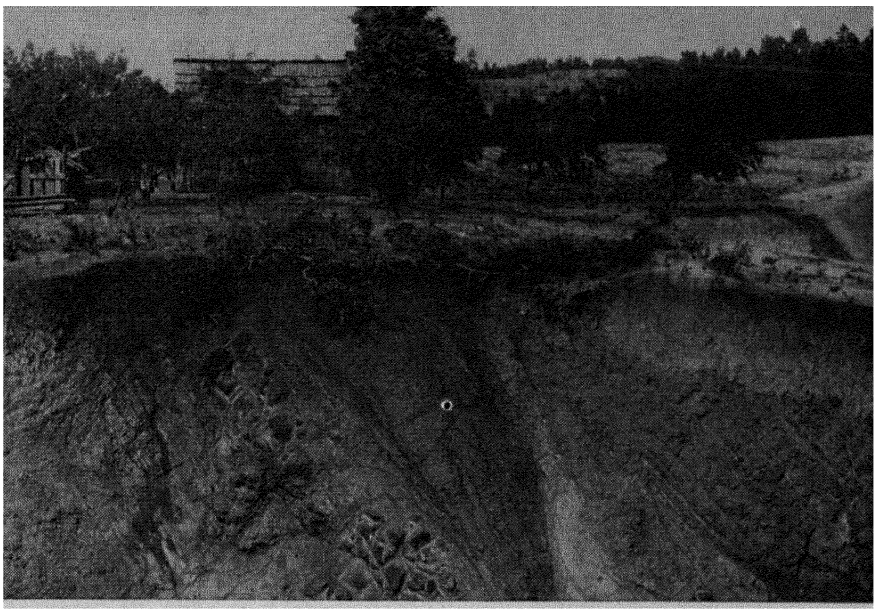


FIG. 220. *Bedrock grading upward into mantle and soil. Weathering has penetrated more deeply into some parts of the bedrock than others. In the middle of the picture weathering along cracks has left isolated boulder-like remnants of unweathered rock. (Photograph by Russell, U. S. Geological Survey.)*

(Fig. 219). Plant roots aid in rock disintegration by growing and enlarging themselves in cracks. Expansion and contraction of a rock's surface caused by temperature changes from day to night may gradually weaken it and aid in its destruction, but the importance of this effect is dubious.

How fast these processes of chemical and mechanical weathering will decompose a rock depends on (1) its composition, (2) how badly it is cracked, (3) its situation, and (4) the climate. Since the agents of weathering can penetrate deeply into a rock along crevices, a badly cracked rock weathers far more rapidly than one whose weathering is confined to the surface. If a rock is situated so that it frequently becomes moist, its decay is hastened; if it is protected from the agents of weathering, particularly from moisture, its disintegration may be very slow.

The influence of climate is strikingly illustrated by the granite obelisk called Cleopatra's Needle, brought from Egypt to New York about 80 years ago. Inscriptions on its sides, fresh after 2,000 years of Egypt's dry climate, have already become partly undecipherable after a few decades in the moist, cold climate of Eastern United States. Limestone shows even sharper differences in rates of weathering between wet and dry climates; a strong and resistant rock in the arid Southwestern part of this country, limestone is easily weathered in the humid Eastern part. In general, chemical weathering is favored by warmth and moisture, while mechanical weathering is most rapid where freezing and thawing are frequent.

Weathering processes clothe the naked rock of the lithosphere with a *mantle* of debris, made up largely of clay mixed with rock and mineral fragments (Fig. 220). The upper part of the mantle, in which rock debris is mixed with decaying vegetable matter, is the *soil*. From a human point of view the formation of soil is by all odds the most important result of weathering.

Stream Erosion

All the processes by which rock material is disintegrated and worn away, and by which its debris is removed, are included in the general term ***erosion.*** Weathering is a part of erosion, preparing rock material for easy removal by the more active erosional agents. Among the active agents, those whose work is more obvious are *streams*, *glaciers*, *wind*, and *waves*. Less apparent is the erosional work of *groundwater*, water in crevices and channels beneath the surface. All these agents are capable of cutting slowly into solid, unweathered rock, but their work is greatly speeded by the disintegration of rocks into the softer material of the mantle.

By far the most important agent of erosion is the running water of streams. The work of glaciers, wind, and waves is impressive locally, but by comparison with running water they play only minor roles in the shaping of the earth's landscapes. Even in deserts, mountain sides are carved with the unmistakable forms of stream-made valleys (Fig. 221).

A stream performs two functions in erosion: the active cutting at the sides and bottom of its channel, and the transportation of debris supplied by weathering and by its own cutting. Its effectiveness in carrying debris depends on its *gradient* or slope and on its volume of water. Its effectiveness in cutting its channel depends on these two factors and also on the amount and kind of debris with which it is supplied. Sand grains, pebbles, and boulders are the tools which a stream uses to dig into its bed. Scraping them along its bottom, ramming them against its banks, a stream can cut its way through the hardest rocks; the rounded

forms of stream channels in hard rock and the smoothly rounded surfaces of the pebbles in stream gravel are testimony to the effectiveness of this grinding and pounding mechanism.

An additional factor of preeminent importance in determining how rapidly a stream will erode its valley is the frequency of violent storms in its neighborhood. Often during the few hours of a heavy rain a stream accomplishes more than in months or years of normal flow. One reason that running water is the dominant erosional agent in deserts is that desert storms, when they do occur, are violent enough to send raging torrents down the normally dry valleys.

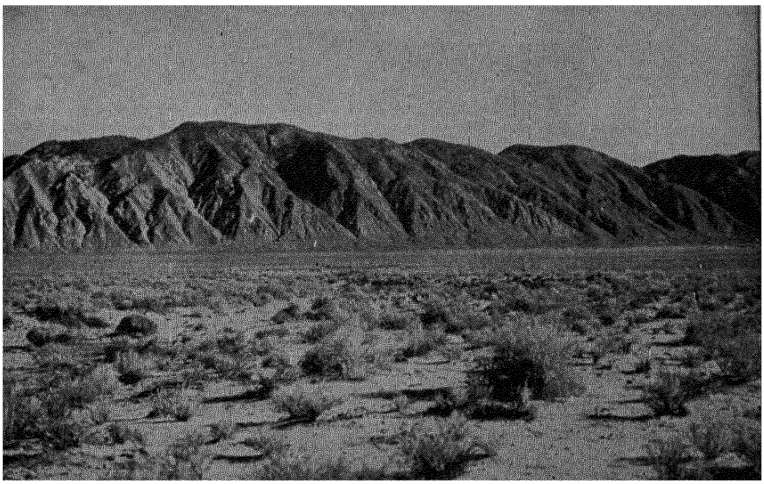


FIG. 221. *The stream-carved slope of a desert range in eastern California. (Photograph by Eliot Blackwelder.)*

The development of a *valley* by stream erosion is admirably illustrated in miniature by the nearest gully cut in the soft material of a hillside. By the temporary stream formed during each successive rain, the gully is deepened, lengthened, and widened. Deepening is accomplished by the downcutting of the stream. Lengthening takes place at the head of the gully, where the stream eats further and further into the hill. Widening is a direct result, not of the stream's activity, but of rainwash and slumping of material on the gully's sides. Thus the stream itself cuts like a blunt knife downward and backward into the hill, while secondary processes widen the gash. The combination of deepening and widening gives the gully its characteristic V-shaped cross section: the V is steep when downcutting is rapid compared with the work of rainwash and slumping, broader when downcutting is slow. As a gully grows older the rate of downcutting slackens, and the processes of widening make its cross section a broader and broader V.

Rivulets on the sides of the gully slowly entrench themselves to form tributary valleys. Each tributary lengthens and widens itself, and from it branch smaller tributaries. So there develops, perhaps in the course of several years, a connected system of gullies with tiny tributaries feeding larger ones, the whole pattern resembling the trunk, branches, and twigs of a tree. If growth of the gully is not stopped artificially or by vegetation, the treelike pattern extends itself until it approaches a part of the hillside drained by another gully system. The remaining surface between the two systems, cut into from both sides, becomes at length a narrow ridge or *divide* between them. Thus in time

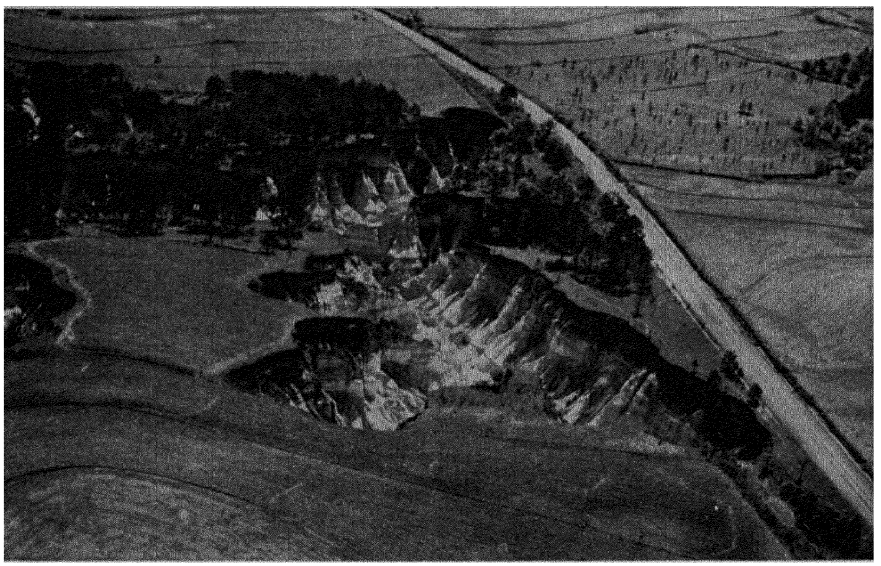


FIG. 222. *Agricultural land ruined by the unchecked development of gullies. (Courtesy of Soil Conservation Service, U. S. Department of Agriculture.)*

stream erosion may carve an entire hillside into treelike patterns of valleys, each valley system separated from its neighbors by dividing ridges. Dissected slopes of this sort are common in arid regions (Fig. 221), and all too often are allowed to develop in cultivated areas (Fig. 222).

The origin of hillside gullies is perfectly obvious, for we can actually observe their growth. The development of larger stream valleys cannot be observed, by reason of the shortness of human life. But large valleys are strikingly similar to gullies: we find streams cutting at their bottoms, the size of a stream depending roughly on the size of its valley; we see the same V-shaped cross sections; we observe tributaries entering main streams from smaller valleys; on a map or from an airplane we note the same treelike patterns separated by sharp divides. It seems obvious that similar processes are at work in gullies and in river valleys. Rivers are

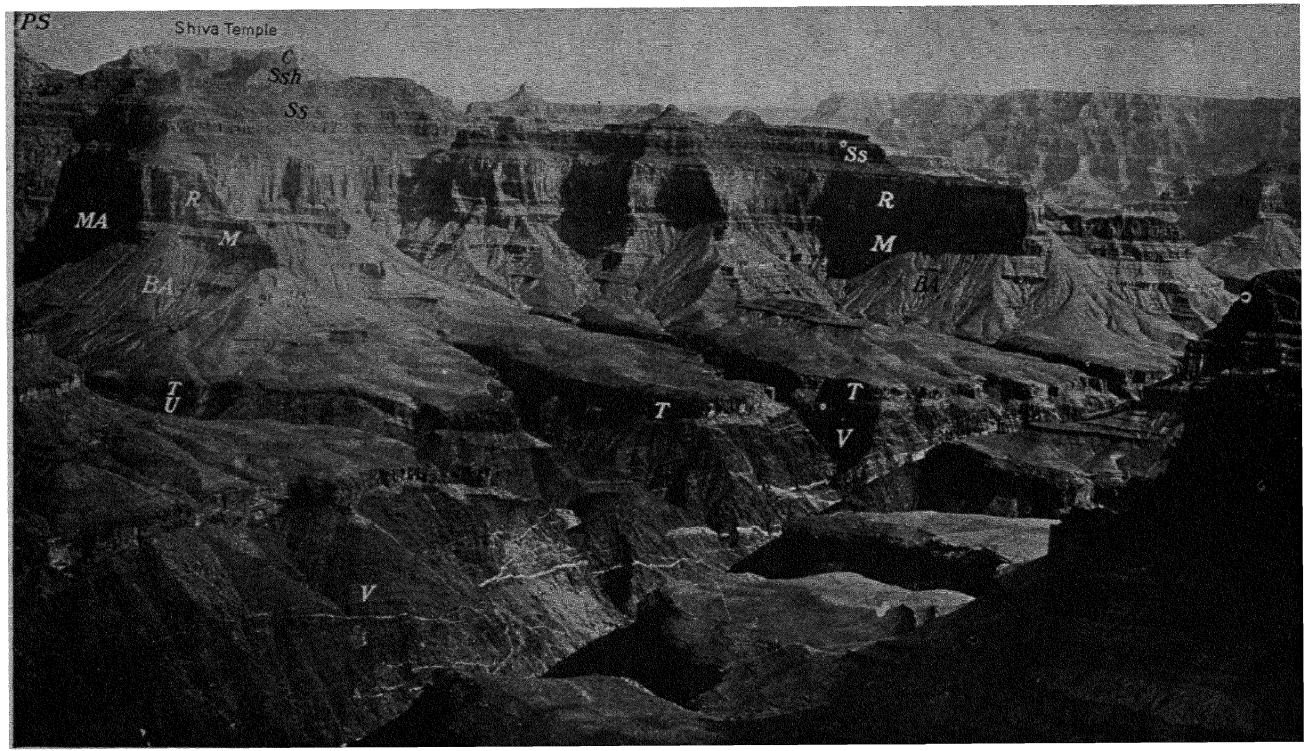


FIG. 223. *The Grand Canyon of the Colorado River. (The letters refer to names of rock formations exposed in the canyon.) (Photograph by N. W. Carkhuff. U. S. Geological Survey.)*

vastly slower in their work than the tiny streams of gullies simply because of the magnitude of their task and the harder materials they must cut.

Not all valleys are made by running water, but for the great majority this origin is certain. The world over, except in Greenland and Antarctica, we find valleys with the characteristic forms produced by stream erosion. Perhaps the grandest monument to the power of this erosional agent is the Grand Canyon of Arizona (Fig. 223). Stand on its brink, look down

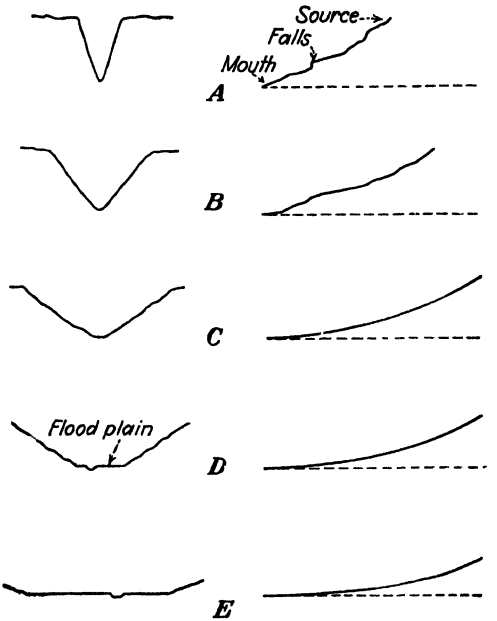


FIG. 224. *Stages in the development of a valley. Cross sections on left, slopes or "long profiles" on right. Compare with Fig. 225.*

into its yawning depths, and the tiny river in the distance seems utterly incompetent to cut so huge a gash. Yet the Grand Canyon is but a gully enormously magnified, its origin betrayed by its shape and by the pattern of its tributaries. In long ages the river has cut its way downward through more than a mile of solid rock.

Landscapes Produced by Stream Erosion

If by some magic we could watch the development of an idealized river, cutting downward through rock of uniform hardness, we should see repeated in slow motion the events which mark a gully's growth. At first the river would dig a deep gorge, its cross section steeply V-shaped; a profile drawn to represent the slope of the river would show a steep channel with numerous rapids and waterfalls (Fig. 224*A*). Presently the

part of the gorge near the river's mouth would be deepened to the level of an adjacent valley, or perhaps nearly to sea level. Below this point the river cannot cut, and downcutting would gradually slacken along its entire course. Its long profile would grow less steep, and its cross section more broadly V-shaped (Fig. 224*B*, 225*A*). Presently, when the slope of the stream becomes a smooth curve (Fig. 224*C*), downcutting would practically cease. From this point the stream would devote its energy to cutting into the sides of its channel, giving its valley a flat floor or ***flood plain*** (Fig. 224*D*, 225*C*). In dry weather the stream would wander over its plain in a sinuous, meandering channel, in very wet weather it would rise out of its channel and overflow the plain. Wider and wider would grow the flood plain, more and more sluggish the stream, lower and lower the sides of its valley.

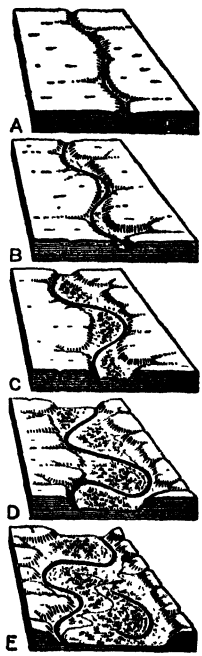


FIG. 225. *Diagrams to show how a stream develops and widens its flood plain. (From Elements of Geography by Finch and Trewartha.)*

During this development of the major valley, tributaries would extend their smaller valleys on either side. Presently the characteristic treelike pattern would be fully developed, separated from the patterns of adjacent streams by sharp divides (Fig. 226). As flood plains develop along the main streams, divides would slowly be lowered by attack from the streams on either side. In the final stages of valley growth, when flood plains are wide and rivers broadly meandering, most of the divides would be obliterated and the remaining ones would be low and rounded (Fig. 227).

So a landscape is altered progressively by the erosional work of streams. Whatever its original form, streams begin their attack by cutting a few deep canyons; valleys grow and tributaries develop until the entire region is dissected into V-shaped valleys and sharp ridges; flood plains form along the main streams and divides are lowered; finally the rivers widen their plains until the entire region is reduced to a nearly flat surface, with a few low hills representing the higher of the old divides. For convenience we may speak of a landscape as *young* when streams are beginning their work and valleys are few and steep, *mature* when tributaries are well developed and most of the area is cut into ridges and valleys, *old* when flood plains are broad and the work of erosion is nearly complete. These are terms without precise definitions, often useful in describing the general appearance of a landscape.

The plateau of northern Arizona, cut into by the Grand Canyon and a few smaller gorges, is a good example of a young landscape. Mature topography (Fig. 228) is common in mountain regions and is especially well shown by the Cumberland Plateau region west of the Appalachians. Simple landscapes in old age are not common; parts of the lower Mississippi Valley show the characteristic broad flood plains, the meandering streams, the low divides, but the topography here is complicated by other geologic processes besides stream erosion.

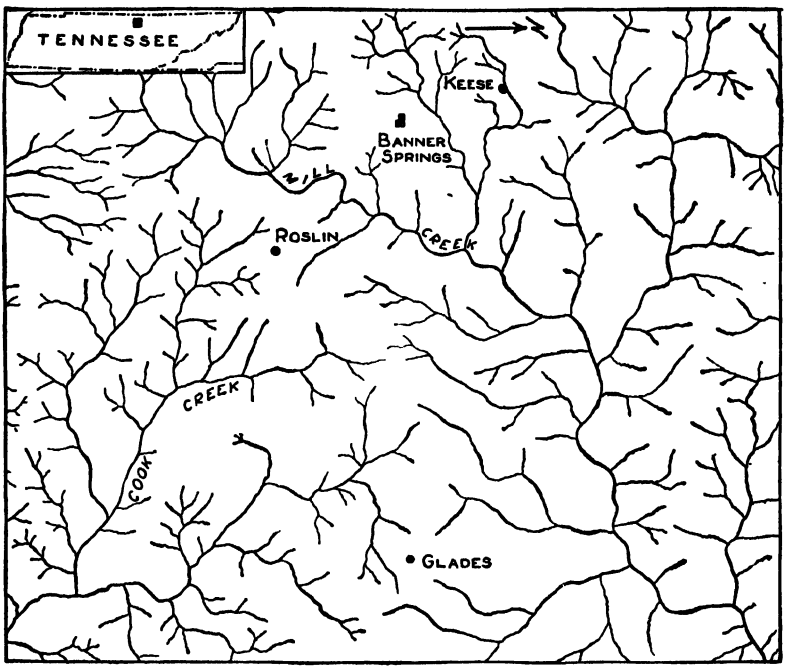


FIG. 226. *Map of streams in an area in northern Tennessee, showing the typical treelike patterns of a fully developed drainage system. (From Geology by Emmons, Thiel, Stauffer and Allison.)*

Actual landscapes seldom show precisely the simple valley shapes and valley patterns just described. Perhaps the commonest reason is the presence of rocks of different hardness: hard rocks in general remain as cliffs and high ridges, while the more easily eroded soft rocks are worn away. In the Grand Canyon, the typical V shape of a young stream valley is modified by steps (Fig. 223) because the stream has cut through horizontal layers of hard and soft rocks; the hard layers form the cliffs, the soft layers the more gentle slopes between. In the Appalachian Mountains treelike patterns of tributaries are not developed, because here alternate hard and soft layers are tilted on end and streams preferentially erode the soft layers (Fig. 229). Many of the fantastic shapes

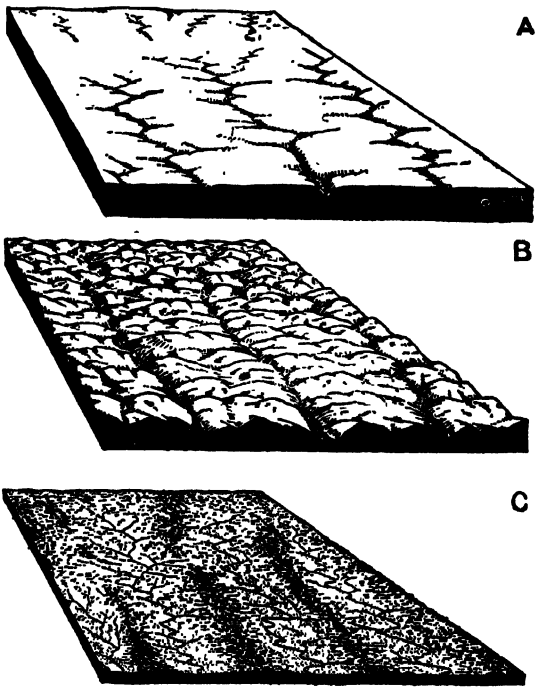


FIG. 227. *Diagram showing the development of a land surface by stream erosion from youth (A) through maturity (B) to old age (C). (From Elements of Geography by Finch and Trewartha.)*

FIG. 228. *Mature topography in southwestern Colorado. (Photograph by E. B. Branson.)*

produced by erosion are due simply to differences in resistance from one rock layer to the next (Fig. 230).

Whatever the valley shapes produced in various stages of landscape development, whatever different kinds of rock may be present, the ulti-

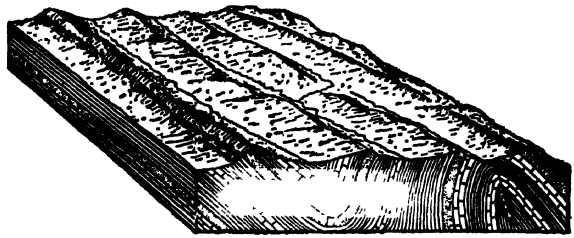


FIG. 229. *Parallel ridges and valleys produced by stream erosion in tilted layers of hard and soft rocks. Soft layers underlie the valleys, hard layers the ridges. Landscapes and rock structures of this sort are typical of the Appalachian Mountains. (After Matthes, U. S. Geological Survey, in Geology by Emmons, Thiel, Stauffer and Allison.)*

FIG. 230. *Erosion of inclined layers of hard and soft rock. Creation Rock, Red Rocks Park, near Denver, Colo. (Courtesy of Denver Tourist Bureau.)*

mate goal of stream erosion is to reduce the land surface to a nearly flat plain approximately at sea level. This goal is fortunately never attained, because other slow geologic processes continually counteract the effects of running water.

Other Agents of Erosion

Glaciers. In a cold climate with abundant snowfall, the snow of winter may not completely melt during the following summer, so that a

deposit of snow accumulates from year to year. Partial melting and continual increase in pressure cause the lower part of a snow deposit to change gradually into ice. If the ice is sufficiently thick, gravity forces it to move slowly downhill. A moving mass of ice formed in this manner is called a ***glacier.***

Existing glaciers are of two principal types. Easily accessible ones—in the Alps, on the Alaskan coast, in the Western United States—are

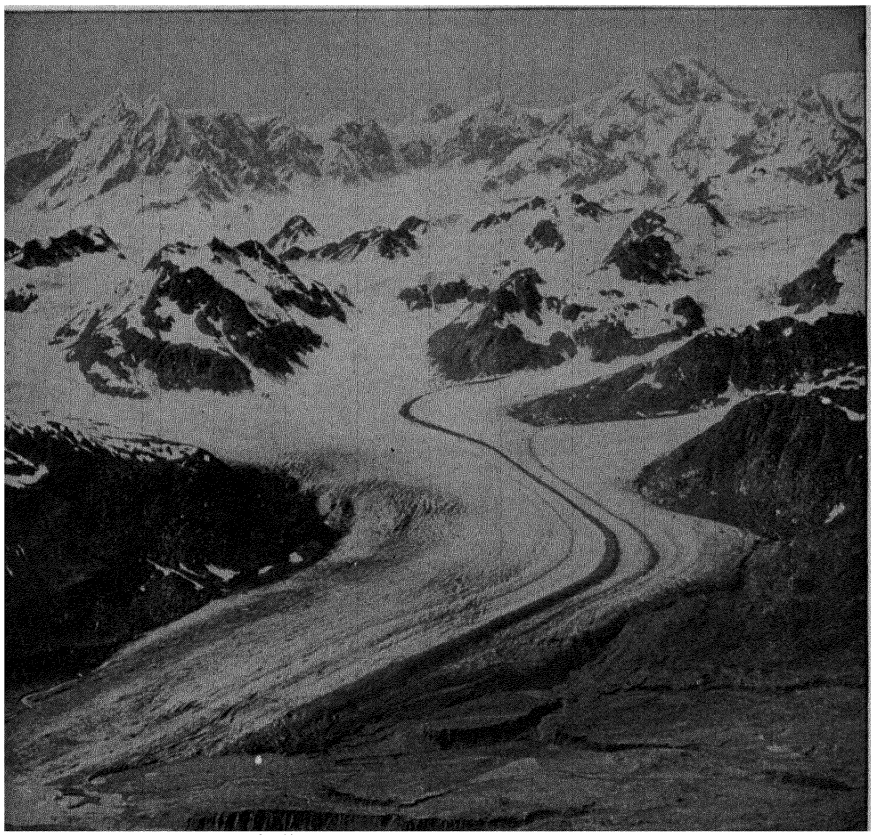


FIG. 231. *Geikie glacier, Alaska, seen from the air.* (*Courtesy of U. S. Forest Service.*)

patches and tongues of dirty ice lying in mountain valleys, called *valley glaciers* (Fig. 231). These move slowly down their valleys, melting copiously at their lower ends, the combination of downward movement and melting keeping their ends in approximately the same position from year to year. Movement in the faster valley glaciers (a few feet per day) is sufficient to keep their lower ends well below timber line. Glaciers of another type cover most of Greenland and Antarctica: huge masses of ice thousands of square miles in area, engulfing hills as well as valleys,

appropriately called *icecaps*. These too move downhill, but the "hill" is the slope of their upper surfaces. An icecap has the shape of a broad dome, its surface sloping outward from a thick central portion of greatest snow accumulation: its motion is radially outward in all directions from its center (Fig. 232).

FIG. 232. *Diagrammatic cross section through the Greenland icecap. Arrows show direction of ice movement. Vertical scale greatly exaggerated.*

Just how a glacier moves is not altogether clear, but it is in part by sliding, in part by internal fracture and healing in the crystals of solid ice. Like a stream, a glacier picks up rock fragments to use as tools in cutting its bed. Some fragments are the debris of weathering which drop on the glacier from its sides, others are torn from its bed when melted water freezes in rock crevices. Fragments at the bottom surface of the glacier, held firmly in the grip of the ice and dragged slowly along its bed, gouge and polish the bedrock and are themselves flattened and scratched (Fig. 233). Smoothed and striated rock surfaces and deposits of debris containing boulders with flattened sides are common near the ends of valley glaciers. Where such evidence of the grinding and polishing of ice erosion is found far from present-day glaciers, geologists have good reason to infer that glaciation was more extensive in the past.

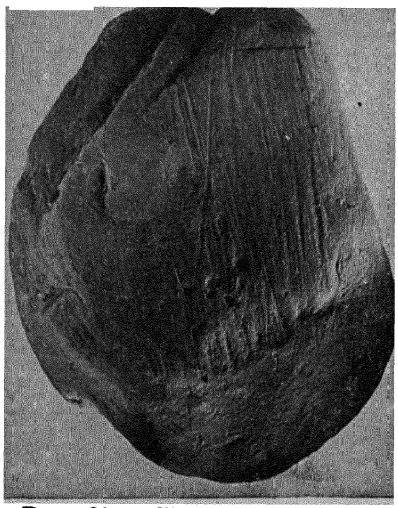


FIG. 233. *Flatiron boulder from a deposit of glacial debris in Nebraska, left by the icecap which once covered the north central states. (Photograph by Alden, U. S. Geological Survey.)*

Valley glaciers form in valleys carved originally by streams, and by long erosion produce characteristic changes in the valley shapes. A mountain stream cuts like a knife vertically downward, letting slope wash, slumping, and minor tributaries shape its valley walls; by contrast, a glacier is a blunt erosional instrument which grinds down simultaneously all parts of its valley floor and far up the sides as well. Effects of this erosion are best seen in valleys which have been glaciated in the past but in which glaciers have dwindled greatly or disappeared. Typically such valleys have U-shaped cross sections with very steep sides, instead

of the V shapes produced by stream erosion (Fig. 234). Their heads are round, steep-walled amphitheaters called *cirques*, in contrast to the small gullies at the heads of stream valleys (Fig. 235). Tributaries often drop into glaciated valleys over high cliffs because glaciers cut their valleys much more actively than do their tributaries; such tributary valleys, left stranded high above their main valleys, are called *hanging valleys*. Lakes and swamps are numerous in glaciated valleys, formed where the glaciers gouged basins in their channels or left deposits of debris as dams. Divides between cirques and between adjacent U-shaped

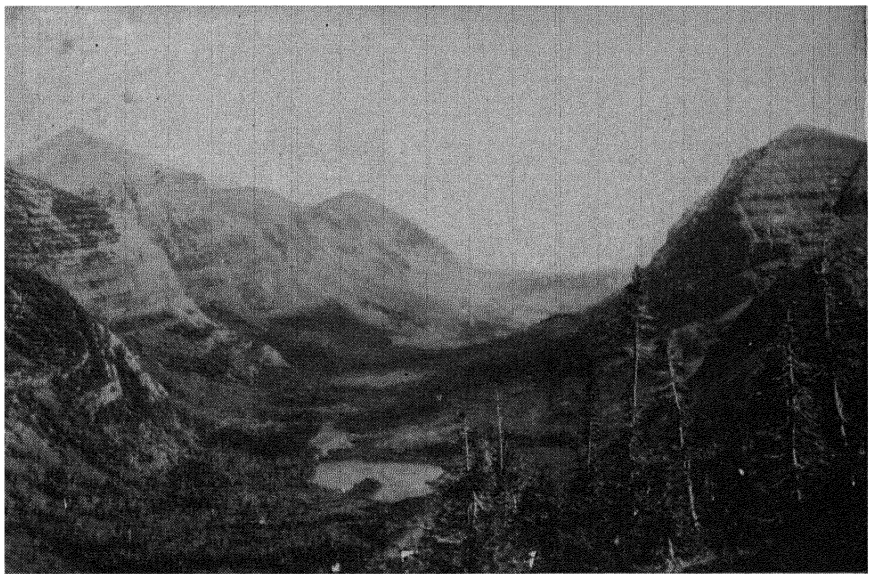


FIG. 234. *A glaciated valley in the Rocky Mountains of Montana, showing the characteristic U-shaped cross section and small lake basins. Note the layered sedimentary rocks of which these mountains are composed. (Photograph by Campbell, U. S. Geological Survey.)*

valleys are often extremely sharp ridges because of the steepness of the valley walls. In general, since valley glaciers produce deep gorges, steep slopes, and knifelike ridges, their effect is to make mountain topography extremely rugged. The earth's most spectacular mountain scenery is in regions (the Alps, the Rockies, the Himalayas) where valley glaciers were large and numerous several thousand years ago.

The influence of icecaps on landscapes is very different from that of valley glaciers. We cannot, of course, observe directly the effect of existing icecaps on the buried landscapes of Greenland and Antarctica, but larger icecaps which once covered much of northern Europe and North America have left clear records of their erosional activity. We shall discuss the evidence for the existence of these glaciers in more detail

later; here we need only note the rounded hills and valleys, the abundant lakes and swamps so characteristic of these regions. Like a gigantic piece of sandpaper, an icecap rounds off sharp corners, wears down hills, fills depressions with debris, leaving innumerable shallow basins which form lakes when the ice recedes.

Glacial erosion is locally very impressive, particularly in high mountains. The amount of debris and the size of the boulders which a glacier

FIG. 235. *Airplane photograph of small cirques at the crest of a high ridge in the Rocky Mountains of Colorado. (Photograph by T. S. Lovering, U. S. Geological Survey.)*

can carry or push ahead of itself is often startling. But by and large, the world over, the erosional work accomplished by glaciers is small. Only rarely have they eroded rock surfaces deeply, and the amount of material transported long distances is insignificant compared with that carried by streams. Most glaciers of today are but feeble descendants of mighty ancestors, but even these ancestors succeeded only in modifying somewhat landscapes already shaped by running water.

Wind. The erosional activity of wind is limited to places where fine material is abundant and unprotected by vegetation—exposed beaches,

arid and subarid lands too intensively farmed, and especially deserts. Fine dust particles can be carried long distances by wind, as the dust storms which plague Kansas and New Mexico in drought years demonstrate all too forcefully. Sand grains are swept along close to the ground, locally grooving and polishing rock surfaces. Fragments larger than sand grains cannot be carried by wind, except rarely in violent local storms. Just how much erosion wind can accomplish is not certain, but compared with streams its effect is probably slight, even in favorable localities. In the desert strong winds blow frequently and rain falls but seldom, yet the sides of desert mountains are carved with the characteristic patterns of stream-eroded valleys.

Waves and Currents. Waves are produced by the friction of wind on open water. The breaking of waves on a beach gives rise to two kinds

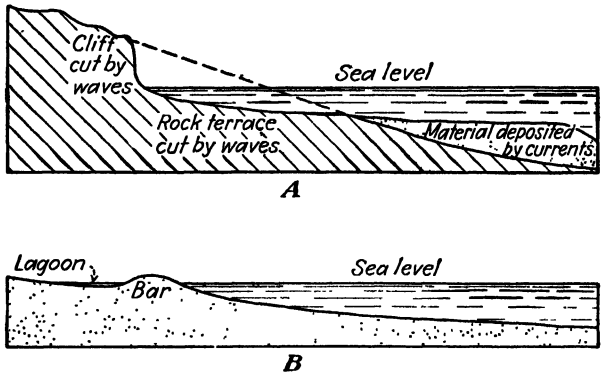


FIG. 236. *Cross sections of shore lines: A, steep, rocky shore; B, low, sandy shore.*

of currents: the *undertow*, a movement of water from a breaking wave returning seaward under succeeding waves; and *long-shore currents*, movements of water parallel to the shore set up by waves which approach the shore obliquely. Cutting of the shore by waves and removal of the debris by currents are responsible for the coast-line scenery so dear to artists and amateur photographers.

Direct erosion by waves on a lake shore is limited to a narrow vertical range approximately at lake level, on the ocean shore to a somewhat wider range determined by high and low tides. Thus wave erosion acts like a horizontal saw, cutting always further inland along the same level. Where mountainous country is adjacent to the coast line this erosion results in steep cliffs, their steepness maintained and their faces continually freshened by the undermining action of the waves. Under water, wave erosion cuts a broad platform, the seaward margin of the platform often being extended by material deposited by currents (Fig. 236*A*). Portions of the shore line formed of hard rock, less easily eroded than

adjacent softer rocks, may be left as projections or as small rocky islands (Fig. 237). On gently sloping coast lines a similar cliff-and-platform profile is produced by wave erosion, but its curves are more gentle; the "cliff" may be simply the seaward slope of a bar or beach (Fig. 236*B*).

Like streams and glaciers, waves erode solid rock effectively only when supplied with rock debris to use as cutting tools. The rounded pebbles of beach gravel show the constant wear to which these tools are subjected. Just as streams do their most effective cutting during severe rainstorms, so waves accomplish the greater part of their erosional

FIG. 237. *The wave-cut cliff at Cape Blanco, Oregon. (From Geology, by Emmons, Thiel, Stauffer, and Allison.)*

work when strong gales drive them against the shore with higher energies than usual.

Groundwater. Much water which falls as rain does not run off immediately in streams but soaks into the ground. All water which thus penetrates the surface is called ***groundwater.*** The soil, the lower mantle, and any porous rocks beneath act together as a huge sponge, taking up great quantities of water into their pore spaces. During and immediately after a heavy rain all available pores in the sponge may be filled, and the ground is then said to be *saturated* with water. When the rain has stopped, water slowly drains away from the upper part of the sponge (hilltops) into adjacent valleys. Thus a few days after a rain porous material in the upper part of a hill has little moisture, while that

in the lower part may still be saturated. Another rain would raise the upper level of the saturated zone, prolonged drought would lower it. This fluctuating upper surface of the saturated zone is called the ***water table.***

Beneath valleys the water table is commonly nearer the surface than under adjacent hills, since water from the saturated zone continually moves outward into valleys. These general relations are shown in Fig. 238. Movement of groundwater in the saturated zone is principally a slow seeping downward and sideward into streams, lakes, and swamps. The motion is rapid through coarse material like sand or gravel, slow through fine material like clay. It is this flow of groundwater which maintains streams when rain is not falling; a stream goes dry only when the water table drops below the level of its bed. A *spring* is formed where groundwater comes to the surface in a more or less definite channel.

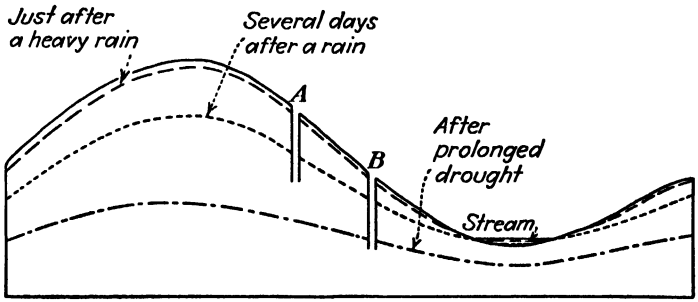


FIG. 238. *Diagrammatic cross section through a hill and valley underlain by porous material, showing positions of the water table just after a heavy rain, several days after a rain, and after a prolonged drought. Well A would be dry during the drought, while well B would still have water.*

Groundwater is the source of all water used by plants and of much water used for drinking and irrigation. Because it is so vitally important to human welfare and because its movements are hidden beneath the ground surface, groundwater has long been a subject of extravagant superstitions. A widespread belief still persists that the best site for a well is indicated by the motion of a willow twig held by a person with occult powers, called a "water witch"; that water moves in immense underground streams; that mountain springs and lakes are fed by water moving upward from the ocean. All such beliefs are baseless. Groundwater moves according to a few simple rules: its source is rain; it moves in a general downward direction; its motion consists of a slow seepage, faster in sand or gravel than in clay, faster in large cracks than in small ones; definite underground streams or pools are rare except in limestone caverns. From these simple facts, together with adequate knowledge of the topography and rock characteristics of a given region, a geologist

can usually make accurate predictions as to the whereabouts and motions of available groundwater.

Although groundwater movement is slow, its erosional activity is by no means negligible. It can accomplish little mechanical wear, but its intimate contact with rocks and soil enable it to dissolve much soluble material. The dissolved substances are in part transported to neighboring streams, in part redeposited at other points in the mantle or bedrock. Dissolved material is responsible for the "hardness" of water from many wells. In regions underlain by limestone, most easily soluble of ordinary rocks, the dissolving action of groundwater makes itself conspicuous by the formation of *caves*. A cave is produced when water moving through tiny cracks in limestone gradually enlarges the cracks by dissolving and removing adjacent rock material (page 363).

The activity of groundwater extends downward for some hundreds or thousands of feet, the depth in a given region depending on the kinds of rock present. Nearly all rocks within a short distance of the surface have sufficient pore space or are sufficiently cracked to permit some circulation, although in massive igneous and metamorphic rocks the amount of groundwater may be very small. But at lower levels cracks become too small and too scarce to permit free movement. Thus deep mines often have abundant water in their upper parts, but so little at lower levels that dust from drilling and blasting becomes a menace.

Sedimentation

Most of the material transported by the agents of erosion is presently deposited to form *sediments* of various sorts. Only substances in solution escape deposition: ions of various salts carried by streams to lakes and oceans may remain dissolved indefinitely. The salt of the sea is an accumulation of material dissolved out of rocks by rain, rivers, and groundwater through long geologic ages. Under some conditions even a part of the dissolved material may form sediments; slightly soluble salts like calcium carbonate precipitate readily, and others appear when evaporation concentrates the water of a salt lake or an arm of the sea.

The ultimate destination of erosional debris is the ocean, and the most widespread sediments accumulate in shallow parts of the ocean near continental margins. But much sedimentary material is carried to the sea in stages, deposited first in thick layers elsewhere—in lakes, in desert basins, in stream valleys. Each of the various erosional agents has its own peculiar methods of depositing its load, and these methods leave their stamp on the character of the deposits formed. Since sediments laid down ages ago often retain many of their original characteristics, an acquaintance with the processes of deposition enables us to infer the probable origin of older deposits. In this way we can reconstruct

past conditions of erosion and sedimentation, and so gain an insight into many chapters of earth history.

A few terms from the following discussion need brief explanation. One important characteristic of a sediment is its *degree of sorting*. This refers to the extent of separation of fine material from coarse—whether boulders, clay, and sand are mixed up together or segregated in different layers. Another feature which often helps to betray a sediment's origin is the kind of *stratification* (layering) it shows (Fig. 223). Some sediments

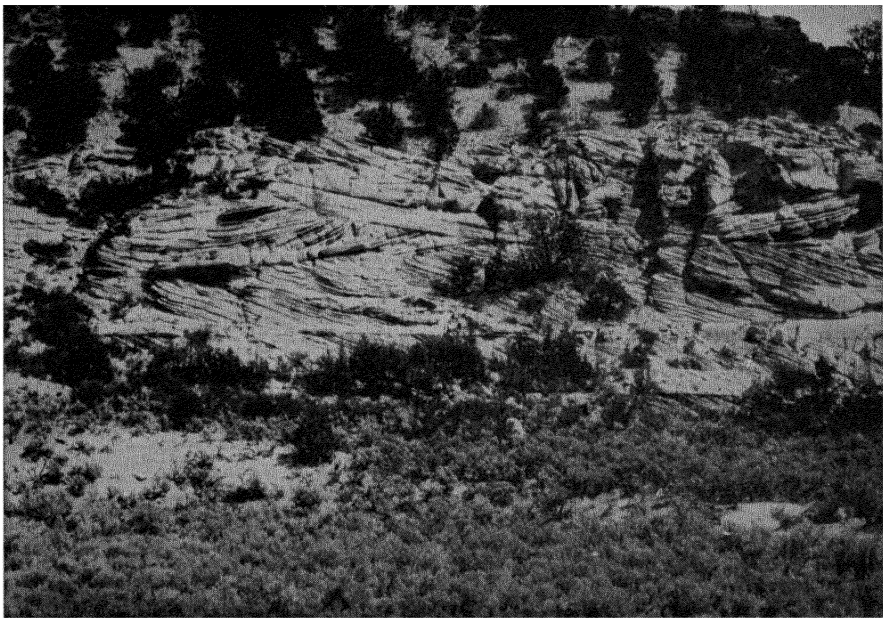


FIG. 239. *Cross-bedding in sandstone. Cross-bedding of this sort, with long sweeping curves and steep slopes, suggests deposition of the sand by wind in shifting dunes. (Photograph by F. G. Tickell.)*

have practically no layering, but retain the same color and texture through great thicknesses; in others each bed or stratum is sharply marked off from those above and below by differences in color and grain size. The layers of stratified deposits are sometimes uniform and parallel over long distances, sometimes show abrupt variations in thickness. Occasionally sediments show a type of layering called *cross-bedding* (Fig. 239), in which thin curved beds lie at moderate angles to the general trend of stratification.

In the following paragraphs sediments laid down by the various agents of erosion will be examined in some detail. We shall find that sedimentation is by no means a mysterious process, but that we can

usually predict from the nature of an erosional agent and the type of material it carries what kinds of sediment it will produce.

Streams. Much of the material carried by streams is delivered to the sea, there to be re-sorted and deposited by waves and wave-formed currents. We shall consider first, however, sediments for which streams themselves are the active depositional agents.

A stream carrying abundant debris must drop part of its load whenever its speed slackens or its amount of water decreases. We may note four common sites of deposition: (1) Debris carried in time of flood is deposited in gravel banks and sand bars when the swiftly flowing waters

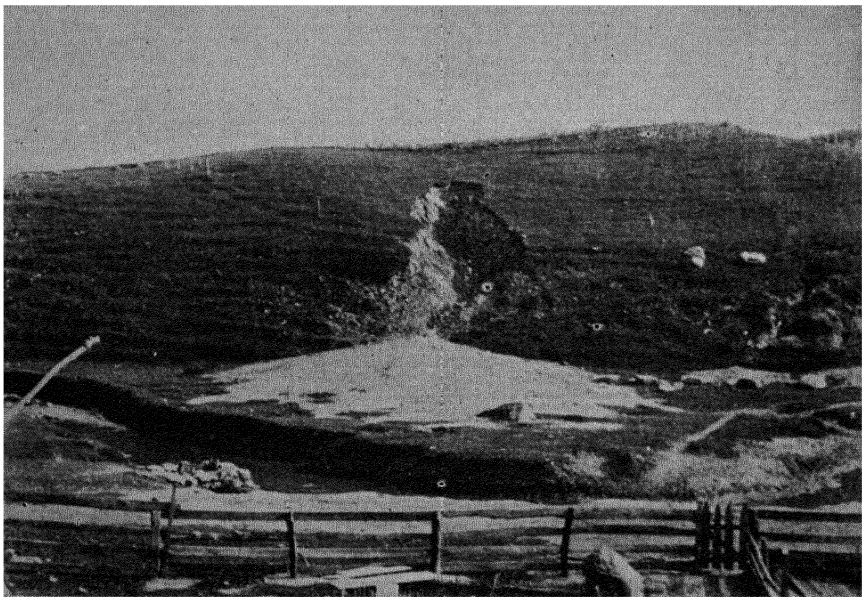


FIG. 240. *A small alluvial fan. (Photograph by Eliot Blackwelder.)*

begin to recede. (2) The flood plain of a meandering river is a site of deposition in occasional floods, when the river overflows its banks and loses speed as it spreads over the plain. In Egypt, for example, the fertility of the soil has been maintained for centuries by the deposit of black silt left each year when the Nile is in flood. (3) A common site of deposition, especially in the Western United States, is the point where a stream emerges from a steep mountain valley and slows down as it flows onto a plain. Such a deposit, usually taking the form of a low cone pointing upstream, is called an *alluvial fan* (Fig. 240). (4) A similar deposit is formed when a stream's flow is abruptly stopped as it flows into a lake or sea. This kind of deposit, built largely under water and with a surface usually much flatter than that of an alluvial fan, is called a *delta.*

Streams can transport fragments of all sizes, from fine clay particles to large boulders. Turbulent stream currents are only fairly good sorting agents, so that individual layers in their deposits commonly contain mixtures of sand and gravel or of sand and clay. Pebbles and boulders are well rounded when a stream has rolled them along its channel for long distances, angular if they have traveled only a little way from their source. The constant shifting of stream courses, the filling of old channels and cutting of new ones, the building and tearing down of gravel banks and sand bars, lead to conspicuously uneven bedding. This irregular bedding—thin layers of gravel, sand, and silt which thicken or pinch out abruptly, often cross-bedded—is the outstanding characteristic of stream deposits (Fig. 241).

FIG. 241. *Stream-deposited sand and gravel, showing the characteristic irregular bedding. (From Field Geology by Lahee.)*

Fragments carried by a swift stream are subjected to continual hammering against one another and against the stream bed, and are constantly exposed to the agents of weathering. In such an environment only the toughest and most resistant minerals can survive. Stream gravels which have not traveled far from their source may contain pebbles of many different rock types, but gravels which a stream has carried for long distances contain hard rocks almost exclusively. Since quartz is the hardest and most resistant of common minerals, material which has been battered down to sand-grain size usually contains an abundance of quartz. Fine material in stream sediments consists chiefly of clay minerals.

Glaciers. The material scraped from its channel by a glacier is in part heaped up at its lower end and pushed forward as a low ridge by the slow ice movement, in part dumped into depressions and spread as a layer of irregular thickness beneath the ice. The pile of debris around the end of the glacier, called a *moraine*, is left as a low ridge of hummocky

topography when the glacier melts back. Moraines in mountain valleys and in the North Central states are part of the evidence for a former wide extent of glaciation (Fig. 242).

FIG. 242. *Part of the hummocky, boulder-strewn surface of a large moraine in Wisconsin, left by the retreat of the icecap which once covered the north central states. (Photograph by Alden, U. S. Geological Survey.)*

FIG. 243. *A deposit of till. (From Geology by Emmons, Thiel, Stauffer and Allison.)*

All of the material deposited directly by ice goes by the name of *till*. The motion of glaciers gives little chance for sorting or stratification, so that till is characteristically an indiscriminate mixture of fine and coarse material with no semblance of bedding (Fig. 243). Huge boulders are often embedded in the abundant fine claylike material which a glacier produces by its polishing action. Typically most of the boulders are

angular, a few rounded and showing the flat, scratched faces produced as they were dragged along the bed of the glacier. Unstratified till is often associated with stream deposits, for the melting of glacial ice furnishes abundant water which sorts and stratifies some of the glacially eroded material.

Wind. Wind is a highly efficient sorting agent, winnowing fine silt and dust cleanly from sand and gravel. Much of the fine material is carried to great heights and for long distances, spreading so thinly over land and sea that it loses its identity. Where wind-blown dust accumulates in sheltered places, it forms a soft, fine, claylike material called *loess.* Great deposits of loess are found in northern China, representing the long accumulation of dust swept from the desert basins of central Asia. Wind piles sand grains on exposed beaches and in deserts into the low, shifting hills called *dunes.* Dune sand is typically free from dust or mud particles, its grains are well rounded, and it usually shows large-scale cross-bedding produced by deposition of successive layers on the curving surfaces of the dunes.

Waves and Currents. Most important of the agents of deposition, since they handle by far the largest amount of sediment, are the wave-formed currents of the sea. Currents deposit not only the materials eroded from coast lines by wave action, but also abundant debris brought to the ocean by streams, wind, and glaciers. Visible deposits of waves and currents include beaches and sand bars, but the great bulk of the sediments brought to the ocean are laid down under water.

Deposition beneath the sea takes place in many ways. Sand, gravel, and clay are dragged outward along the bottom by the undertow, dropping wherever the current becomes too weak to carry them. Much very fine material carried in colloidal suspension by large rivers like the Mississippi is deposited on entering the sea because the electric charges of the colloidal particles are neutralized by the ions of sea water (page 436). Some salts, notably calcium carbonate, are deposited as chemical precipitates when sea water becomes locally oversaturated. In places living organisms are so abundant that their shells, when the organisms die, become an important part of the material deposited.

These depositional processes operate chiefly in the shallow parts of the ocean bordering the continents, out to depths of 2,000 to 3,000 ft. In shallow arms of the sea, like Hudson Bay and the Baltic Sea, active deposition may take place over the entire bottom, but little material from land ever reaches the deeper parts of the oceans.

Coarse fragments in marine sediments are commonly well rounded, since before deposition they are used by waves in their battering of the shore. The constant back-and-forth shifting of loose material on the bottom by waves and currents provides a good sorting mechanism, so

that the grain size in any one layer of sediment is fairly uniform. Often, however, there is a very gradual change in grain size away from shore, since outgoing currents become progressively feebler; if fragments of all sizes are available, coarse gravel is dropped near shore where currents are strong, sand farther out, and clay in still deeper water. Since the sea floor is nearly flat, and since marine deposition as a rule takes place uninterruptedly for long periods of time, marine sediments are characterized by even, parallel beds often of considerable thickness (Fig. 223). Cross-bedding is of minor importance and is always on a small scale; it suggests deposition on beaches and sand bars. Fossils are more abundant in marine deposits than in any other type.

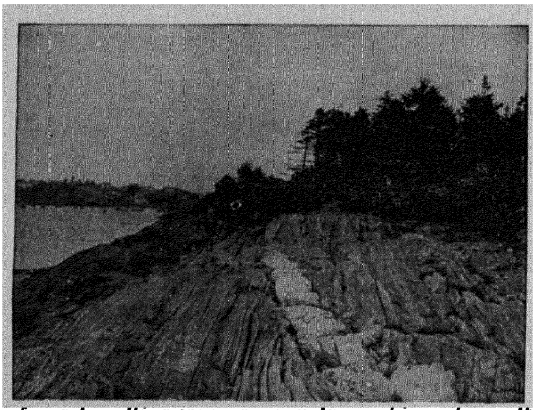


FIG. 244. *A vein of quartz cutting wave-worn metamorphic rocks on the Maine Coast. (From Elements of Geography by Finch and Trewartha.)*

The feebler currents of an inland lake also sort out the debris supplied by streams and by wave erosion and spread it in even layers over the lake bottom. Commonly the currents are too weak to transport any but the finer particles out to deep water, so that lake deposits are characterized by thin, uniform beds of clay. In arid regions evaporation of lake water causes precipitation of various salts (calcium carbonate, calcium sulfate, sodium carbonate), which are mixed with the clay or in separate layers. Deposits of shallow lakes and swamps in humid regions often contain partially decayed plant material in great abundance.

Groundwater. Deposition by groundwater is almost entirely chemical precipitation from solution. The precipitation may be brought about by evaporation of the water, by mixing of groundwater from different sources, by a lowering of temperature, or by the escape of a gas when pressure is reduced. One of the most important geologic functions of groundwater is the deposition of material in the pore spaces of sediments, which helps to convert the sediments into rock. Much dissolved material is deposited along open cracks to form *veins* (Fig. 244), common in all

kinds of rocks. The commonest vein minerals are quartz and calcite, but some veins contain commercially valuable minerals. More spectacular examples of groundwater deposition are the "rock icicles" or stalactites (Fig. 180, page 364) that hang from the roofs of limestone caverns, and the colorful deposits often found around hot springs and geysers.

Sedimentary Rocks

Sediments buried beneath later deposits are gradually hardened into rock. The formation of rock is often a complex process, completed only after slow changes have gone on for thousands or even millions of years. One important change in a sediment is compaction, the squeezing together of its grains under the pressure of overlying deposits. Some recrystallization may accompany compaction; the calcite crystals of limy sediments, in particular, grow larger and interlock with one another. Chemical changes brought about by circulating groundwater are largely responsible for the hardening of many sediments: the grains of coarse sediments are cemented by material precipitated from solution in groundwater, and some sediments have much of their original material dissolved away and replaced by other substances. The profound changes which groundwater can accomplish are strikingly illustrated by petrified wood, in which the original organic compounds have been removed, molecule by molecule, and replaced by silica—the whole process taking place so gradually that the finest details of wood structure may be preserved.

The most common cementing materials with which groundwater binds together the grains of sediments are silica, calcium carbonate, and hydrated ferric oxide. The latter betrays its presence by the red, yellow, and brown colors of many sandstones and conglomerates.

The hardening of sediments by cementation, compaction, and recrystallization produces the ordinary varieties of sedimentary rock. Gravel beds are cemented into tough conglomerates, sand deposits into sandstones. Layers of clay become shale, and precipitates of calcium carbonate become limestone. The origin of chert is more obscure: its texture and banding suggest that it was formed as a gel of colloidal silica which slowly hardened by losing water, but why and under what conditions such a gel would accumulate remain awkward questions.

The structures of a sediment—such as stratification, cross-bedding, fossils—are commonly preserved through the hardening process, although delicate structures may be partly obliterated. Hence examination of a sedimentary rock often reveals a good deal about conditions under which the original sediment was laid down. For instance, a hard sandstone in thick, even layers, containing impressions of fish skeletons or clamshells, was almost certainly a marine deposit. A sandstone in thin beds, interstratified with shale and conglomerate and showing strong cross-bedding,

was probably laid down by a stream. A sandstone free from clay and gravel, made up of well-sorted rounded grains, and cross-bedded in large, sweeping curves, represents hardened dune material.

Sedimentary rocks are peculiarly important in geology because they contain material which was deposited *at or near the earth's surface.* In the nature and structure of this material is preserved a record of the changing surface conditions of past time. If we read the record with sufficient insight and imagination, we find spread out before us a panorama of earth history: seas that once spread widely over the land, the advance and retreat of ancient glaciers, the winds and torrential streams of long-vanished deserts. Dimly we can see even the living creatures which inhabited lands and seas of the past, for in sedimentary rocks are entombed abundant fossil remains of plants and animals. While igneous and metamorphic rocks reveal in their structures something about conditions in the earth's interior, sedimentary rocks tell us the more varied and interesting history of surface landscapes.

Questions

1. List the chief products formed by long chemical weathering of granite.
2. In what sort of climate would a quartzite with numerous tiny cracks weather rapidly?
3. Which would you expect to be more effective in the tropics, mechanical weathering or chemical weathering? Which on a high mountain peak?
4. What color would you expect in mantle material derived from the weathering of basalt in a warm, humid climate?
5. What is the evidence that streams are the principal agent of erosion on the earth's surface?
6. Suppose that a nearly flat area underlain chiefly by clay stands near the seashore, its surface about 100 ft above sea level. Describe the stages in the erosion of this area by streams.
7. After long erosion in a region of temperate climate underlain by the following types of rock, which types would you expect to find just beneath the surface (*a*) of hills, (*b*) of valleys? marble, quartzite, basalt, shale, granite.
8. Compare the mountain landscapes produced by stream erosion in the mature stage and by valley glaciers.
9. In general, would you expect the debris deposited by a glacier to show more or less chemical decay than stream sediments?
10. The effects of past glaciation in the Sierra Nevada of California extend to much lower elevations in valleys on the west side of the range than in valleys on the east. Suggest a possible reason.
11. In sand derived from the attack of waves on granite, what mineral or minerals would you expect to be most abundant?
12. Imagine two small hills separated by a stream valley, one consisting chiefly of sand and gravel, the other chiefly of clay. Show by a diagram the position of the water table in each (*a*) immediately after a heavy rain, (*b*) several days after a heavy rain, (*c*) after prolonged drought. Indicate by vertical lines the depth to which a well must penetrate from the top of each hill to maintain a supply of water during drought.

13. What are the chief characteristics of (*a*) the material in an alluvial fan at the base of a short, steep slope; (*b*) sediments laid down a short distance from shore along a steep, rocky seacoast; (*c*) material in a sand dune; (*d*) the material in a moraine?
14. What is the probable origin of the following sedimentary rocks?
    a. A thick, evenly bedded limestone.
    b. A conglomerate with well-rounded boulders and numerous thin beds of sandy and clayey material, showing conspicuous cross-bedding.
    c. A sandstone consisting of well-sorted, well-rounded grains of quartz, showing conspicuous large-scale cross-bedding.
15. What characteristics of a sedimentary rock might indicate an arid climate at the time the original sediment was deposited?
16. Compare the surfaces of earth and moon with respect to (*a*) topography, (*b*) erosional agents.
17. Why are hot-spring deposits thicker than deposits around ordinary springs?

CHAPTER XXXVI

Vulcanism and Diastrophism

EROSION and sedimentation are processes of leveling, processes by which the higher parts of the earth's surface are worn down and the lower parts filled with sediment. If their work could be carried to completion, these processes should ultimately reduce all the continents to plains beneath the level of the sea. The simple fact that continents still exist, with many rugged mountains and deep valleys, is good evidence that the work of erosion and sedimentation is opposed by other processes capable of elevating some parts of the earth's surface and depressing other parts.

The processes by which irregularities of the earth's surface are maintained may be grouped under two heads—processes of ***vulcanism,*** which involve the movement of liquid rock, and processes of ***diastrophism,*** which include all movement of the solid materials of the crust. Vulcanism and diastrophism are not wholly independent; for movement of liquid rock often causes considerable distortion of adjacent rock layers, and major diastrophic movements are often accompanied by volcanic activity.

Volcanoes

Molten rock, usually called ***magma*** while it is beneath the surface and ***lava*** when it appears at the surface, in places finds a way to the surface through fissures or cylindrical openings. Such openings are called ***volcanoes.*** Because the material which escapes from a volcano accumulates as solid rock near the orifice, most volcanoes in the course of time build up mountains of characteristic shape—roughly conical, steepening toward the top, with a small depression or *crater* at the very summit (Fig. 245). From a few volcanoes liquid rock escapes almost continuously, but the greater number are active only at intervals.

A volcanic eruption is one of the most awesome spectacles in all nature. Usually a few hours or a few days beforehand there is warning in the form of earthquakes—minor shocks probably caused by the movement of gases and liquids underground. An explosion or a series of explo-

sions begins the eruption, sending a great cloud billowing upward from the crater (Fig. 246). In the cloud are gases from the volcano, water droplets (since water vapor is a prominent volcanic gas), fragments of solid material blown from the crater and the upper part of the volcano's orifice, dust and larger solid fragments representing molten rock blown to bits and hurled upward by the violence of the explosions. Gas continues to issue in great quantities, and explosions recur at intervals. The cloud may persist for days or weeks, its lower part glowing red at night. Activity gradually slackens, and presently a tongue of white-hot lava spills over the edge of the crater or pours out of a fissure on the mountain slope. Other flows may follow the first, and explosive activity may continue with diminished intensity. Slowly the volcano becomes quiescent, until only a small steam cloud above the crater suggests its recent activity.

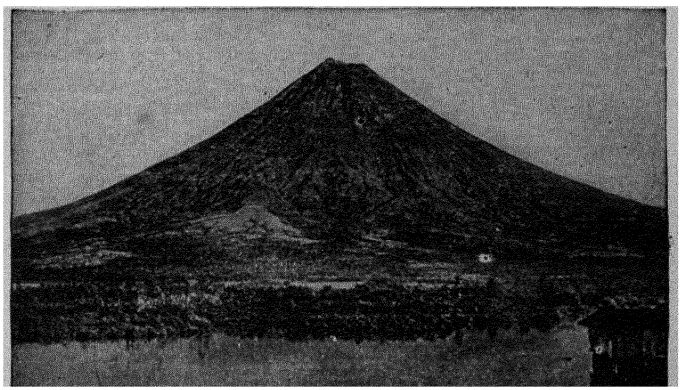


FIG. 245. *Mayon, a remarkably symmetrical volcanic cone in the Philippine Islands (7,616 ft high). (Chester H. Reeds, American Museum of Natural History.)*

Not all eruptions by any means follow this particular pattern. Volcanoes are notoriously individualistic, each one having some quirks of behavior not shared by others. In one group of volcanoes the explosive type of activity is dominant, little or no fluid lava appearing during eruptions. Cones of these volcanoes, built entirely of fragmental material ejected in a solid or nearly solid state, are very steep sided. Good examples are found in the West Indies, in Japan, and in the Philippines. Other volcanoes, like those of Hawaii, have eruptions characterized by quiet lava flows with little explosive activity. Mountains built by these volcanoes are broad and gently sloping, quite different from the usual volcanic structure. The most common kind of volcano is neither wholly of the "explosive" type nor wholly of the "quiet" type, but has eruptions in which both lava flows and gas explosions figure prominently (Fig. 247).

The chief factors which determine whether an eruption will be largely quiet or largely explosive are the viscosity of the magma and the amount

FIG. 246. *The volcano Ngauruhoe, New Zealand.* (*Copyright National Geographic Society. Reproduced with permission.*)

of dissolved gas it contains. A magma is a complex mixture of the oxides of various metals with silicon dioxide, usually containing an abundance of gas dissolved under pressure. Like most silicate melts it is exceedingly

viscous; with rare exceptions, molten lavas creep slowly downhill like thick sirup or taffy (Fig. 248). The viscosity depends in large measure on chemical composition, magmas with high percentages of silica being in general more viscous than those with large amounts of metallic oxides. Gas content also affects viscosity, magmas with little gas being the most viscous. If the magma feeding a volcano has a high gas content and is

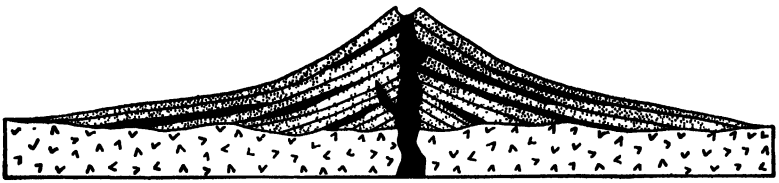


FIG. 247. *Diagrammatic cross section of a volcano, showing lava flows (solid black) alternating with beds of tuff and volcanic breccia.*

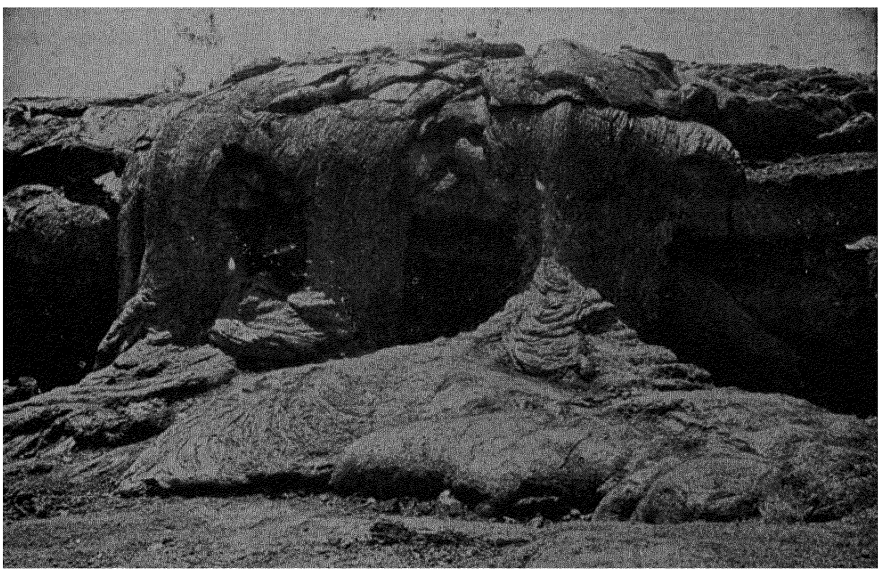


FIG. 248. *Lava which ... cascaded over a small cliff. Kilauea, Hawaii.* (*Photograph by Mendenhall, U. S. Geological Survey.*)

still siliceous enough to be fairly viscous, eruptions will be largely explosive; basic magmas with relatively small amounts of gas are required for quiet eruptions.

The gaseous products of volcanic activity include carbon dioxide, nitrogen, water vapor, hydrogen, compounds of sulfur, and minor amounts of the halogens and metallic compounds. Most of these mix at once with the atmosphere. Much of the water vapor condenses, giving rise to the torrential rains which often accompany eruptions.

Molten lava solidifies to form one or another of the various volcanic rocks. All of these are characterized by their fine grain size, since lava flows cool rapidly. Basalt is by far the commonest volcanic rock and forms the largest flows; coming from a basic magma, it is fluid enough to flow long distances. Rhyolite, the most siliceous of ordinary lavas, forms small, thick flows. Rhyolitic lava is sometimes so viscous and cooling is so rapid that crystallization does not take place, and the natural glass *obsidian* is formed. All kinds of volcanic rocks frequently show rounded holes left by expanding gas which was trapped during the final stages of solidification; viscous, siliceous lavas are sometimes so filled with gas cavities that the light, porous rock *pumice* is formed.

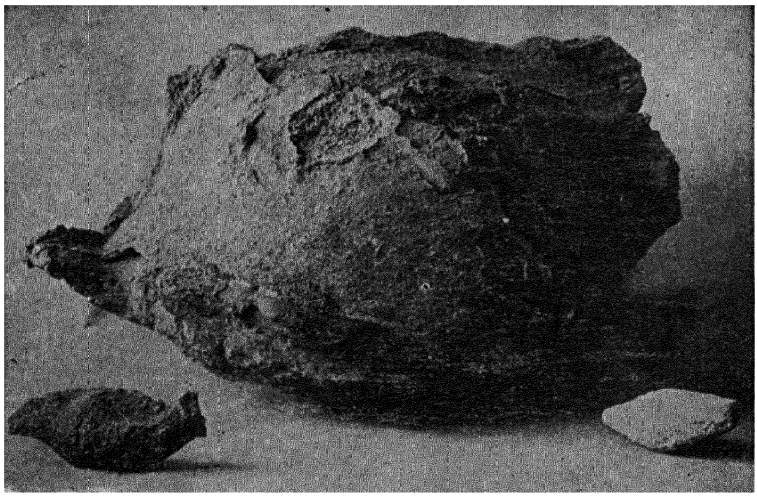


FIG. 249. *Volcanic bombs of basalt, about one-third natural size. (From Petrography and Petrology by Grout.)*

During explosive eruptions much of the liquid rock is blown into fragments which solidify immediately or during their flight through the air. The fragments range in size from fine dust particles to huge "volcanic bombs" several feet in diameter (Fig. 249). Most of this material settles on the slopes of the volcano, though the finer dust may be carried long distances by wind. Deposits of the finer material may be cemented to form the rock *tuff*, and deposits of the coarser material may form a kind of conglomerate called *volcanic breccia*.

Lava flows and thick falls of volcanic dust may do considerable property damage during an eruption but are seldom responsible for much loss of life. The really destructive feature of volcanic activity is a phenomenon often associated with explosive eruptions: part of the cloud above the crater becomes so choked with solid debris, and both debris and gas are in such a wild state of commotion, that the mixture is capable of

moving down the mountainside like a liquid. Within this rapidly moving cloud are hot, poisonous gases and solid fragments of all sizes in a maelstrom of violent movement. No living thing can survive in its path. Such

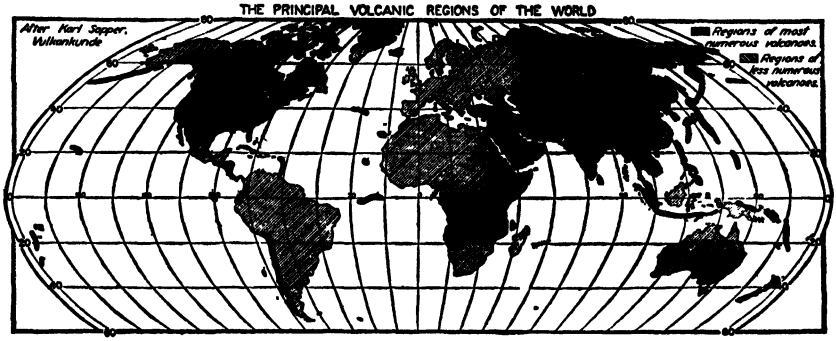


FIG. 250. (*From Elements of Geography by Finch and Trewartha.*)

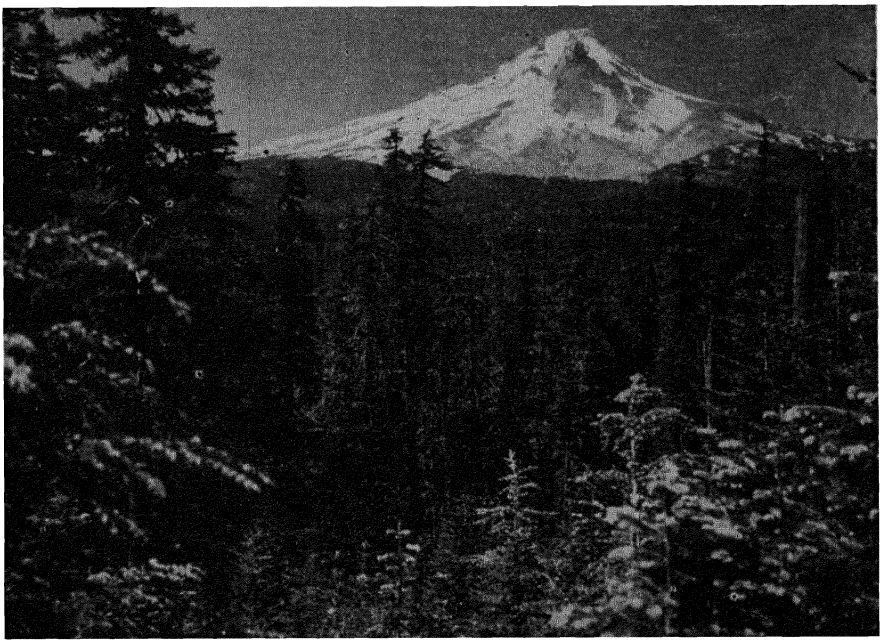


FIG. 251. *Mt. Hood, an extinct volcano in the Cascade Range of western Oregon. The shape of the mountain has been somewhat altered by valley glaciers.*

a cloud from the volcano Mt. Pelée in 1902 completely destroyed the city of St. Pierre, snuffing out the lives of its 25,000 citizens in a few seconds.

Active volcanoes of the present are found around the borders of the Pacific Ocean, on some of the Pacific Islands, in Iceland, in the Medi-

terranean region, in the West Indies, and in East Africa (Fig. 250). In many other parts of the world volcanoes have been active in the past. Where volcanoes have become extinct in the recent geologic past, we find evidence for their former activity in isolated, cone-shaped mountains, in lava flows, and in hot springs, geysers, and steam vents. The great mountains of the Cascade range (Fig. 251) in the Western United States —Rainier, Baker, Hood, Shasta—are old volcanoes, some of their lava flows so recent that vegetation has scarcely gained a foothold on them (Fig. 252). In regions where volcanoes have been dead for a longer period,

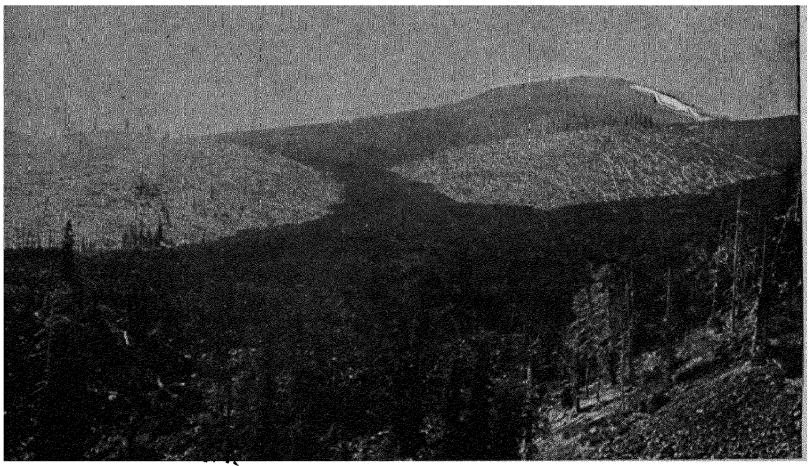


FIG. 252. *A lava flow partly filling old valleys in the Cascade Mountains of Oregon. No active volcanoes exist in this region today, but the recency of the flow is indicated by its lack of vegetation. (Photograph by H. W. Fairbanks. From Physiography of Western United States by Fenneman.)*

erosion may have removed all evidence of the original mountains and left only patches of volcanic rocks to indicate former igneous activity.

Intrusive Rocks

Magma which rises into the rocks of the earth's crust but does not reach the surface solidifies to form *intrusive* bodies of various kinds. Because cooling in these bodies is slower than at the surface, intrusive igneous rocks are in general coarser grained than volcanic rocks. We find such coarse-grained rocks exposed at the surface only when deep erosion has uncovered them long after their solidification.

The igneous origin of volcanic rocks is evident enough, for we can actually watch fluid lava freeze to solid rock. But no one has ever seen a rock like granite in its original liquid state; the evidence that it is an igneous rock must be entirely indirect. Nevertheless, the following facts

establish an igneous origin beyond any reasonable doubt: (1) Granite shows the same relations among its minerals that a volcanic rock shows under the microscope: the separate grains are intergrown, those with higher melting points showing by their better crystal forms that they crystallized a little earlier (Fig. 213, page 479). (2) In some small intrusives every gradation can be found between coarse granite and a rock indistinguishable from the volcanic rock rhyolite, whose igneous origin is established by direct observation. (3) Granite is found in masses which cut across layers of sedimentary rock and from which small irregular pockets and stringers penetrate into the surrounding rocks; sometimes blocks of the sedimentary rocks are found completely engulfed by the granite (Fig. 253). (4) That granite was at a high enough temperature

FIG. 253. *Schist (dark) intruded by granite (light). (Photograph by E. B. Branson.)*

to be molten is shown by the baking and recrystallization of the rocks which it intrudes. These same types of evidence apply equally well to the other intrusive rocks.

Intrusive bodies are classified and named according to their shapes and sizes. Here we need consider only two out of the many recognized varieties, *dikes* and *batholiths*.

A ***dike*** is a sheetlike mass of igneous rock intruded along a fissure (Fig. 254). The largest dikes have thicknesses in thousands of feet, but the commoner ones range from a few inches to a few tens of feet thick. Any kind of igneous rock may occur in a dike: rapid cooling in small dikes may give rocks like volcanic types, while slow cooling in larger ones often gives coarse-grained rocks. Dikes may cut any other kind of rock. They are frequently associated with volcanoes, some of the magma apparently forcing its way into cracks rather than ascending through the central

orifice. In regions of intrusive rocks dikes are often found as offshoots of larger masses (Fig. 255).

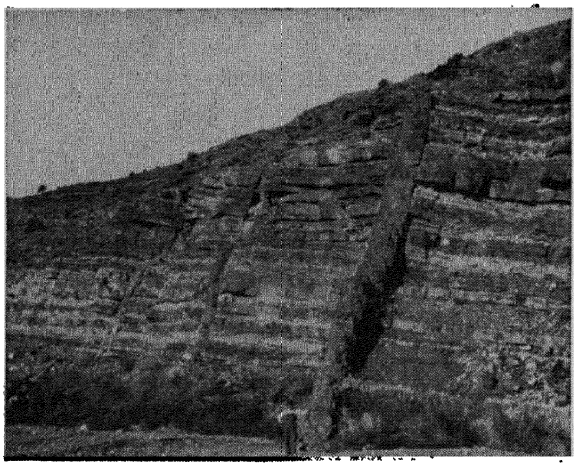


FIG. 254. *Three dikes of intrusive igneous rock cutting sedimentary beds. The man at the base of the largest dike gives the scale. Note that the beds on either side of the large dike do not match, suggesting that the dike was intruded along a fault (page 532) rather than a simple crack. (Photograph by Darton, U.S. Geological Survey.)*

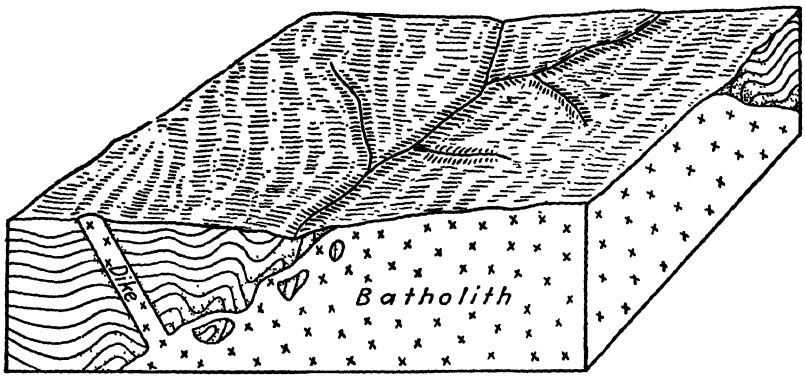


FIG. 255. *Diagram showing part of a batholith and a dike intruded into a series of sedimentary beds. Since it solidified, the igneous rock has been uncovered and partly removed by erosion. The dots indicate zones of thermal metamorphism at the intrusive contacts.*

Batholiths are very large bodies of intrusive rock which appear to extend downward indefinitely. Visible exposures of batholiths cover hundreds or thousands of square miles; the great batholith which forms the central part of the Sierra Nevada in California, for instance, is some 500 mi long and in places over 100 mi wide (Fig. 256). Just how far down into the crust batholiths actually extend we have no means of knowing, for all we ever see of these enormous masses is the upper part which has

been exposed by erosion. Granite is the principal rock in batholiths, although many have local patches of diorite and gabbro. Batholiths are always associated with mountain ranges, either mountains of the present or regions whose rock structure shows evidence of mountains in the distant past. This association, however, does not mean that batholiths are the *cause* of mountain ranges; we shall find presently that the intrusion of a batholith is merely one incident in mountain building.

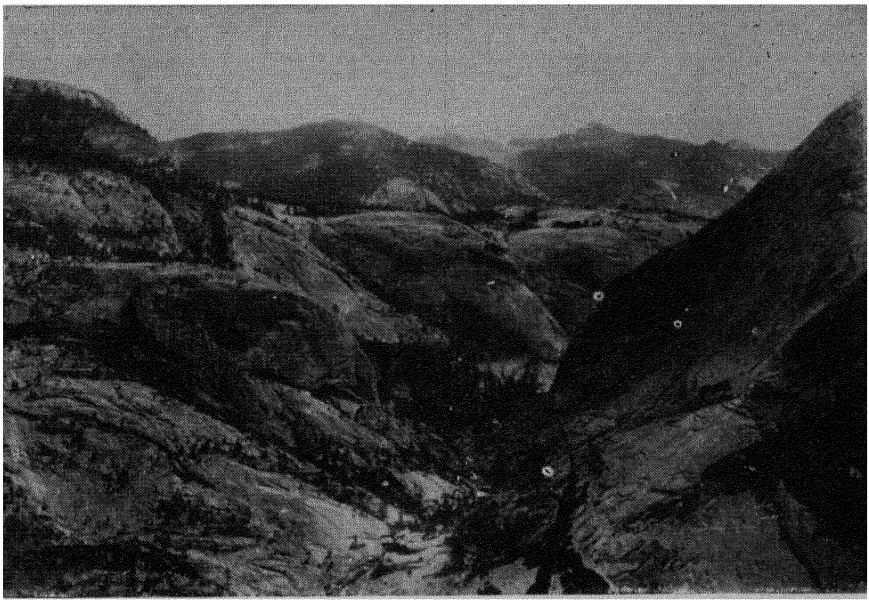


FIG. 256. *Mountain landscape developed by erosion in a granite batholith. Erosion has removed the rocks which originally covered the batholith and is now cutting deeply into the granite itself. The smooth, rounded surfaces indicate that glaciers have aided erosion in the past. Sierra Nevada of California.* (*Photograph by Gilbert, U. S. Geological Survey.*)

Batholiths are so large and remain hot for so long a time that they have a profound effect on rocks along their contacts. Mineral grains in the intruded rocks grow larger and lock together, minerals stable only at low temperatures disappear and new minerals are formed from them. Soft sedimentary rocks become hard and change their appearance completely; limestone changes to marble, sandstone to quartzite, and shale to a hard rock called *hornfels*. These metamorphic changes, produced by heat at the border of an igneous intrusion, are included in the term *contact metamorphism* or *thermal metamorphism*. The metamorphic rocks produced are typically unfoliated, since they have not been subjected to movement or directed pressure which would orient the newly formed minerals.

The magma which forms a batholith, like the magmas which come to the surface in volcanoes, contains a large quantity of dissolved gases. As the magma solidifies its gases are in part driven out into the surrounding rocks, in part concentrated under ever-increasing pressure in the residual, uncrystallized liquid. In the last stages of crystallization this liquid is largely hot water, containing in solution besides gaseous material many soluble, low-melting substances concentrated from the original magma. At length pressures become great enough to force the liquid out into cracks, not only in the nearly frozen batholith but in near-by rocks as well. Like ordinary groundwater, the hot liquid deposits some of its dissolved material as it goes, forming *veins.** The commonest material of these veins is quartz, but the hot water associated with batholiths sometimes has concentrates of commercially valuable metals and their compounds. Many important deposits of gold, silver, and copper have been formed in this way.

Unsolved Problems of Vulcanism

The subject of vulcanism bristles with unsolved problems, of which we shall note briefly two of the most important.

First, where does liquid rock come from? In the days when the earth was believed to be a molten sphere with a thin solid crust an answer to this question was easy: volcanoes and batholiths were formed where material from the fluid interior had worked its way into the crust. But today we have reliable measurements of the earth's rigidity and the speed of earthquake waves which prove that the earth is not liquid but largely solid, and that no really large mass of liquid can exist anywhere in the interior except possibly near the earth's center. The magma which forms a batholith or feeds a volcano must be a relatively small mass of liquid generated locally. So the problem is to discover some process in the earth's crust which can produce a body of liquid rock in one isolated region.

A possible answer is local release of pressure. We know that both pressure and temperature increase downward in the crust; it may well be that rock material at lower levels is at a temperature above its melting point and is kept from melting by the enormous pressures. So if some sort of diastrophic movement, some kind of readjustment in the upper part of the crust, could reduce the pressure locally, a part of the rock at depth might liquefy. This hypothesis is plausible, and it accounts neatly for the common association of batholiths and volcanoes with large-scale diastrophic disturbances, but it remains far from proved.

* Note the distinction between dikes and veins: a *dike* is molten rock which has filled a fissure and solidified, while a *vein* consists of material deposited along a fissure from solution in water.

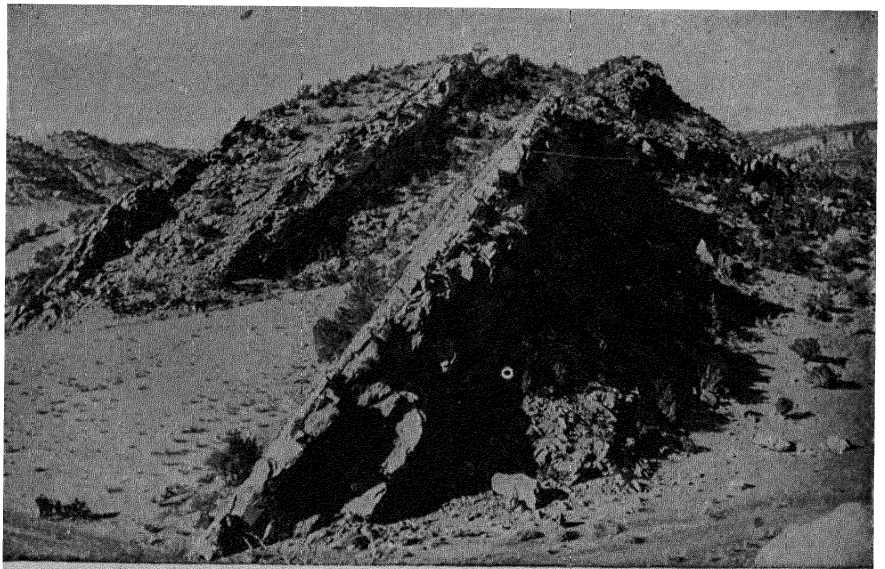
A second knotty problem is the manner by which liquid rock works its way upward through the crust. All kinds of igneous phenomena give evidence that magma is formed at lower levels than those where we now find it, and that its motion in the liquid state is dominantly upward. The motion is not hard to understand for the magma which forms dikes or that which escapes from volcanoes: the liquid rock moves toward a zone of lower pressure along fissures, probably widening the fissures as it moves. But the migration of the magma which forms batholiths is not so readily explained.

Certainly the magma does not simply move into open cavities, for large cavities at such depths in the crust can neither form nor be maintained. Sedimentary layers are often cut off sharply at a batholithic contact (Fig. 255), suggesting that emplacement of the magma is accomplished at least in part by removal of the preexisting rock. Perhaps the removal is accomplished by melting the rock, perhaps by the detachment and sinking of huge blocks through the liquid. At other batholithic contacts layers in the adjoining rocks have been squeezed and contorted into tight folds, which apparently mean that the batholith made room for itself in part by compressing and pushing aside the intruded rocks. Another possible way for a batholith to work its way upward would be to lift or dome up the rocks above it; this would be difficult to demonstrate, since we do not know that a batholith exists until erosion has worn away most of its cover. Just what part is played by these several processes in the movement of large bodies of magma remains an unanswered geologic riddle.

Evidences of Diastrophic Movement

Terra firma, the solid earth, has come to be a symbol for stability and strength. On foundations of rock man confidently anchors his buildings, his dams, his bridges. The massive rock of mountain ranges seems strong enough to withstand any conceivable force which might be exerted upon it. Yet even casual observation shows at once how naïve are our simple notions of the earth's stability. Entombed in the strata of high mountains we find shells of marine animals, shells which can be there only if solid rock formed beneath the sea has been lifted high above sea level. Sedimentary rocks, which must have been originally deposited in horizontal layers, are found tilted at steep angles or folded into arches and basins (Fig. 257). Other layers have broken along cracks, and the fractured ends have moved apart (Fig. 258). Despite superficial appearances, there must be gigantic forces in the crust capable of lifting, bending, crumpling, and breaking even the strongest rocks.

Nor can we conclude that these forces have racked the earth's crust only in some remote prehistoric age and that our planet since then has



IG. 257. *Steeply tilted sedimentary beds near Gallup, N. M. (Photograph by N. H. Darton, U. S. Geological Survey.)*

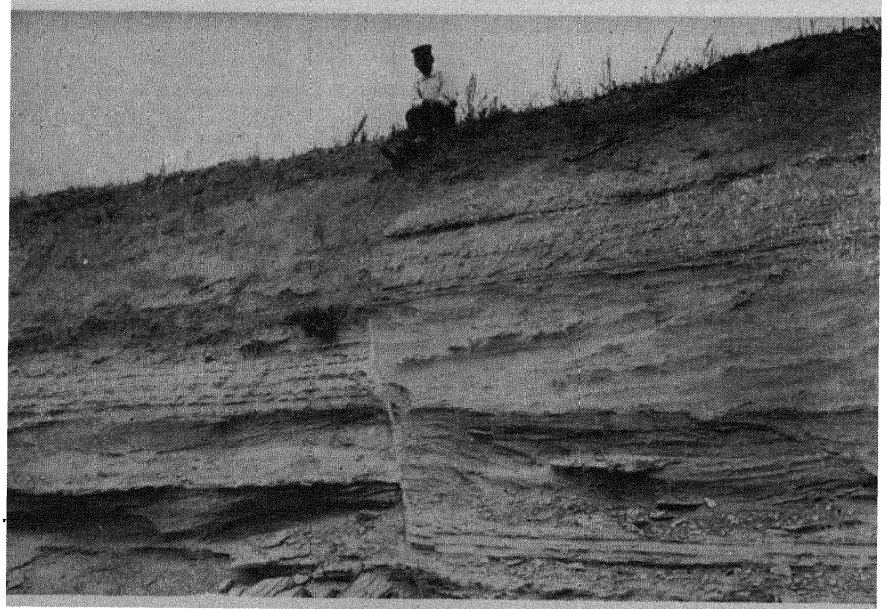


FIG. 258. ***A small fault. The sedimentary beds have been displaced along a fracture. (Photograph by Darton, U. S. Geological Survey.)***

acquired its apparent stability. Careful observation shows clearly that today as in the past slow diastrophic movements are changing the earth's landscapes.

Major earthquakes, for instance, often give evidence of permanent displacements in the crust: cracks open in the ground, and the material on one side of a crack may shift up, down, or sidewise with respect to the other side. At Yakutat Bay, Alaska, the shoreline was raised 40 ft during

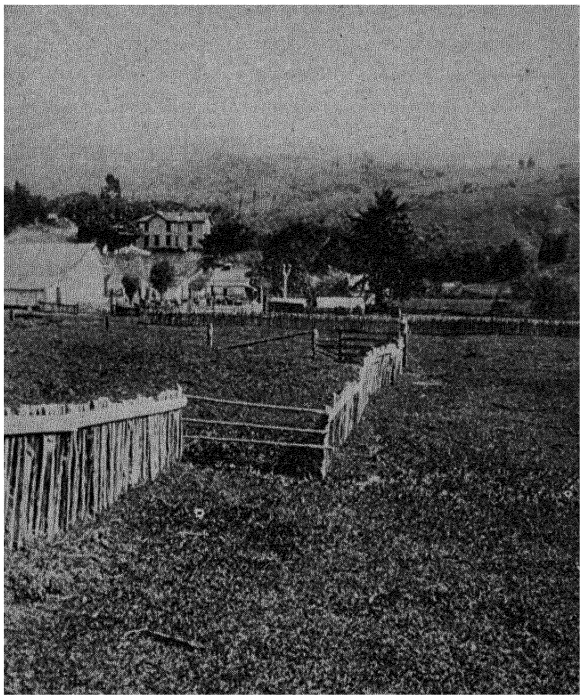


FIG. 259. *This fence near San Francisco was straight before the 1906 earthquake. During the quake it was broken and displaced sidewise. The gap has been temporarily repaired.*

an earthquake in 1899. In the San Francisco quake of 1906 fences and roads were displaced sidewise along the line of fracture, some by as much as 20 ft (Fig. 259). These are relatively small movements, of course, but a succession of earthquakes in the same region may in time produce large displacements.

Shore-line features often show evidence of recent shifts of sea level. One of the most striking is the ruined temple of Jupiter Serapis, near Naples, built in Roman times (Fig. 260). When the base of the temple was excavated a century or so ago, the bottoms of the three columns left

standing were found riddled with holes made by a boring clam that lives in the near-by Mediterranean Sea. The temple was obviously built on dry land, and stands today on dry land, but the clam borings show clearly that at some time during the past 2,000 years this region has been several feet below sea level.

In Scandinavia accurate records of the shore line kept for over a century show that the land is steadily rising out of the Baltic Sea, in places by several millimeters per year.

Elevation of a coast line in the not too distant past is often shown by a wave-cut cliff and terrace high above the present shore (Fig. 261). Such a feature does not necessarily mean movement within historic time; but geologically the rising of the land or sinking of the sea must be fairly recent events, for stream erosion will quickly obliterate cliff and terrace from the landscape. Recent sinking of the land with respect to sea level is most clearly shown by long, narrow bays filling the mouths of large stream valleys. A body of water like Chesapeake Bay (Fig. 262), for instance, could not be formed by wave erosion, since wave attack normally straightens a coast line rather than deeply indenting it; the shape of the bay suggests that it lies in a stream-carved valley whose lower part has been submerged beneath the sea.

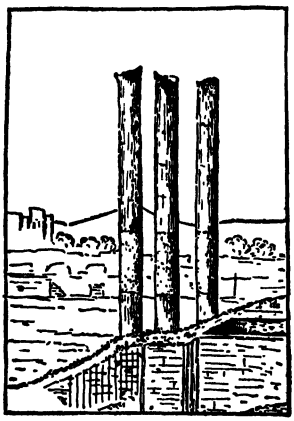


FIG. 260. *Ruins of the temple of Jupiter Serapis near Naples. (From Geology by Emmons, Thiel, Stauffer and Allison.)*

From evidence of this sort, evidence which could be multiplied a thousand times, we must conclude that the earth's solid outer shell is not as substantial as it seems, but that it has yielded and is yielding to powerful forces of deformation.

Kinds of Diastrophic Movement

Three kinds of movement in the solid crust have been suggested in the preceding discussion: (1) movement involving the slippage of rocks along a fracture or *fault;* (2) movement resulting in the formation of *folds;* (3) broader, regional movements of uplift, subsidence, or tilting.

A ***fault*** is any surface along which movement has taken place. All kinds of rock are cut by numerous fractures, but only those fractures which show definite evidence that one side has moved with respect to the other are called faults. In an outcrop a fault is commonly recognized as a fairly straight line against which sedimentary layers and other structures end abruptly (Fig. 258). Near the fault, layers may be bent

or crumpled, and along the fault itself streaks of finely powdered material may have developed from friction during movement.

Faults are distinguished by the direction of relative movement along them. Three important kinds are illustrated in Figs. 263, 264, and 265. (1) A *normal fault* is an inclined surface along which the rocks above the surface have slipped down *with respect to* those below the surface. Note that this kind of movement requires that the strata fill a greater hori-

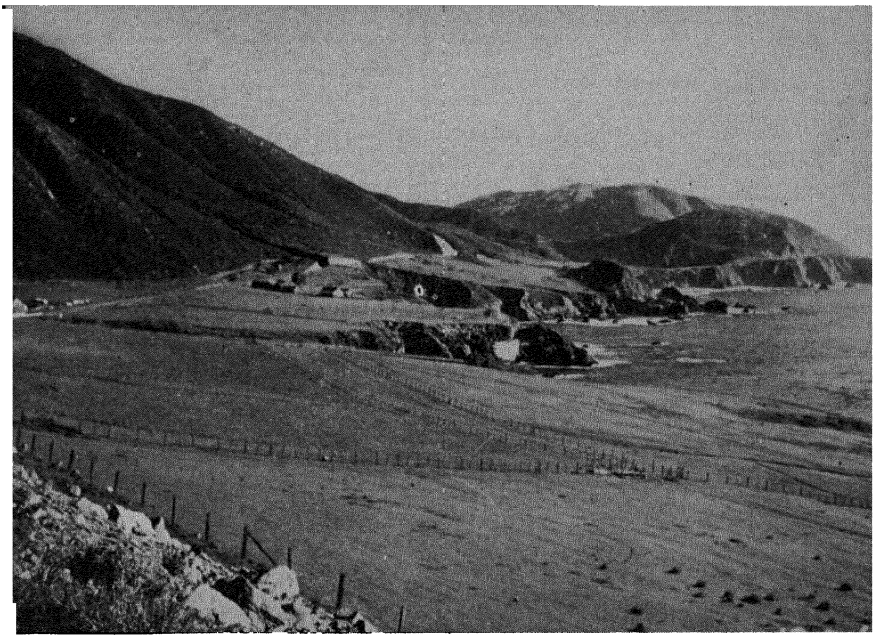


FIG. 261. *An elevated beach near Monterey, Calif. The former shore line was at the base of the steep, brush-covered slope. Both the wave-cut cliff and the terrace have been modified somewhat by stream erosion since the uplift. (Photograph by Eliot Blackwelder.)*

zontal area after faulting than before; in other words, normal faults would be expected in parts of the crust where the chief forces are of a *tensional* or "stretching-apart" nature. (2) A *thrust fault* is an inclined surface along which the rocks above the surface have moved up *with respect to* those below the surface. Thrust faulting decreases horizontal area, hence presumably is the result of *compressional* forces. Note the words "with respect to" in the definitions of normal and thrust faults: actually the lower block may have moved rather than the upper, but the name of the fault applies simply to relative movement; usually in practice it is impossible to tell whether one side or both sides were actually in motion. (3) A *strike-slip fault* is a vertical or nearly vertical

surface along which one side has moved approximately horizontally with respect to the other side.

Movement along faults usually takes place in a series of small sudden displacements, with intervals of years or centuries between successive jerks. The immediate topographic effect of displacement along a thrust fault or normal fault is the production of a small cliff (Fig. 266). Erosion attacks the cliff at once, and may obliterate it before the next movement.

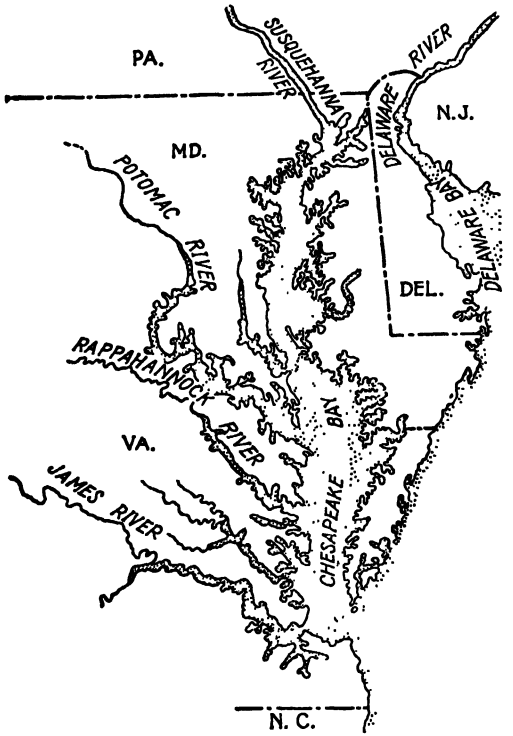


FIG. 262. *Drowned valleys on the Atlantic Coast. (From Introduction to Geology by Branson and Tarr.)*

If successive movements follow each other fast enough, erosion may not be able to keep pace with diastrophism, and a high cliff may be formed only slightly modified by erosion. Cliffs of this sort are called *fault scarps.* Good examples of scarps produced by normal faults are the steep mountain fronts of many of the desert ranges in Utah, Nevada, and eastern California (Fig. 221, page 493). A more deeply eroded scarp produced by thrust faulting is the eastern front of the Rocky Mountains in Glacier National Park.

The most famous strike-slip fault in this country is the San Andreas fault of California, movement along which caused the San Francisco

earthquake. Along much of its course the line of this fault is represented by a straight valley (Fig. 267). The valley is not a direct result of faulting but is due to rapid stream erosion in rock material fractured and pulverized by the fault movement.

Folds always result in a shortening of the crust (Fig. 268), hence are in general produced by compressional forces. Brittle rocks yield to compression by thrust faulting; less brittle and deeply buried rocks yield by folding. Not uncommonly major thrust faults are associated with intense folding. Two names often used to designate types of folds are *anticline*, referring to an arch or a fold convex upward (Fig. 269), and *syncline*, referring to a trough or a fold convex downward. In regions of intense folding, anticlines and synclines follow one another in long series.

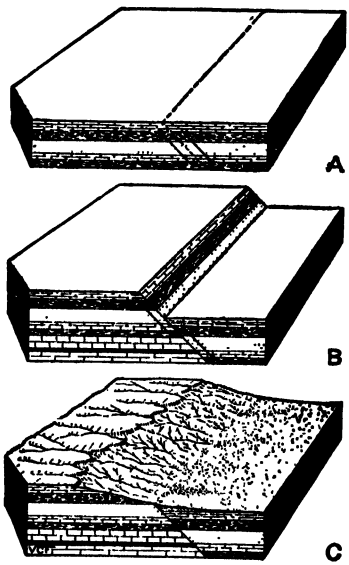


FIG. 263. *Diagram to show the development of a normal fault. A, strata before faulting; B, after faulting, erosion assumed negligible; C, modification of the fault scarp by stream erosion. Compare Fig. 221, p. 477. (From Elements of Geography by Finch and Trewartha.)*

Folding apparently takes place by slow, continuous movement, in contrast to the sudden displacements along faults. Sometimes folding produces hills and depressions in the landscape directly, but more commonly erosion keeps pace with folding and obliterates its direct topographic effects. Indirectly folds affect topography by exposing tilted beds of varying degrees of resistance to the action of streams, so that characteristic long, parallel ridges and valleys develop, like those of the

FIG. 264. *Diagram of a thrust fault. The dotted lines show the amount of material removed by erosion. (From Elements of Geography by Finch and Trewartha.)*

Appalachian Mountains (Fig. 229, page 500). In these mountains as in many others the actual folding is very ancient, the present ridges being due entirely to deep erosion after successive uplifts of the stumps of the old folds

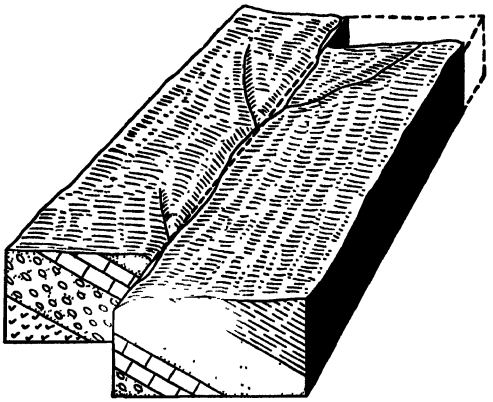


FIG. 265. *Diagram of a strike-slip fault.*

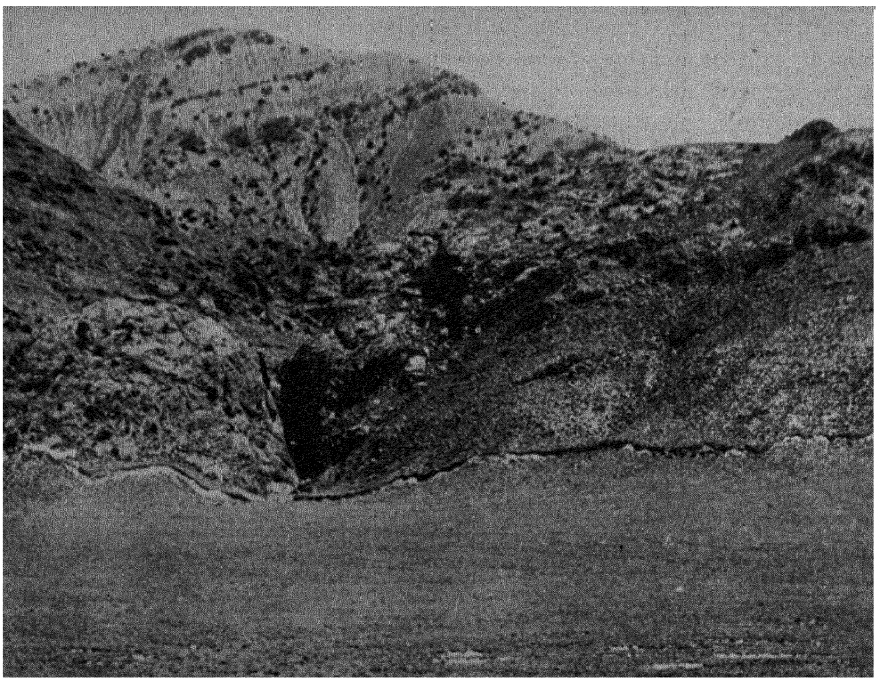


FIG. 266. *The small cliff just below the mountain front is a fresh scarp about 16 feet high formed during an earthquake in 1915. The mountain front itself is a scarp produced by a succession of older movements and modified by erosion. Sonoma Range, Nevada. (Photograph by Eliot Blackwelder.)*

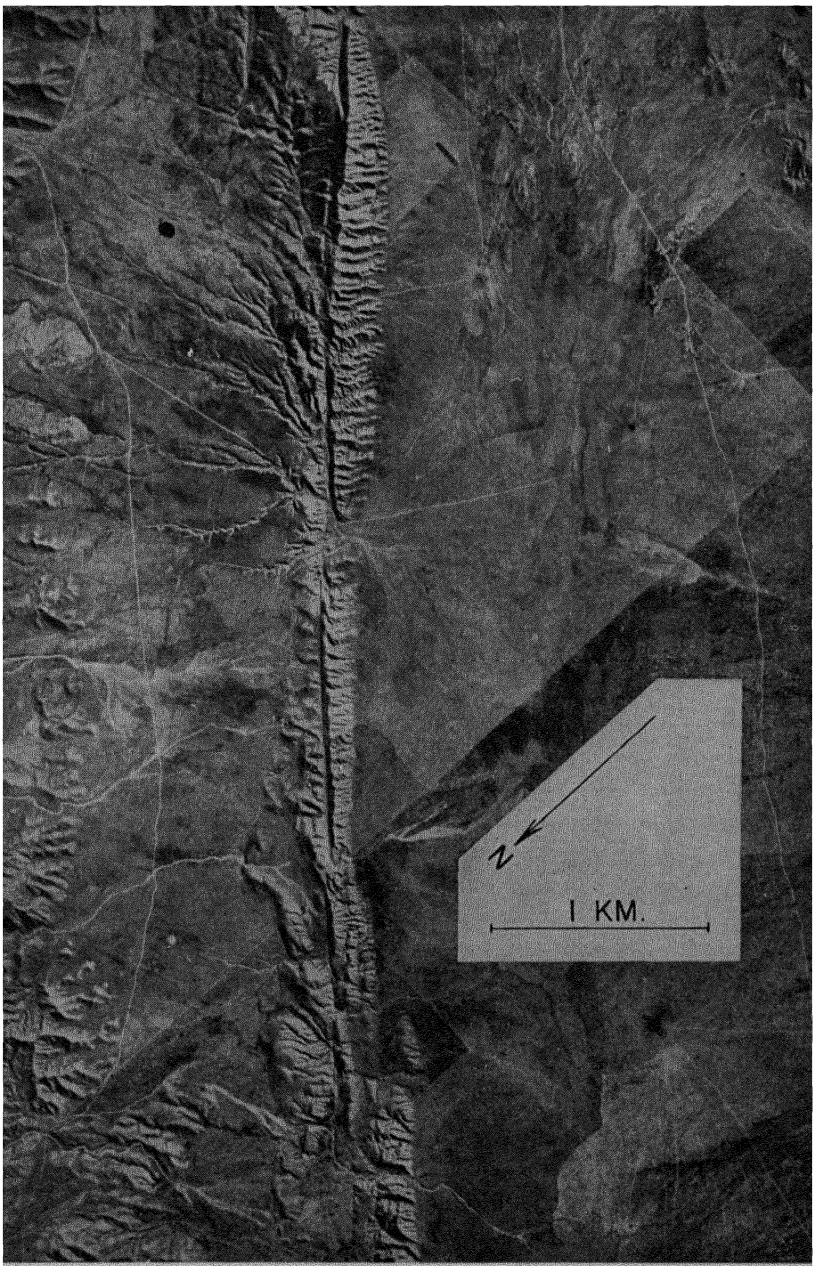


FIG. 267. *Airplane photograph of part of the long, straight valley produced by rapid erosion of pulverized material along the San Andreas fault, California. (Courtesy of the Barnsdall Oil Company and the Fairchild Aerial Survey.)*

Regional uplift and subsidence may involve whole continents or large parts of them. It may be a vertical movement; it may involve some tilting; it may mean broad, large-scale bending or warping. Just how the movement takes place is not clear; there may be faulting at the edges of the moving block, but the presence of faults usually cannot be demonstrated. The movement in general is imperceptibly slow.

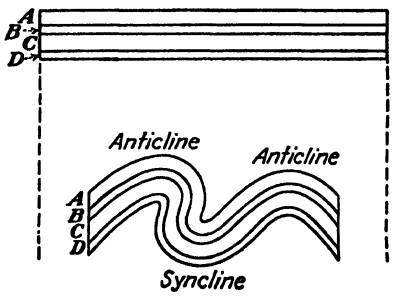


FIG. 268. *Cross sections showing effects of folding in horizontal strata.*

Subsidence creates a basin of deposition, in which marine sediments accumulate if the land drops sufficiently low, otherwise river and lake sediments. Uplift increases the vigor of stream erosion: an old-age landscape raised a few hundred feet is cut into by youthful streams, and the long history of valley and landscape development begins anew. Along shore lines recent uplift and subsidence are recorded in the raised beaches and drowned valleys mentioned in the last section.

In the landscapes and sedimentary rocks of this country a multitude of regional movements are recorded. To mention only a few of the more

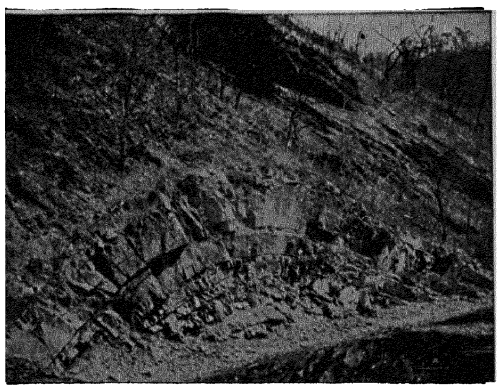


FIG. 269. *Part of a buried anticline which has been exposed in a stream valley. (U. S. Geological Survey.)*

obvious ones: Across the Mississippi Valley lie thick, horizontal beds of marine sandstones and limestones, recording a time when this part of the continent lay beneath the sea and a later time of uplift when the sea retreated. Shapes and patterns of river valleys in the Appalachians and the Rockies show that these great mountain systems have been uplifted repeatedly since the original mountains were formed. The thick accumula-

tions of sediment on and near the Mississippi Delta and in the central valley of California suggest that these are regions where subsidence is in progress.

Earthquakes

Earthquakes, most dreaded and most destructive of natural phenomena, are exceedingly rapid vibratory motions of rock near the earth's surface. A single shock usually lasts no more than a few seconds, but in that time may take a terrific toll of property and human life. The rapidity of the vibrations rather than the actual amount of motion is responsible for the damage: rigid man-made structures are shaken to pieces because they are unable to follow the fast back-and-forth motions of the underlying rock. Unlike most volcanic eruptions, earthquakes come without

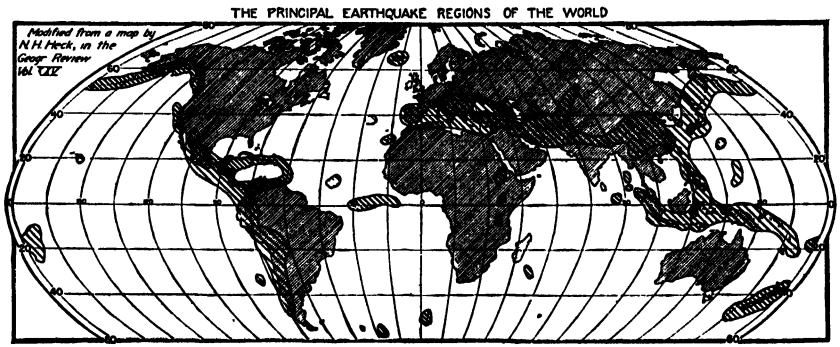


FIG. 270. (*From Elements of Geography by Finch and Trewartha.*)

warning. Usually the first shock is the most severe, disturbances of gradually slackening intensity following at frequent intervals for days or months afterward. A major earthquake may be felt over an area of many thousands of square miles, but its destructiveness is limited to a much smaller area. Small quakes shake some part of the crust daily, but destructive shocks are commonly separated by intervals of months or years.

Earthquakes are produced in a number of ways—by landslides, artificial explosions, movement of magma before and during volcanic eruptions—but by far the greater number have a common cause in the sudden dislocation of solid rock along faults. The rapid motion acts like an explosion, sending vibrations in all directions into the surrounding material. We have already (page 447) discussed the ability of these vibrations to travel through the earth and the information which they yield concerning the earth's internal structure.

Regions in which severe earthquakes are comparatively frequent include the mountain chains fringing the Pacific and a broad belt extend-

ing from China across southern Asia into the Mediterranean basin (Fig. 270) Major earthquakes have occurred sporadically elsewhere, but the greater number have been concentrated in these zones. In or close to the earthquake belts lie most of the world's active volcanoes, a fact which suggests that there may be some remote connection between the two phenomena.

Geologists will do a great service to humanity when they learn to predict the occurrence of earthquakes with some show of accuracy. Present knowledge of the subject is far too meager for any attempt at prediction. All that is known for certain is that the next earthquake may affect any part of the world, but the probability of its occurrence is far greater in the earthquake belts. In recent years very accurate surveys along active faults have indicated that slight movements occur between quakes; there is some hope that a study of such movements and the strains they set up may eventually lead to a method of earthquake prediction.

Dynamothermal Metamorphism

Pressures resulting from diastrophic movement, especially at depths where temperatures are high, may produce conspicuous changes in the rocks undergoing deformation. The pressures are in general *directed* pressures, more intense in one direction than in others. In response to such pressures new minerals develop, particularly those which can grow in long needlelike forms or those like mica which grow in flat, platy crystals. The needles and flat plates grow in directions determined in part by the maximum pressure, in part by structures within the rock, hence arrange themselves in roughly parallel layers to give *foliated* rocks like slates, schists, and gneisses. Because these rocks show the effect of both heat and pressure, the process of alteration is called *dynamothermal metamorphism*, in contrast to *thermal* metamorphism produced by heat alone at the borders of igneous intrusives. Foliated rocks produced by dynamothermal metamorphism are abundant in any region which has undergone intense folding followed by deep erosion; parts of New England and much of northeastern Canada are good examples.

Shale is particularly susceptible to dynamothermal metamorphism, since its chief constituents, the clay minerals, alter readily to mica even at relatively low temperatures and pressures. The first step in the alteration of shale is slate, whose shiny cleavage surfaces betray the parallel arrangement of myriads of tiny mica flakes. More intense heat and pressure convert slate into coarser grained mica schist; under extreme conditions part of the mica is changed to feldspar and the rock becomes a gneiss. Pure limestone and pure quartz sandstone subjected to dynamothermal metamorphism do not give foliated rocks, since their composition does not permit

the formation of minerals with platy or needlelike crystals; as in thermal metamorphism, these rocks change to marble and quartzite, respectively. Volcanic rocks commonly become schists and under extreme conditions gneisses. Granite may change directly to gneiss. These various changes are often clearly evident in the field, a layer of shale, for instance, altering within a few thousand feet to slate and then to schist as the region of greater deformation is approached.

Rocks which have undergone dynamothermal metamorphism often show by their crumpling and complex folding that there has been much movement not only between separate layers but within the solid rock itself (Fig. 271). Original structures such as sedimentary bedding and fossils have small chance of preservation. Parts of the rock may show

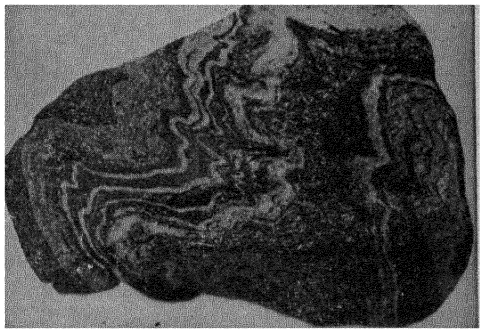


FIG. 271. *Crumpled gneiss. One-half natural size.* (*From Introduction to Geology by Branson and Tarr.*)

streaks of finely granulated material produced by the crushing of one mineral grain against another; by bringing particles of different substances into intimate contact, such crushing doubtless speeds up the chemical reactions involved in the production of new minerals. Despite the intense deformation, however, these rocks ordinarily *show no evidence of melting* during metamorphism. The processes involved are chemical reactions between solid particles, aided by crushing and probably by small amounts of water, but not, except locally, by actual fusion of the rock material.

This is an important point. Dynamothermal metamorphism is in general a process taking place several miles below the surface, and its products accordingly give us valuable information about conditions at those depths. We find that rocks here are hot enough and under sufficient pressure to be readily deformed; they will crumple and flow like cold pitch squeezed slowly between the jaws of a vise, and their mineral composition may be profoundly altered, but they remain for the most part solid.

Causes of Diastrophism

What are these irresistible forces which can twist and break the strongest rock? Why should there be tensional forces in one part of the crust, as shown by normal faults, while in another part compression is indicated by folding and thrust faults? Where do the forces originate which can raise and lower continental masses vertically? Why have not forces in the crust long since reached equilibrium? With questions like these we reach an impasse. Geologists know a great deal about what forces in the crust can do, but why these forces exist remains one of the great unanswered problems of physical science.

A partial explanation of vertical movements is given by the hypothesis of *isostasy*. This hypothesis rests on the well-established idea that rocks below a depth of a few miles are hot enough and under sufficient

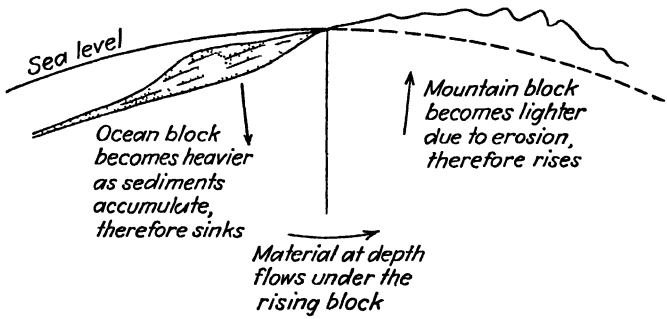


FIG. 272. *The hypothesis of isostasy.*

pressure so that they will flow if subjected to any considerable difference in pressure. Hence if the weight on a large segment of the earth's surface —a mountain range, a broad valley, an ocean basin—is markedly increased, it should sink, forcing the rock material beneath it to flow out under adjacent segments where the pressure is smaller. Thus the pressures beneath large blocks of the crust must all be substantially the same, any differences which come to exist being slowly adjusted by rock flow at depth. This means, of course, that the rock material of mountain ranges must have a lower density than the material below ocean basins, since mountains project farther from the earth's center—a conclusion confirmed by gravity measurements and speeds of earthquake waves. Now a mountain range is continually undergoing erosion, and the eroded material is largely deposited in adjacent valleys or oceans. The mountain block accordingly becomes lighter and the valley block or ocean block becomes heavier as the weight of sediment increases. Eventually the difference in pressure beneath the mountain range and the adjacent blocks becomes too great for the material at depth to withstand, so that the mountain range rises and the basins of deposition sink until the

isostatic equilibrium is restored (Fig. 272). The situation may be duplicated on a small scale with a large cork and a small cork floating in water: if material is removed from the top of the large one and placed on the small one, the former will rise a little out of the water and the latter will sink.

The hypothesis of isostasy accounts well enough for the long life of mountain ranges and for the repeated vertical uplifts which they undergo, but it offers no explanation for the folding and faulting which form a range initially. All horizontal movement in the crust and probably much vertical movement as well must have some other cause.

An attractive explanation for horizontal compressional forces has long been the supposed shrinking of the earth's interior. Presumably ever since the earth's beginning its interior has been radiating heat out through the crust into space. This loss of heat must lead to a decrease in volume, to which the outer shell would accommodate itself by wrinkling and thrust faulting, much as the skin of an orange wrinkles when the interior loses water. Provided the loss of heat is great enough, this mechanism would provide a sufficient force for much diastrophic activity, but grave doubts have recently been expressed by physicists that the rate of heat loss is anywhere near adequate. Measurements and calculations of the heat loss are still not accurate enough to give a final answer to the problem.

Tidal forces, forces due to horizontal sliding of whole continents, changes in the position of the earth's axis have all been suggested as possible causes of diastrophic movement, but none of them has been shown to provide forces of sufficient intensity.

Questions

1. Which of the following rocks might you expect to find (*a*) in lava flows, (*b*) in dikes, (*c*) in a batholith?
   rhyolite marble andesite obsidian
   diorite basalt granite conglomerate
2. What kind of rock would you expect to find as the chief constituent of lava flows from a volcano whose eruptions are dominantly of the explosive type? What kind in flows from a volcano of the quiet type? Why?
3. What characteristic topographic features do active volcanoes produce? From what topographic features could you conclude that volcanoes had once been active in a region where actual eruptions have long since ceased?
4. List all the evidence you can for each of the following statements:
   a. Granite is an igneous rock.
   b. Mica schist is a rock which has been subjected to nonuniform pressure.
   c. Compressional forces exist in the earth's crust.
   d. Diastrophic movement is going on at present.
5. Suppose that you are studying a series of horizontal layers of shale and sandstone at some distance from the border of a batholith If you should follow the layers toward the batholith, what changes in the rocks would you expect to observe?

What other evidences of igneous activity might you find before you reached the actual contact?

6. Suppose that you find a nearly vertical contact between granite and sedimentary rocks, the sedimentary beds ending abruptly against the granite. How could you tell whether the granite had intruded the sedimentary rocks or had moved up against them after its solidification by diastrophic movement along a fault?
7. When stream erosion has been active for a long time in a region underlain by folded strata, what determines the position of the ridges and valleys? Explain how a valley may be formed at the crest of an anticline.
8. Figure 273 is a diagrammatic cross section through a part of the Appalachian Mountains. Point out (*a*) an anticline, (*b*) a syncline, (*c*) a thrust fault.

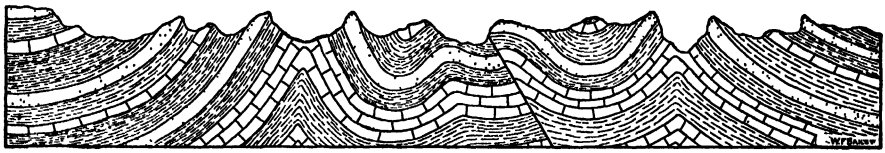


FIG. 273. *Cross section through part of the Appalachian Mountains. (From Introduction to Geology by Branson and Tarr.)*

9. Where would you expect to find the wider zone of thermal metamorphism, at the contact of a dike or a batholith? Why?
10. Indicate whether each of the following rocks is formed by thermal or dynamo-thermal metamorphism (or either). Indicate also from what kind of igneous or sedimentary rock each might be produced.
    mica schist quartzite slate marble gneiss
11. Distinguish between the foliation of a metamorphic rock and the stratification of a sedimentary rock.

CHAPTER XXXVII

The Law of Uniform Change

For several chapters now we have been discussing the materials of our planet and the natural processes by which those materials are altered. There is nothing fundamentally new in either the materials or processes. Elements and compounds, solids, liquids, and gases, forces, pressures, and temperature changes—these are all the familar subject matter of chemistry and physics. Only the point of view is different: here we consider substances as they occur in nature rather than the carefully purified materials with which a chemist deals, natural events rather than the controlled processes of laboratories. This point of view is essentially a geological one.

Geology, the earth science, is in large part an application of physics, chemistry, and biology to the earth. To pursue further our study of rocks would require knowledge of the chemistry of silicates and the physics of crystal structure. To learn more about the earth's interior would require detailed study of wave motion in solids and the behavior of materials under high temperatures and stresses. To delve into the lore of fossils would necessitate an extensive acquaintance with some branches of biology. In all of its many aspects geology leans heavily on other sciences.

But in one important sense geology is more than an application of other fields of knowledge: it deals, more than any other science, with the problem of past time. A geologist is concerned not alone with relationships among processes and substances in the present-day earth, but with the remote origins of earth materials and the changes which they have undergone. Concerning a mountain, a chemist might inquire: "What are its rocks made of? How is their composition changing on exposure to the atmosphere?" A physicist would be more curious about the strength of its constituent rocks, or about changes in the force of gravity near the mountain. To a biologist the mountain's chief attraction might be the plant and animal assemblages he would find at different elevations.

But a geologist would regard all these matters as secondary to the questions: "Why is the mountain here? How can its past history account for its present peculiarities of shape? When was the mountain formed, and what sort of landscape preceded it?"

Already in several discussions we have had occasion to mention the earth's great age and to explain landscapes and rock structures by slow processes acting through long ages. These ideas about past time, so fundamental in modern geology, have been generally accepted among scientists for less than a century. In this chapter we shall go back into history to see how concepts of the earth's past have changed and how modern ideas have been justified.

The "Catastrophic" Hypothesis

The emphasis of geology on events in the earth's past hindered early development of the science, for the simple reason that scientific thinking along such lines quickly ran afoul of firmly established ideas. The book of Genesis, for instance, tells the story of the earth's beginning with beauty and simplicity and an uncomfortable preciseness as to dates. So carefully are events dated in the biblical account that a learned seventeenth-century theologian, Bishop Ussher, found it possible to compute that the creation of our planet from formless void took place at 9 o'clock in the morning of October 12, 4004 B.C. Now the most casual dabbling in geology shows that changes in climate and sea level clearly recorded in the present landscape cannot possibly have occurred in the space of a few thousand years; yet the best efforts of many early investigators were nevertheless spent in trying to make their findings square with a literal interpretation of Genesis.

Even apart from biblical preconceptions, many of the notions current among educated people two centuries ago regarding the earth's history seem to us fantastic. Today it is commonly accepted that most valleys are formed by stream erosion; in the eighteenth century they were often ascribed to the downward slipping of fragments of the earth's crust. Mountains were believed to have risen by gigantic upheavals in some dim, chaotic stage of the earth's beginning. References to ages of universal flood and universal fire are frequent in literary works.

Newton's spectacular success, near the close of the seventeenth century, in reducing the complex motions of the solar system to order and simplicity proved a mighty spur to eighteenth-century scientists to find some similar order and simplicity in the phenomena of the earth's surface. Lacking Newton's genius and faced with a far more difficult problem, these lesser philosophers too often let imagination supplant patient observation. One who stood out was a German, Abraham Gottlob Werner. Werner was primarily a mineralogist and teacher, a teacher of

such amazing talent that scholars flocked from all Europe to his small provincial school at Freiberg, there to imbibe not only knowledge but enough of the master's burning enthusiasm to make them ardent supporters of his doctrines when they returned to their own countries. For his researches in mineralogy and for the enthusiasm he inspired, geology owes Werner a great debt, but his misguided theory regarding the earth's history was a stumbling block in the establishment of geology as a logical science. Near Freiberg, Werner found granite overlain by folded, somewhat metamorphosed rocks, these in turn overlain by flat sedimentary beds. Untraveled, deaf to the reports of others, Werner considered this sequence world-wide. Each of the three types of rock, he imagined, was deposited by a universal ocean, granite precipitating first and the flat upper beds last. Thus all rocks, in Werner's system, were sedimentary rocks; and the history of the earth consisted of three sudden precipitations from a primeval ocean, followed by the disappearance of most of the water.

Another man of altogether different stamp who greatly influenced geology at the beginning of the nineteenth century was the French biologist Georges Cuvier. Primarily an anatomist, Cuvier extended his work from animals of the present to the creatures whose fossilized shells and bones he found in great abundance in the rocks near Paris. This was the first careful, critical study of fossils, the first attempt to catalogue the precise similarities and differences between living forms of the past and present. Cuvier was too careful a scientist to indulge his fancy in unlicensed speculation, but his work on fossils led him to one important theory concerning the earth's past: in successive rock layers of the Paris region he found distinct assemblages of animals, different from each other and from the present animals of that region; he concluded that each assemblage appeared on the earth as the result of a special creation, and that each was destroyed by a universal cataclysm before the next creation. Thus Cuvier regarded the earth's history as a succession of catastrophes, separated by intervals of more normal conditions.

The theories of Cuvier and Werner have this much in common: both suppose that the history of our planet has been punctuated by tremendous events which have no counterpart in the natural processes which we observe today. No modern ocean has ever suddenly precipitated great masses of rock over its entire basin, as did Werner's universal ocean, nor has any modern animal assemblage been suddenly created or destroyed. Because both theories involved events, or "catastrophes," which apparently violated the natural laws of the present world, their central idea came to be known as *catastrophism*. A similar idea, of course, pervades all the older notions of world-wide fire and flood, of the formation of mountains and valleys in a time of chaotic movement in the earth's crust.

Hutton and Lyell

First to combat the catastrophic ideas was a Scotsman, James Hutton, greatest of the eighteenth-century philosophers who sought to imitate Newton by establishing a complete and coherent "theory of the earth." More clearly than his predecessors, Hutton recognized the erosive work of streams in carrying material from the land into the sea. Opposed to the wearing down of the land, Hutton believed, was the formation of rock from ocean deposits by heat and pressure, and the gradual reelevation of this new rock above sea level. Thus the cycle was continuous: rocks were formed from deposits in the sea, raised up to make land, worn away by streams, and again deposited in the ocean. For every step of this process Hutton believed he could find evidence in the present-day world. The biography of the earth needed no catastrophes, but, as far back as the rocks gave evidence, consisted of repetitions of the same cycle. "In the phenomena of the earth," Hutton says in concluding his book, "I see no vestige of a beginning, no prospect of an end."

In detail, Hutton's views are often as erroneous as Werner's. Especially did he go astray in that part of his cycle which deals with the formation of rocks. More traveled than Werner, Hutton paid more heed to volcanic phenomena, hence to the part which heat and pressure play in rock formation. Granite and basalt he recognized as igneous rocks, formed by solidification from the liquid state rather than as precipitates from solution in sea water. But his idea carried him too far: all rocks, even such sedimentary rocks as limestones and sandstone, owed their hardness to pressure, partial melting, and infiltration of hot liquids. He made no clear distinction between igneous and metamorphic rocks, simply regarding heat and pressure beneath the earth's surface as essential for the formation of all rocks.

Publication of Hutton's work precipitated a bitter feud with the disciples of Werner. Hutton died (1797) before the fight had fairly begun; but even had he lived, his retiring nature and ponderous style of writing would have made him a poor match for the brilliant German. Happily the battle was undertaken by Hutton's friend and pupil, John Playfair, whose writing was as lucid and persuasive as his master's was dull and obscure. Playfair's great work, *Illustrations of the Huttonian Theory of the Earth*, cites one clear observation after another supporting Hutton's ideas, argument after argument reducing Werner's claims to absurdity. Werner's many disciples in England of course struck back, and the quarrel continued for more than a decade. In the end neither side could claim full victory; rather, from a combination of the opposing views emerged a clear understanding of the difference between rocks which were once liquid and rocks formed from sedimentary deposits.

Perhaps the greatest contribution of Hutton and Playfair to geology was their emphasis on the importance of direct observational evidence. After Hutton came no more philosophers content to fabricate grandiose speculations about the earth's past; instead there came to the fore, especially in England, a group of hard-headed observers who realized that what geology needed most was a careful gathering of facts.

An extreme example of the dominance of fact-gathering over theory in the early nineteenth century is William ("Strata") Smith, a quarryman and surveyor, whose lack of education was partly compensated by immense patience and acuteness of observation. In the wandering over England which his business made necessary, Smith studied carefully the fossils of different localities. Like Cuvier, Smith observed that fossil assemblages differed from one rock stratum to another; and he presently discovered, as Cuvier had not, that the sequence of fossil assemblages is the same in different localities, so that each stratum can be followed from place to place, even if it is partly covered with soil and vegetation, by means of the fossils it contains. Eventually Smith succeeded in following several different strata over the whole of England and Wales. Sometimes the layers were flat, sometimes arched or tilted, but always they followed one another in regular order, and always they could be recognized by their fossils. To show this sequence of strata, Smith colored a map of England to indicate where each layer appeared at the surface. This was the first geologic map. Smith drew no theoretical conclusions from his important observation, but it has since become the foundation of geologic time reckoning, and his method of mapping is now employed by geologists the world over.

FIG. 274. *Charles Lyell* (1797–1875).

During the long controversy with Werner, Hutton's more far-reaching principle, that events of the earth's past can be adequately explained by forces at present in operation, remained all but forgotten. Catastrophism in one form or another was the dominant belief, even among Hutton's defenders. But by 1830 a great mass of evidence had accumulated to make catastrophism untenable, at least in its more extreme ideas. The time was ripe for a man of fresh outlook, with a mind capable of digesting the accumulated observations and reading from them a defense of Hutton's principle.

Hutton's new champion was Sir Charles Lyell (Fig. 274), one of the greatest of geological thinkers. Trained as a lawyer but from boyhood consumed with a passion for natural history and geology, Lyell made it

the business of his life "to explain the former changes of the earth's surface by forces now in operation." Lyell reemphasized Hutton's point that processes which seem incapable of altering the landscape appreciably may produce great changes if given sufficient time. If the earth's history is confined to a few thousand years, then catastrophism is indeed necessary; but if geologic time extends back for millions of years, the slow processes which we find at work today are sufficient to account for the earth's eventful past. In extensive travels through Europe and America Lyell sought always for positive data regarding geologic changes of the present—the shifting of stream courses, the building of deltas, advances and retreats of shore lines, the outpouring of lava and ash from volcanoes. Then from rocks and landscapes he read the slow accumulation of these changes through the long past—rivers now cutting far below their former channels, coasts with remnants of old beaches high above the present shore, immense lava flows where no volcanoes exist today. Lyell was the first to distinguish metamorphic rocks as a group, and so to understand clearly the cycle of changes by which rocks are formed, destroyed, and re-formed. He continued the work of Cuvier and his English followers in studying the succession of animals whose remains are entombed in the rocks of France and England. Nowhere did he find evidence of past changes brought about by agencies other than those in the world about him. Lyell's immense accumulation of facts and observations was overwhelming; within a few years after his *Principles of Geology* was published, his friends, at first skeptical, had accepted his views.

One last step was supplied in 1859, when Darwin introduced the theory of evolution into biology. Not only changes in the inorganic world of rocks, but, according to Darwin, changes in living things as well, could be explained in terms of processes operating in the present world. Thus the distinct assemblages found by Cuvier in the Paris basin were not special creations but stages in a continuous line of development. Lyell, though by this time well over sixty, grasped at once the importance of Darwin's work and became one of his earliest and most active supporters. The theory of organic evolution knocked the last prop from the outmoded idea of catastrophism.

The Modern Concept of Uniform Change

The law of uniform change is simply Lyell's thesis that *past changes of the earth's surface are adequately explained by processes now in operation.* This principle is as fundamental to geology as the law of conservation of energy is to physics or the periodic law to chemistry.

Proposed at first as an antidote to extreme catastrophism, Lyell's earliest concept of uniform change was extreme in the opposite direction, but in later years he modified his position considerably. Modern geologists

interpret the law very broadly. They hold with Lyell's basic tenet that processes now in operation are sufficient to account for changes in the past, but they find good evidence that these processes have not always operated with their present intensity. For example, climates of the world have been colder at some periods, so that glaciers were more widespread; volcanic activity has at times been much greater than at present; mountains have been fewer and lower, so that stream action has been less effective; widespread desert conditions and periods of intense cold have at times wiped out whole races of animals and hastened the development of other forms. Thus our modern concept is a sort of compromise with catastrophism: although present-day processes can adequately explain the past, certain combinations of these processes have at times made the face of the earth very different from its present aspect.

Tacit acceptance of uniform change in future chapters should not obscure the fact that this law, like all scientific laws, is no more than a generalization based on long and careful observation. Just as the law of conservation of energy rests on the circumstance that all known physical processes conform to it (except nuclear changes), so the principle of uniform change depends on the failure of long searches to find any good evidence for an opposing view.

Many people outside of scientific circles still prefer to believe in some form of catastrophism. Such a belief cannot be proved wrong; clearly, a sufficient number of local catastrophes and special creations would be capable of accounting for the present form of the earth's surface. But if advocates of this view were to consider *all available evidence* (as ordinarily they do not), they would find necessary an enormous number of assumptions. They must assume catastrophes of different intensities and different durations, occurring haphazardly at different times in different regions, they must assume innumerable special creations of organisms very much like earlier organisms, but differing in minute details; to make their hypothesis complete, they must find some reasonable assumption to explain the occurrence of events so different from those in our experience, and further to explain why these extraordinary events have left no clear record of themselves in the rocks. Now such a body of assumptions is of course permissible; a scientist would object to the hypothesis not because it is *wrong* in any absolute sense, but because it is so absurdly complex. The idea of uniform change is accepted in scientific circles because it involves the fewest awkward assumptions, hence gives the *simplest* explanation for all available facts.

One conspicuous exception to the idea of uniform change is the cosmic event in which our solar system was formed. Nowhere in the universe do we observe planets being formed at present, or even any process which might reasonably be expected to form planets. We can, of course, preserve

the letter of the law by reasoning that the sun's family could have developed by a collision of two or three stars acting under forces now in operation; but the event was certainly a catastrophe in the sense that it was an extremely violent and unusual occurrence, changing the sun's history abruptly and profoundly. How far back in the past toward this catastrophic event the law of uniform change can be carried without great modification is problematical. Presumably in its earliest stages the earth's surface was shaped by processes very different from any now in evidence All we can say for certain is that the oldest rocks now exposed show a clear record of the action of geologic processes very similar to those of the present.

CHAPTER XXXVIII

Interpreting the Rock Record

MUCH as the archeologist tries to read early human history from fragments of old inscriptions and the ruins of ancient cities, so the geologist seeks to interpret the earth's past from the record preserved in rocks and landscapes. Events of the recent past have left abundant evidence in present landscape features: from moraines, lakes, and U-shaped valleys we learn of the spread and retreat of ancient glaciers; wave-cut cliffs and terraces above the sea suggest recent elevation of the land; hot springs and isolated, cone-shaped mountains show past volcanic activity. Earlier episodes are recorded more dimly in the rocks. A geologist finds a bed of salt or gypsum buried beneath other strata, and he knows that the region must once have had a desert climate in which a lake or an arm of the sea evaporated; from a layer of coal he reconstructs an ancient swamp in which partly decayed vegetation accumulated; a limestone bed with numerous fossils suggests a clear, shallow sea in which grew clams, snails, and other hard-shelled organisms. As the long history is carried farther and farther back, the evidence becomes always more fragmentary and the geologist's reconstruction of the earth's surface always more vague.

With the knowledge we have gained of present-day geologic processes as a background, bearing in mind that these processes have operated in the past much as they do at present, let us see how it is possible to piece together the episodes of our planet's history.

Dating Folds, Faults, and Intrusives

Historical geology poses two fundamental problems: (1) to arrange in order the events recorded in a single outcrop or in the rocks of a single small region; and (2) to correlate events in this region with events in other parts of the world so as to give a connected history for the earth as a whole. We begin with the first of these problems.

Some of the principles used in reading the history of a small area are so simple that they scarcely need comment:

1. In a sequence of sedimentary rocks, the uppermost bed is the youngest and the lowermost the oldest. Thus in Fig. 275, bed *A* must have been deposited before the others and bed *E* after those below it. A not very common exception to this rule is a sequence of strata overturned by intense folding.

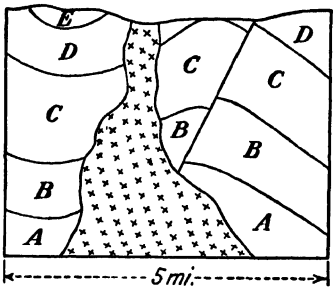


Fig. 275. *Diagrammatic cross section showing folded sedimentary rocks intruded by granite and displaced along a fault.*

2. Diastrophic movement resulting in folding or tilting is later than the youngest bed affected by the deformation. Thus the strata of Fig. 275 were obviously not folded until after the deposition of bed *E*.
3. Diastrophic movement along a fault is later than the youngest bed cut by the fault. Faulting in Fig. 275 could not have occurred before the deposition of bed *D*.
4. An intrusive igneous rock is younger than the youngest bed which it intrudes. The granite shown in Fig. 275 is younger than bed *D*. (This assumes that the "age" of an igneous rock refers to the time at which it solidified; as magma the rock material may have existed long before the intruded sediments were laid down.)

Obvious as these statements sound, their application in regions of complexly folded and faulted strata requires no little ingenuity. The problem is especially difficult in regions where much of the rock structure is hidden by later sediments or vegetation.

Unconformities

A structure like that shown in Figs. 276 and 277 requires further attention. Here the lower, tilted sedimentary beds are cut off abruptly by an uneven surface, on which rest the upper horizontal beds. An irregular surface of this sort, separating two series of rocks, is called an *unconformity*.

At first glance the surface looks as if it might be a fault. But unconformities do not show the characteristic minor features of faults—no distortion of beds, no finely granulated material, no grooved and polished surfaces showing the grinding of one rock against another. An unconformity, moreover, can usually be traced from outcrop to outcrop over an area of many square miles, with the same bed always occurring just above it. Evidently some other explanation than faulting is called for.

Presumably the lower tilted or folded beds were deposited first, in order from the lowest to the highest. After deposition diastrophic movement deformed them, and subsequently the higher part of the series

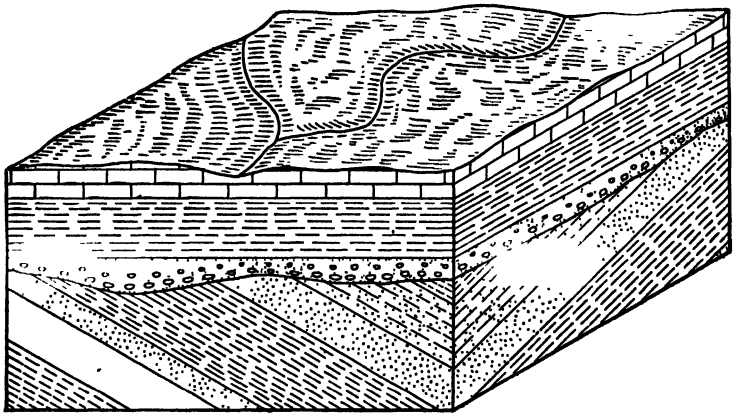


FIG. 276. *Diagram of an unconformity.*

FIG. 277. *An unconformity separating [illegible] tilted layers below. The cliff is about 20 ft. high. Gaspé Peninsula, Quebec.*

was removed. The only conceivable agent besides faulting which might have accomplished this removal is erosion. After erosion, conditions changed so that sedimentation began anew, and the horizontal strata were deposited on the old eroded surface.

Thus an ***unconformity*** is interpreted as *a buried surface of erosion.* It always records at least three geologic events: (1) diastrophic movement resulting in uplift and exposure of the older rocks; (2) a period of erosion when the land surface was above sea level; (3) a change of conditions resulting in deposition of sediment on the eroded landscape. Usually the third event involves subsidence, lowering the eroded surface either beneath the sea or to a level where stream deposition can occur. An unconformity is most easily recognized when the lower beds are tilted or folded as in the example of Fig. 277, but if the original diastrophic movement is simply a vertical uplift, the lower beds may be horizontal and therefore parallel with those above. It is not necessary that the lower series should consist of sedimentary layers, for the eroded surface may equally well be carved on igneous or metamorphic rocks. The one essential feature of an unconformity is that it must represent a surface formed by the processes of erosion and buried beneath later deposits.

Unconformities are important in historical geology for several reasons. For one thing, they make possible the approximate dating of past diastrophic movements: obviously the movement responsible for an unconformity must have occurred *after* the latest rocks of the older series and *before* the oldest of the upper layers. Secondly, unconformities tell us something about the distribution of land and sea at different periods of the earth's past, for an unconformity always means that dry land must have existed during the period of erosion that formed it. And thirdly, unconformities are important in a negative sense: they indicate gaps in the geologic record, times when no deposits were forming in particular regions. An unconformity tells us that a region was above sea level, but all details of the region's history for that period are lost.

The Formation of Mountains

Mountains can form in a number of ways. Some are accumulations of lava and fragmental material ejected by a volcano. Some are small blocks of the earth's crust elevated along faults. But the great mountain ranges of the earth, like the Appalachians, the Rockies, the Alps, and the Himalayas, have a much longer and more complex history involving sedimentation, folding, faulting, igneous activity, repeated uplifts, and deep erosion. The formation of a mountain range is a major event of earth history, and it leaves an indelible record in the rocks which can be read long after the range itself has vanished.

A careful observer would note one conspicuous difference between the sedimentary rocks exposed in most mountain ranges and the corresponding rocks under adjacent plains: as each layer is followed from the plains toward the mountains, its thickness increases enormously (Fig. 278). Now in general the sedimentary rocks of both plains and mountains

are formed from deposits which accumulated in shallow seas or on low-lying parts of the land—that is, on a surface not very far above or below sea level. For successive beds to be laid down the land must have been slowly sinking, the accumulating deposits keeping the surface at roughly

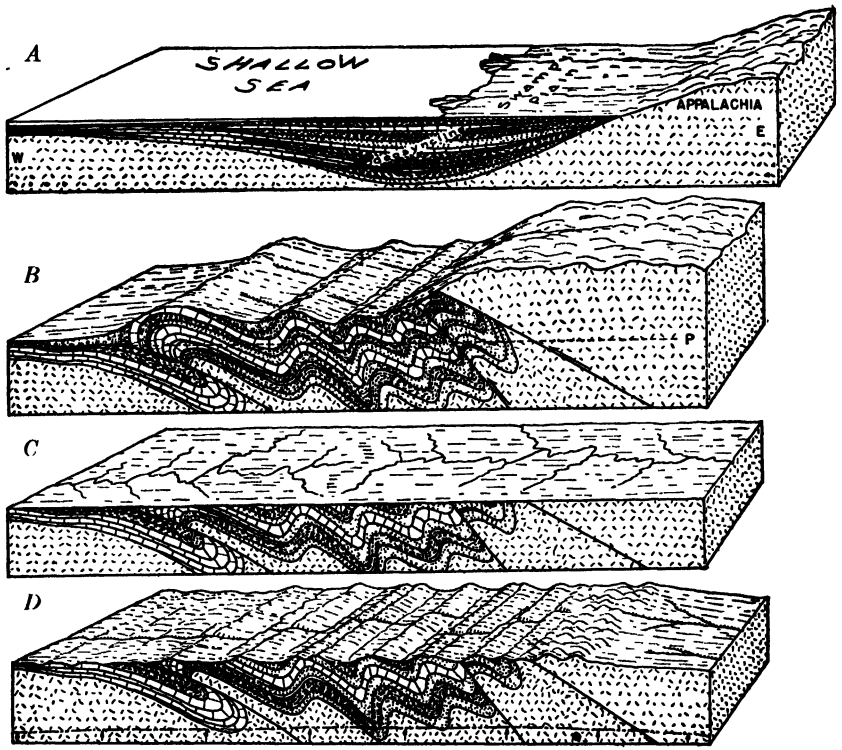


FIG. 278. *Four stages in the evolution of the present Appalachian Mountains. A, sediments accumulating in the Appalachian geosyncline; B, folding and thrust-faulting of rocks in the geosyncline (the form and height of the surface are hypothetical; it is not known to what extent erosion kept pace with the uplift); C, the original mountains worn down to a nearly level plain by stream erosion; D, renewed erosion of the folded strata following vertical uplift, producing the parallel ridges and valleys of the present landscape. (Reprinted by permission from Textbook of Physical Geology by Longwell, Knopf and Flint, published by John Wiley & Sons, Inc.)*

the same level. Since the thickness of sediments in the present mountain region is so much greater, this part of the surface during sedimentation must have been sinking more rapidly than adjacent areas. In other words, the thickness of sedimentary layers in a mountain range suggests that the region where the range now stands was once a broad subsiding basin in which sediments were rapidly accumulating.

A huge sinking basin of this sort is called a ***geosyncline.*** A good example of a small geosyncline in the present world is the central valley

of California. This is a flat-bottomed valley about 500 mi long and close to 100 mi wide, most of its surface just above sea level and at present receiving sediments brought down by rivers from the mountains on either side. Oil wells drilled in the rocks beneath the valley show that the strata are almost entirely soft sedimentary rocks formed from shallow-water marine and lake deposits. The deepest well has penetrated 15,000 ft of such material, and calculations indicate that parts of the valley must have at least 10,000 ft more. Thus the floor of the valley has been slowly sinking for millions of years while these sediments piled up, the sedimentation approximately keeping pace with subsidence. Sinking troughs like this one but on a grander scale must once have existed where now stand the Appalachians and the Rockies. The Appalachian geosyncline reached from the mouth of the St. Lawrence River to the Gulf of Mexico, and the still larger Cordilleran geosyncline extended from Mexico to the Arctic Ocean.

Another conspicuous feature of the sedimentary rocks in mountain ranges is their complex structure. They are folded into large anticlines and synclines, and often are cut by huge thrust faults and minor normal faults. Thus the long sinking of a geosyncline is evidently terminated by a period of diastrophism in which the thick pile of sediments is subjected to intense compressional forces. The compression raises part of the folded layers above the sea, and erosion begins to wear down the exposed beds as folding continues. In this manner the geosyncline is transformed into a mountain range (Fig. 278).

Where erosion has cut deeply into the rocks of a mountain range, much of the old sedimentary material is found converted into such metamorphic rocks as slate, schist, and marble. Since the earlier rocks of the geosyncline were deeply buried beneath later sediments, it is only reasonable that intense diastrophic forces should induce dynamothermal metamorphism on a large scale. Also brought to light by long erosion in a mountain range are intrusive igneous rocks—dikes of various kinds and huge batholiths of granite, often forming the highest central part of the present range because of their resistance to erosion (Fig. 256, page 527). Since these igneous masses intrude the folded sedimentary and metamorphic rocks, and since the igneous rocks usually show evidence of only minor deformation, we must infer that they were intruded in the last stages of mountain building, when diastrophic movement had nearly ceased. Sometimes lava flows and other volcanic products suggest that the intrusive activity at depth was accompanied by volcanic eruptions at the surface.

During and after the diastrophic movement which forms a mountain range, erosion shapes its surface features. As more and more material is removed from the range, the isostatic balance between it and adjacent

segments of the crust is disturbed, until at length the balance is restored by an uplifting of the mountain block. This leads to an increase in the rate of erosion, further removal of material from the mountains, and eventually another upward movement. The later history of a mountain range is punctuated by these successive vertical uplifts, perhaps in part due to some other cause than isostasy. Thus the features of most present-day mountain landscapes are due, not primarily to the compressional forces which folded and faulted their rocks, but to long erosion in regions subjected to periodic uplifts.

When a mountain system has finally succumbed to the processes of erosion and is worn down to a region of low hills or a plain, evidence of its former existence is still preserved in the rocks. All the original folded and faulted sedimentary layers may have disappeared, leaving exposed only metamorphic rocks intruded by igneous masses; but these show clearly, by their intense folding and crushing, that they once formed the roots of a mountain range. Whenever a geologist comes across numerous outcrops of contorted schists intruded by dikes and batholithic masses, he knows that mountains once existed in that region.

The History of the Grand Canyon

For an example of the piecing together of geologic history in a particular region, let us turn to the Grand Canyon of northern Arizona, where the Colorado River has cut a mile-deep gash into the earth's crust (Fig. 223, page 495).

The essential features which an observant visitor would see in the rocks of the canyon are shown diagrammatically in Fig. 279. In the upper part are massive, nearly horizontal sedimentary layers, responsible for the sculptured cliffs and the brilliant hues which have made the canyon so famous. Near the top of the steep inner gorge where the river is now cutting, the lowermost of these beds rests on an uneven surface which bevels a series of tilted sedimentary strata. The tilted beds in turn are separated by an irregular surface from a still lower series of dark-colored schists and gneisses, complexly folded and intruded by dikes and irregular masses of gray granite.

The oldest rocks in the canyon are evidently the schists and gneisses of the inner gorge. The history of these early rocks is obscure, for metamorphism has all but obliterated their original structures. Probably they were once sedimentary layers interbedded with lava flows, lying in an approximately horizontal sequence. In a period of diastrophism they were folded and metamorphosed; at the same time or later they were intruded by granite. These are events which accompany mountain building, so we may picture a range of mountains here at some distant time in the earth's past.

The uneven surface which planes off schist and granite alike is an unconformity, representing an immensely long period of erosion when the ancient mountains were reduced to a nearly level plain. Sinking of the land or the rising of other mountains near by at length made this plain a basin of deposition, and beds which now form the tilted series were laid down—originally, of course, in a more nearly horizontal position. These beds contain no fossils, so we find it difficult to reconstruct accurately the conditions under which they were formed. Because the rocks are mostly fine-grained sandstone and shale, because each stratum is relatively thin and often shows irregularities of bedding, we may picture as a probable site of deposition the flood plain of a large river. Following

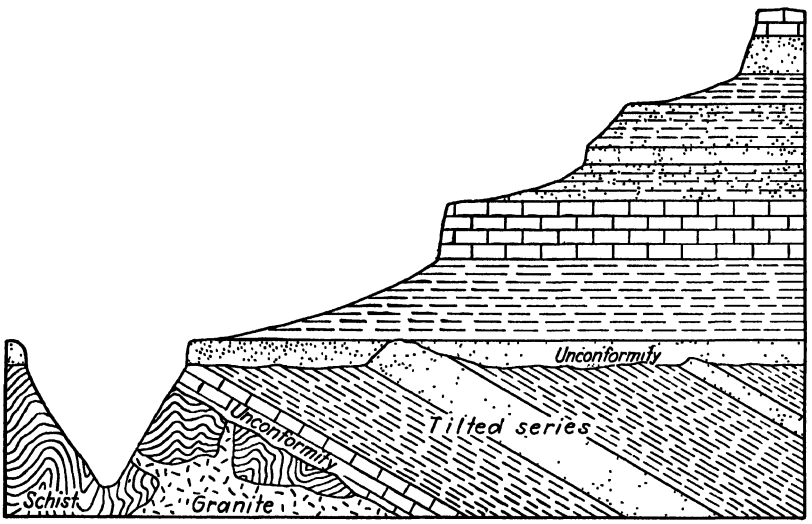


FIG. 279. *Diagrammatic cross section of the Grand Canyon.*

the deposition of these layers they were tilted in another period of diastrophism, and then long erosion reduced the land to the nearly level surface of the second unconformity, on which the horizontal strata rest.

From here on the record becomes clearer. Many of the horizontal beds contain fossils, and from these or other structures the conditions of their deposition can be inferred with confidence. The thick limestone at the top of the canyon and some of the massive, well-sorted sandstones are typical marine deposits. Some thin shale beds show tracks of land reptiles, and so probably represent ancient mudflats along a river or a beach. A sandstone layer shows the rounded grains, good sorting, and large-scale cross-bedding of windblown sand in an arid climate. Transitions between some layers are gradual, suggesting that deposition was continuous in spite of changing conditions. Other layers are separated by distinct

unconformities, showing that at times the land was elevated sufficiently for erosion to take place. We need not follow the history in detail, but careful study of the different layers would make possible an accurate reconstruction of changing conditions through a long part of earth history.

The last geologic events recorded at the canyon are elevation of the land high above the sea and erosion by the Colorado River, two events which probably took place simultaneously.

In few places are rocks of many ages exposed as magnificently as at the Grand Canyon, but patient investigation makes possible the working out of similar histories even where exposures are poor. The histories differ widely in detail, but often the general pattern is much like the one just described: fragmentary glimpses of episodes in the far distant past, separated by enormously long periods of erosion whose separate events have left no trace; a fairly connected story of seas and rivers and deserts recorded in more recent rocks; and finally, events of the immediate past clearly suggested in the present landscape.

Geologic Time

We turn now to the second big problem of historical geology—correlation of events in different regions. Suppose, for instance, that in addition to the history of northern Arizona we have traced the sequence of geologic events in New England. How could we tell what events in the two regions happened at the same time? When a shallow sea was depositing limestone in the Grand Canyon region, what sort of a landscape existed in New England?

An answer might be possible if various rock layers could be followed continuously from one region to the other. In general this is impracticable, since a given layer when followed for a long distance is usually either cut off by erosion or concealed by later deposits.

The problem would be solved even more convincingly if geologic episodes could be dated in terms of years. If we knew, for instance, that a rock layer in Arizona and another in Vermont were both 50 million years old, we could compare events directly. Unfortunately no method has been discovered for such accurate dating of most rocks.

Only one means is available for estimating geologic time in years, and this method can be applied to only a very few rocks. The method is borrowed from the physicist's work on radioactive substances. It involves measurement of the amount of various radioactive elements present in a rock, and depends on the fact that radioactive decay goes on at a constant rate regardless of extreme conditions of temperature and pressure (page 277). Thus if a certain amount of a radioactive element is present in a rock at the time of its formation, after a million years a known fraction of the element will have decayed, regardless of what other elements it is

combined with or what high temperatures and enormous pressures the rock undergoes.

The method of calculation may be illustrated with a rock which contains a compound of uranium. This element disintegrates by a series of steps into other radioactive elements such as radium and polonium, the series ending with unradioactive lead. Each transformation in the series takes place by the emission from a radioactive atom of either an electron or a helium nucleus. Thus the end products of the complete decay of uranium are the two elements lead and helium. Since the rate of decay of uranium itself is extremely slow compared with that of the other radioactive elements, the intermediate elements will be present in relatively small quantities. For approximate calculations the end products only need be considered.

The necessary data are the amounts of lead and undecayed uranium or the amounts of helium and undecayed uranium. The amount of either lead or helium gives by simple calculation the quantity of uranium which must have decayed, and this figure added to the amount still undecayed gives the original quantity present. Comparison of the initial amount with the amount which has decayed tells immediately the age of the rock in years, for the rate of decay is accurately known. Since helium is a gas and escapes easily from a rock if much of it is produced, the calculation is based on lead if the quantity of uranium present is large. If only minute amounts of uranium are present, however, the helium produced is trapped and may be more accurately determined than the lead. Results of the two methods agree fairly well.

Time measurements with the "radioactive clock" tell us that manlike creatures appeared on the earth about 2 million years ago, that rocks with the first fossil remains of mammals are about 200 million years old, that animals with hard shells first became abundant about 500 million years ago. The oldest rocks whose ages have been determined are intrusive rocks from Karelia in Russia, roughly 1,800 million years old. These intrude metamorphosed sedimentary rocks which must be still older, but how much older is unknown.

Figures like these are extremely valuable, for they give us an accurate idea of the immense reaches of time involved in geologic processes. But unfortunately rocks with sufficient radioactive material to make the measurements possible are scarce. Only a handful of exact age determinations have been made, for the rocks of a few isolated localities. In the general problem of correlation, measurements of radioactive substances are not very helpful.

Correlation by Means of Fossils

For establishing relationships among the rocks of different regions, and for arranging beds in sequence when their order is not obvious,

geologists use wherever possible the method discovered by William Smith —comparison of the fossils entombed in different layers.

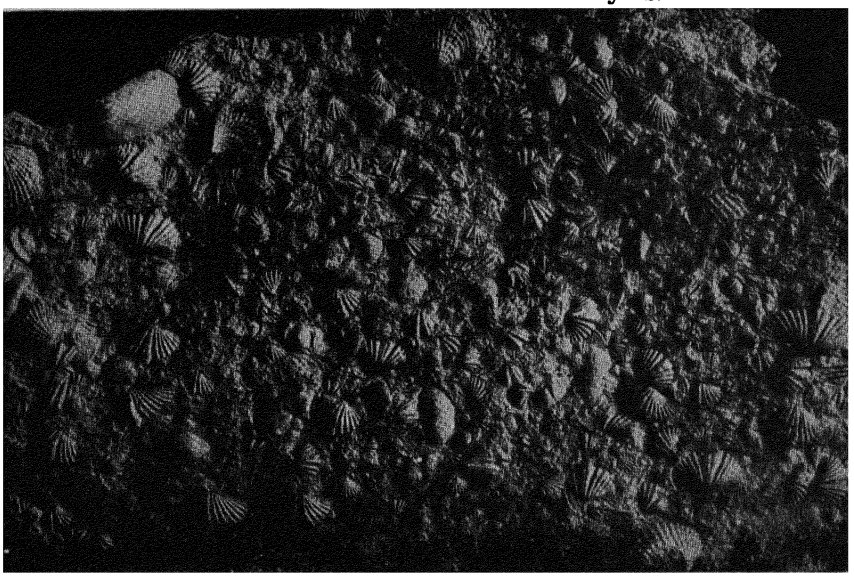


FIG. 280. *A piece of limestone containing abundant fossil shells of molluscs and brachiopods ("lamp shells"). (Photograph by Hardin, U.S. Geological Survey.)*

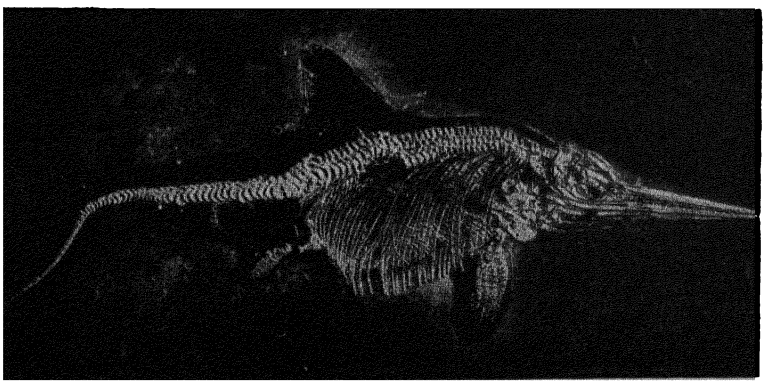


FIG. 281. *A fossil ichthyosaur (an extinct marine reptile) showing not only the skeleton but by a carbonaceous residue the outlines of the body, paddles, and fins. (American Museum of Natural History.)*

Fossils *are the remains or traces of organisms preserved in rocks.* The commonest fossils are the hard parts of animals, such as shells, bones, and teeth (Figs. 280, 281). Very rarely entire animals are preserved, soft parts as well as hard parts: ancient insects have been preserved in amber, and woolly mammoths of the Ice Age are sometimes found in the frozen

soil of Siberia. Plant fossils are relatively scarce, for plants do not contain easily preserved hard parts. The structure of tree trunks is sometimes beautifully shown in *petrified wood*, which is wood whose original organic materials have been replaced by silica deposited from solution in groundwater. Incomplete decay of buried leaves and wood fragments produces black, carbonaceous material which sometimes preserves the original organic structures (Fig. 282); coal is a thick deposit of such material. Occasionally fine sediments preserve impressions of delicate structures like leaves, feathers, skin fragments, even when all trace of the original materials has vanished. Some fossils are merely trails or footprints left in soft mud and covered by later sediments (Fig. 283).

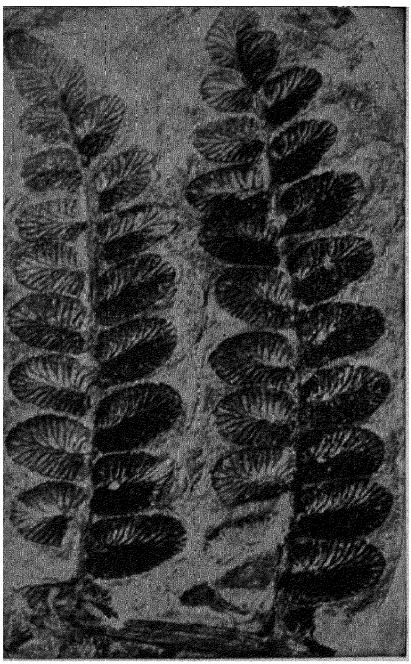


FIG. 282. *Part of a fossil fern. The organic structures are preserved as impressions in the fine-grained rock, and a little carbonaceous residue remains of the original plant material. (Charles Butts, Alabama Geologic Survey.)*

A moment's reflection will show how very slim are the chances of any modern plant or animal to become a fossil for future geologists to unearth. Consider an animal which dies on land: its body is in part eaten by other creatures, in part decays; presently only a few dry bones remain, and even these succumb at length to the slow action of the atmosphere. Dead animals in lakes or the ocean disappear in similar fashion; chemical decay is less rapid under water, but scavengers are more numerous. Dead trees slowly rot, and their material returns to the soil and air. Smaller plants and soft-bodied animals like insects and spiders vanish even more quickly. The only creatures with any chance of preservation as fossils are those fortunate enough to die in a place where sediments are rapidly accumulating. Preferably the spot should be under water, for decay is rapid even after burial if air can circulate.

Conditions necessary for preservation have been much the same throughout geologic history. Chemical decay, bacteria, and scavengers have quickly disposed of most of the organisms which have lived on the earth. Only special conditions of burial occasionally permitted the survival of fossil groups. In general these conditions were best realized on the

floor of shallow seas, where life is abundant and deposition of sediment is sometimes rapid. Our picture of marine life in the past is accordingly far more complete than our picture of the organisms which lived on land, but even the marine record is fragmentary. Thick marine strata all too often contain no fossils at all, and the fossils which do occur are frequently broken and poorly preserved.

Although the fossil record is so very far from complete, careful study reveals a great deal about living things of the past. One important fact which emerges from such a study is that groups of organisms show a progressive change in form from those entombed in ancient rocks to those of the most recent strata. In general the change is from simple forms to more complex forms; in general also the change is from forms

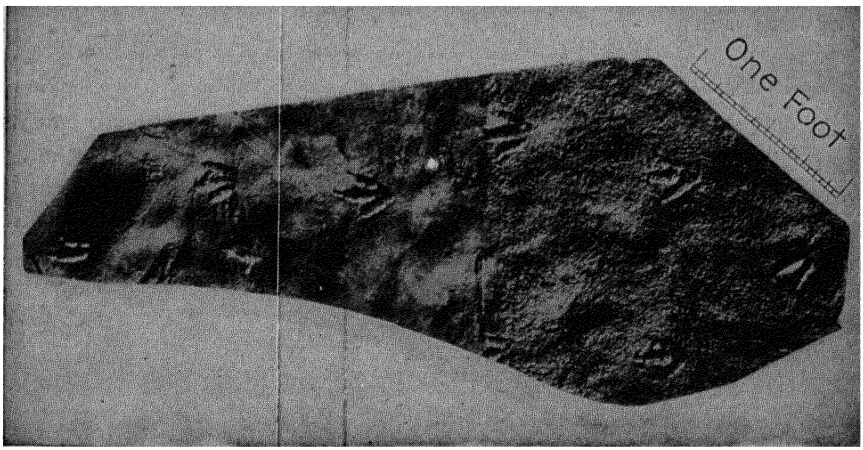


FIG. 283. *Tracks of a dinosaur (an extinct reptile) preserved in fine sandstone. (Photograph by Walter E. Corbin, Florence, Mass.)*

very different from those in the present world to creatures much like the ones we find today. These observations are a part of the factual basis for Darwin's theory that life has evolved by a continuous development from simple forms to the complex organisms of the present.

Because plants and animals have changed continuously through long ages, rock layers from different periods can be recognized by the kinds of fossils they contain. This fact makes possible the arrangement of beds in a time sequence, even when their relationships are not directly visible, and also provides a means of correlating the strata of different localities. If, for example, fossil snail shells and clamshells are found in a rock layer in New York which are exactly similar to fossil shells from a stratum in the Grand Canyon, the two layers must be of approximately the same age. Suppose that above the layer in New York is an unconformity, while at the Grand Canyon continuous deposition is recorded into a higher

layer with a different group of marine fossils; then we could infer that in this later time the New York region was a land area undergoing erosion, while northern Arizona was still covered by the sea. Thus fossils are an all-important tool in historical geology for linking together the events of distant regions.

Fossils are useful not only in tracing the development of life and in correlating strata, but in helping us to reconstruct the environment in which the organisms lived. Some creatures, like barnacles and scallops, live only in the sea, and it is probable that their close relatives in the past were similarly restricted to salt water. Other animals can exist only in fresh water. On land some organisms prefer desert climates, others cold climates, others warm and humid climates. Thus many details about the environment in which a rock was formed are often revealed by its fossil organisms.

Eras and Periods

Fossils make possible a chronological arrangement of geologic events over the entire earth. Enough of these events can be accurately dated by measurements of radioactive decay so that good estimates can be made for dates of the others. But before discussing the succession of events in detail, we need some method of dividing the long history into parts. Just as we find it convenient to divide human history into ancient, medieval, and modern periods, so in reading the geologic record we need time divi sions to which the various episodes can be referred.

Divisions between major periods in human history are placed at times of great change in methods of government, in culture, in ways of life. So in geologic history the chief divisions are separated by times of rapid and profound changes—changes in landscape, in climate, in types of organisms. These times of change, called *revolutions*, were periods of mountain building in some parts of the earth. Much of the surface of the continents was above the sea, undergoing erosion, so that revolutions are marked in the geologic record by widespread unconformities. Sediments laid down under desert conditions and sediments deposited by glaciers suggest that revolutions were accompanied by extremes of climate. Evolutionary development was speeded up during revolutions: many kinds of organisms could not survive the climatic changes and the shifts in position of land and sea, and those which did survive underwent changes in structure which made them better adapted to the new conditions.

Revolutions divide the latter part of geologic time into three major divisions called *eras*. The era in which we are now living, the Cenozoic era, began about 60 million years ago (according to measurements of radioactive decay). Before that came the Mesozoic era, which lasted 140

million years, and the Paleozoic era, which lasted about 300 million years. The geologic record of events before the Paleozoic era is so dim that geologists are not agreed about the proper division of this early time into eras. Although time before the Paleozoic makes up three-fourths of all geologic history, we shall do best to consider it as a single long division, *Pre-Cambrian time*—just as in human history we might lump together the longest but least known part of man's development in a chapter called Prehistoric Man.

The three later eras are divided into shorter time intervals called *periods*. When the divisions were first set up, geologists tried to separate periods, like eras, on a basis of widespread diastrophic movement. In limited areas such a division is possible, and chapters of the sedimentary record can be conveniently separated by unconformities marking times of uplift and erosion. But generally divisions set up in one area cannot be extended to adjacent areas, because the original diastrophic movements were local; rocks of two periods may be separated by a marked unconformity in one region, while another region shows a gradual transition between them. Probably the most satisfactory basis for separating periods (and perhaps even eras as well) is changes in groups of fossil

TABLE XXVII. DIVISIONS OF GEOLOGIC TIME

Dates at which eras began	*Eras*	*Periods*	*Dominant forms of animal life*
60 million years ago	***CENOZOIC***	Quaternary Tertiary	Mammals
	Laramide Revolution		
200 million years ago	***MESOZOIC***	Cretaceous Jurassic Triassic	Reptiles, ammonites
	Appalachian Revolution		
		Permian Pennsylvanian Mississippian	Amphibians
500 million years ago	***PALEOZOIC***	Devonian	Fishes
		Silurian Ordovician Cambrian	Marine invertebrates
	Killarney Revolution		
2,000 million years ago	***PRE-CAMBRIAN TIME***		Primitive invertebrates in latter part, probably one-celled organisms in early part

organisms. Just what changes should be used are still disputed among geologists, so that boundaries of many periods are not entirely definite.

Periods and eras, after all, are simply human attempts to subdivide a continuous history. They have all the imperfections to which any classification of natural phenomena is subject, but their use is justified by convenience. The commonly accepted periods and eras are listed in Table XXVII.

Questions

1. In Fig. 275, did movement on the fault occur before or after intrusion of the granite?
2. Draw cross sections to show (*a*) an unconformity representing a surface of erosion developed on granite, (*b*) an unconformity separating two series of strata which are both horizontal.
3. Could the diagonal line in Fig. 275 represent an unconformity tilted by later movement, rather than a fault? Why or why not?

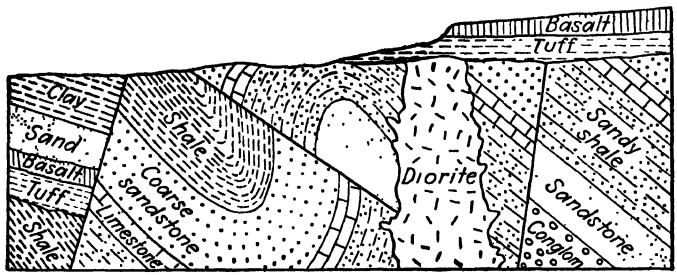


FIG. 284.

4. Figure 284 represents a cross section of a region about 10 mi wide. On the figure label the following:
   a. The youngest rock.
   b. The oldest rock.
   c. An unconformity.
   d. An intrusive contact.
   e. A syncline.
   f. An anticline.
   g. A fault on which movement occurred before the intrusion.
   h. A fault on which no movement has occurred since the basalt was extruded.
5. In order from the earliest to the latest, list as far as possible the geologic events recorded in the cross section of Fig. 284. Indicate the events which cannot be accurately dated.
6. In parts of Colorado and Wyoming a long period during which marine sediments were deposited in a subsiding basin ended with intense diastrophic movement at the close of the Mesozoic era. A mountain range was formed, which in the early part of the Cenozoic era was worn down by erosion to a nearly level plain. On parts of this plain stream and lake sediments were deposited. Describe the rocks and rock structures which you would expect to find in this region.
7. What are the chief assumptions involved in the determination of the age of a rock by measurements of its uranium and lead content?

CHAPTER XXXIX

Earth History

THE story of two billion years would be a long time in the telling if every detail were recounted. It would be doubly long if the evidence for each reconstructed scene were presented and evaluated. In one short chapter we can do no more than touch the high points of the story and suggest in broadest outline the evidence behind it.

Two kinds of events make up this history: on the one hand such physical changes as the spread and retreat of seas, the rise of mountain ranges, outpourings of lava and ash, the advance and recession of glaciers; on the other hand, changes in the organic world from primitive forms to the complex plants and animals of today. Like plot and subplot in a novel the two series of events are closely interwoven, for the course of organic evolution has been determined in large measure by changes in physical environments. The kind of evidence from which both histories are reconstructed has been discussed in the last few chapters. We must keep in mind that at best this evidence is fragmentary; although some pages of the earth's biography can be read clearly, others are almost illegible and many have been lost altogether.

Before the Paleozoic

The earth's beginning is a problem which has long concerned geologists, astronomers, and physicists. Modern theories all begin with a cosmic catastrophe in which the earth's material was torn from the sun or a companion star (page 96), but the nature of the catastrophe and details of the earth's formation from star matter remain unsettled.

The most important geologic issue is whether our planet began its existence as a molten globe or as an essentially solid accumulation of planetesimals. Into this long controversy we cannot go, but much scientific ink has been spilled in able defenses of both sides. Supporters of the solid-earth idea, for instance, point out that the rock record shows no evidence for any appreciable cooling of the earth during geologic

time; the oldest known rocks were laid down as sediments in a sea much like present-day seas, and rocks not quite so old prove the existence of glaciers and therefore of cold climates very early in earth history. Proponents of a molten earth counter by pointing to the concentration of heavy, metallic material near the earth's core and lighter, silica-rich material in the crust; this sort of stratification according to density, similar to the stratification of molten iron and slag in a blast furnace, would be expected from the cooling of a liquid but not from a heterogeneous accumulation of solid fragments. Expert opinion at present favors the molten-earth idea, but the planetesimal hypothesis still has stanch supporters.

Assuming a molten earth at the beginning, we may picture in imagination the earliest days of our planet. There would be a time of rapid cooling, and the formation of a solid crust. Repeatedly fragments of the crust would be engulfed by liquid material from the interior and a new crust would form. At length the crust would cool sufficiently for rain to reach the surface, and the work of erosion would begin. Igneous activity on a grand scale would persist for millions of years, but ultimately a large part of the earth would become solid. Gradually the sequence of events would settle into the familiar cycle of erosion, sedimentation, diastrophic uplift, and volcanic eruptions.

With the earliest Pre-Cambrian rocks the story leaves the realm of speculation. These most ancient rocks are schists and gneisses with interbedded layers of marble and quartzite, metamorphic rocks evidently derived at least in part from sediments. So the visible record of the earth's history begins with a time more than 1,800 million years ago, when the cycle of erosion and diastrophism was already well established.

Pre-Cambrian rocks are exposed at the surface over a broad area covering most of eastern Canada and adjacent parts of the United States. This immense region of ancient rocks, one of the largest in the world, has stood above sea level for most of the 500 million years since the beginning of the Paleozoic era. Smaller areas of Pre-Cambrian rocks are found in many parts of the country, particularly in the cores of mountain ranges where repeated uplifts and deep erosion have combined to expose them. In the Grand Canyon the Colorado River has cut through more than 5,000 ft of Paleozoic strata to reveal the older rocks at their base.

In the Pre-Cambrian rocks of the Lake Superior region, and again at the bottom of the Grand Canyon, a conspicuous unconformity separates highly metamorphosed schists and gneisses below from less metamorphosed sedimentary and volcanic rocks above (the lower, tilted unconformity, page 560). On the basis of this unconformity many geologists divide Pre-Cambrian time into two eras, "Archeozoic" and "Protero-

zoic." There is no certainty, however, that the unconformities in the two widely separated areas represent the same time interval; furthermore, some geologists in this country and many in Europe recognize three major time divisions before the Paleozoic rather than two. We shall not enter this controversy, but we should keep in mind that the Pre-Cambrian record includes unconformities representing very long periods of uplift and erosion.

Although in places the later Pre-Cambrian strata are practically unaltered sedimentary and volcanic rocks, these old rocks in general show considerable metamorphism and often intense deformation. This is hardly surprising in view of their great age and the deep erosion they have undergone. Careful study of even the highly metamorphosed rocks usually reveals something about their origin, and the less-altered strata often give surprisingly detailed records of their history. Among Pre-Cambrian rocks we find varieties formed in nearly all the environments recorded in later geologic ages. The sedimentary beds are in part stream deposits, in part marine. In late Pre-Cambrian strata we even find rocks of glacial origin—coarse, angular conglomerates containing smoothed and striated boulders, resting on grooved rock surfaces—indicating at least two distinct periods of glaciation in this early stage of earth history. The volcanic rocks include all types, with basalt flows then as now the most common. Intrusive rocks are represented in great abundance and variety. Evidently geologic processes a billion years ago were not very different from those in the modern world.

Unlike rocks of later ages, Pre-Cambrian rocks are almost devoid of fossils. This does not mean that life was absent, for rare finds of structures produced by algae, by sponges, and probably by worms prove that primitive organisms did exist (Fig. 285). The scarcity of fossils is probably due to lack of animals with preservable shells rather than to any actual scarcity of living things. Indirect evidence for the presence of abundant life in Pre-Cambrian seas comes from: (1) thick beds of limestone, which presumably require the action of organisms for their deposition, and (2) occasional layers of graphite, most reasonably explained by the metamorphism of organic debris. Where and how life began on the earth we do not know, but marble and graphite in very old Pre-Cambrian rocks suggest that primitive sea-dwelling forms existed nearly 2 billion years ago.

The scarcity of fossils coupled with widespread metamorphism and extensive deformation makes the correlation of Pre-Cambrian events very difficult. Some headway can be made by comparing rock compositions, degrees of metamorphism and deformation, extent of intrusive activity, but these make possible only a very general outline of Pre-

Cambrian history. Like the later history these early chapters are largely a record of successive periods of sedimentation, separated by intervals of erosion and at times by mountain building. In the later Pre-Cambrian lava flows were especially numerous and widespread, and glaciation was extensive in two distinct periods. Pre-Cambrian time ended with the

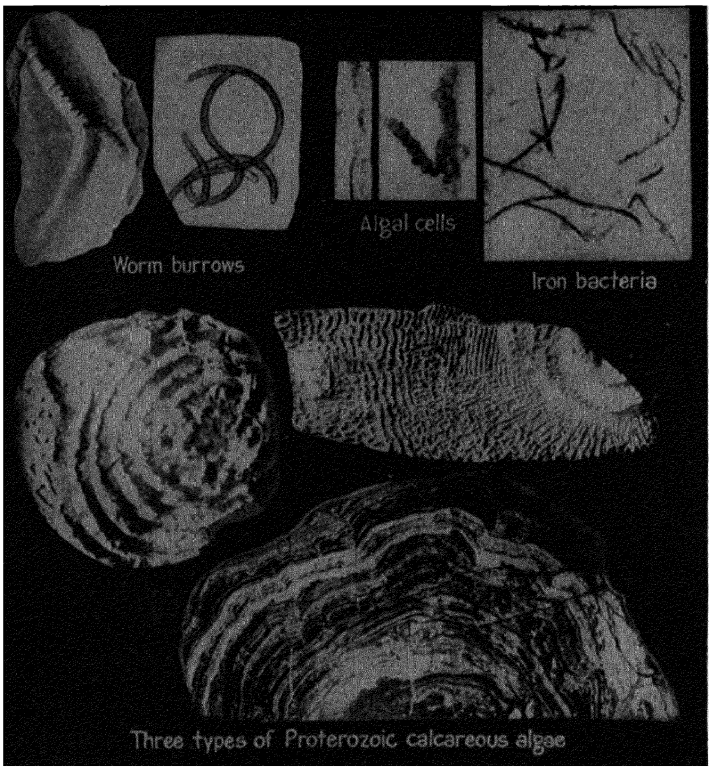


FIG. 285. *Some examples of organic remains from late Pre-Cambrian rocks. (From Historical Geology by R. C. Moore.)*

uplift of continents and long erosion, forming the unconformity which everywhere separates these rocks from Paleozoic strata.

Pre-Cambrian rocks have yielded mineral wealth in abundance. The great iron ore deposits of the Lake Superior region were formed from iron-rich sediments of late Pre-Cambrian age by weathering and concentration by groundwater. The native copper deposits of northern Michigan are associated with Pre-Cambrian basalts. The fabulously rich gold, silver, and nickel ores of eastern Canada were formed during the solidification of Pre-Cambrian intrusive rocks.

Early Paleozoic

Wherever Paleozoic strata are found resting on older rocks, they are separated by a marked unconformity. This unconformity is well exposed in the Grand Canyon (the upper, horizontal unconformity, page 560) and at many other localities on this continent. It indicates a time when all the continents stood well above sea level, so that material eroded from them was carried into the oceans. It is most unfortunate that sediments deposited during this interval are not somewhere exposed, for the unconformity means a gap of several million years in the geologic record. The gap comes at a most critical time in geologic history, since strata above the unconformity contain many fossils while those below have only the faintest signs of living creatures.

Thus the Paleozoic era begins in mystery. But once the era has fairly begun, its history is remarkably complete. No longer need there be the doubts and the vagueness which characterize Pre-Cambrian correlations; for Paleozoic strata are widely exposed, and their wealth of fossils makes possible correlation of rocks and events from one side of the earth to the other.

Sedimentary rocks belonging to the first three periods of the Paleozoic era underlie much of the Mississippi Valley and appear also in the mountains on either side of the continent. In the Central States the beds are nearly horizontal, in the mountains complexly folded and faulted. In the Central States unconformities separate many of the beds, showing that deposition was often interrupted by periods of uplift and erosion; in the mountains, although some unconformities are present, thicker strata preserve a much more continuous sedimentary record. For the most part these beds are sandstones, shales, and limestones formed from the sediments of shallow seas.

From these facts we infer that North America in the early Paleozoic bore little resemblance to the present continent. To the northeast was a highland of Pre-Cambrian rocks, perhaps with erosional remnants of late Pre-Cambrian mountains still standing. At either side of the continent was a broad subsiding trough or geosyncline (Fig. 286), covered with a shallow sea during most of this time. Between the geosynclines stretched a low plain, parts of which were submerged at intervals by spreading of the shallow seas. To complete the picture we must imagine a high land mass east of the Appalachian geosyncline and another west of the Cordilleran geosyncline, for the distribution of sediments in the geosynclines shows that much material came from those directions.

The chief physical events of these periods were advances and retreats of the shallow sea which covered the geosynclines most of the time and

parts of the continental interior occasionally. At one time nearly 65 per cent of the continent was under water (Fig. 287). In the intervals between periods the continent was uplifted so that even the geosynclines were partly or wholly dry. In general this was a time when the earth's crust was remarkably stable, its movements consisting simply of minor ups and downs of the continent as a whole. The only important exception

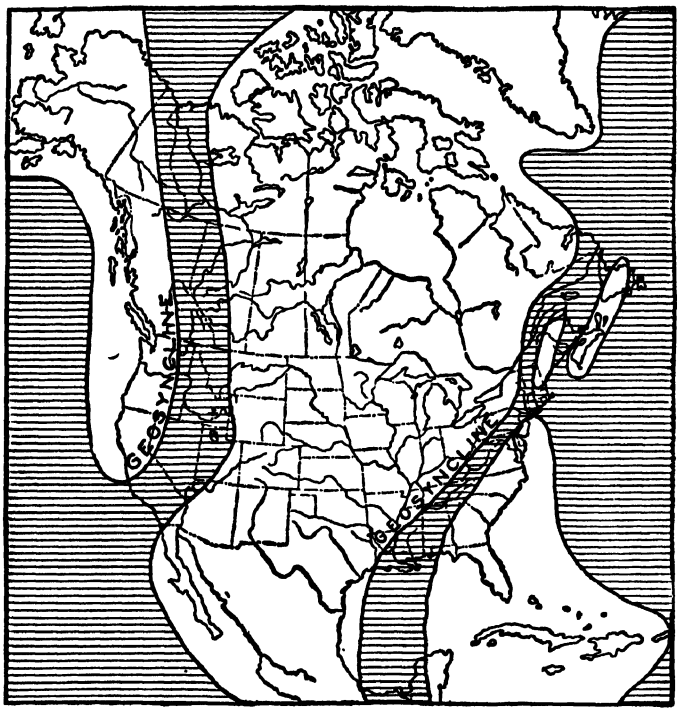


FIG. 286. *Map of North America showing the location of the Appalachian geosyncline (right) and the Cordilleran geosyncline (left). (White areas, land; shaded areas, sea.) (From Geology by Emmons, Thiel, Stauffer and Allison.)*

was a mountain-building disturbance in New England toward the end of the Ordovician period.

The nature and distribution of fossil animals suggest that the early Paleozoic climates were mild over the entire continent, without the marked zoning into hot and cold regions that we find today. One exception was an interval of desert conditions recorded by beds of salt and gypsum in the Silurian strata of New York.

In western Europe the Silurian period ends with the rise of a mountain range extending from Scotland northeastward along the Scandinavian peninsula. This diastrophic activity makes a convenient division of the

Paleozoic era into an early part consisting of the first three periods and a later part extending from the Devonian through the Permian. No important mountain building occurred in North America to mark this break,

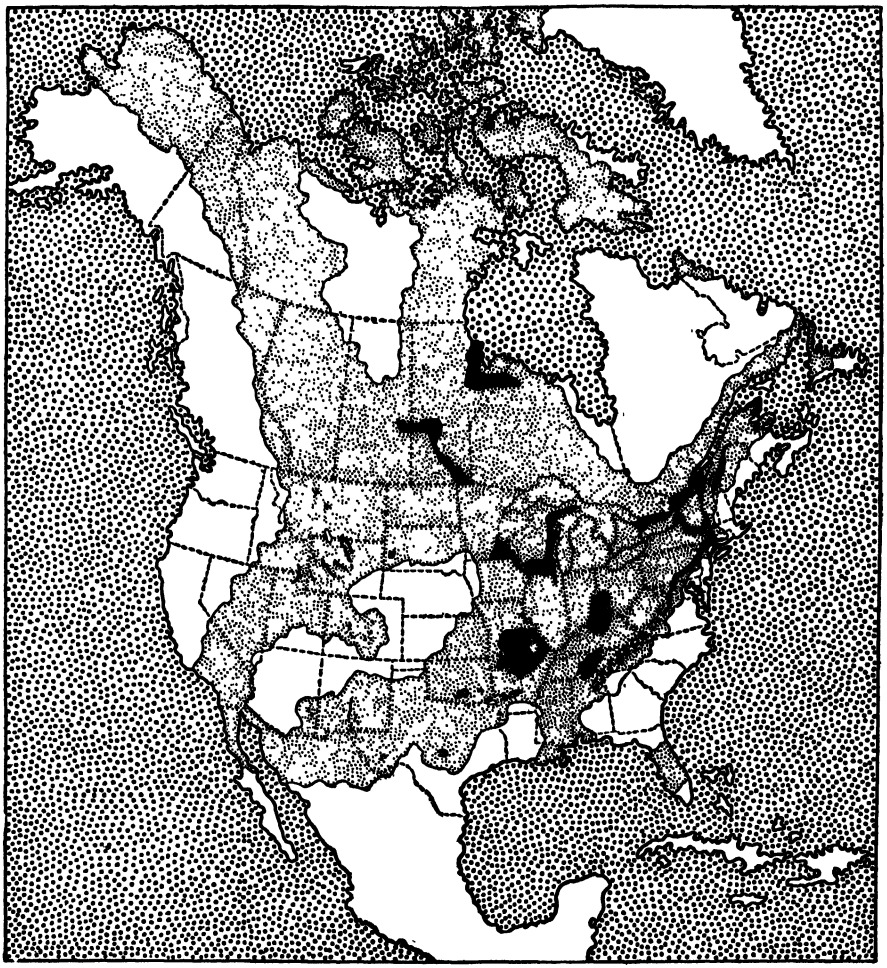


FIG. 287. *Extent of shallow seas (fine stippling) during part of the Ordovician period. Areas where Ordovician rocks are exposed at the surface shown in black. (From Introduction to Geology by Branson and Tarr.)*

but the separation is nevertheless convenient because the physical and organic events of the two parts of the era were quite different.

The animals whose fossils appear in the earliest Paleozoic beds were *invertebrates*, creatures without internal skeletons but with external shells of calcium carbonate, silica, or chitin. All the major groups of the invertebrates are represented, some by organisms which must have had complex internal structures. It seems impossible that such varied and

complex forms could have evolved during the erosion interval at the end of Pre-Cambrian time, but the odd fact remains that fossils of these creatures are not found in late Pre-Cambrian rocks. A possible way out of the dilemma is the hypothesis that fairly complex invertebrates were indeed present in the seas of late Pre-Cambrian time, but that they had not yet mastered the trick of making hard shells which could be preserved.

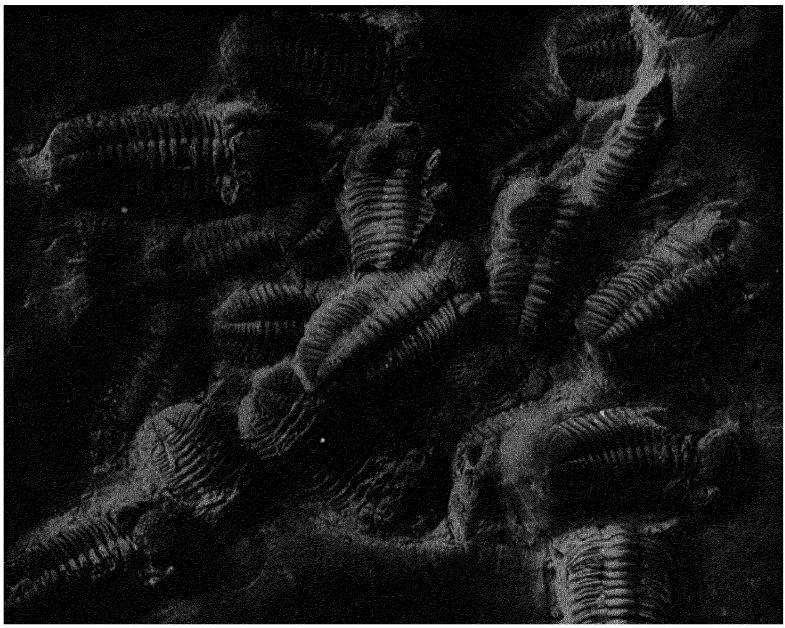


FIG. 288. *Trilobites, about one-half natural size.* (*U. S. National Museum.*)

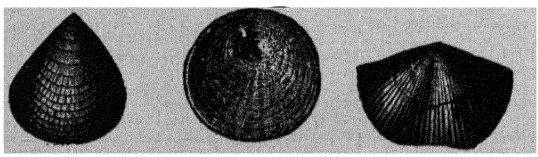


FIG. 289. *Cambrian brachiopods, about one-third natural size.* (*After Walcott.*)

This idea would allow a long period of evolutionary development during Pre-Cambrian times, the only radical advance immediately before the Cambrian being the acquisition of shells.

So far as we know the Cambrian invertebrates were all sea-dwelling animals. The dominant creatures were animals called *trilobites*, sluggish, segmented beasts distantly related to modern sow bugs (Fig. 288). *Brachiopods*, with shells somewhat like those of clams but very different internal structures, were also numerous (Fig. 289). Snails, sponges, and

several other forms were present in smaller numbers. With this assemblage must have lived seaweeds and a variety of soft-bodied animals which had no preservable hard parts.

Animals of similar types, but in general different species, inhabited the seas of Ordovician and Silurian times. A new group, the *cephalopods*, creatures related to the modern pearly nautilus, joined the dominant trilobites and brachiopods in the Ordovician. True corals for the first time became numerous. The earliest records of vertebrates are skeletons of primitive fishes which lived in the Ordovician.

What the lands of early Paleozoic times looked like we can only surmise. Probably primitive plants had gained a foothold even in the Pre-Cambrian, but the oldest definite fossils of land plants (primitive fungi and mosses) are Silurian. The first animal which succeeded in adapting itself to an air-breathing existence, so far as available records show, was a relative of the modern scorpion which came out of the sea in the late Silurian.

Late Paleozoic

In contrast to the wide seas and the minor amount of diastrophic movement in the early Paleozoic, rocks of the later periods reveal a time of restricted seas and of diastrophic activity which began as local mountain-building disturbances and culminated in the Appalachian Revolution which closed the era.

Sedimentation, largely in shallow marine waters, continued through most of this time in parts of the two major geosynclines. Occasionally seas spread over the continental interior, but never widely as in earlier periods. A peculiar type of sedimentation characterized part of the Mississippian and Pennsylvanian periods: marine deposition alternated with nonmarine, so that thin beds of sandstone, shale, and limestone follow one another in a regular sequence repeated time after time. The nonmarine parts of these cycles, especially in the Pennsylvanian, often contain layers of coal, indicating times of widespread, low-lying swamps with abundant vegetation. These are the great coal deposits of the central Mississippi Valley and the Appalachian region. So abundantly was coal formed at this time both in this country and in Europe that the Mississippian and Pennsylvanian are frequently lumped together as the *Carboniferous* period.

The mild climate of early Paleozoic periods persisted over most of this continent until mid-Permian times. But with continental uplift and the rise of mountain ranges at the end of the era came radical climatic changes. Broad deserts formed in the lee of mountains; icecaps and valley glaciers developed in many parts of the world. An extremely arid Permian climate in Texas and New Mexico is indicated by thick deposits of salts,

formed by evaporation of a restricted arm of the sea—not only gypsum and sodium chloride, but far more soluble salts of potassium and magnesium as well. Similar Permian deposits in Germany have long supplied potash to the rest of the world, but development of the recently discovered Texas and New Mexico beds should supply this country's future needs. Glacial deposits of late Paleozoic age have been found in Massachusetts, but the principal glaciers of this time were in the southern hemisphere.

Volcanic activity in the Devonian period is shown by lava flows in the extreme eastern part of Canada, and flows of later periods are found in British Columbia and the northwestern part of this country. Minor mountain-building disturbances occurred in New England during the late Devonian, in the Appalachian region and Oklahoma at the end of the Mississippian, in west Texas toward the close of the Pennsylvanian. But these and earlier Paleozoic disturbances were dwarfed by the Appalachian revolution which ended the Permian period. At this time of intense diastrophic activity, affecting many other parts of the world besides this continent, the sediments which had accumulated for more than 300 million years in the Appalachian geosyncline were crumpled, fractured, and uplifted into a mountain chain which must have rivaled any modern range in height and grandeur. At this time also the entire continent was uplifted, and shallow seas disappeared except for small areas.

Marine life in the late Paleozoic shows many changes from that in earlier seas, but was still far different from marine life of the present (Fig. 290). Trilobites lost their place of prominence; they declined steadily in numbers and variety, and became extinct at the end of the era. Brachiopods and cephalopods were still numerous, but not as prominent as in earlier periods. Clams and snails increased in numbers and show considerable evolutionary development. Corals built widespread reefs in the middle Devonian, but thereafter were not conspicuous. Starfish and sea urchins were not common, but some of their distant relatives (crinoids and blastoids) which today are extinct or rare were extremely numerous. Fishes were far more abundant than in earlier periods and show a greater variety of form.

In late Paleozoic rocks we find for the first time abundant evidence of land-dwelling organisms. In the coal swamps of Pennsylvanian times grew dense forests of primitive plants—huge fernlike trees, enormous horsetails, primitive conifers (Fig. 291). A modern man wandering through such a forest would find no bright-colored flowers, no grasses, few plants at all familiar except possibly some of the ferns and mosses. In and near these primeval forests lived a great variety of animals: scorpions, land snails, primitive insects of many kinds.

FIG. 290. *Restoration of Devonian marine life. A large armored fish in the background; sharks and other fishes in the foreground; crinoids (the flowerlike creatures) in the foreground; brachiopods, clams, and cephalopods on the sea floor. (From Introduction to Geology by Branson and Tarr.)*

Of most interest to us as humans, since they are early members of our own family tree, are the land-living vertebrates of the late Paleozoic. Fossil *amphibians*, oldest of the land vertebrates, appear first in Devonian

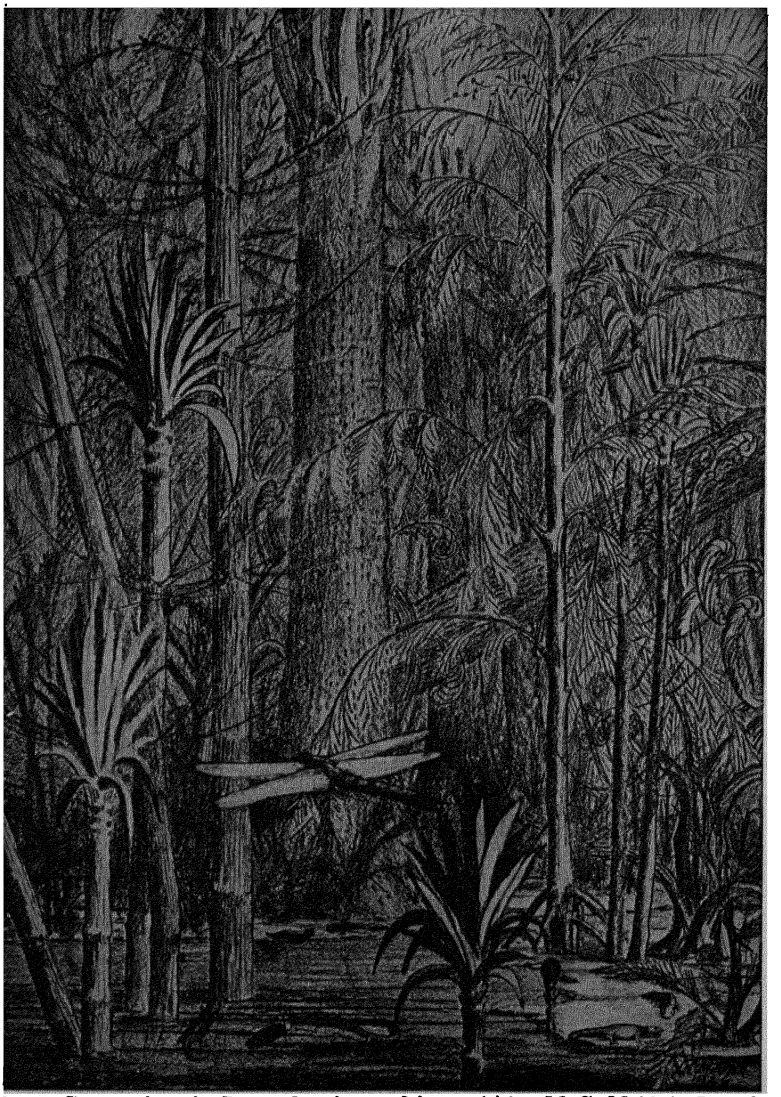


FIG. 291. *Restoration of a Pennsylvanian coal forest. (After M. G. Mehl, in Introduction to Geology by Branson and Tarr.)*

rocks (Fig. 292). These are relatives of modern frogs and salamanders, sluggish creatures which laid their eggs and spent the early part of their lives in water. Their body structure and their dependence on water suggest that they were descended from fishes, but the complete line of

development cannot be traced. In Pennsylvanian rocks appear fossils of *reptiles*, animals which at first looked much like their amphibian ancestors but which had the great advantage of being able to lay their eggs on dry land. The dry climates at the end of the era wrought havoc with the amphibians, but the reptiles, no longer dependent on water for hatching

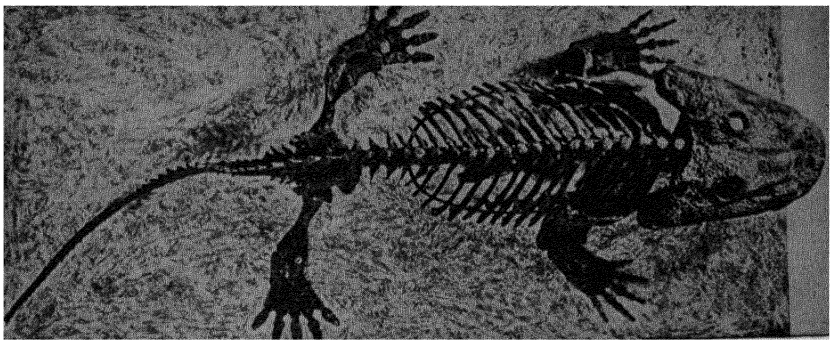


FIG. 292. *Skeleton of a Permian amphibian. Large individuals of this species range up to nearly 10 ft long.* (*American Museum of Natural History.*)

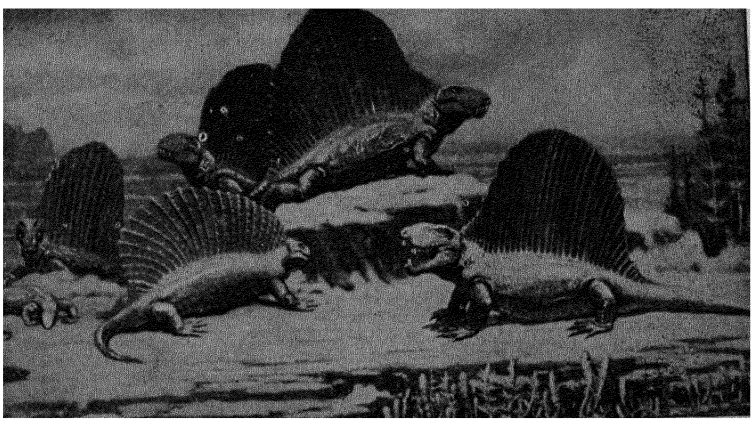


FIG. 293. *Restoration of Permian reptiles from northern Texas.* (*C. R. Knight, Field Museum of Natural History.*)

their eggs, multiplied rapidly and developed a great variety of species (Fig. 293). During the Permian the reptiles became the dominant creatures of the land.

The great revolution which ended the Paleozoic, with its harsh climates, its dwindling water supplies, its rapid changes in environment, was a time of profound change in the organic world. Many kinds of plants and animals were unable to cope with the new conditions, and others

survived only because rapid evolutionary development made them better adapted to a changing world. So in the earliest Mesozoic rocks we find that many of the common Paleozoic organisms are missing, and that both land and water are inhabited by a host of new forms.

Coal and Oil

From Paleozoic rocks man extracts a great number of useful materials. In many regions the rocks themselves make excellent building stone. Iron is extracted from Silurian ores in Alabama. Important lead and zinc ores occur in the Mississippian rocks of southwestern Missouri and eastern Kansas. Strata containing pure quartz sand supply material to the glass industry, and several Paleozoic limestones furnish raw material for the manufacture of lime and cement. Salt and gypsum deposits in New York, Michigan, Kansas, Texas, and New Mexico are important sources of these materials. The potash deposits of Texas and New Mexico will be increasingly important in the future. But the most important economic products won from Paleozoic rocks are coal and petroleum, materials so vital to American life and industry that we may well give their geologic setting a moment's attention.

The origin of coal is evident from the fossils it often contains: it forms from plant material which accumulates under conditions where complete decay is prevented. The only possible situation where large amounts can accumulate under such conditions is in swamps, so that a bed of coal nearly always implies an ancient swamp. Coal has formed in swamps from the Devonian to the present, but at no time have conditions been so favorable for large-scale accumulation as in the Pennsylvanian period. These special conditions were apparently broad swamps almost at sea level, periodically submerged so that partly decayed vegetation was covered with thin layers of marine sediments. So long did this favorable environment persist in the Appalachian region and the Mississippi Valley that the United States need have no fear about its coal supply for many thousand years.

The formation of coal begins with slow bacterial decay, chiefly of the cellulosic material of plants. Taking place largely under water and in the absence of air, this decay results in a gradual removal of oxygen and hydrogen from the cellulose and a concentration of carbon in complex compounds of unknown composition. These are the compounds which on heating break down to give hydrocarbons with ring structures in their molecules, like benzene and naphthalene, which are so valuable in chemical industry (page 407). Aiding decay in coal formation is the action of heat and pressure resulting from burial beneath later sediments.

The origin of petroleum is more obscure, for two good reasons: fossils cannot be preserved in a fluid, and oil often migrates long distances

from the place where it forms. Because petroleumlike substances are sometimes found in and near fossil shells, because oils resembling petroleum can be prepared artificially from organic material, and because petroleum is commonly found in shallow-water marine sediments, most geologists believe that it has an organic origin. Probably both plant and animal matter contributes to its formation, the substances involved being largely proteins, fats, and waxes rather than cellulose. From these materials slow bacterial decay in the absence of air produces the characteristic hydrocarbons of petroleum, those with "straight-chain" molecules like butane and octane. Natural gas, usually associated with petroleum and probably formed by the same type of decay, consists of the lighter hydrocarbons.

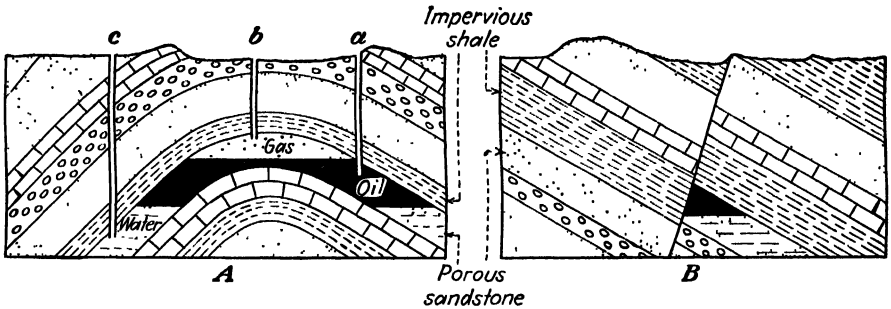


FIG. 294. *Diagrammatic cross sections showing two common types of structural trap in which oil accumulates. A is a trap formed by an anticline, B a trap formed by a fault. In both cases oil in a porous layer is prevented from moving upward by an impervious layer. Note that a well drilled at a would strike oil, one drilled at b would strike gas, one at c only water.*

Both gas and oil, like groundwater, can migrate freely through such porous rocks as loosely cemented sandstones and conglomerates. Wherever they may be formed, they often find their way into porous beds, and it is from these beds that they are obtained by drilling. Migration along a porous bed may be induced by a number of factors—gravity, pressures due to compaction or diastrophic movement, gas pressure, movement of groundwater. Since both oil and gas are lighter than water, they may be displaced by groundwater and so move upward to the surface to form oil seeps. In general, oil becomes available in large quantities only when it is trapped underground by impervious material. Two common kinds of trap are illustrated in Fig. 294, in one of which oil is trapped under an anticline and in the other against a fault.

The bulk of this country's oil production comes from Paleozoic rocks, although more recent strata contribute a considerable amount. Big oil fields tap Paleozoic reservoirs in Texas, Oklahoma and Kansas, in parts of the central Mississippi Valley, in Pennsylvania, and in Wyoming.

Although petroleum resources cannot compare with coal resources, this country faces little danger of shortage for many years to come.

Mesozoic

The earliest Mesozoic sediments were laid down about 200 million years ago, a long time by ordinary reckoning. But the earth was already very old. Some 350 million years had elapsed since the beginning of the Paleozoic, and more than a billion and a half years since the oldest known rocks in the Pre-Cambrian. All the time included in the Mesozoic and Cenozoic eras is hardly more than the last tenth of the earth's recorded history.

Sedimentary rocks of early Mesozoic age are scarce in North America. Marine deposits, thick and widespread in Europe, are restricted on this continent to the Pacific coast. Land deposits of this time were laid down in narrow troughs formed by faulting along the east front of the Appalachian Mountains; we find these beds today as red sandstones and shales lying unconformably on Paleozoic and Pre-Cambrian rocks in a belt extending from Nova Scotia to South Carolina. Other early Mesozoic continental deposits are the thick, cross-bedded, red and white sandstones which form the steep cliffs of Zion Canyon in southern Utah and give the brilliant colors to the Painted Desert of Arizona. Rocks of late Mesozoic times are more widespread, both marine and continental beds appearing in the Rocky Mountains, the Great Plains, and along the Gulf coast; continental deposits are found on the Atlantic seaboard and marine deposits along the Pacific coast.

From this distribution of rocks we may reconstruct the chief Mesozoic scenes. At first almost the entire continent stood above sea level. Probably the Atlantic coast was farther east than at present, so that the new-formed Appalachians were some distance from the sea; the remainder of the continent except for a few low ranges was a broad plain. The old highlands east of the Appalachians and west of the Cordilleran geosyncline had disappeared. Streams from the eastern Appalachians carried sediments into narrow valleys along the mountain front. Arid conditions in the southwest made possible the heaping up of wind-blown sand. Then toward the middle of the era shallow seas along the Pacific coast became more prominent. From the Arctic Ocean a sea invaded the Cordilleran geosyncline, at the time of its widest extent spreading out over the plains states and connecting with the Gulf of Mexico (Fig. 295). Fluctuations of this sea led to the formation of inland basins where river and lake deposits were formed. Finally at the end of the era the sea withdrew and left the continent once more above water.

The Mesozoic, like the early Paleozoic, was for the most part a time of crustal stability interrupted only by minor uplifts and subsidences.

One local disturbance broke into this peaceful picture: toward the end of the Jurassic, folding and the intrusion of granite batholiths formed a mountain range on the site of the present Sierra Nevada of California and lesser ranges to the north and south.

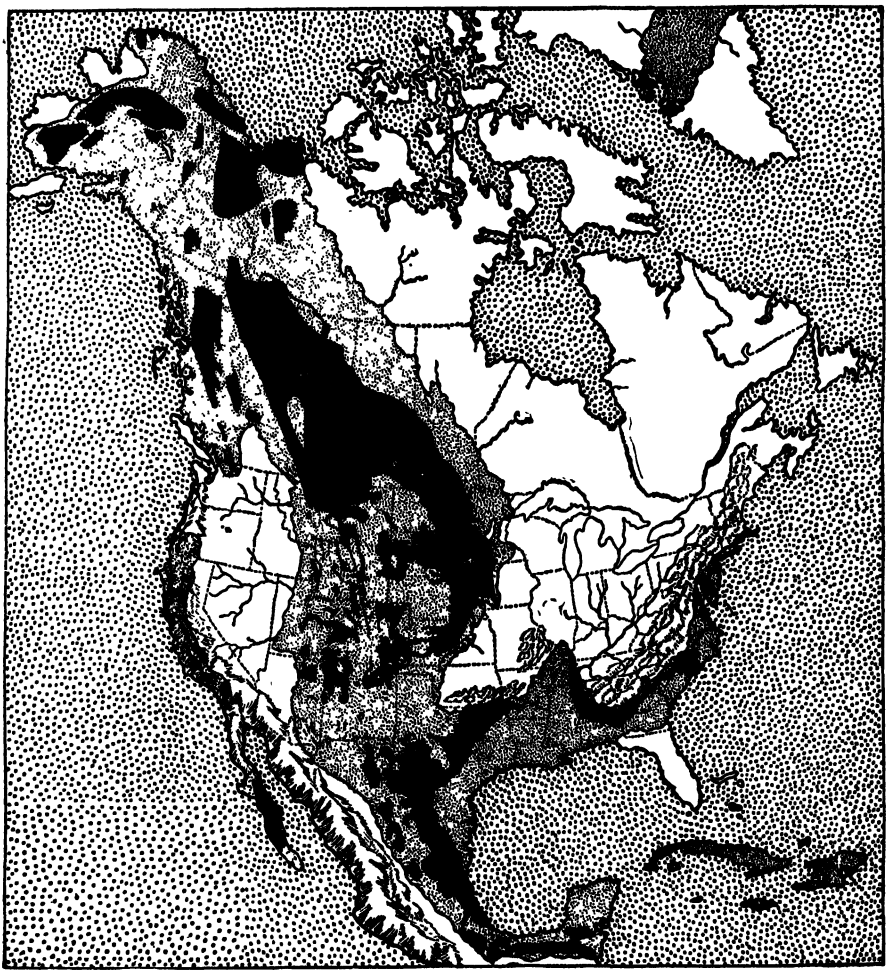


FIG. 295. *Cretaceous seas (fine stippling) at about maximum extent. Cretaceous outcrops shown in black. (From Introduction to Geology by Branson and Tarr.)*

Volcanic activity during the early Mesozoic is recorded in basalt flows and intrusive sheets in the red Triassic sediments east of the Appalachians. Volcanic materials are also abundant in Triassic beds of British Columbia and the Northwestern states. Eruptions in California and adjacent regions accompanied the mountain building at the end of the Jurassic.

Climates of the Mesozoic were in general mild, although arid conditions in the Southwestern states are recorded for at least part of the era. When the late Mesozoic sea was widespread, equable climates extended to far northern latitudes, for fossil palms and breadfruit trees are found in the Cretaceous rocks of Greenland.

The creatures that swam and crawled and floated in Mesozoic seas were much more modern in appearance than their Paleozoic ancestors. Trilobites had disappeared, and brachiopods were represented by only a

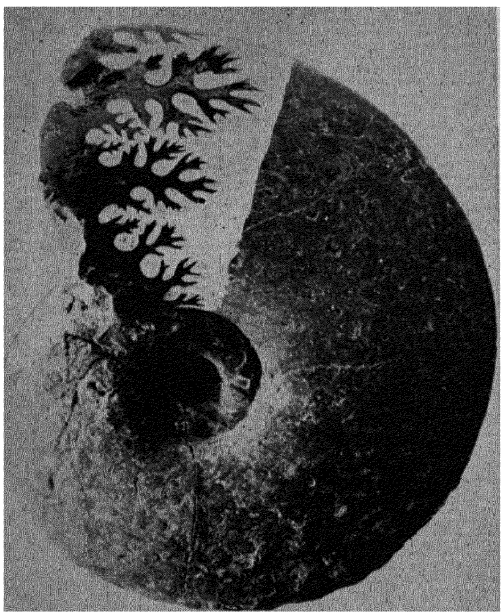


FIG. 296. *An ammonite. On part of the shell the complex structural pattern is outlined with white paint. About one-third natural size. (Photograph by Siemon W. Muller.)*

few simple forms like those of the present. Fishes, oysters, and lobsters looked much like their modern descendants. One prominent race of animals which have no close relatives today was a branch of the cephalopod line called *ammonites*, for the most part creatures with flat coiled shells, some of the larger ones attaining diameters of over a yard (Fig. 296).

On the Mesozoic lands there developed a group of reptiles which includes some of the largest animals the earth has seen—the race of *dinosaurs*, descendants of the few primitive reptiles which survived the Appalachian revolution. Early in the Triassic this reptile line developed an amazing variety of species, adapted for all manner of different habitats

(Figs. 297, 298). Some were carnivores, their bodies designed for pursuing and eating other animals. Some were herbivores, with jaws and digestive organs adapted for a vegetarian diet. Active forms were adapted for life on open plains, more sluggish forms for life in swamps. Some developed bony armor for protection, while others depended on speed to escape

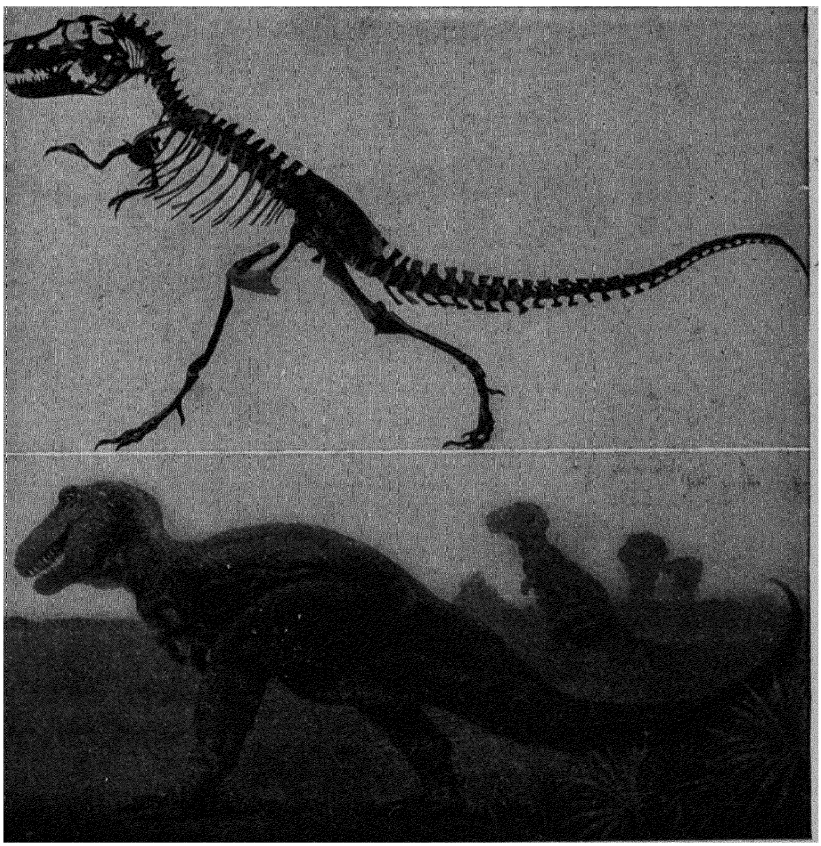


FIG. 297. *The largest of the flesh-eating dinosaurs, Tryannosaurus rex. (Length about 47 ft.) (Skeleton in the American Museum of Natural History. Restoration by C. R. Knight, Field Museum of Natural History.)*

their enemies. In all these directions development continued through the Mesozoic, some of the species becoming marvelously specialized for particular environments. Not all the dinosaurs by any means grew to large size, but the biggest ones were enormous beasts over 80 ft in length and weighing more than 40 tons.

Meanwhile other reptilian stocks had invaded the air and the sea. Of ocean-going reptiles the long-necked *plesiosaurs* and the fishlike *ichthyosaurs* (Fig. 299) were the most prominent. Those which took to the

air, the *pterosaurs* (Fig. 300), developed wing membranes stretched from greatly elongated claws. Some of the pterosaurs were far larger than any modern bird, attaining a wingspread of nearly 30 ft.

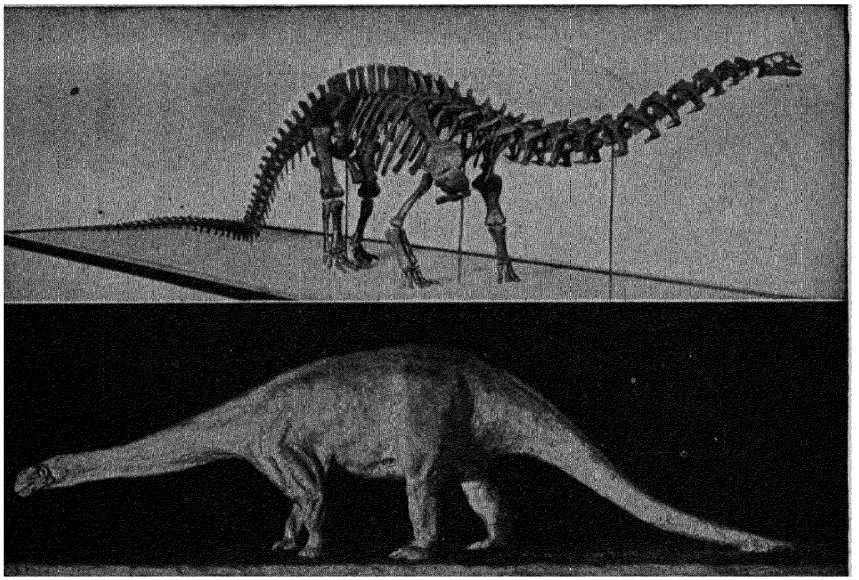


FIG. 2[illegible] *[illegible]osaurus, one of the largest of the herbivorous dinosaurs. Length about 70 ft, estimated weight 38 tons. (American Museum of Natural History.)*

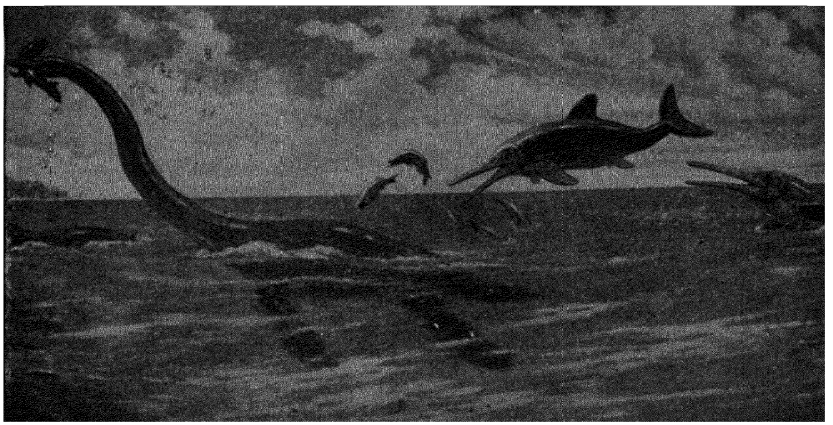


FIG. 299. *Plesiosaur (left) and ichthyosaurs. (C. R. Knight, Field Museum of Natural History.)*

The reptiles were by all odds the most spectacular land animals of the Mesozoic or any other age, but meanwhile other land organisms were undergoing important developments. Flowering plants appeared in mid-

Mesozoic and with them a host of modern-looking insects adapted for helping in the pollination of flowers. The first true birds, with feathered wings rather than membranes, developed from reptilian ancestors in the Jurassic. Sometime in the Triassic appeared the first *mammals*, tiny creatures probably descended from a group of small Permian reptiles. All during the Mesozoic the mammals remained small and inconspicuous, but in several respects they represented an evolutionary advance over the reptiles: they were warm-blooded, hence better able to cope with changes of temperature; they had bigger brains relative to

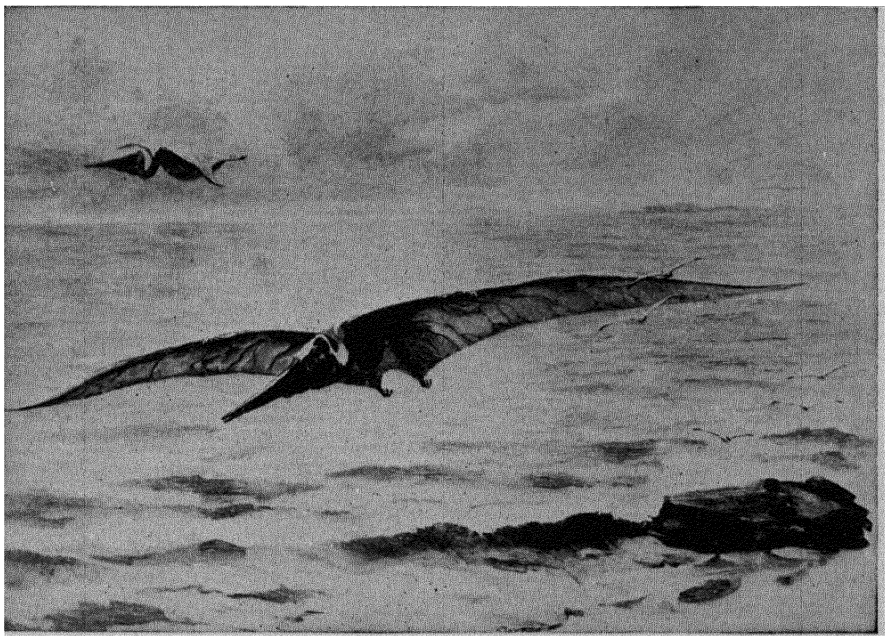


FIG. 300. *Pterosaurs.* (*U. S. National Museum.*)

their body size; and they cared for their young after birth, so that some of the experience of one generation could be passed on to the next. These traits enabled the mammals to survive the drastic climatic changes at the end of the Cretaceous, while the highly specialized reptiles could not.

Like the Paleozoic, the Mesozoic era closed with a time of intense diastrophic activity, this time centering in the Cordilleran geosyncline. The sediments of this great trough, which had accumulated intermittently since earliest Cambrian times, were folded, thrust eastward, and uplifted high above sea level. At the same time other parts of the continent were raised vertically; in particular the old Appalachian range, worn down by the end of the Cretaceous to a nearly level plain, was warped upward so that erosion could begin anew. This period of disturbance, called the

Rocky Mountain or *Laramide revolution,* like earlier revolutions was accompanied by extremes of climate and by rapid shifts in the position of land and sea. Many groups of Mesozoic animals succumbed during this period, and others underwent rapid evolutionary development to keep pace with the changing environment. The reptilian groups which for more than 100 million years had ruled land, sea, and air were annihilated in a body; not a single dinosaur, plesiosaur, ichthyosaur, or pterosaur appears in Cenozoic rocks. The only reptiles which survived the destruction were such minor groups as the snakes, lizards, and turtles, whose descendants have an inconspicuous place in the present world.

Economic products in Mesozoic rocks are scanty compared with those of earlier eras. Important are oil in Texas and in Mexico, and gold in quartz veins associated with Jurassic batholiths. Quartz veins of the Sierra Nevada batholith and stream gravels derived from them were the lure that started the mad scramble of the forty-niners to California.

The Tertiary Period

Tertiary and *Quaternary* are terms which hark back to a common belief among early geologists that the earth's rocks could be divided into four major groups, Primary, Secondary, Tertiary, and Quaternary. The first two terms are no longer used, and the last two have changed their significance. Today they refer to the two periods of the Cenozoic era: the Tertiary lasting from the end of the Mesozoic to the time when ice sheets formed in the northern hemisphere, and the Quaternary from that time to the present.

In many ways the Cenozoic is markedly different from preceding eras. In the first place it is much shorter, only 60 million years compared with 150 million years for the Mesozoic and 350 million for the Paleozoic. Secondly, it is a time when the continents stood for the most part well above sea level. No longer did shallow seas spread widely; in North America, marine beds are found only in narrow strips along the Pacific coast and on the Atlantic coast from New Jersey south to Yucatan. The locally thick Tertiary beds east and west of the Rocky Mountains are river, lake, and wind deposits made in continental basins. Thirdly, climates during much of the Cenozoic had a diversity like those of the present: the distribution of plants and animals shows that instead of widespread moderate climates as in other eras, Cenozoic continents had zones of distinct hot climates, cold climates, humid and dry climates.

A fourth characteristic of Cenozoic times is widespread volcanic activity. From the Rockies to the Pacific coast lava flows and tuff beds testify to the former presence of volcanoes, some of which have only recently become extinct. In the mid-Tertiary immense flows of basalt inundated an area of nearly 200,000 sq mi in Oregon, Idaho, and Washing-

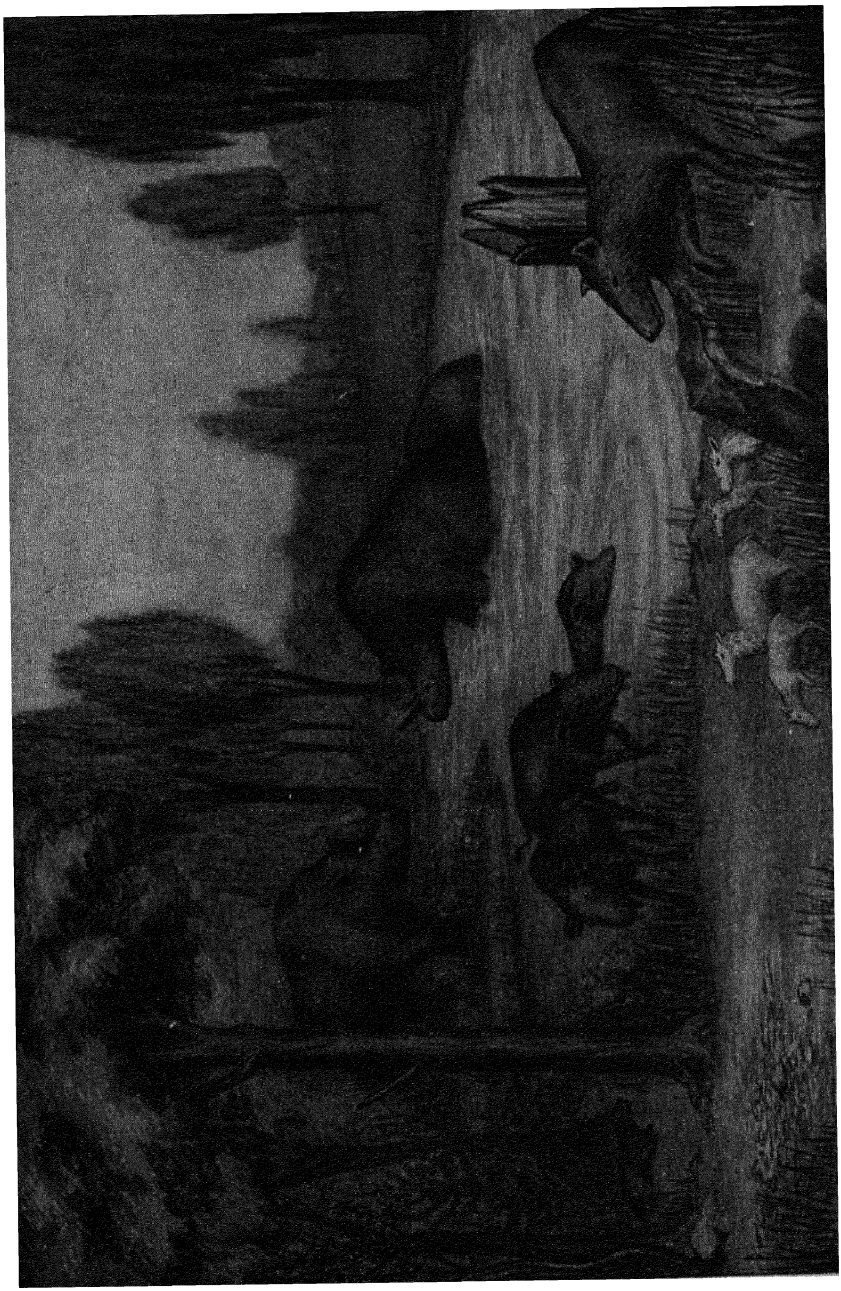


FIG. 301. *Early Tertiary mammals.* (*From Introduction to Geology by Branson and Tarr.*)

ton, some of the flows today forming the somber cliffs of the Columbia River gorge.

Finally, the Cenozoic was a time of almost continuous diastrophic disturbance, in contrast with the long periods of crustal stability in previous eras. Movements associated with the Rocky Mountain revolution lasted well into the Tertiary. In mid-Tertiary the Alps and Carpathians of Europe and the Himalayas of Asia were folded and uplifted. Toward the end of the Tertiary the Cascade range of Washington and Oregon was formed, and other mountain-building movements began

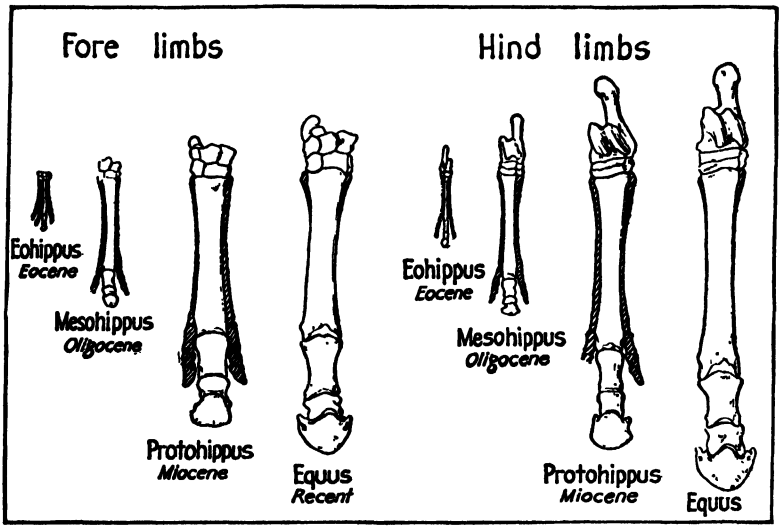


FIG. 302. *The lower part of the limbs of Tertiary and modern horses, showing the progressive enlargement of the middle toe and the gradual disappearance of the side toes. This specialization makes possible greater speed in running. (From Historical Geology by Moore.)*

around the border of the Pacific which have continued to the present day. Mountain ranges which had been folded earlier—the Appalachians, the Rockies, the Sierra Nevada—were repeatedly uplifted during the Cenozoic, and erosion following these uplifts has shaped their present topography.

The inhabitants of Tertiary seas were much like those of the present. Ammonites had disappeared with the dinosaurs as the Mesozoic ended, and even in the lowermost Tertiary beds we find fossils of clams, snails, sea urchins, crabs, and fishes not very different from modern forms.

Like the reptiles at the start of the Mesozoic, the mammals in the early Tertiary evolved quickly from a few primitive forms into species adapted for many different modes of life. Carnivores like cats and wolves, herbivores like horses and cattle, sluggish armored beasts like rhinoceroses, agile

creatures built for speed like deer and rabbits—ancestors of all these modern forms roamed the early Tertiary landscape (Fig. 301). A few mammals, like the whales and porpoises, became adapted to life in the sea; another line, the bats, developed wings. By the middle Tertiary mammals dominated the earth as the reptiles had before them.

Some of the mammalian lines have left fossils in sufficient abundance so that their development can be traced in detail from Mesozoic forms. In general, evolutionary changes in the Tertiary involve an increase in size, an adaptation of tooth structure for special diets, a specialization of limb structure for various postures and modes of life, and an increase in brain size. The modern horse, for example, is descended from a tiny rabbitlike creature of the early Tertiary; succeeding generations increased in size and brain capacity, their original five toes decreased to one powerful hoof, their legs became strengthened for speed, their teeth developed into efficient grinding instruments for a diet of grass (Figs. 302, 303).

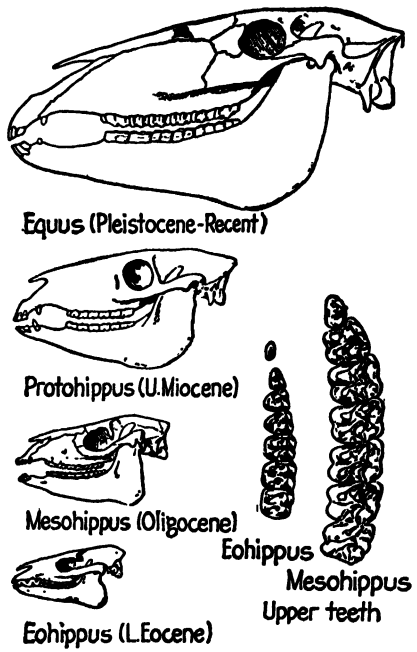


FIG. 303. *Stages in the evolution of the skull and teeth of the horse. (From Historical Geology by Moore.)*

Side by side with the mammals developed modern birds, modern insects, and the deciduous trees of modern forests. As the end of the Tertiary approached, both the physical and the organic world assumed more and more closely their present aspect.

Important economic materials from Tertiary rocks are gas and oil from marine beds in California and along the Gulf coast, gold in deposits formed by Tertiary igneous activity (including the famous "bonanzas" of Nevada), and gold from placers or stream gravels formed by the erosion of earlier vein deposits.

The Ice Age

The Quaternary period of the Cenozoic has lasted for about 2 million years. This is roughly one-thirtieth of the entire era. The 60 million years which make up the Cenozoic represent in turn about one-thirtieth of the total time since the earliest Pre-Cambrian sediments were laid down. If the

geologic record encompassed a thousand days instead of 2 billion years, the Quaternary would cover little more than a single day. About this last day we have a great deal of information, for its events are very close to our own time.

During the Pleistocene epoch, which makes up all but the last 30,000 years of the Quaternary, great icecaps formed in Canada and northern

FIG. 304. *Map of North America showing the maximum extent of Pleistocene glaciers. (W. C. Alden, U.S. Geological Survey.)*

Europe, and valley glaciers advanced in high mountains elsewhere. This was but the latest in a series of glacial periods which have punctuated earth history, but since these particular glaciers have taken part so directly in shaping present landscapes, the Pleistocene is often referred to simply as *the* Ice Age.

Why the climate at this particular time became cold enough for glaciers to form remains an unsolved riddle. Perhaps high mountain ranges and general continental uplift were responsible; perhaps great quantities of volcanic dust in the atmosphere dimmed the sun's rays;

perhaps the sun itself was shining less brightly. These and other theories have been argued at length, but none is entirely satisfactory.

Although our knowledge of ultimate causes is meager, we can find evidence to decipher the movements of the glaciers in extraordinary detail. This evidence comes from moraines that mark the limit of glacial advance at different times, from other material deposited beneath the glaciers as they moved, from scratched and grooved rock surfaces, from deposits in glacial lakes, from changes in valley forms and river courses.

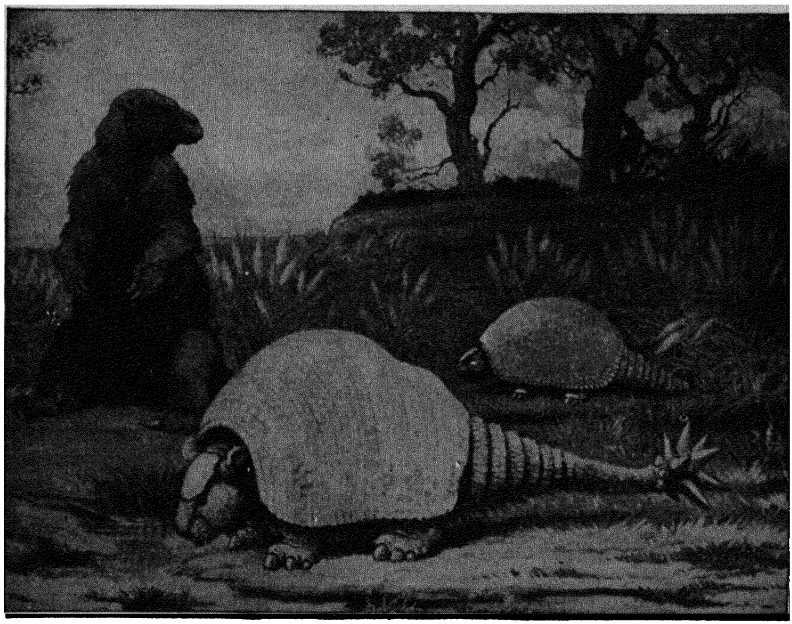


FIG. 305. *Three mammals which became extinct during the Pleistocene. Giant grouna-sloth, left; Daedicurus, foreground; Glyptodon, right. (C. R. Knight, Field Museum of Natural History.)*

Glacial deposits on this continent show that ice spread outward from three centers of accumulation in Canada, the ice front in its farthest advance reaching the Missouri River on the west and the Ohio River to the east (Fig. 304). Throughout this area the numerous lakes and swamps, the rounded hills, the abundant boulders of Pre-Cambrian rocks spread over the surface of eroded Paleozoic strata, would suggest the work of glaciers even to a casual observer. Closer study reveals fresh, almost unweathered deposits overlying older, deeply weathered glacial material which sometimes contains plant fragments. Thus the ice must have spread not once but several times, melting back between the periods of advance sufficiently for vegetation to flourish. Four different times of ice advance can be distinguished, with long interglacial periods between: during at

least one of these interglacial times the ice disappeared completely and the climate became warmer than it is today.

The changing climates of the Pleistocene proved a severe ordeal for the mammals. In the present world mammals are still dominant, but in numbers and diversity of species they have declined markedly since the later Tertiary. Of the many animals which became extinct during the Ice Age perhaps the most notable are the saber-toothed tiger, the ground sloth, the mammoths, and the mastodons (Fig. 305).

Against this background of shifting ice fronts and a declining mammalian population was played the drama of early human history. Unfortunately man and his nearest relatives, the monkeys and apes, have left but a scant record of their evolutionary progress during the Cenozoic.

FIG. 306. *Restorations of early men. Pithecanthropus erectus, the Java apeman, left; Neanderthal man, middle; Cro-Magnon man, right. The last two were predecessors of modern man in Europe.* (*American Museum of Natural History.*)

Probably our ancestors were too clever to let themselves be buried often in the stream gravels and swamps and tar pits which entombed less wary mammals; dying on land, their bodies decayed or were eaten instead of being preserved as fossils. Of the few fossils which we can claim as probable direct progenitors, the oldest is a small ratlike primate of the early Tertiary. By the mid-Tertiary descendants of this creature had attained a height of some 2 ft, a face more like that of modern apes, and a vastly improved brain. Probably sometime in the late Tertiary man's evolutionary branch split off from that of the apes and monkeys, for crude stone implements are sometimes found in uppermost Tertiary beds. But true human fossils have not been discovered in deposits older than early Pleistocene.

Even early in the Ice Age human creatures had spread widely over the earth, for their remains have been found in Java, in China, and in England. The later record is most complete in Europe, where stone

implements, burial sites, drawings on caves, and skeletal fragments give a fairly connected history. Unfortunately the history does not show a continuous development but rather a succession of races, each flourishing for a time and then being supplanted by another (Fig. 306). Our own immediate ancestors did not appear in Europe until the retreat of the last ice sheet, some 30,000 years ago. Of their previous history we know nothing. Since the successive invasions apparently came from the East, we shall probably have to wait for further exploration of Central Asia before we can trace more fully man's development in the Ice Age.

Because the Cenozoic era is so brief, because it is a time of mountain building, of continental elevation, of climatic change, there is some question as to whether it should be considered an era at all—at least in the sense that we regard the Mesozoic and Paleozoic as eras. Events in the Cenozoic seem important and the times involved very long simply because the events are so close to us compared with those of earlier times. From a longer perspective the whole Cenozoic may perhaps be better regarded as a part of the Rocky Mountain revolution.

Whatever the answer to this problem, it is certain that we live in a time which, geologically speaking, is highly unusual. The widespreading shallow seas of earlier ages are nearly absent; mountains stand high above sea level; large parts of the earth are desert: temperatures vary widely from one part of the earth to another. Similar conditions we find in the past only during major revolutions. If we are living in a time of revolution, the future may hold more diastrophic activity, continued volcanic eruptions, perhaps another advance of glaciers. Or perhaps we live just at the close of the revolution, so that the future should bring a time of milder climates, of gradually lowering mountains, of seas spreading far inland.

Whatever the earth's future may be, prospects for our own race seem bright. Man is a newcomer on the earth: his 2 or 3 million years of existence are very short compared to the life histories of other animals. Further, man is still adaptable to new conditions of diet and climate, so that he should be able to survive even fairly rapid geologic changes. His brain gives him many advantages, notably an ability to modify his environment if climatic conditions become too extreme. Barring unforeseen catastrophes, it should be many millions of years before man follows the trilobite, the dinosaur, and the saber-toothed tiger into extinction.

Questions

1. In what parts of North America would you expect to find
   a. Early Paleozoic strata in horizontal layers?
   b. Continental deposits of Tertiary age?
   c. Large exposures of pre-Cambrian rocks?

d. An angular unconformity between Mesozoic and Paleozoic strata?
e. An angular unconformity between Jurassic and Cretaceous strata?

2. In rocks of what era or eras would you expect to find fossils of the following?

a. Plesiosaurs.	*e.* Pterosaurs.	*i.* Clams.
b. Horses.	*f.* Elephants.	*j.* Apes.
c. Brachiopods.	*g.* Ferns.	*k.* Birds.
d. Trilobites.	*h.* Ammonites.	*l.* Insects.

3. What are the probable major steps in the evolution of the vertebrates from fishes to mammals? At what times did the different forms develop?
4. What is the evidence that the earth's surface was not appreciably warmer in the Proterozoic than it is today?
5. In the New England region a local mountain-building disturbance occurred at the end of the Ordovician period. On what sort of evidence would such a conclusion be based?
6. What characteristics of the Cenozoic era suggest that it may be a continuation of the Rocky Mountain revolution?
7. How could you tell from an examination of rocks at the surface that you were on top of a partly eroded anticline? Suppose that a well drilled into the crest of the anticline strikes gas: where would you drill a second well in order to strike oil?
8. At what other times besides the Pleistocene are there records of extensive glaciation? What sort of evidence of these earlier glacial periods would you expect to be preserved?
9. With rocks of what ages are large deposits of each of the following found in North America? (*a*) gold, (*b*) coal, (*c*) iron, (*d*) oil.
10. Describe the landscape in the Appalachian Mountain region of eastern Pennsylvania as it appeared (*a*) in the early Paleozoic, (*b*) during the Pennsylvanian period, (*c*) at the beginning of the Mesozoic.
11. At what time in geologic history
    a. Was most of the Mississippi Valley covered by a shallow sea?
    b. Did a mountain range exist in the Grand Canyon region?
    c. Were extensive mountains first formed in the Rocky Mountain region?
    d. Were the North Central States covered by an icecap?

Suggestions for Further Reading—Part V

On physical geology:

LONGWELL, C. R., A. KNOPF, and R. F. FLINT: *Textbook of Geology*, Part I, John Wiley & Sons, Inc., New York, 1939. A standard elementary text.

EMMONS, W. H., G. A. THIEL, C. R. STAUFFER, and I. A. ALLISON: *Geology*, McGraw-Hill Book Company, Inc., New York, 1939. A standard elementary text.

CRONEIS, C., and W. C. KRUMBEIN: *Down to Earth*, University of Chicago Press, Chicago, 1936. More elementary and more popularly written than the two preceding books. Contains a section on historical geology.

On historical geology:

MOORE, R. C.: *Historical Geology*, McGraw-Hill Book Company, Inc., New York 1933. A standard elementary text.

SCHUCHERT, C., and C. DUNBAR: *Textbook of Geology*, Part II, John Wiley & Sons, Inc., New York, 1941. A standard elementary text.

SNIDER, L.: *Earth History*, D. Appleton-Century Company, Inc., New York, 1932. More elementary and more popularly written than the preceding books.

On weather and climate:

TREWARTHA, G. T.: *An Introduction to Weather and Climate*, McGraw-Hill Book Company, Inc., New York, 1937. A standard elementary text.

PETTERSSEN, S.: *Introduction to Meteorology*, McGraw-Hill Book Company, Inc., New York, 1941. Modern techniques of weather forecasting. Nonmathematical, but requires a background of elementary physics.

On the history of geology:

GEIKIE, A.: *The Founders of Geology*, The Macmillan Company, New York, 1905. A series of short biographies of the great geologists.

FENTON, C. L., and M. A. FENTON: *The Story of the Great Geologists*, Doubleday & Company, Inc., New York, 1945. Well-written biographies, especially good for American geologists of the past century.

PART VI

STARS AND GALAXIES

We have subjected our small planet to a pretty thoroughgoing inspection. The gases of the atmosphere, the water of the oceans, the solid rock of the crust, and the metallic material of the deep interior have all been passed in review. We have watched the slow chemical decay and mechanical disintegration of rocks, the wearing away of soil and rock by streams, waves, wind, and glaciers, the piling up of rock debris in low basins and in the shallow margins of the oceans. We have seen the leveling work of erosion and deposition undone by the buckling and fracturing of solid rock and by the outpouring of liquid rock from volcanoes. From successive rock layers we have read the story of 2 billion years of earth history.

In present-day changes of the earth's surface and through all its long past we find at work physical and chemical processes similar in nature to laboratory processes which we have discussed in earlier chapters. Behind these processes we see the atoms, molecules, ions, and electrons of which the earth's matter is composed; behind them, too, we see the radiant energy from the sun in which the earth is constantly bathed. The tiny particles and the energy transformations of the physicist's laboratory give us a basis for interpreting, at least in outline, all the varied phenomena of the earth in simple terms.

There remains for our consideration the universe beyond the earth. Can we extend to the universe, as we have to all parts of this planet, the ideas of modern physics regarding the ultimate building blocks of matter?

We shall not stop to discuss the other planets and satellites of the sun's family. These are solid bodies like the earth, with no light of their own, with diameters at most no more than a dozen times the earth's. We should find peculiar conditions of temperature and pressure on the planets, to be sure, but nothing in the nature or behavior of matter fundamentally different from our experience on the earth.

We shall look beyond the planets to the great universe of which the solar system is but a tiny part. We shall find that matter in the universe is largely concentrated in the huge aggregates called stars, bodies of an altogether different sort from planets and satellites. Temperatures in the stars range up to millions of degrees, pressures up to hundreds of millions of atmospheres. Under these conditions matter behaves quite differently from the matter of our everyday acquaintance. Familiar compounds do not exist; even atoms become unstable; collisions of atomic nuclei, which we bring about with such trouble and expense in our laboratories, are a normal occurrence in stars. But we shall find that the basic physical and chemical principles which we have derived from terrestrial experiments are powerful enough to give us some understanding of the stars even out to the far corners of the universe.

CHAPTER XL

The Sun

IN EARLIER chapters we have discussed the solar system in considerable detail. Members of this orderly group of circling planets are isolated from each other by distances which seem enormous by any ordinary standard. Separating the earth and the sun are 93 million miles, separating Pluto and the sun 3½ billion. In spite of these distances telescopes show us most of the planets as clear disks and on several reveal a good deal of surface detail.

Distances among the stars make the solar system shrink to insignificance. Separating us even from the nearer stars are distances measured not in millions or in billions, but in *trillions* of miles. Light reaches us from the sun in 8 min, but light from the nearest star has been on its way for almost four years. If an airplane could fly from the earth to the sun, maintaining a steady speed of 200 mi/hr, its journey would require fifty years; but in order to reach the nearest star it would have to travel 12 million years. An airplane which had set out for this star at the beginning of the Ice Age would now be only well started on its journey.

These appalling distances present an immediate and serious difficulty in any attempt to study the universe. They mean, to begin with, that *stars appear simply as points of light even in the largest telescopes.* We can never hope to see the surface detail of a star, as we can that of a planet.

One exception to these statements is the star which forms the center and the source of energy for our planetary system, the sun. This one star is close enough for detailed examination. Fortunately the sun appears to be an average sort of star, so that a study of its surface should give us a reasonably accurate idea of what other stars would look like if we could inspect them from a vantage point 93 million miles away.

In this chapter we shall begin with a discussion of the instruments which have made stellar astronomy possible, and then see what these instruments can tell us about the nearest star of all.

Telescope, Camera, and Spectroscope

Modern stellar astronomy begins toward the end of the eighteenth century, with the work of Sir William Herschel. For nearly a century astronomers had been chiefly concerned with refining and elaborating

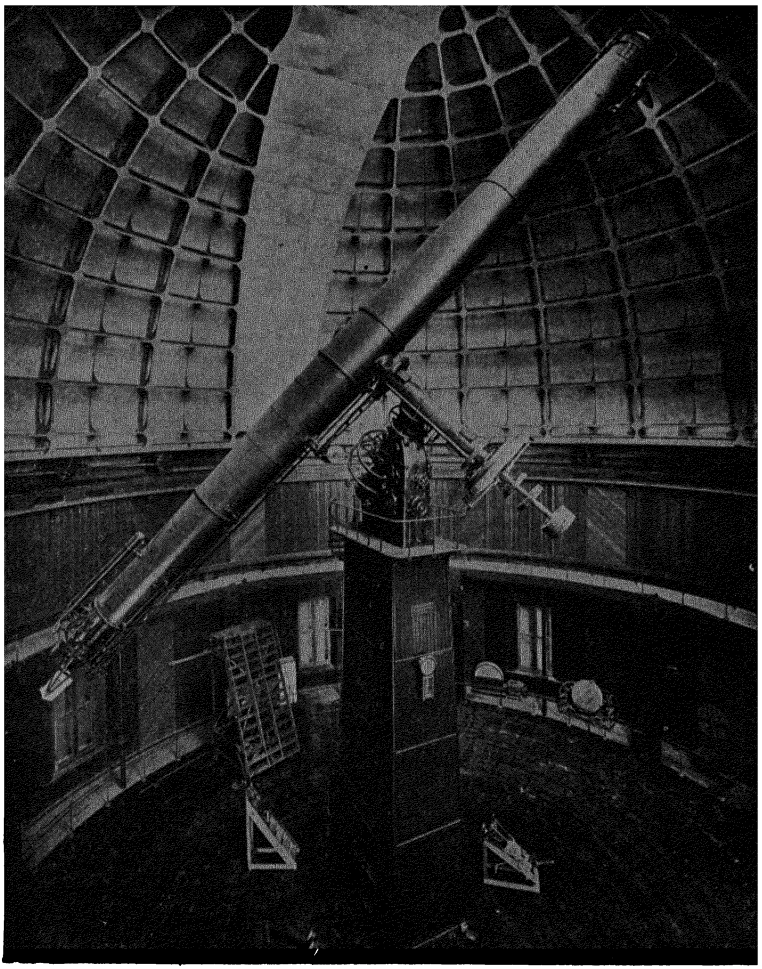


FIG. 307. *The refracting telescope of the Lick Observatory, Mt. Hamilton, California. The lens, mounted at the upper end of the tube, has a diameter of 36 in.*

Newton's great synthesis of the solar system. Herschel, the same Englishman who accidentally discovered the planet Uranus, sought to find in the stars something of the same regularity of structure and orderliness of motion which Newton and his predecessors had found in the sun's family of planets. Like a pioneer in any branch of science, Herschel began

with patient observation and classification: he spent many years in cataloguing, counting, and observing the motions of stars in all accessible parts of the sky. From this study he was able to deduce a structure for the universe which in all essentials is the same as the one which modern astronomers believe to be correct.

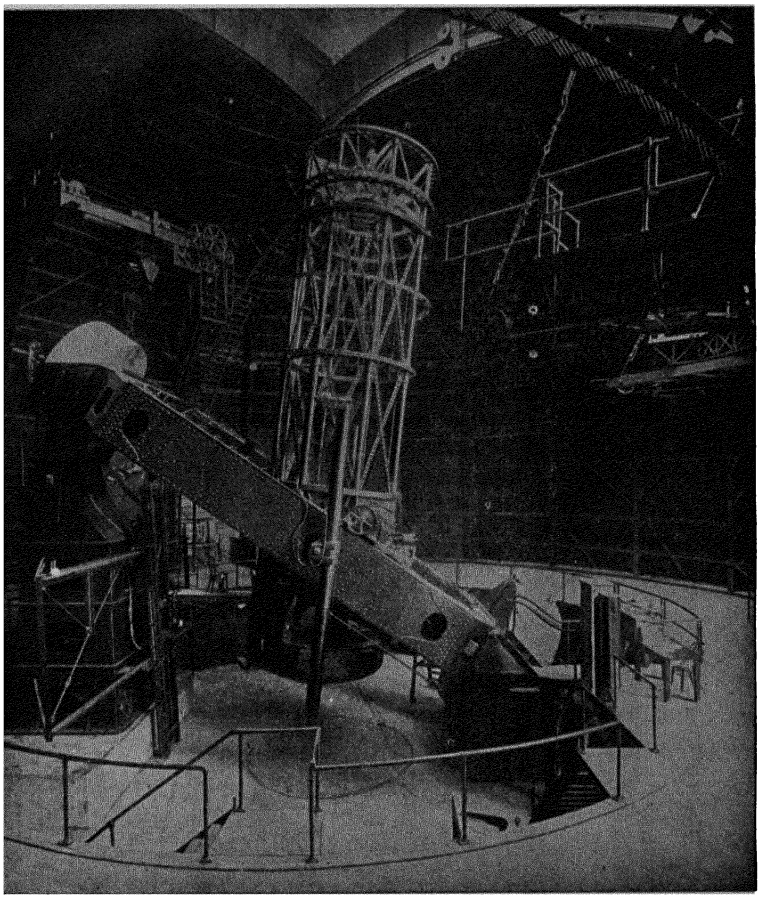


FIG. 308. *The 100-in. reflecting telescope of the Mt. Wilson Observatory, Mt. Wilson, California. The great mirror is mounted at the lower end of the tube.*

In Herschel's time just as in our own day the telescope was the basic astronomical instrument, and his success rests largely on the improvements which he introduced in telescope construction. Herschel was the first to build and use a large *reflecting* telescope, an instrument in which light is reflected from a concave mirror instead of being refracted through a lens. All the larger modern telescopes used in studying the stars are of the reflecting type (Figs. 307, 308).

In stellar astronomy the chief purpose of a big telescope is not to secure great magnification, for no obtainable magnification can make the stars appear larger than points of light. The great advantage of big mirrors and big lenses is their light-gathering power: more light from a given object can be brought to a focus by a large surface than by a small one. Thus faint objects which would otherwise be invisible are revealed by a large telescope, and more light from other objects is available for study.

Within a century after Herschel's time another extremely valuable instrument was introduced into stellar astronomy, the camera. The camera is used with a telescope, light collected by the lens or mirror falling on a photographic plate rather than on the astronomer's eye. Photographic plates have a great advantage over the eye in that they do not become fatigued: the longer a plate is exposed to faint light, the clearer the resulting image will be. A telescope with camera attached can be trained on the same star for hours, or if need be for several nights, so that the image of the star may be made as dark as is desired. Objects much too faint for the eye to detect are faithfully recorded by the camera, so that with its aid we can see more objects and more distant objects than would otherwise be possible. Photographic plates have the further advantage that their records are permanent, so that positions of stars as we see them today can be compared with those photographed years ago.

The most valuable item in the modern astronomer's equipment is an instrument whose extraordinary usefulness has become apparent only in the last half-century. This is the spectroscope, the same instrument which has contributed so much to our knowledge of atomic structure. Perhaps it is not right to single out the spectroscope as more important than the telescope and camera, since it is nearly always used in combination with these instruments, but to the spectroscope we owe a vast amount of information concerning the stars which telescope and camera alone could not possibly give.

A spectroscope is designed to break up light into its separate wave lengths, as droplets of water in the atmosphere break up sunlight into the colors of the rainbow. The spreading apart of the different wave lengths is accomplished in one of two ways: (1) by means of a prism, which refracts short waves more than long waves (page 265); or (2) by means of a grating of fine, closely spaced lines on a polished metal surface, which causes light waves reflected from it to interfere. The band of colored light, or *spectrum*, produced by the prism or grating is focused on a photographic plate and so recorded for comparison with other spectra. Since each wave length can thus be singled out from the others, the spectroscope makes possible an exhaustive analysis of the light which reaches us from a star.

Modern astronomers have available a variety of other instruments, but the most widely useful remain the telescope, the camera, and the spectroscope.

Information from Spectra

Spectra are so all-important in stellar astronomy that they deserve not only a review of Chap. XXII but a few additional comments here.

The spectrum of a star is not at all impressive. If photographed in natural colors, it generally consists of a rainbow band crossed by a multitude of fine dark lines. Ordinarily the colors are not reproduced, so that the spectrum shows simply black lines on a light gray background. Usually two photographs of an iron-arc spectrum are taken on the same plate, to make possible a direct comparison of lines in the star's spectrum with known lines of the iron spectrum. Thus stellar spectra usually look like the one in Fig. 309, in which the middle strip is a star's spectrum and the bands on either side are iron spectra.

At first glance it does not seem that a few black lines on a photographic plate can get us very far in understanding the stars. But each of those lines has its own story to tell about the conditions which produced it, and an expert can piece together the information from different lines into a remarkably accurate picture of an entire star. Some of the types of information directly obtainable from spectra are outlined in the paragraphs below; we shall find later that this list is far from an exhaustive one.

1. *Structure.* A spectrum of dark lines on a continuous colored background is the type which we have previously (page 307) called an *absorption spectrum.* It is produced when light from a hot object passes through a cooler gas: atoms and molecules of the gas absorb light of certain wave lengths—wave lengths which they would emit if they themselves were hot—and so leave narrow gaps in the band of color. Thus a star which has this kind of spectrum (and nearly all of them do) reveals at once something of its structure: it must have a hot, incandescent interior surrounded by a relatively cool gaseous atmosphere.

The continuous background of the spectrum unfortunately tells us little about the star's interior. Continuous spectra are produced by any incandescent solids, liquids, or highly compressed gases—any form of matter in which the atoms are close enough to interfere with each other's motions. A heated gas at low or moderate pressures, in which each atom can emit its own characteristic radiation without interference from its neighbors, gives a discontinuous spectrum consisting of isolated bright lines (page 267). Thus we know that the interior of an average star cannot be a gas at moderate pressures, but whether it consists of solid, liquid, or highly compressed gas the continuous spectrum does not tell us.

2. *Temperature.* One valuable datum derived from the continuous background of a star's spectrum is the temperature of its "surface"—*i.e.*, the part from which we receive light.

When an object—say a piece of iron—is heated slowly in the laboratory, it gives out at first infrared radiation, the invisible "heat" radiation which you could detect by holding your hand near the iron. As the temperature rises, the iron begins to glow dull red; now a part of its radiation is visible light, although most is still in the longer wave lengths beyond the red. With further heating the iron glows brightly, first red and then orange; thus more and more of its radiation is in the visible range, and the maximum intensity of the radiation moves toward lower wave lengths. At very high temperatures the light becomes yellow and then white; here the maximum intensity is in the middle of the visible range, so that all colors are sufficiently represented to combine into white. Higher temperatures cannot be obtained in the laboratory, but they should make the light progressively bluer as the shorter wave lengths become more and more prominent. Thus the color of a piece of heated iron is a rough measure of its temperature. The measure can be made more exact by an examination of the spectrum and determination of the wave length where the radiation is most intense.

Other substances on heating behave similarly to iron. Except for a correction depending on the kind of surface exposed, for any material the wave length of greatest intensity of radiation depends only on temperature. This means that star temperatures can be estimated simply from color: blue stars are hotter than white stars, and these in turn are hotter than yellow or red stars. It means further that the temperature of a star can be accurately determined by finding in what part of its continuous spectrum its radiation is most intense.

3. *Composition.* Since each element has a spectrum consisting of lines with characteristic wave lengths, the elements present in a star's atmosphere may be identified from the dark lines in its spectrum. In principle the method is very simple: one need only determine the wave length of each line in the spectrum, either from a wave-length scale or by comparison with the iron-arc lines, and then compare these wave lengths with those produced by various elements in the laboratory.

4. *Condition of matter.* In practice the identification of lines in a star's spectrum is beset with difficulties, since the wave lengths and intensities of the lines emitted or absorbed by an element depend on such conditions as temperature, pressure, and degree of ionization. These difficulties prove to be blessings in disguise, however, for once the lines are identified they reveal not only what elements are present in a star's atmosphere but something about the physical conditions in which the elements exist.

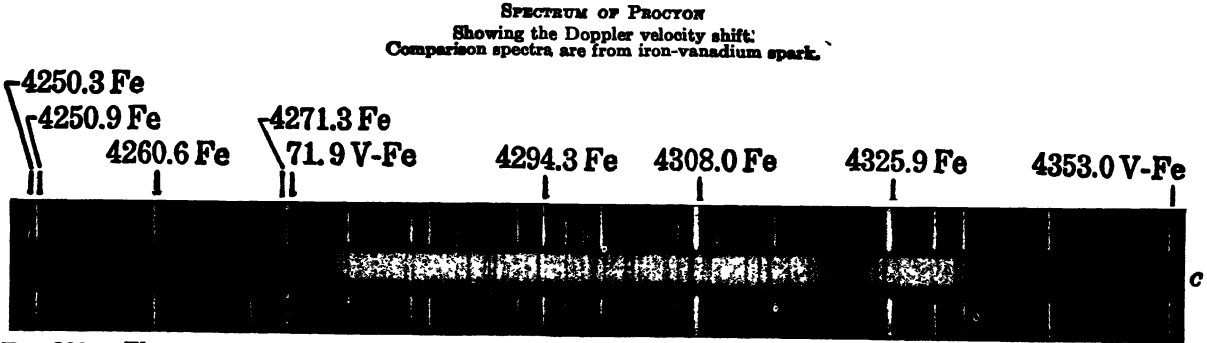


FIG. 309. *The spectrum of the bright star Procyon (middle strip), with comparison spectra on either side. Note that many dark lines in the star's spectrum nearly correspond in position to the bright iron lines, but that between them is a slight shift. This shift is toward the violet end of the spectrum, indicating that Procyon and the earth are approaching each other. The amount of the shift indicates a relative velocity of about 30 km (18 mi) per second. (V. M. Slipher, Lowell Observatory, Flagstaff, Ariz.)*

Chemical compounds also have spectral lines of recognizable wave lengths, so spectra provide a means of determining how much of the matter in a star's atmosphere is in the form of compounds rather than elements.

A serious limitation on the use of spectral lines for determining the composition and physical conditions of a star's outer layers is imposed by the absorption of light in the earth's atmosphere. A small amount of ozone in the stratosphere effectively blocks out all ultraviolet radiation except for a small range of wave lengths just below the violet—probably a fortunate circumstance for man's health, but a calamity to astronomers since the lines of several elements occur only in the ultraviolet region. Isolated wave lengths elsewhere in the spectrum are obscured by other atmospheric gases.

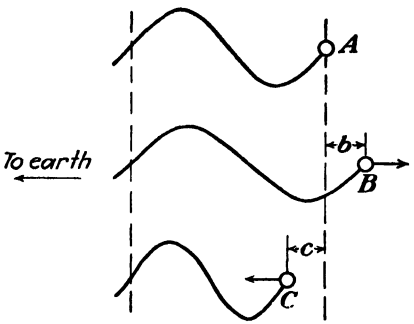


FIG. 310. *Diagram to show the effect of a star's motion on the wave lengths of its light. Star A is stationary, star B is receding from the earth, star C is approaching the earth. Distance b shows how far B has moved during the emission of a single light wave, hence how much the wave length is increased. Distance c shows the decrease in wave length due to approach.*

5. *Motion.* A star moving away from the earth has a spectrum in which each line is shifted slightly toward the red end, whereas a star moving toward us has a spectrum with lines shifted toward the violet end (Fig. 309). The amount of the shift makes possible a calculation of the speed with which the star is approaching or receding.

The shift is easily explained (Fig. 310). Imagine a star emitting light of only one wave length. If the star is receding, it moves a short distance between the emission of each wave and the next succeeding one. Thus each wave starts from a point a little farther away than the last one, and it appears to us that the distance between waves is a little greater than it would be if the star were motionless. Now a greater distance between waves, or a longer wave length, means a slight change in color toward the red. So if the star were at first motionless and then started to recede, the single line in its spectrum would shift toward the red end, the amount of shift depending on the rate of motion. Similar reasoning applies to an approaching star: the wave length of its light would be shortened because the star moves a short distance toward us between each wave and the next.

In the radiation from a real star all wave lengths are shifted by the same amount toward one end or the other. The dark lines shift as well as

the continuous background, since wave lengths which are absorbed are affected just as much as those which are not. Note that *this spectral shift records only motion of approach or recession;* motion of a star across the line of sight makes no change in its spectrum.

Solar Statistics

Let us turn now to the sun, to see what telescope and spectroscope can show us about the one star available for close-range examination. We have already discussed the sun briefly in Chap. II, but for the sake of completeness we shall repeat a few of the statistics given there.

The sun weighs as much as 333,000 earths, and its huge volume would hold the matter contained in 1,300,000 earths. Enormous as these figures sound, we shall find them rather smaller than average among the stars. Combination of the figures for mass and volume gives an average density for the sun of 1.4 g./cc. This is little more than a quarter of the earth's average density and about half the average density of ordinary rocks. Such comparisons suggest at once that matter in the sun is very different from that which makes up our planet.

We recall from earlier pages (page 27) the conspicuous features of the sun when seen through a telescope: the darkening near the edge, and the sunspots which move slowly across the disk as the sun rotates. We recall further that both the darkening at the edge and the fact that different parts of the sun rotate with different speeds suggested that the outer part of the sun must be fluid.

Temperature measurements prove beyond question that at least a large portion of the sun is gaseous. The temperature of the sun's visible surface, determined both from the spectrum and from direct measurements of the amount of radiation reaching the earth, turns out to be around 6000°C. This temperature, higher than any which can be maintained in the laboratory, is sufficient to vaporize all known substances. Even high pressure cannot prevent vaporization of ordinary matter at such a temperature.

Since the spectrum of the sun is a typical absorption spectrum, an atmosphere of relatively cool gas must surround the visible part of the sun. If the part from which light is emitted is also gaseous, what meaning can we give the term "surface of the sun"? Evidently it is simply the lower limit to which we can see, the level beyond which the gas is so compressed and emitting so much radiation that it becomes opaque. There is no sharp break between the sun's atmosphere and its interior, as there is on the earth; both atmosphere and interior are gaseous, the one grading into the other.

Of the thousands of dark lines in the sun's spectrum (Fig. 311) many can be identified with those of elements known on the earth. The remain-

ing lines are in all probability produced not by unknown elements but by familiar elements under peculiar conditions of temperature and pressure. In all, about sixty of the ninety-odd known elements have been detected in the sun's atmosphere, and the remaining ones would probably be found if it were possible to examine the far ultraviolet part of the spectrum. Some of the elements are in their normal states, others are ionized. Lines of only a very few extremely stable compounds are recognizable; temperatures in the parts of the atmosphere where most of the light is absorbed are high enough to decompose nearly all molecules into atoms.

Although conditions on the sun are so different from those on the earth, the elementary substances which make up the two bodies appear

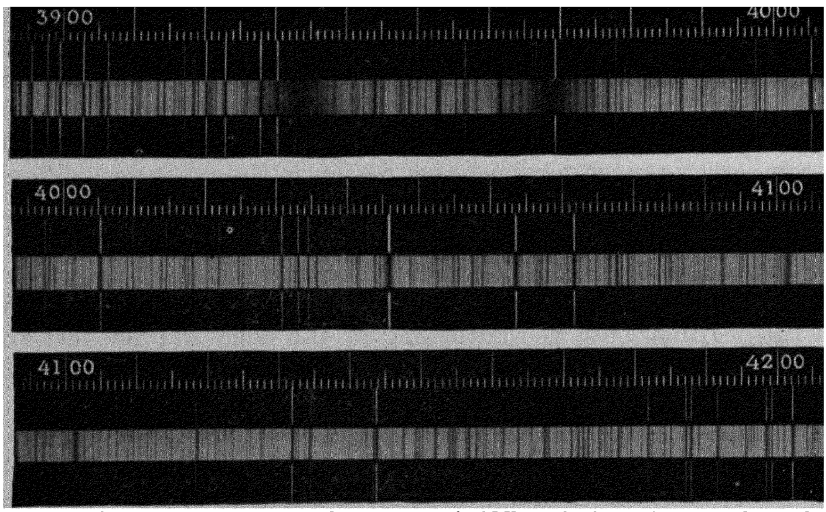


FIG. 311. *A small part of the sun's spectrum (middle strip in each of the three photographs) with comparison spectra of the iron arc. Note that many of the iron lines correspond with dark absorption lines in the sun's spectrum. (Courtesy of Mt. Wilson Observatory.)*

to be the same. Even the relative amounts of different elements are similar, except for a greater abundance of such light elements as hydrogen and helium on the sun. At the low temperatures prevailing on the earth, most of the elements have combined to form compounds; in the sun's terrific heat the elements are present mostly as individual atoms, many of them ionized. The general similarity in composition between earth and sun is good support for the idea that the earth's material was once a part of the sun or a similar star.

Sunspots, Prominences, Corona

The nature and origin of sunspots (Fig. 312) have long been astronomical mysteries. The spectroscope has made possible a partial explanation, but several important questions remain unanswered.

Spectra of large spots show that temperatures within them are some 2000° lower than normal sun temperatures. Even so, sunspots are sufficiently hot to be brightly luminous; they appear black only because we see them against a more brilliant background. Because of the lower temperature, spectra of sunspots show the presence of more compounds than normally exist in the sun's atmosphere.

A further hint as to the nature of sunspots comes from photographs like that shown in Fig. 313. This picture, taken with the aid of a spec-

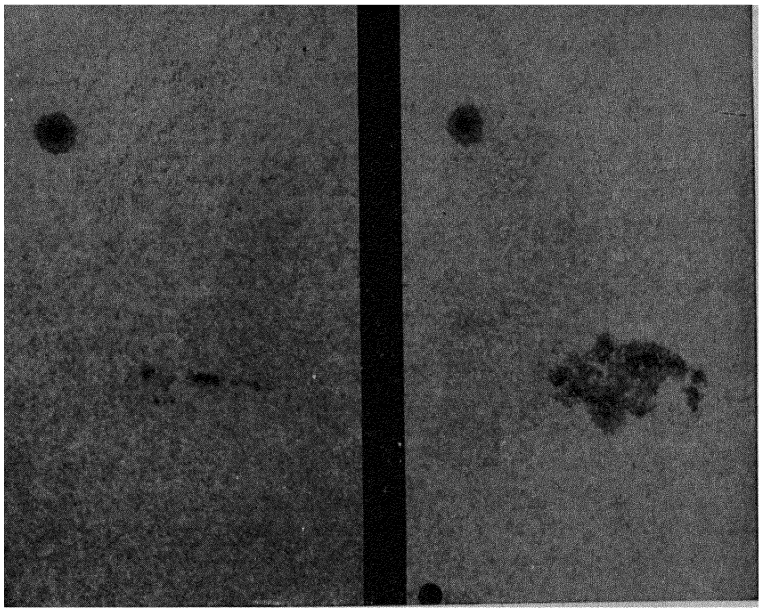


FIG. 312. *Development of a sunspot during 24 hr, Aug. 18 and 19, 1917. The small black circle in the right-hand picture shows the size of the earth on the same scale. (Courtesy of Mt. Wilson Observatory.)*

troscope, shows the sun's surface near a pair of spots when photographed with the light of a single element, in this case hydrogen. It shows, therefore, the distribution of hydrogen alone at a certain level in the sun's atmosphere. Its important feature is the suggestion of spiral motion in the neighborhood of each spot.

Detailed spectroscopic study confirms the hint given by these pictures that sunspots consist of gas in rapid, spiral motion. The gas moves in general outward from the sun's interior, expanding and thereby cooling itself as it spirals out into regions of lower pressure. Thus sunspots are gigantic whirlpoollike disturbances in the gases of the sun's outer layers, in many respects similar to the cyclonic storms of the earth's atmosphere.

Another perplexing problem connected with sunspots is the periodic change in their number. Approximately every eleven years the number of

visible sunspots reaches a maximum, declining thereafter until six or seven years later practically no spots appear, and then rising once more to the maximum. An effect of the sunspot cycle on the earth's weather seems fairly well established, years of sunspot maxima corresponding approximately with years of abundant rain. Other terrestrial influences of sunspots are an increase in the brilliance of the aurora borealis, or "Northern Lights," during sunspot maxima, and an increase in the intensity of the

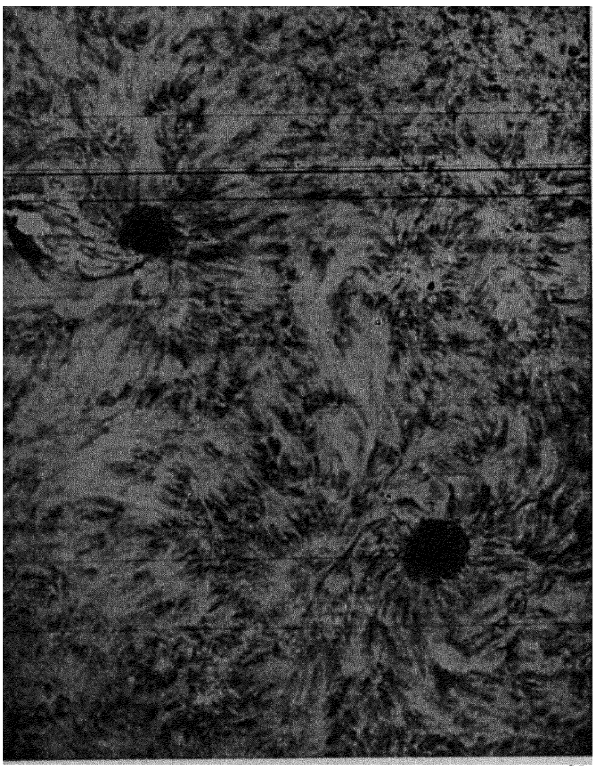


FIG. 313. *Photograph of two spots showing whirlpoollike structure. Note that whirls about the two spots seem to be in opposite directions. (Courtesy of Mt. Wilson Observatory.)*

magnetic storms which sometimes play havoc with radio and telegraph communication.

The outrushing gases from sunspots are in part carried high above the surface to form pinkish, flamelike *prominences* (Fig. 19, page 28). These clouds of luminous gas are best seen at the edge of the sun during a total eclipse, when the moon hides the brilliant disk. With the spectroscope, however, it is possible to photograph and study them at any time. Spectra of the prominences show that the principal gases present are hydrogen, helium, and ionized calcium; the lines of helium in these spectra

led to the prediction of this element's existence several years before it was found on earth. What makes prominences luminous is not clear, but it is probably some sort of excitation by electrically charged particles shot out from the sun.

During a total eclipse of the sun a wide halo of pearly light surrounds the dark moon (Fig. 20, page 29). This is the sun's *corona*, an exceedingly diffuse outer atmosphere consisting of various kinds of atoms, molecules, and electrically charged particles. So faint is the corona that it becomes visible only when the sun's disk is completely hidden, but spectroscopically it can be studied at other times.

The Source of the Sun's Energy

Here on the earth, 93 million miles from the sun, a surface 1 sq cm in area exposed to the vertical rays of the sun receives an average of nearly 2 cal of heat energy per minute. Adding up all the energy received over the earth's surface gives a staggering total, yet this is but a tiny fraction of the sun's total radiation. So prodigious is the output of energy at the sun's surface that a layer of ice 12 yd thick completely surrounding the sun would be melted in less than a minute. What possible source can we imagine for this huge and seemingly inexhaustible supply of energy?

The answer that first suggests itself is combustion, for fire is the only familiar source of energy that seems at all comparable to the sun. Yet a moment's reflection will show how impossible any sort of burning would be. The sun is too hot to burn: burning implies the combination of other elements with oxygen to form compounds, but on the sun nearly all compounds are decomposed by the terrific heat. Even if burning were chemically possible, the heat obtainable from the best fuels known would be woefully inadequate to maintain the sun's temperature.

The sun's energy must somehow be produced by processes taking place in the interior. Although the sun's interior cannot be observed directly, theoretical considerations make possible some shrewd guesses about conditions which exist there. Pressures must be high even at moderate depths, simply because of the weight of overlying material. Temperatures must increase rapidly toward the interior, since a continuous flow of energy is supplied to the surface to make good the prodigious losses by radiation. Mathematical analysis and a few reasonable assumptions lead to an estimate of 20,000,000°C for the temperature and 1 billion atmospheres for the pressure near the sun's center.

The behavior of matter even under these fantastic conditions is not impossible to picture. Temperatures of millions of degrees cannot be approached in our laboratories, but we can supply atoms with corresponding amounts of energy by subjecting them to short-wave ultraviolet radiation, X rays, and bombardment with tiny particles. Absorption

of energy in such processes makes atoms lose not only their outer electrons, as in ordinary ionization, but electrons from deep within their electron clouds. At temperatures of several million degrees, light atoms would lose all their electrons, heavy atoms all but those in their innermost shells. Thus matter in the sun's interior probably consists of broken-down atoms—free electrons in great numbers, and positive nuclei surrounded by a few electrons or none at all.

These atom fragments are in a state of wild commotion, moving far more rapidly than the molecules of an ordinary gas. Since both electrons and nuclei are very small in comparison with the atom's diameter, the moving particles are separated by distances relatively great compared with their size, even in the very dense matter near the sun's center. Hence the aggregate of electrons and stripped atoms in the sun's interior fulfills the essential requirements of a gas—particles far apart and rapidly moving. In this sense the sun is gaseous throughout, but the "gas" is vastly different from any that we know on earth.

Collisions between particles in the sun's atom-fragment gas must be frequent and violent. Electrons, protons (hydrogen nuclei), alpha particles (helium nuclei), and heavier nuclei continually rush headlong into one another. These collisions duplicate on a huge scale the "bombardment" experiments of modern physical laboratories, in which particles given out by radioactive elements or speeded up electrically are used to bombard atoms. One result of bombardment experiments is the demonstration that part of the mass of two colliding particles may be converted into energy, disappearance of a small amount of mass producing relatively enormous quantities of energy (page 296). Now in the sun's interior conditions are ideal for such energy-producing collisions—not as events affecting rare, isolated atoms, like those in our laboratories, but as commonplace events occurring many times a second in every cubic centimeter of the sun's material. *This mechanism, the partial conversion of mass into energy during collisions of electrically charged particles, provides an adequate source for the sun's energy.*

The energy-producing reaction in the sun is the conversion of hydrogen into helium. This takes place not directly but by a series of steps in which carbon nuclei are bombarded by a succession of hydrogen nuclei (protons). Each step can be duplicated on a small scale in the laboratory; for each, the energy required and the energy given out can be measured. For the entire process, the energy available per helium atom corresponds to the difference in mass between 4 hydrogen atoms and 1 helium atom ($4 \times 1.008 - 4.002$, or 0.030 units on the atomic-weight scale). Every 4 g. of helium produced mean the liberation of

$$E = m \times 9 \times 10^{20} = 0.030 \times 9 \times 10^{20} = 3 \times 10^{19} \text{ ergs of energy}$$

[Eq. (29), page 296]

In a sense the sun is like a huge uranium-graphite pile (page 300), since it produces energy by using up matter. But details of the energy-producing reactions are very different: In a pile heavy atoms are split by bombardment with neutrons, while in the sun energy is produced by collisions between nuclei of light elements.

Every second, by the hydrogen-helium reaction, the sun loses more than 4 million tons of matter; every 50 million years it loses an amount equal to the earth's mass. Yet the sun's supply of hydrogen is so enormous that even these appalling losses are trivial by comparison. The amount of matter lost in all geologic history is not enough to have changed the sun's radiation appreciably—which confirms geologic evidence that the earth's surface temperature has remained approximately constant. Nor need we fear that the sun's hydrogen supply will be seriously depleted for at least 10 billion years in the future.

Questions

1. A photograph of a star cluster, like that on page 638, shows a great many more stars than can be seen by direct visual observation of the cluster through the same telescope. Explain.
2. Describe the procedure you would use to determine whether or not gold is present in the sun's atmosphere.
3. Arrange the following types of stars in a sequence of decreasing surface temperatures: yellow stars, blue stars, red stars, white stars.
4. What part of a star produces the continuous background of its spectrum? What part produces the dark absorption lines? For what part of a star can the composition be determined?
5. What evidence suggests that the sun is almost wholly gaseous?
6. Suggest two pieces of evidence that sunspots are cooler than their surroundings.
7. What is the evidence that sunspots have at least an indirect influence on the earth?
8. Suppose that you examine the spectra of two stars, and find that lines in one are displaced slightly toward the red end when compared with those in the other. What conclusion can you draw?
9. If the earth were moving toward a star instead of the star toward the earth, would lines in the star's spectrum be shifted toward either the red or the violet? Explain
10. Theoretically, would the motion of the earth in its orbit produce any appreciable shift in the lines of the sun's spectrum? In the lines of a star's spectrum?
11. What is the evidence that the sun's radiation has not changed appreciably for at least 2 billion years?
12. Which of the various particles in the sun's interior (electrons, protons, alpha particles, stripped atoms) would you expect to be most abundant? Why?
13. On what kind of laboratory experiments are inferences about the sun's interior based?

CHAPTER XLI

The Stars

THE "myriads of stars" visible to the naked eye on a clear night would add up, if you took the trouble to count them, to a couple of thousand at most. A small telescope would increase the number enormously, and a large one often records several thousand stars on a single photograph. Each of these is an incandescent sphere like our sun, separated from its neighbors by immense reaches of empty space.

As we set out to explore this universe of stars, let us keep in mind that no one of them is ever seen as more than a point of light, even in the most powerful telescopes. Probably many stars, like the sun, have spots, prominences, coronas, possibly even planets, but we cannot prove their existence by direct observation. Nearly all our information about the stars is based on studies of their motion and on spectroscopic analysis of the feeble light they send us.

Double Stars and Variable Stars

We shall consider first two odd kinds of stars which attract the attention of every amateur astronomer who owns a small telescope.

Some of the points of light which appear to the eye as single stars are revealed in the telescope as actually two stars close together. Such close pairs, or ***double stars,*** are very numerous, although a small telescope is capable of separating only a few. The closeness of a pair may be illusory, two widely separated stars just happening to lie nearly on a line with the earth. But the components of most double stars are actually close together in space, close enough for a strong gravitational attraction between them. The attraction is often shown by the fact that one star follows an elliptical orbit around the other (Fig. 314). More properly we should say that each star moves in an elliptical orbit around the center of gravity of the pair.

Frequently a star which appears single even in a large telescope turns out to be double when examined spectroscopically. Only one spectrum

can be obtained for the star, of course, but if photographed on successive days the spectrum shows small periodic shifts in some lines, periodic doubling of others. The periodic shifts and the doubling suggest that two separate stars are responsible for the lines. On certain days one of the stars is moving toward the earth, so that its lines are shifted slightly toward the violet end of the spectrum, while the other star is moving away from the earth, so that its lines are shifted toward the red end (page 594). On other days neither star is moving directly toward or away from the earth, so that all lines are single. Motion to produce these

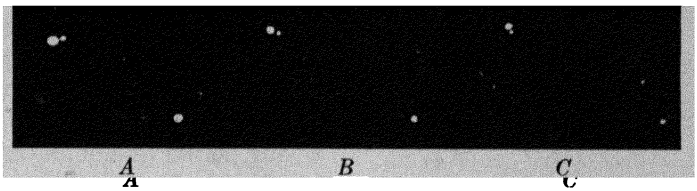


FIG. 314. *Photographs showing the slow orbital motion of a double star. A, 1908; B, 1915; C, 1920. In 12 yr the fainter star has moved about a quarter of the way around its bright companion. (Courtesy of the Yerkes Observatory.)*

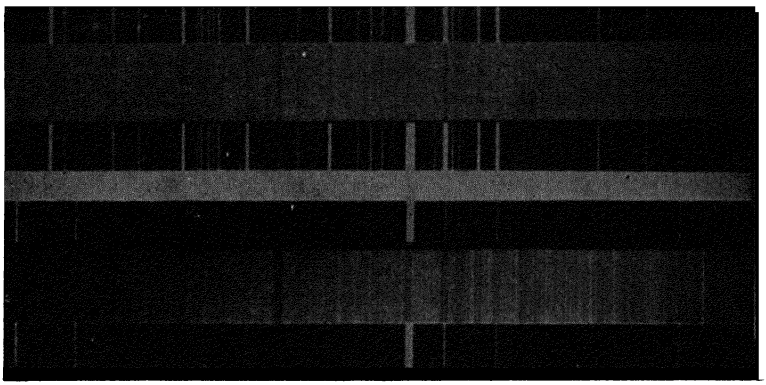


FIG. 315. *Part of the spectrum of one of the stars in the Big Dipper. This is a double star, its components having similar spectra and moving in orbits which we see approximately edgewise. Hence the spectrum sometimes shows single lines (above) and sometimes double lines (below). (Courtesy of the Yerkes Observatory.)*

effects is possible if the stars revolve in orbits which we see nearly edgewise (Figs. 315, 316).

The spectroscope makes possible the detection of double stars whose components are so close together as to be almost touching. These stars move rapidly, completing a revolution in a few days or hours. Pairs which the telescope shows as separate stars are many millions of miles apart and require years to complete each revolution.

A second peculiar type of star is one whose light does not remain constant but varies in brightness. Some of these ***variable stars*** show

wholly irregular fluctuations, but the greater number repeat a more or less definite cycle of changes. A typical variable grows brighter for a time, then fainter, then brighter once more, with irregular minor fluctuations during the cycle. Periods separating times of maximum brightness range all the way from a few hours to several years. Maximum brightness for some variables is only slightly greater than minimum brightness, while for others it is several hundred times as great. Since the sun's radiation changes slightly during the sunspot cycle, we may consider it a variable star with an extremely small range in brightness (4 per cent or less) and a long period (about eleven years).

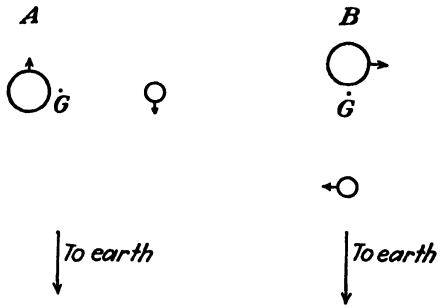


FIG. 316. *How motion of a double star causes periodic shifts in spectral lines. Each star revolves about the center of gravity of the system, G. In the position shown by A, one star is approaching the earth and one is receding, so that lines in their spectra are shifted in opposite directions. In the position shown by B the motion of each is across the line of sight, so that their spectral lines will have normal wave lengths.*

The light changes in a few variable stars are simply explained: the stars are actually double stars whose orbits we see edgewise, so that one component periodically eclipses the other. But the fluctuations in most variables cannot be accounted for so easily. Perhaps the appearance of numerous spots at intervals dim their light; perhaps they are pulsating, expanding and contracting so that their surface areas change periodically. Perhaps the irregular variables are passing through or behind clouds of gas and fine particles. But none of these suggestions is entirely satisfactory.

Stellar Distances

With this brief glimpse at the variety of stars which even a small instrument makes visible, let us see what spectroscopes and larger telescopes can tell us about the characteristics of stars in general. We shall take up these characteristics one by one, and later try to find relationships among them.

Aristotle long ago pointed out that if the earth revolves around the sun, the stars should appear to shift in position, just as trees and buildings shift in position when we ride past them. Since he could detect no shift, Aristotle concluded that the earth must be stationary. Another possible explanation for the lack of apparent movement among the stars, urged by some of the Greeks and later by Copernicus, is that the stars are too far away for the movement to be detected. When the Copernican theory

of the solar system had become well established and astronomical instruments had been improved, many observers tried vainly to find the small shift in position which should result from the earth's motion. An undoubted shift for one star was finally discovered in 1838 by the German astronomer Bessel, and in the following years several others were found. The shifts are so exceedingly small that the long failure to detect them is not strange.

Bessel's discovery made possible the direct measurement of distances to the nearer stars. The method is a simple one, similar to that used by surveyors in finding the distance to an inaccessible object. In Fig. 317 suppose that E_1 and E_2 represent positions of the earth on opposite sides of its orbit and that S is a relatively near star which we see against a background of more distant stars, TT'. If S is observed when the earth is

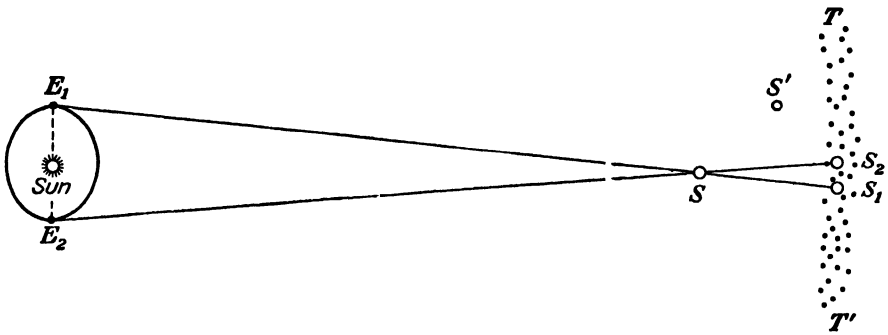


FIG. 317. *Finding the distance of a star by direct measurement of its parallax. Angles greatly exaggerated.*

at E_1, it appears to be in the position S_1 among the background stars. Six months later, when the earth has moved to E_2, the star will appear to be at S_2. The angular distance S_1S_2 is the star's apparent shift in position, or *parallax*,* due to the earth's orbital motion. Accurate sighting on the positions S_1 and S_2 makes possible measurement of the angle E_1SE_2. From this angle and the known distance E_1E_2 (the diameter of the earth's orbit, 186 million miles), the desired distance to the star, E_2S, may be found by simple trigonometry.

The parallax, or apparent shift in position, of more distant stars is smaller, as lines drawn through S' from E_1 and E_2 will demonstrate. It is large enough to be measurable for only a few thousand of the nearer stars. Even for these the measurement is difficult: measuring the parallax of the closest one is equivalent to measuring the diameter of a dime seen from a distance of nearly four miles.

* As used in astronomical calculations, parallax is defined as half the angle E_1SE_2 rather than the shift in position S_1S_2.

The distance to the nearest star (besides the sun, of course) is about 24×10^{12} (24 million million) mi. To give such figures meaning, they are often expressed in terms of *light-years*, a light-year being the distance which light would travel in a year's time. Since light covers 186,000 mi every second, the light-year is an enormous distance, about 6×10^{12} mi. The nearest star's distance is equal to about 4 light-years—which means that we see the star not as it is today, but as it appeared four years ago.

Only about forty stars are within 16 light-years of the solar system. Such distances between stars are typical for much of the visible universe. This means that space is appallingly empty, far more empty even than the solar system with its tiny, isolated planets.

Direct measurements of parallax are possible only for distances up to about 300 light-years. Several indirect methods are available for finding distances to stars farther away than this, the most useful one depending on the measurement of the *intrinsic brightnesses* of stars by means of a spectroscope.

The ***apparent brightness*** of a star is the brightness as we see it from the earth; it expresses simply the amount of light which reaches us from the star. The ***intrinsic brightness*** is the real brightness, a number expressing the total amount of light which the star radiates into space. The apparent brightness of a star depends on two things: its intrinsic brightness, and its distance from us. A star which is actually very bright may appear faint because it is far away, and one which is actually faint may have a high apparent brightness because it is close. Now if the apparent brightness of a star is accurately measured, and if its distance is known, the intrinsic brightness can be calculated; the calculation is simply a matter of figuring out how bright an object would have to be at the known distance to send us the observed amount of light. Since the apparent brightness is easily determined for any star, intrinsic brightnesses are known for all stars whose distances can be measured.

Now suppose we reverse the problem: if both the apparent brightness and the real brightness of a star are known, it should be possible to calculate its distance. The calculation now involves finding out how far away an object of known brightness must be placed in order to send us the amount of light we observe. Since this calculation is easily accomplished, we can find the distance to a star beyond the 300-light-year limit *provided some other method is available for determining the star's intrinsic brightness.*

Such a method for finding intrinsic brightness, depending on the use of the spectroscope, was discovered by the American astronomer W. S. Adams, director of the Mt. Wilson observatory. Studying the spectra of the nearer stars, for which intrinsic brightnesses are known, Adams observed that the relative intensities of certain lines (for stars of any one type) seemed to depend on the intrinsic brightness. That is, the spectrum of a bright

star showed a certain relation among the intensities of its lines, and the spectrum of a faint star showed a different relation. So definite was the connection between relative intensities of the lines and intrinsic brightnesses of the stars that Adams found it possible to predict and check the brightness of a star simply by examination of its spectrum. *Assuming that the relationship would hold for more distant stars,* Adams could then use spectra to find intrinsic brightnesses and therefore distances to these stars.

Note the implications of this spectroscopic method. It depends first of all on accurate direct measurements of distance for a great many nearer stars. Intrinsic brightnesses for these stars are calculated, and a relationship is found between the brightnesses and certain characteristics of their spectra. Finally, the relationship is assumed to hold for stars so distant that a direct check is no longer possible.

The spectroscopic method is applicable to any star bright enough to give a good spectrum (except for comparatively uncommon stars of certain spectral types for which Adams' relation does not hold). By its use stellar distances have been determined up to several thousand light-years.

Stellar Masses

A direct determination of mass is possible only for stars noticeably affected by gravitational forces. Of course every star is attracted to some extent by all other stars in the universe, and its motion is controlled by these combined forces; but most stellar distances are so great that the attraction between any two stars is not measurable. Only in double stars do we find stars sufficiently close together for gravitational effects to be perceptible, and only for these can accurate measurements of mass be made.

Masses of double stars are found by an application of the laws of gravitation and centrifugal force, in a calculation similar to one that we previously used to find the sun's mass (Prob. 8, page 89). Suppose that the shape and size of the orbit followed by each component of a double star has been determined, as well as the time required by each component to complete a revolution. Then the attractive force between them can be expressed in terms of Newton's constant (known), the two masses (unknown), and their distance apart (known). For each component, this force must be balanced by the centrifugal force of its orbital motion; the centrifugal force is expressed in terms of its mass (unknown), the average radius of its orbit (known), and the time it requires to complete a revolution (known). By setting the attractive force equal to the centrifugal force for each star, we obtain two equations in which the masses are the only unknown quantities and from which they may be readily computed.

Many double stars are so distant or so close together that complete information about their orbits is not obtainable. For some of these, however, sufficient data can be secured to determine the combined masses of the two components, and the mass of each can then be at least roughly estimated from its brightness.

The oddest fact which emerges from measurements of star masses is that all stars seem to contain roughly the same amount of matter. No star is known with a mass smaller than one-tenth that of the sun, and only a few have masses greater than ten times that of the sun. In view of the enormous variations among the stars in brightness, in diameter, and in density, it seems extraordinary that the range of masses should be so small.

Temperatures, Diameters, Densities

The temperature of a star is determined from its spectrum, by finding in what part of the spectrum the star's radiation is most intense (page

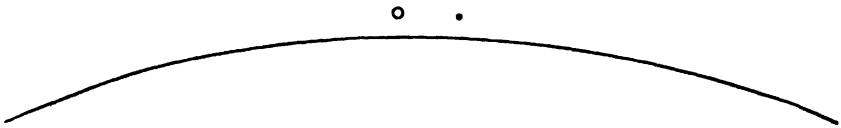


FIG. 318. *The range of stellar sizes. The broad curve represents a portion of the circumference of Antares; the small circle is the sun; the dot is a large white dwarf. On this scale the smaller white dwarfs would be microscopic dots.*

608). This measurement gives the temperature of the star's "surface"—the part from which radiation is emitted. Measured surface temperatures of a few very hot stars range up to 30,000°C, but the great majority have temperatures between 3000 and 12,000°. Probably many stars have temperatures below 3000°, but unless they are relatively close their radiation is too feeble for us to detect. Like the sun, stars must have enormously high internal temperatures to maintain their surface radiation.

The surface temperature of a star is intimately related to its color. In general, the hottest stars are blue-white, those of intermediate temperatures white or yellow, and the coolest ones red.

Measurements of a star's temperature and its intrinsic brightness make possible an estimate of its diameter. Since the temperature determines the intensity of radiation from the star's surface, a measurement of temperature (plus a few reasonable assumptions about the nature of the surface) gives a value for the amount of radiation emitted from every square centimeter or square inch of the star's area. The intrinsic brightness is a measure of the total radiation from the star's entire surface. We need only divide the total radiation by the radiation per square centimeter to find the number of square centimeters in the star's surface area, and from the area the diameter and volume are easily computed.

A more direct method of measuring stellar diameters (which would take us too far afield to explain) is available for the very largest stars. Results obtained by this means check well with estimates from temperatures and intrinsic brightnesses.

The diameters of stars, in contrast to their masses, have an enormous range (Fig. 318). The smallest ones, like the faint companion of the bright star Sirius, are little bigger than the earth. The largest ones, like the giant red star Antares in the constellation Scorpio, have diameters of more than 300 million miles. Antares is so huge that if the sun were placed at its center, the four inner planets could pursue their normal orbits inside the star with plenty of room to spare.

If the mass and volume of a star are known, calculation of the average density means simply dividing one by the other. Like the volumes, densities vary greatly from star to star. Giant stars like Antares have densities less than one-thousandth that of ordinary air—densities corresponding to a fairly good vacuum here on the earth. At the other extreme are the incredible densities of small stars like the companion of Sirius, densities so great that a cubic inch of their substance would weigh more than a ton.

Stellar Motions

Rates of motion of stars moving toward or away from the earth are found by measuring slight shifts in the wave lengths of lines in their spectra (page 610).

Motion across the line of sight is followed by direct observation. The great distances of the stars make their apparent movements exceedingly slow, so slow that we commonly refer to the stars as "fixed stars." Yet the motion is sufficient to have caused perceptible changes in the shapes of some constellations during the few thousand years since accurate observations began (Fig. 319).

Most stars, of course, are not moving either directly along or directly across our line of sight, but at an oblique angle to it. For such a star the spectrum reveals a certain speed of approach or recession, and direct observation shows motion in a certain direction across the line of sight. If the distance of the star is known, the actual speed of this latter motion may be calculated. Then the star's real motion may be found by adding vectors (page 65) corresponding to the two observed velocities. Such calculations show that most stars are moving with speeds of several miles per second.

Since other stars are moving, we might suspect that the sun is traveling also. Such a motion should reveal itself in apparent movements of the stars: if the sun is moving toward a certain part of the sky, stars in that direction, on an average, should appear to be approaching us and to

be radiating out from a point, much as trees in a forest seem to approach and spread out when we drive toward them. Average stellar motions of this sort are observed in the neighborhood of the constellation Hercules, while in the opposite part of the sky stars are apparently receding and coming closer together. Careful study of these motions indicates that the sun and its family of planets are moving toward Hercules at a speed of about 12 mi/sec.

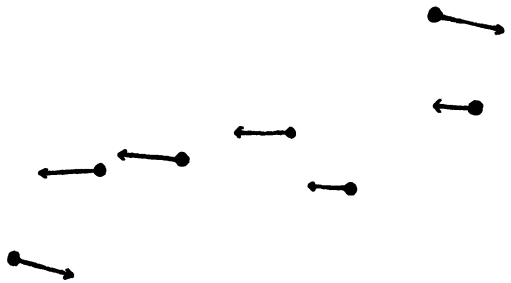


FIG. 319. *Motions across the line of sight of stars in the Big Dipper. At the end of 100,000 years the stars will have moved to the points of the arrows, so that the dipper shape will be completely lost. (From Elements of Astronomy by Fath.)*

Stellar Spectra

We have found spectra useful in obtaining many different kinds of information about the stars—their distances, their temperatures, their sizes, their motions. Let us now turn our attention briefly to the spectra themselves.

Superficial examination of stellar spectra reveals at once two important facts: (1) nearly all stars have absorption spectra like the sun's, which imply relatively cool atmospheres around hot interiors; and (2) practically all lines in the spectra can be identified with lines of elements known on the earth. Thus we know that matter throughout the visible universe is made up of the same kinds of particles as matter here on the earth, and furthermore that most of the matter in the universe occurs in large aggregates built on a single general pattern. This uniformity of material and structure among the stars is perhaps the most amazing single discovery of stellar astronomy.

Examined more carefully, stellar spectra show considerable variety. Some have relatively few lines, others have many; some have only sharp lines, others have diffuse bands; some have prominent lines of hydrogen, others prominent lines of certain metals. Comparison of large numbers of spectra shows that nearly all can be arranged in a single sequence, depending on the intensities of different lines (Fig. 320). Lines which are prominent in spectra at one end of the sequence grow less intense in successive spectra, and other lines become prominent; then these become faint and

still others grow conspicuous. Thus between two distinct types of spectra there are gradual changes, and the changes follow one another in a single regular order.

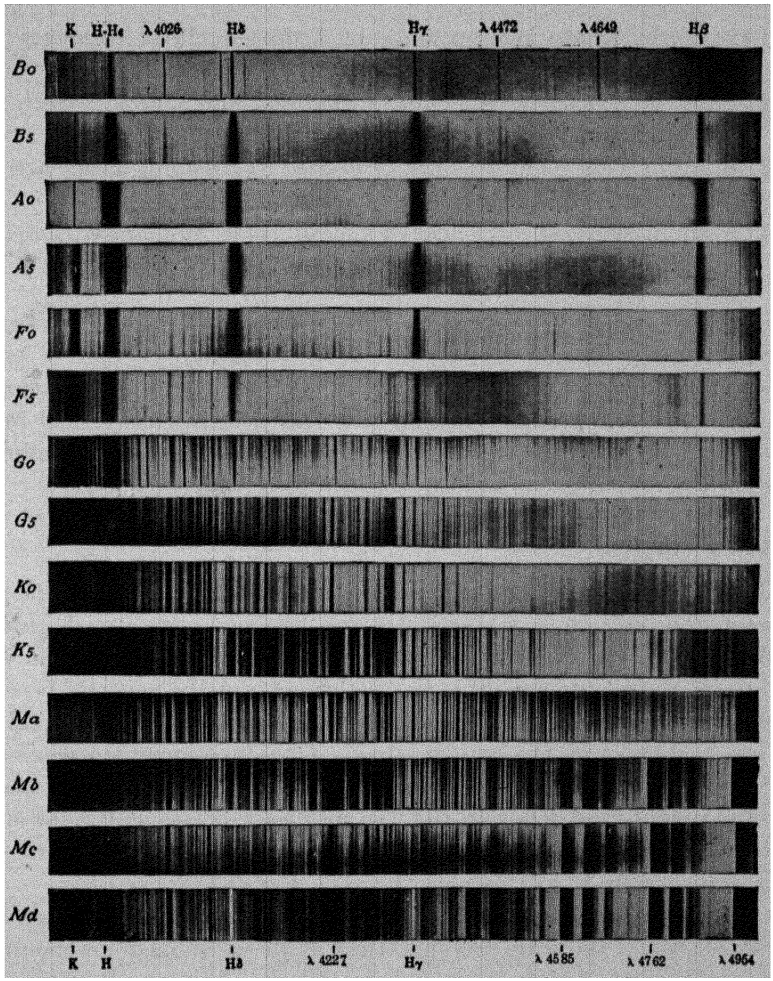


FIG. 320. *Types of stellar spectra. The most prominent lines in the first two spectra belong to hydrogen and helium; in the next four, to hydrogen and ionized calcium; in the next four, to ionized and neutral metals; in the last four, to neutral metals and compounds. (Photographed by Rufus at the Observatory of the University of Michigan.)*

Each spectrum shows the composition and the physical state of a certain layer in a star's atmosphere. Differences in spectra therefore represent differences in the make-up of this layer. These differences may be due to slight variations in temperature and pressure in various stars rather than to any fundamental difference in composition, but they

nevertheless make possible a convenient classification of the stars. We need not go into technical details of the classification, but roughly the spectral sequence distinguishes the following groups:

1. Stars in whose spectra lines of *hydrogen* and *helium* are most prominent.
2. Stars in whose spectra lines of *hydrogen* and *ionized metals*, especially calcium, are most prominent.
3. Stars in whose spectra lines of *ionized and neutral metals* are most prominent. The sun is a star of this type.
4. Stars in whose spectra lines of *neutral metals* and *simple compounds* are most prominent.

There are all gradations from stars of one group to stars of the next group.

Study of the continuous backgrounds of the different types of spectra shows that this grouping of stars is closely related to temperature. In general, the hydrogen and helium stars are the intensely hot, bluish-white stars; those in the second and third groups are white or yellowish; those in the fourth group are the comparatively cool red stars.

Russell's Diagram

The American astronomer Russell discovered a peculiar relationship between the position of a star in the spectral sequence and its intrinsic brightness. This relationship is shown by the graph in Fig. 321, on which each point represents the intrinsic brightness and the spectral type of a single star. Intrinsic brightnesses are plotted along the vertical axis and spectral types on the horizontal axis. Obviously most stars belong to the "main sequence," a considerable number to the "red giant" class at the upper right, and a few to the "white dwarf" class at the lower left. (The names "giant" and "dwarf" refer to very large and very small stars, respectively.)

On this same diagram we may represent at least roughly certain other characteristics of the stars. Temperature, as we mentioned a moment ago, varies with spectral type, so that hot stars are shown at the left of the diagram, cool stars at the right. In general, the brighter stars are those with both greater mass and greater volume, so that large, heavy stars are found at the top of the diagram and smaller, lighter stars at the bottom. Densities are commonly greater for the small stars, so that densities at the bottom of the diagram are greater than those at the top. (These correlations are very general; they do not mean, for example, that all stars with absolute magnitude +5 would have the same mass and density, or that all stars of spectral type A_0 would have precisely the same temperature.)

Stars at the upper end of the main sequence are large, hot, massive bodies, with prominent lines of hydrogen and helium in their spectra. Stars at the lower end are small, dense, and reddish, with low enough temperatures so that compounds form a considerable part of their

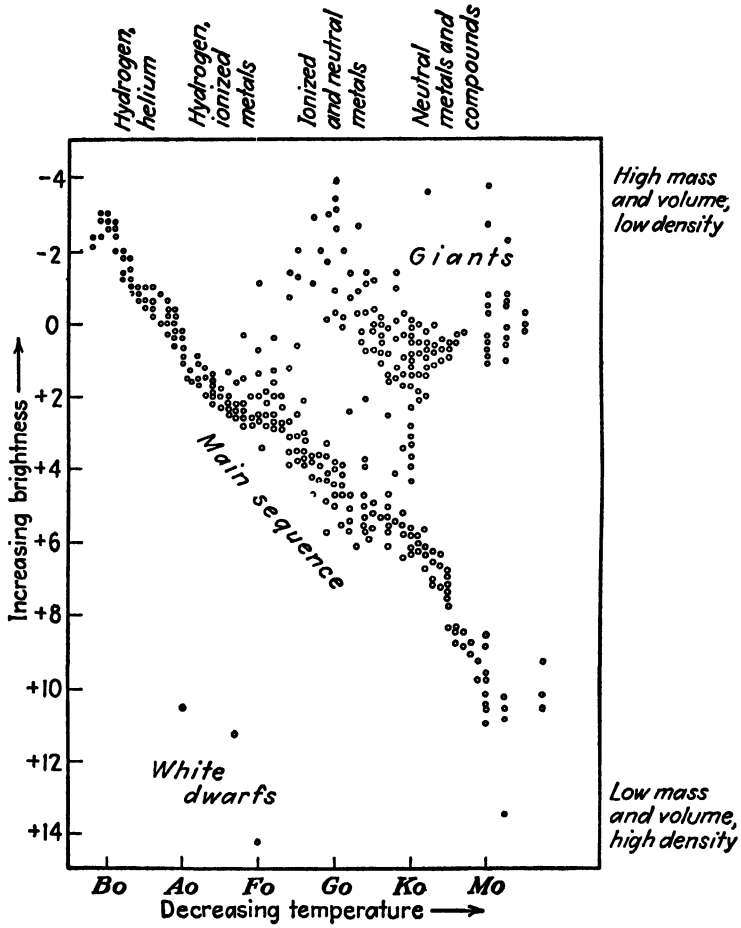


FIG. 321. *Russell's diagram, showing the distribution of stars according to spectral type and intrinsic brightness. (Numbers along the vertical axis refer to absolute magnitude, the astronomer's measure of intrinsic brightness; it is a backward sort of measure, low numbers indicating bright stars and high numbers faint stars. The letters at the bottom of the diagram are technical designations of spectral types.)*

atmospheres. In the middle part of the main sequence are average stars like our sun, with moderate temperatures, densities, and masses, rather small diameters, and spectra in which lines of metallic elements are prominent. The majority of stars show this definite relationship among their principal characteristics.

To the *red giant* class belong the huge, diffuse stars like Antares, with very low densities and diameters up to several hundred million miles. These stars have low surface temperatures, as their reddish color indicates, but their enormous surfaces make them nevertheless very bright.

The position of the *white dwarfs* in Russell's diagram suggests their peculiar combination of properties: intensely hot surfaces, but small total radiation. These properties imply that their surface areas, and hence their volumes, must be exceedingly small (page 624). Most white dwarfs are about the size of a large planet, and a few are no bigger than the earth. But their masses are not correspondingly small: for those which are components of double stars, calculation shows masses roughly the same as that of the sun. Now to get the sun's mass in the volume of a planet would require squeezing, squeezing so drastic that the density of the material must increase to a fantastic figure in the neighborhood of 100,000 g./cc. A pinhead of such matter would weigh nearly a pound here on earth; a cupful would weigh several tons.

Densities like this seem incredible, but they have been checked by enough methods to leave little doubt of their correctness. The only possible explanation is that atoms in these stars have partially collapsed: instead of ordinary atoms with electrons following wide orbits around their nuclei, white dwarfs must have electrons and nuclei packed closely together. The properties of matter in such a state we can only surmise, for nothing like it has ever been found or prepared on earth.

Only a small number of white dwarfs have been discovered, but their scarcity may be more apparent than real; they are so faint that none but the nearer ones can be seen even in large telescopes. Enough have been found in recent years to suggest that the universe probably contains great numbers of these peculiar stars.

Stellar Evolution

So striking a series of relationships as that shown in Russell's diagram demands some attempt at explanation. Are the stars in different parts of the diagram perhaps in various stages of development? Does the mass or the density of a star perhaps exert some controlling influence over its temperature and the composition of its atmosphere? Astronomers have not reached full agreement about answers to such questions. But in general terms the diagram finds a satisfactory explanation in recent speculations about the life history of a star.

A star shines for two reasons: (1) it is a large aggregate of matter in a fairly small region of space, and (2) it contains abundant hydrogen. A body of this sort *cannot help* shining, by the conversion of its hydrogen into helium. If it were not shining, the mutual gravitation of its particles would cause it to contract; energy liberated by the contraction would

presently heat the interior sufficiently to start the hydrogen-helium transformation; thereafter the radiation would maintain itself as long as any hydrogen remained. A star does not shine because some occult force has started it shining; it shines because it has a certain mass and a certain composition. If we could somehow build a star by heaping together sufficient matter of the right composition, it would start to shine of its own accord.

The only condition under which the material of a star would not shine would be if it were scattered through so large a region of space that mutual gravitation among its particles was very feeble. So we may imagine as the most reasonable initial stage in a star's history a time when its matter was an irregular mass of cool, exceedingly diffuse gas.

Gravitation in such a mass would ultimately concentrate it into a smaller space. The gradual contraction would heat the gas, much as the gas in a bicycle pump is heated by compression. At length the temperature would grow high enough for hydrogen to be converted into helium and the mass would begin to glow brightly. During these early stages we imagine the star to be still highly diffuse, very large, and relatively cool. Perhaps the red giants in the upper right-hand corner of Russell's diagram are stars in this early part of their careers.

Once the conversion of hydrogen to helium (by collisions of protons with carbon nuclei, page 616) is well started, the star settles down to a comparatively stable middle age lasting several billion years. Its temperature very gradually rises, and its mass grows somewhat smaller as a part of it is continually converted into radiation. How hot the star becomes depends on its original mass, stars with large masses reaching very much higher temperatures than small ones.

Ultimately the star's hydrogen supply must become so small that rapid production of helium can no longer be maintained. At this point in its career the star again begins to contract under the influence of gravity, its energy now coming chiefly from the contraction rather than from nuclear reactions. For stars with comparatively small mass the shrinking leads presently to the white dwarf stage, in which the atoms themselves have partly collapsed. As a slowly contracting dwarf the star may remain luminous for many billions of years, but it must at length exhaust all available energy and grow dark. The old age of stars with large mass cannot be so clearly predicted: there seems to be nothing to prevent their contracting indefinitely, until perhaps they split apart into two or more white dwarfs of normal mass.

Since a star covers the early part of its evolution fairly rapidly, and since the contraction to the white dwarf stage is also fairly rapid, we might expect to find most visible stars in the long middle parts of their careers. These stable, middle-aged stars are those which lie along the main se-

quence of Russell's diagram. Stars at the lower end of the sequence are those with small mass, whose hydrogen supply is insufficient to give them a high temperature; those at the upper end are very heavy stars, whose enormously rapid production of helium accounts for their high surface temperatures and great intrinsic brightness. Stars are scarce between the main sequence and the red giants, and between the main sequence and the white dwarfs, since these parts of the diagram represent stages in which a star's development is rapid.

Our own sun, lying about two-thirds of the way down the main sequence and abundantly supplied with hydrogen, is still a comparatively "young" star. It should remain at about its present size for many billions of years, its temperature gradually increasing. Life on earth will ultimately become impossible—not, as was once thought, because the sun will cool off, but because the sun is growing hotter. After a very long time the sun will exhaust its hydrogen, and will shrink into the old-age condition of a tiny white dwarf. Only after billions of years as a dwarf will it finally grow cold.

This reconstruction of stellar evolution is, of course, highly speculative. It cannot be checked by direct observation, simply because human life is so short in comparison with a star's life. We can justify this picture of a star's history only by its agreement with modern ideas about energy production in stars and by its ability to explain the concentration of stars in the main sequence of Russell's diagram.

Questions

1. Give all the evidence you can in support of each of the following statements:
   a. The sun is a star.
   b. Most stars have relatively cool atmospheres surrounding hot interiors.
   c. Distances to the stars are very great.
   d. Some stars which appear to be single even in a telescope are actually double stars.
   e. The sun is moving with reference to the other stars.
2. What methods may be used for determining the intrinsic brightness of a star? What assumption is involved in the spectroscopic method?
3. How is a star's diameter estimated from measurements of temperature and intrinsic brightness?
4. For what stars is a direct determination of mass possible? On what characteristic of these stars does the determination depend?
5. Would it be possible for a star to have any other shape than that of a sphere or spheroid? Could a star, for instance, be shaped like a cube or like a corkscrew? Why or why not?
6. What data are needed for the determination of a star's average density? How would you expect the density to change from the surface layers to the interior of a star?
7. What lines would you expect to find most prominent in the spectrum of (*a*) a blue-white star (temperature 10,000 to 12,000°C)? (*b*) A star whose surface temperature is around 3000°? (*c*) A star whose surface temperature is about 6000°?

8. What are the chief characteristics of an average star (*a*) in the upper left-hand corner of Russell's diagram? (*b*) In the lower left-hand corner? (*c*) In the upper right-hand corner? (*d*) In the middle of the main sequence?
9. Calculations indicate that at present the sun's temperature is slowly increasing, that is, that its hydrogen is being transformed into helium at a faster and faster rate. Describe the probable stages in the sun's history up to the present.
10. What is the evidence for the enormous average densities ascribed to white dwarfs?
11. Sirius, the "Dog Star," is a bluish-white star with a very great intrinsic brightness. What, if anything, can you conclude from these facts about each of the following?
    a. Its temperature.
    b. Its average density.
    c. The principal lines in its spectrum.
    d. Its position in Russell's diagram.

in number should follow a regular law: when the distance at which we can barely see stars of a certain brightness is increased 10 times, the total number of stars visible should increase 1,000 times; when the distance is increased twentyfold, the number of stars should be 8,000 times as great, and so on.

Counting stars in this manner shows the expected rate of increase for relatively small distances, but a rapid falling off in the rate of increase

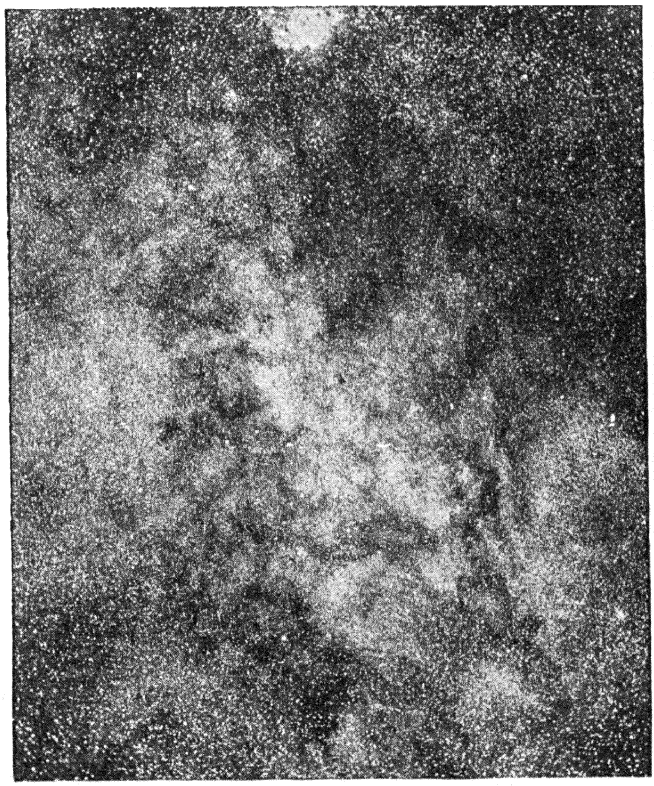


FIG. 322. *Part of the Milky Way in the constellation Sagittarius. (Photograph by Barnard at the Mt. Wilson Observatory.)*

for large distances. This means that a rocket traveler moving away from the sun would presently find the stars thinning out, and at length would come to a region where they are absent or exceedingly scarce. The stars do not continue indefinitely into space; rather, our sun is part of a huge aggregate of stars with more or less definite limits. We shall find later that there are other similar aggregates far beyond the limits of this one, each separated from the others by vast reaches of empty space. Such aggregates are called ***galaxies,*** and the particular one to which our sun belongs we possessively term "our" galaxy.

CHAPTER XLII

The Nebulae

THE band of misty light we call the Milky Way forms a continuous ring around the heavens. As the earth swings in its orbit, we see different parts of the ring, but a section near the south pole remains hidden as long as we stay in the northern hemisphere.

Examined with a telescope, the Milky Way is an unforgettable sight. Instead of a dim glow we see countless myriads of stars (Fig. 322), stars as thick as the sand grains on a beach, but so faint and far away that the naked eye cannot distinguish them. In other parts of the sky the telescope reveals many stars not visible to the eye, but nowhere else in such unbelievable numbers.

This observation tells us at once that the stars are not uniformly distributed in space, but concentrated more thickly in some regions than in others. Let us see what light the distribution of the stars can shed on the great problem of the structure of the universe.

The Shape of Our Galaxy

Suppose that a swift rocket could carry us out through space in any direction and to any distance we wished. Would we find ourselves surrounded by stars no matter how far we traveled, or would we at last come to a part of space in which stars are absent?

A telescope is a good substitute for a rocket in answering this question. Suppose we examine a small region of the sky with telescopes of greater and greater light-gathering power; or better, suppose we use a single powerful telescope and leave photographic plates exposed to the same part of the sky for longer and longer times. An increase in either the light-gathering power of the telescope or in the exposure time enables us to see objects farther out in space. Now if the stars continue indefinitely into space in about the same numbers as we find them near the sun, then each increase in the distance our instruments can penetrate should lead to an increase in the number of stars visible. Furthermore, this increase

The falling off in the number of stars at increasing distances is least rapid for sections of the sky near the Milky Way, most rapid for sections at right angles to it. In the direction of the Milky Way we seem to be looking out through a far greater thickness of stars than we find in other parts of the sky. These facts tell us something about the shape of our galaxy: it must be a relatively thin, flat structure, shaped like a thin watch or discus (Fig. 323).. From our vantage point deep within the galaxy we look along the plane of the watch toward its edge and see the thick mass of stars in the Milky Way; when we look out through the

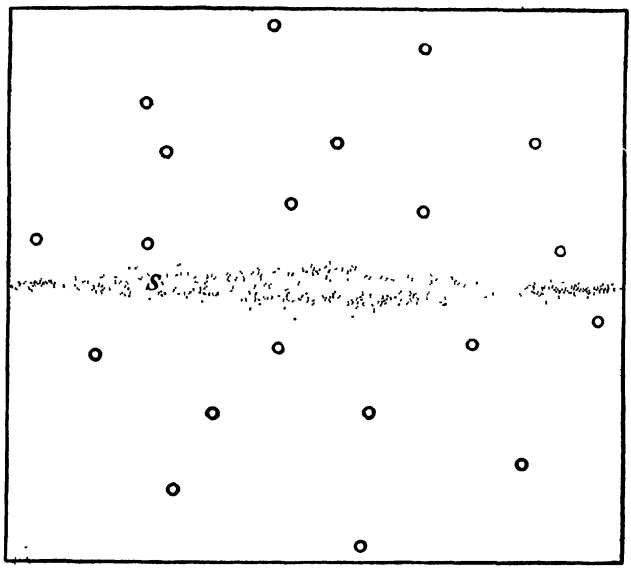


FIG. 323. *Diagrammatic edgewise view of our galaxy. The dots represent stars and the circles represent globular clusters. The sun's position is indicated by the letter S. For an approximate scale, remember that the greatest length of the galaxy is about 100,000 light-years.*

face or the back of the watch, we see the relatively small number of stars at right angles to the Milky Way.

Since the earth appears to be nearly in the plane of the Milky Way, the sun must be nearly in the central plane of the galaxy. This conclusion is borne out by accurate counts of stars in opposite directions at right angles to the Milky Way: in these directions the increase in the number of stars begins to fall off at about the same distance from us, indicating that we should have to travel about as far to reach the front of the watch as to reach its back. Our position relative to the edges of the watch cannot be so readily determined from star counts because of the great distances involved and because parts of the Milky Way are hidden by dark obscuring masses. Studies of the motion of the sun and nearer stars indicate,

however, that we are probably about two-thirds of the distance out from the center of the galaxy toward one edge. The center is believed to lie in the direction of the densest part of the Milky Way in the constellation Sagittarius.

Stars in the galaxy show a general slow motion of rotation about the center of the system—as they must, to prevent the galaxy from gradually collapsing under the gravitational attraction of its parts. Superposed on this rotation is a random motion, stars moving in all directions with a variety of speeds, much like the molecules of a gas. The comparison with motion in a gas is an accurate one, except that the gas must be exceedingly tenuous to resemble the emptiness of the galaxy: we must imagine a gas more rarefied than the best vacuum we can produce on earth, so rarefied that collisions between its molecules are extremely uncommon.

The size of the galaxy is difficult to determine, for stars near its borders are much too faint for distances to be determined even with a spectroscope. From star counts, from the distances of Cepheid variables (which we shall discuss in a moment), and from other lines of evidence, astronomers have estimated that the galaxy is roughly 100,000 light-years across and between 10,000 and 15,000 light-years thick. Included in this enormous system is a total of somewhere near 100 billion stars.

Among these 100 billion stars the one to which our destiny is bound has a very humble place. Surrounded by countless similar bodies, the sun is distinguished neither by size, by temperature, nor by motion. Not even its position is significant, for it is far from the center of the system. Perhaps it is unusual in possessing a family of planets, but of this we cannot be sure. Our exaggerated opinion of the sun's importance comes simply from our accidental nearness to it: if we could choose the star which we wished to live near, we could find millions of others equally or better able to supply us with light and heat.

Star Clusters

Within the galaxy the stars are not uniformly distributed, but roughly segregated into groups or swarms, the stars in each swarm staying relatively close together and moving as a unit. To one such local swarm belong the sun and most of the brighter visible stars. More interesting than these loose star groups are the ***globular clusters,*** objects associated with our galaxy but lying beyond the limits of the main watch-shaped aggregate.

To the naked eye the largest globular clusters are barely visible on clear evenings as faint patches of light. Through a telescope they are spectacular aggregates of stars, roughly spherical in form, bright and dense near the center and thinning out toward the edges. About 100 of these objects have been discovered. In photographs of one of the largest,

the great cluster in Hercules (Fig. 324), more than 50,000 stars have been counted. These are only the very brightest stars, since the cluster is so far away that faint ones cannot be seen; estimates place the probable total number of stars at close to a million.

Distances to these clusters and to other remote parts of the galaxy cannot be measured by spectroscopic determinations of intrinsic brightness, because good spectra for such faint stars are impossible to obtain. The best method of estimating these enormous distances is based on observations of the fluctuations in brightness of certain variable stars

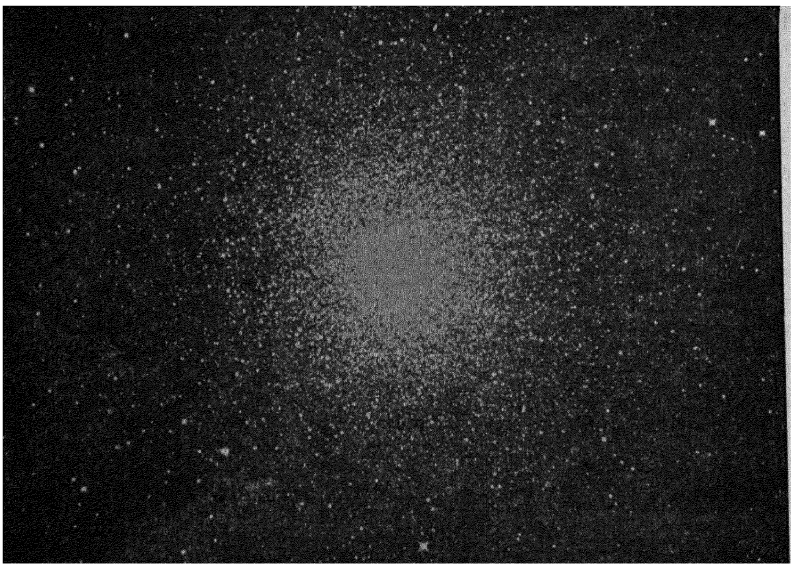


FIG. 324. *The great globular star cluster in Hercules.* (*Photographed by Ritchey at the Mt. Wilson Observatory.*)

called *Cepheid variables.* These are variables of short period which take their name from a typical example in the constellation Cepheus.

The usefulness of Cepheid variables in measuring distances was discovered by a study of the Cepheids in a single large cluster (not a globular cluster, but an irregular cluster called the Small Magellanic Cloud, visible in the southern hemisphere). The Cepheids of this cluster showed a curious relationship between their apparent brightnesses and the periods during which their light fluctuated: thus a relatively bright Cepheid required a long time (two to three months) to complete its cycle of changes, while a faint Cepheid required only a few days. Now the cluster is so very far away that all its stars may be considered as approximately at the same distance from the earth; hence differences in apparent brightness must be due solely to differences in intrinsic brightness.

Therefore the observed relationship among the Cepheid variables must be due to some connection between their periods of variation and their actual brightnesses.

Values for the intrinsic brightnesses of stars in the cluster could not be determined directly, because of their immense distance. But for other Cepheids closer to the earth an estimate of intrinsic brightness was possible; by combining figures for these closer Cepheids with data from the cluster, it was possible to establish a definite numerical relationship between brightness and period of variation. If we assume that this relationship holds for Cepheids elsewhere in the universe, then the intrinsic brightness of any Cepheid may be found simply by measuring its period of variation. Comparison of intrinsic brightness with apparent brightness then gives the approximate distance. This method can be extended to much greater distances than determinations with the spectroscope, since the period of a Cepheid can be measured regardless of its faintness, but the method is handicapped by its limitation to a single peculiar type of star.

Many Cepheid variables have been found in globular clusters, and measurements of their periods give approximate distances to the clusters. Strictly, of course, distances are measured only to particular stars, but the clusters are so far away that the distance to any one star gives a sufficiently accurate value for the distance to an entire cluster. The nearest clusters appear to be about 20,000 light-years from us, while the farthest ones are more than 100,000 light-years away. Light from the great Hercules cluster (Fig. 324) travels for 33,000 years before reaching our eyes; we see the cluster not as it would look today, but as it appeared toward the end of the Ice Age. If observers on a planet somewhere within the cluster are watching the earth with a strangely powerful telescope, they see nothing of our present civilization, but only hairy savages migrating northward as the icecaps recede.

Knowing how far away a cluster is, we can estimate the average distance which separates its stars. This distance is about 1 light-year, so that stars in a cluster, particularly toward its center, are considerably more closely packed than those near the sun. Even so, stars in a cluster have sufficient room to move about so that collisions between them are infrequent.

In comparison with the great watch-shaped aggregate of stars to which the sun belongs, the globular clusters are relatively small objects. They seem to be definitely a part of the galaxy, for they are grouped close to the main aggregate on either side (Fig. 323). They form, so to speak, the outposts of the galaxy: between them are regions of space with few or no stars, and beyond them on all sides are still vaster regions of emptiness.

Galactic and Spiral Nebulae

The telescope reveals two kinds of objects in the sky besides stars and members of the solar system. Both kinds appear as faint patches of

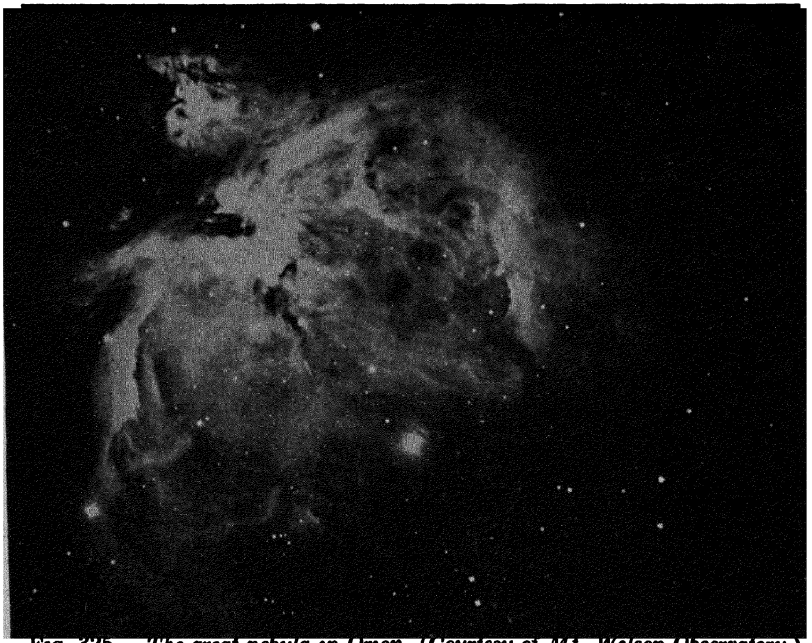


FIG. 325. *The great nebula in Orion.* (*Courtesy of Mt. Wilson Observatory.*)

FIG. 326. *A dark nebula.* (*Photographed by Duncan at the Mt. Wilson Observatory.*)

diffuse light, this appearance giving them their early name, *nebula* (the Latin term for cloud). It is unfortunate that the same name is still applied to both, for they are altogether different in nature.

Galactic nebulae are irregular masses of diffuse material within our galaxy. Some appear as small luminous rings or disks surrounding stars; some take the form of intertwining lacy filaments; many are wholly irregular in outline. The brightest one, the great nebula in Orion (Fig. 325), is barely visible with the naked eye, but most of them are so faint that long exposure of a photographic plate is necessary to bring out details of structure. Probably these nebulae are great masses of rarefied gas and tiny solid particles, shining only because they reflect light from near-by stars or because electrically charged particles and high-frequency quanta from the stars excite their matter to luminescence.

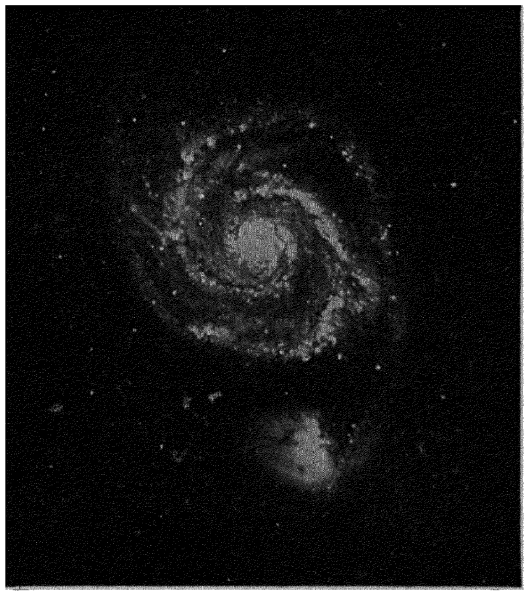


FIG. 327. *Spiral nebula in Canes Venatici.* (*Courtesy of Mt. Wilson Observatory.*)

Similar masses, apparently at a distance from any stars and therefore nonluminous, sometimes reveal their presence as dark patches obscuring the light of stars beyond (Fig. 326).

Spiral nebulae, much the more interesting of the two kinds, are not a part of our galaxy but lie far beyond its limits. Only one of these nebulae, the spiral in Andromeda, is clearly visible without a telescope, as a small, hazy patch of light. Photographs taken with large telescopes and long exposures are needed to show details. Such photographs show many varieties of spirals, from fuzzy, spherical objects practically without structure to distinct, flat spirals like Fourth of July pinwheels. Definite spiral structures with two curving arms radiating from a brighter nucleus are the most common type. Of these we see different ones from all angles: some show us the flat face of the spiral (Fig. 327), some an oblique view

(Fig. 328), some the thin edge (Fig. 329). The general shape of the average nebula appears to be a thin, circular disk with some thickening at the center.

Good spectra of the spiral nebulae are difficult to obtain, because of their extreme faintness. The brighter ones give spectra closely resembling ordinary star spectra—dark absorption lines on a continuous background, most of the lines identifiable with those of familiar elements. This strongly

FIG. 328. *Spiral nebula in Andromeda. (Photographed by Wilson at the [illegible] Observatory.)*

suggests that the nebulae are aggregates of stars, a suggestion borne out by photographs taken with the 100-in. telescope on Mt. Wilson which show the outer portions of the two largest spirals resolved into separate stars. Quite possibly the more diffuse spirals and the central parts of others may be masses of gas and small particles, but the chief components of the principal nebulae are stars.

The exceeding faintness of the few stars which can be made out in the brightest spirals implies that these objects are very remote. Accurate

estimates of the distances to a few of the nearer ones were made possible when Hubble at Mt. Wilson discovered that some of their stars are Cepheid variables. By measuring the apparent brightnesses and periods of fluctuation of these variables, Hubble obtained distances of about 900,000 light-years for the two nearest nebulae. Estimates of distances to other nebulae cannot be based on so sure a foundation as the periods of Cepheid variables, but the faintest ones which have been photographed cannot be much nearer the earth than 500 million light-years. This means that the light from these objects which darkens our photographic plates has been moving through space ever since the beginning of the Paleozoic era.

FIG. 329. *Spiral nebula seen edgewise. (Courtesy of Mt. Wilson Observatory.)*

Spiral nebulae are exceedingly numerous. Hubble has estimated that within a distance of 500 million light-years there are at least 100 million of these objects large enough to be recorded by our telescopes.

"Island Universes"

Herschel, the first to study the spiral nebulae intensively, suggested that they might be other galaxies of stars beyond our own, "island universes" in the empty sea of space. In his day this suggestion was little more than an imaginative hypothesis, but it has been abundantly justified by later work. Let us examine the many points of resemblance between spiral nebulae and our galaxy, to see why astronomers today speak of "island universes" so confidently.

In *shape* the resemblance is evident, for star counts show that our galaxy has the same flat, watchlike form so characteristic of the spirals. There is even some evidence, from the structure of the Milky Way, that our galaxy may have the usual two curved arms. The arms are difficult to detect with certainty from an observation point so nearly in the galaxy's central plane.

The discovery of *globular clusters* near some of the brighter nebulae is further evidence for the similarity of their structure with that of the galaxy. Star clusters would be too small and faint for detection except around the closest nebulae.

In *size*, the spiral nebulae seem to be uniformly smaller than our galaxy. The largest one which can be accurately measured, the Andromeda nebula, has a diameter of about 65,000 light-years, compared with an estimated 100,000 light-years for our galaxy. Yet the sizes are of the same order of magnitude, and it does not seem unreasonable that one galaxy might be considerably larger than others.

Rotation of some of the spirals about their centers, similar to the rotation of the galaxy, has been detected both by direct observation and spectroscopically. In general the rotation is in the direction toward which the spiral arms point.

That the *composition* of at least a great many spirals is similar to that of our galaxy is suggested by the resemblance between nebular spectra and stellar spectra and by the presence of stars in high-power photographs of the brightest spirals. If some of the spirals consist largely of gas and tiny particles rather than stars, it may be that they are simply in an earlier stage of development than our galaxy.

These many points of similarity build up a strong argument for the idea that spiral nebulae are other galaxies, or "island universes." (The term "galaxy" is preferable to "island universe," since by "universe" we generally refer to the sum total of matter which we can perceive.)

Thus we picture the universe as made up of galaxies of stars, each one isolated in space, separated from its nearest neighbors by distances clos to a million light-years. In all directions, in unbroken succession, thes galaxies extend to the farthest parts of the universe which our instruments can penetrate. Not only is the earth an undistinguished planet circling an undistinguished star; even the great galaxy which includes the sun is distinguished from millions of others only by its slightly larger size.

Through all this vast array of uncounted suns and unimaginable distances runs a strange uniformity of material and structural pattern. The elements of our laboratories are the elements of the spiral nebulae, the sun generates energy by a process repeated in billions of other stars, the form of our galaxy recurs again and again in the nebulae. Everywhere

we find the same ultimate particles of matter, the same kinds of energy, the same patterns of structure. We can study at firsthand but a tiny fragment of the universe, yet so ordered and uniform is the whole that from this fragment we can extend our knowledge wherever our instruments enable us to see.

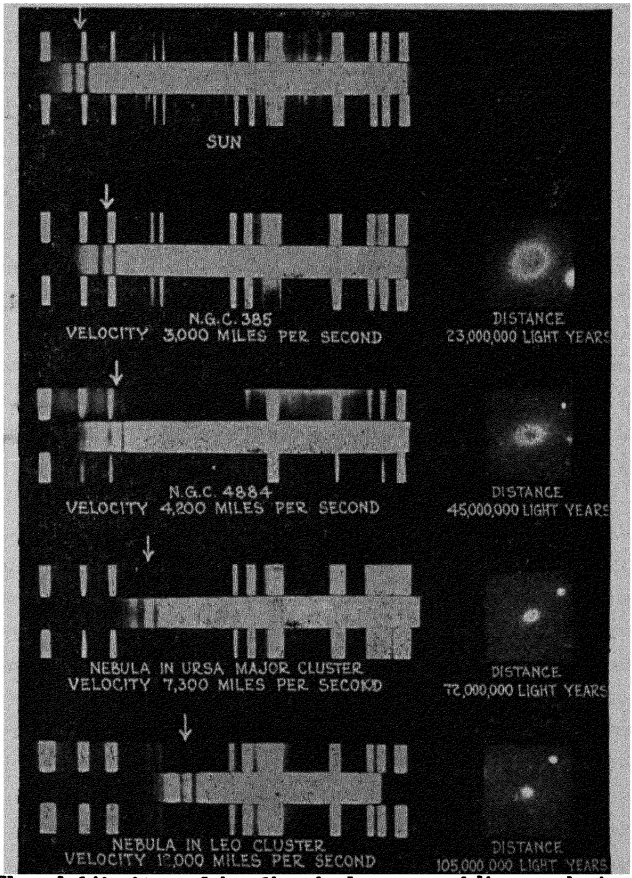


FIG. 330. *The red shift of two calcium lines in the spectra of distant galaxies. The increase of the shift with increasing distance is clearly shown. (Photographed by Humason at the Mt. Wilson Observatory.)*

The Expanding Universe

One curious feature of nebular spectra we have not mentioned. The lines in all but a few of the spectra are shifted toward the red end, the amount of the shift increasing with the distance of the nebulae from us. This displacement is illustrated in Fig. 330, which shows two of the absorption lines of calcium (indicated by arrows) in the spectra of several nebulae. Each nebular spectrum is shown between two comparison

spectra, so that the shift of the two lines toward the red (toward the right in this picture) for nebulae at different distances is clearly evident.

A "red shift" in the spectrum of a star we have interpreted as meaning motion of the star away from the earth. If we use the same explanation for nebular spectra, we must infer that nearly all the spirals are receding from us. Even the few nebulae which are moving toward us probably owe their apparent motion to the movement of the sun as our galaxy rotates; if this movement in our galaxy is allowed for, the other galaxies without exception appear to be retreating. Not all astronomers are agreed that the red shift definitely proves this motion, but no other satisfactory explanation has been offered.

Accepting outward motion of the galaxies as probable, we may use the amount of the red shift to calculate its speed. The calculations give enormous figures: several hundred miles a second for some of the nearer galaxies, many thousand miles a second for the more distant ones. The speeds are in general directly proportional to the distances of the nebulae from us, each increase of a million light-years in distance leading to an increase in speed of roughly 100 mi/sec.

It would look at first as if our galaxy has some strange repulsion for all other galaxies, forcing them to move away from us with ever-increasing speed. But it seems hardly probable that we occupy so important a place in the universe. If we make the more reasonable guess that all the galaxies are drifting away from each other, spreading apart like the fragments of a bursting rocket, an observer on our galaxy or on any other would get this illusion of his neighbors fleeing in all directions. The universe, in other words, seems to be rapidly expanding, its component galaxies moving ever farther apart.

If we have interpreted the red shift correctly, then at some time in the remote past the galaxies must have been very much closer together than at present. We cannot know, of course, how long the present speeds of recession have been maintained, but on the assumption that the speeds have not varied greatly we may calculate backward to a time when the galaxies were close enough for their materials to intermingle. The time turns out to be about 2 billion years ago.

Perhaps it is only coincidence that this figure is so close to the age of the earth as measured by the amounts of radioactive materials in rocks, and so close also to the age of the solar system as measured by radioactive materials in meteorites. But the general correspondence of these numbers hints at a possibility that the sun's family of planets dates from a time not many billion years in the past, when some cosmic cataclysm set the spiral nebulae receding from each other. Perhaps the galaxies and even the stars themselves were formed in earlier stages of this same tremendous event.

Questions

1. What is the evidence for each of the following statements:
   a. The sun is part of an aggregate of stars having roughly the shape of a thin watch?
   b. The sun is nearly in the central plane of this watch-shaped aggregate?
2. Compare the motions of (*a*) stars within a galaxy, and (*b*) the galaxies themselves, with the motions of molecules in a gas.
3. What are the chief differences between (*a*) a globular cluster and a spiral nebula? (*b*) A galactic nebula and a spiral nebula?
4. Why may greater distances be measured by means of Cepheid variables than by determining intrinsic brightness spectroscopically?
5. What is the evidence that the spiral nebulae are not a part of our galactic system?
6. What is the evidence that the spiral nebulae are galaxies of stars?
7. Suppose that you were trying to demonstrate the rotation of a spiral nebula spectroscopically. Which would you choose for your study, a nebula of which you could see the "full-face" view or one turned edgewise toward you? Why?
8. To what conclusions about nebular motions does the "red shift" lead?

CHAPTER XLIII

Frontiers of Physical Science

WE HAVE now explored all the major provinces of physical science. In each province we have recognized the same goal: to explain, to simplify, to see behind complex events an underlying order and uniformity. We have found, too, the same method of attack: careful observation and experiment, then the formulation of hypotheses and laws, at length the bold combination of laws into more general laws and broad theories—with continual checks on each new generality by further observation. We have seen one part of the world after another yield to this attack. Behind the apparent motions of the sun and planets we have found the orderly arrangement of the solar system; the infinitely complex behavior of ordinary matter we have interpreted by means of tiny particles; from rocks and from careful scrutiny of present landscapes we have read much of the earth's history; and finally, from starlight we have learned something about the earth's relationship to the rest of the universe.

We have kept pretty much to the well-trodden paths of science, to those parts of each province where repeated observations and well-tested laws give us confidence in methods and conclusions. Only rarely have we ventured out toward the frontiers of present knowledge: when we watched the physicist bombard atomic nuclei, for instance, or when we joined the geologist in his speculations about the earth's interior, or when we followed the astronomer 500 million light-years into space.

We should not have to go far to find other frontiers of physical science. We might watch the physicist experimenting with the wave properties of electrons and other tiny particles; we might watch the chemist trying to understand the effect of catalysts on the speed of chemical reactions; we might watch the astronomer seeking a reason for the fluctuations in brightness of variable stars. Or we could venture out on the industrial frontiers, where chemists are fabricating new types of molecules for

improved plastics and safer anesthetics, where physicists and engineers are trying to improve the transmission of images by television, where geologists are using new electric and magnetic methods of deciphering hidden rock structures in the search for metals and petroleum. Or we might look for the patient workers who undertake less spectacular problems: a more accurate measurement of the wave lengths of certain spectral lines, a careful determination of the energy involved in a chemical reaction, a study of the amount of ionization at different levels in the atmosphere, a careful mapping of the rocks and rock structures in a small area. In universities and colleges, in the laboratories of modern industry, in lonely observatories, and on expeditions to the far corners of the earth, we could find hundreds of men and women trying to penetrate a little beyond the present limits of our knowledge of the natural world.

Before taking leave of physical science, we shall make one more brief excursion to its far outposts. Of the many points on the frontier which we might choose, we shall select a few which can give us a further insight into fundamental scientific concepts.

Relativity

One of the most searching of modern inquiries into the basic ideas of physics is the theory of relativity, which we owe to the German physicist Albert Einstein.

To approach the meaning of relativity, we begin with some very simple experiments. Suppose that you are on a moving train and that I am observing you from the station platform. We shall suppose that your car has transparent walls, so that I may see all details of your experiments, and that both of us are equipped with the necessary apparatus for making velocity measurements of various kinds. First of all we shall imagine that you roll two balls along a level table, one toward the front end of the car and the other toward the rear, and that both of us attempt to measure the speed of each ball. Your results will indicate that the balls are moving with the same speed, say 20 mi/hr. But to me their speeds will seem very different. The ball which you rolled forward would have not only the speed given it by your arm, but in addition the speed of the train; if the train's speed is 20 mi/hr, I would see the ball moving 20 + 20 or 40 mi/hr. The second ball would appear to me motionless, for you have given it just enough speed backward to overcome the train's motion forward. The speed of each ball is evidently *relative* to the position of an observer: relative to your position both speeds are 20 mi/hr, while relative to mine they are 40 mi/hr and 0 mi/hr, respectively.

All this is commonplace enough. A moving observer must always report the motions of other objects differently than a stationary observer,

since his own motion is added to other motions. But now we face an awkward question: Which is right, your measurement of the balls' speed or mine? If you are sufficiently modest, you will perhaps concede that my measurements show the "real" motion, since I am connected with the stationary earth while you are obviously moving. But if you are more obstinate you will question my claims: after all, the earth itself is moving, so why should measurements from a station platform be any more valid than measurements from a moving train? Perhaps the "real" motion of the balls could only be seen by an observer on the sun, who would add to your observations not only the speed of the train but the speed of the earth as it rotates and moves in its orbit. Yet the sun itself is moving with respect to the stars, so why should an observer there be any more correct than one of us? We might refer the problem to an observer at the center of our galaxy; but even he would have to admit that he is moving with respect to other galaxies. We are driven to conclude that there is no "real" motion of the balls—or perhaps better, that the motion seen by any observer is just as "real" as that seen by any other. Motion has no meaning unless it is referred to an observer; if an object is alone in a universe of empty space, there is no possible way of determining whether it is in motion or not.

If you should try other simple experiments in your moving railroad car while I look on, we would find that your observations and mine were in good agreement. In some of your measurements I would have to make a correction for the speed of the train, but with this correction I would find that the ordinary laws of physics hold as well on the moving car as on the station platform. No experiment gives us any basis for concluding that your measurements are any more "right" than mine, or vice versa. Einstein phrased this result in general terms: *all laws of nature are the same in all systems moving uniformly relative to each other.*

This statement leads to no difficulty until we try experiments which involve the speed of light. To see how the trouble with light arises, let us first try an experiment with another form of wave motion, sound. Suppose that you stand in the middle of your moving car and speak a few words; suppose further that both of us have means for determining when the sound of your voice reaches the two ends of the car. You would find that the sound moved in each direction with its normal speed, about 1,100 ft/sec, and that sound waves reached the two ends at the same instant. I would find, of course, that the sound waves moving forward were traveling somewhat faster than normal because of the train's motion, and that those moving backward were somewhat slower. I would agree with you, however, that sound waves reach the two ends simultaneously, since the front end of the car is moving away from the source of the sound and the rear end toward the source just fast enough to make up for the differences

in speed. There is nothing new here: the motion of sound waves is exactly analogous to the motion of the two balls.

Now suppose that you switch on a small light in the middle of the car. If we undertake the difficult measurement of the speed of light toward the front and rear of the car, would you again expect to find equal speeds in the two opposite directions? Should I expect to find one speed increased and the other decreased by the train's motion? These are difficult questions to answer before trying the experiment. Our results for sound waves depended on the fact that the air molecules which transmitted the sound were being carried along by the train. Now light is not transmitted by any material particles, but consists of changes in an electromagnetic field in empty space. Rather than try to predict how such disembodied waves would behave, let us try the experiment. We find, amazingly, that both of us get the same figure for the speed of light in the two directions. Neither motion of the light source nor motion of the observer seems to affect the speed of light at all.

Historically, our hypothetical experiment puts the cart before the horse. During most of the nineteenth century light was believed to consist of wave motion in a medium called the "ether," just as sound is wave motion in air or in other materials. The existence of the ether became questionable when measurements showed that the speed of light was not influenced by the direction in which the light moved or by the speed of its source.

Einstein generalized this strange property of light in the statement: *the speed of light* in vacuo *is the same in all systems moving uniformly relative to each other.*

This statement and the one on page 650 are the foundation for the *special theory of relativity*, which Einstein published in 1905.

Consequences of Relativity

The two generalizations of the special theory lead immediately to some odd facts about light and about the behavior of ordinary objects at high velocities.

Consider again our measurements of the speed of light in a moving railroad car. From inside the car, you see light traveling with equal speeds from a source in the middle toward the two ends; naturally, you find that light waves reach the two ends simultaneously. From outside the car, I also see light traveling in opposite directions with the same speed. But while the light is traveling, I see the rear end of the car *approaching* the backward-moving light waves and the front end of the car *receding* from the forward-moving waves. Hence I report that light reaches the rear end before it reaches the front end. Two events which are simultaneous from your point of view seem from my point of view to be separated by a

short interval of time. At the instant when you claim that each light ray is striking an end of the car, I claim that one ray has already struck and the other has not yet arrived.

Again we ask, which observation is right? And again the answer must be, *both* are right. Like velocity, the concept of *simultaneity* is relative to the motion of the observer. This means a revision in our cherished notions of past, present, and future: for an event which seems to be happening *now* from my point of view may be an event of the past to a second observer, and may still be in the future to a third. Our different conclusions about the same event depend simply on our motions relative to each other.

Let us try a new and somewhat more fantastic experiment. Suppose that we point a flashlight upward and switch it on momentarily. We may imagine its light, a short train of electromagnetic waves, traveling rapidly outward into space. Suppose that just as the light is flashed you set out in the same direction in a rocket, at a speed say of 180,000 mi/sec, while I as usual remain safely on the ground. If we both measure the speed of the retreating waves, we must both get 186,000 mi/sec, since motion of the observer does not affect the speed of light. But how is this possible? We are both using the same kind of clocks to measure time and the same sort of meter sticks to measure distance. How can you possibly maintain that the light flash is moving away from you at 186,000 mi/sec, when to me it is perfectly obvious that your rocket is remaining close behind?

Our measurements can both be right only if your motion has somehow changed the characteristics of your measuring instruments. The required changes are a decrease in the length of your meter stick and a slowing down of your clock: if your meter stick is shortened, then the distances you measure will seem longer than they should be, and if your clock runs slow the times you measure will be abnormally short. If the changes are large enough, your measurement of 186,000 mi/sec for a speed which to me seems only 6,000 mi/sec becomes understandable. You, of course, would not be aware of the changes; your meter stick and your clock would seem perfectly normal. You would feel, in fact, that my instruments had somehow altered, so that my measurements were at fault rather than yours.

The shortening of a meter stick at high velocities means that *length* also must be relative to the motion of the observer. The change in length is only *in the direction of the relative motion;* the meter stick on your rocket would appear to me to have its normal length if you turned it sidewise to your motion, but to shrink when you turned it into the line of motion. The contraction is not as amazing as it sounds: if we remember that a meter stick is made up of electric charges, and that electric charges in motion exert strong forces on one another, some change in shape due to

very rapid motion seems plausible. A contraction of this sort was, in fact, predicted from Maxwell's electromagnetic theory some years before Einstein's work.

The slowing down of clocks at high speeds is a consequence of an increase in the inertia, or mass, of their moving parts. *Mass*, like length, is relative to the motion of the observer: not only clocks, but all objects on your moving rocket would appear to me to be heavier because of your motion.

The variability of mass and length with motion violates some other long-sacred ideas. Ordinarily we regard these quantities as strictly constant, and use them as constants in describing motions and forces. At all ordinary speeds this usage is correct enough: theoretically a car going 50 mi/hr is heavier than one at rest, but the difference is too minute for any conceivable measurement. Only at very high speeds does a contraction in length and an increase in mass become noticeable; not until the speed of a moving object reaches 161,000 mi/sec is its length cut in half and its mass doubled. Beyond this speed the changes are rapid, the length dropping toward zero and the mass growing ever larger as the speed of light is approached.

Because such enormous speeds are necessary for any change to be apparent, these predictions of the relativity theory are difficult to confirm experimentally. No object of ordinary size can be given a speed nearly sufficient for the changes to be measurable. But electrons shot out as beta rays from the nuclei of radioactive atoms do provide one good check: these tiny particles move at speeds approaching that of light, and their masses show the variation with speed which the theory predicts.

Another conclusion of the special theory which can be tested experimentally is that mass may be converted into energy, the amount of energy (in ergs) equivalent to m grams of mass being given by

$$E = mc^2$$

where c is the speed of light (in centimeters per second). This predicted relation is borne out in nuclear bombardment experiments, in uranium-graphite piles, and in the atomic bomb (pages 295–303).

The Fourth Dimension

In the special theory of relativity, *events* play an important role—for instance, the arrival of a light ray at a certain place at a certain time. An event is fully described only when both the place of occurrence and the time of occurrence are specified. Thus we say, "The Declaration of Independence was signed in Philadelphia on July 4, 1776." If we said only that the signing took place in Philadelphia or if we merely named the date of signing, the description of the event would be incomplete.

To designate the place where an event occurs requires three numbers: thus to locate Philadelphia with respect to the earth's center, we might give its latitude, its longitude, and its distance from the center. Three numbers are necessary simply because we regard space as three-dimensional—as having length, breadth, and thickness. Time, on the other hand, may be designated by a single number, like the number of hours since noon or the number of days since the birth of Christ. To describe an event, then, requires a minimum of four numbers, three referring to space dimensions and one referring to time.

Now in the equations of relativity Einstein uses the numbers referring to time in the same manner as the numbers referring to space. In effect, he regards time as a sort of fourth dimension: events are located in a unified "space-time" rather than separately in space and in time. We have already noted one instance of the intimate connection between space and time in our discussion of simultaneity, when we found that concepts of present, past, and future depend on motion through space.

Thus *time* is Einstein's famous fourth dimension. There is nothing particularly strange or difficult in the idea: it is merely a change in point of view which makes certain natural phenomena easier to describe.

Gravitation

The special theory of relativity is an attempt to generalize the equations describing natural laws, so that they will apply to the experiments of any observers moving *uniformly* with respect to each other, regardless of their speed. A more difficult undertaking is embodied in the *general theory of relativity*, which Einstein published in 1915. Here his problem was to modify the equations for natural laws so that they would apply to any part of space, no matter how that part of space might be moving with respect to an observer. The special theory refers only to uniform motion; the general theory includes accelerated and rotational motion as well.

A simple example will show something of the increased complexity which accelerated motion introduces. Suppose that you had lived all your life in a large box falling freely toward the earth from somewhere far out in space. The box and everything within it would be moving with uniform acceleration, the speed increasing steadily as it approaches the earth. Since you are moving with the box, you would not be conscious of any gravitational pull toward the earth: your feet would not press down on the floor, and if you tried to let a ball drop it would hang motionless in mid-air. Obviously, from my position on solid earth I would see things in your box very differently from the way you do. If I watched you drop a ball, I would say that the ball thereafter approached the earth with a steadily accelerated motion; you would say that the ball

remained motionless with respect to your box, and that how it might move with respect to the earth was no concern of yours. I would say that the ball was impelled downward by a force; you would say that you find no evidence of a force whatever. One of us sees accelerated motion and the action of a force; the other sees no motion and no force. Einstein would tell us that we were both right, as usual. Not only the motion, *but the force itself*, is relative to our points of view.

Thus an extension of relativistic ideas to accelerated motion involves a new description of gravitation. In this description Einstein dispenses with a mysterious gravitational force associated with matter; he substitutes a complex curvature in space-time near every material object, the distortion being greatest near large masses. This curvature is not in the familiar three dimensions of space, not even in a fourth dimension, but in *six additional dimensions*. Such a curvature can be handled mathematically, but it is impossible to visualize in any terms that our minds can grasp. The important thing is that Einstein refers the motion of objects near large masses to a *structure* in the surrounding space, rather than to a *force* in the ordinary sense. Maxwell accomplished much the same thing in electricity by centering attention on the structure of electric and magnetic fields rather than on the forces exerted by charges and magnetic poles.

Einstein's picture of gravitation as a distortion in space-time does not in any sense prove that Newton was wrong. Einstein simply reexamines the whole problem from a new and broader point of view, and suggests a different formulation. Einstein's equations give the same results as Newton's for all ordinary masses, suggesting slightly different results only for very strong gravitational fields, like that near the sun. Three predictions from the theory for strong fields—slight irregularities in the orbit of Mercury, the deflection of starlight passing close to the sun, and displacements of lines in the sun's spectrum—have been verified by observation, suggesting that the theory does actually give a truer description than Newton's for gravitational effects near large masses.

Einstein's great contribution to physical science is perhaps less the theory of relativity itself than his demonstration that ordinary concepts of time, distance, and mass must be modified at high velocities and that ordinary ideas about gravitation must be changed to apply to large masses. The theory of relativity is one possible suggestion as to how these changes should be made. It is amazingly successful in some applications, particularly to large-scale phenomena; but attempts to apply it consistently to particles within the atom have so far met with failure. The theory is no more of a final formulation than was Newton's: rather, it is one more step in the attempt of scientists to base their description of the universe on fewer and simpler fundamental assumptions.

Fundamental Concepts

The theory of relativity changes our older ideas about time, distance, mass, and force so radically that perhaps we should scrutinize some of these concepts more carefully than we have done hitherto.

Let us begin with the familiar concept *distance,* and to make the discussion definite let us consider the distance between two marks on the floor. Of course we know that this distance is not as definite as it seems, since to a moving observer it would be shorter than it appears to us; but ignoring such matters for the moment, let us see what precise meaning we can give, from our own particular point of view, to the expression "distance" as applied to those two marks on the floor. The problem sounds like a futile one: everyone knows so instinctively what distance means that its definition might well be left to the dull expressions of dictionary makers. We think of expressions like "interval," "amount of space," "separation," but these are little more than synonyms.

If the question is asked differently, What is the numerical value of the distance? we can find an answer immediately. All we need do is to lay a meter stick along the line connecting the points, and see how many times the meter stick will fit into the space between them. Now does distance have any real meaning outside of this simple operation? Do we not gain our earliest impressions of distance in childhood by seeing how many fingers or how many steps will fit between two points? Do not our instinctive estimates of distance in later life depend on such childhood experiences and on subsequent experiences with rulers?

Perhaps a philosopher can find a deeper meaning in distance, but as commonly used the word signifies no more than *that which can be measured with a meter stick (or any other unit of length).*

Such a definition leads immediately into difficulties. The distance to the moon, for instance, obviously cannot be measured with a meter stick. What, then, do we mean by the statement that our satellite is 240,000 mi away? This figure is derived by sighting on the moon from two points at opposite ends of a diameter of the earth, reading the angles between the diameter and the rays of light coming to us from the moon, and then putting these angles and a figure for the diameter into an equation which contains the moon's distance as an unknown. This elaborate procedure involves at least two assumptions: (1) that rays of light follow straight lines from the moon to the earth and (2) that the equation is valid. The straightness of light rays can be checked on a small scale in our laboratories, and the equation can be verified for small triangles; but the extension of the assumptions to the moon is not altogether safe. This particular distance can be verified in enough other ways, however, so that we feel considerable confidence in its correctness.

The same sort of procedure gives us the distances to the nearer stars. Again we must trust to the straightness of light rays and the correctness of a formula over distances far larger than any for which a direct check is possible. Here the extension is greater and our checks on the results fewer and less accurate.

For the farther stars we depend on a measurement of intensities of spectral lines, and for star clusters and the nearer nebulae we use the fluctuations in brightness of Cepheid variables. With each step into more remote regions of space we get farther from our fundamental concept of distance as something measured with a meter stick; assumed relationships are stretched farther and farther from regions where they can be accurately checked; distance becomes a term in complex equations, a quantity plotted on graphs. When we speak of a distance of 500 million light-years between galaxies, we may well wonder if we are talking about the same sort of quantity as the distance between two marks on the floor.

Similarly for small distances: we get the separation of ions in a crystal lattice from a formula involving X-ray wave lengths, and the diameter of an electron comes from formulas still farther removed from anything resembling a direct measurement with a meter stick. Again it does not seem quite sure that a quantity so indirectly measured has the same meaning as ordinary distances.

We could meditate in similar fashion about many another fundamental concept of science. We gain our knowledge of mass by measurements of weight or inertia and then extend the term to stars and galaxies on the one hand, to atoms and electrons on the other, whose weight and inertia convey no direct impression to our minds at all. Time is something that we measure with clocks, or with other changes that we experience in the few years of our life span, yet we do not hesitate to speak of a billion years hence or a billion years in the past. Temperature is something that we feel with our skins and measure more accurately with thermometers, yet we use the term in places where neither sensations nor ordinary thermometers would give any measurements at all.

This is not to discredit the accurate measurements of extraordinary distances, masses, and temperatures which modern physics and astronomy have made possible, but merely to emphasize an important fact: that such feats involve an enormous extension of concepts which are based on the response of our minds to simple measurements of familiar objects. There is, of course, no escape from this procedure; we must either make the necessary extensions of familiar concepts or else confess our inability to comprehend anything beyond our immediate surroundings. And of course the equations and the assumptions used in making the extensions are checked as often and as fully as possible. Yet a skeptic may legitimately doubt that quantities used in such grossly unfamiliar cir-

cumstances mean quite the same thing as quantities of the same name in everyday life.

Perhaps we can find here a reason for the odd fact that on the frontiers of science we so often encounter mathematical expressions whose meaning our minds cannot visualize. Thus in relativity we encounter equations which contain such familiar quantities as length, mass, and time, but which suggest a structure of space which our brains cannot grasp. Another equation of this theory sets intangible energy equal to tangible mass. Probings within the atom reveal a strange confusion of particles and waves which cannot be represented by any sort of a model that the mind can visualize. Now perhaps in such applications of science to masses and distances which are enormously small and enormously large, we have stretched these quantities so far that they no longer have the meaning which our minds originally gave them. Perhaps, in other words, we have reached here frontiers of the mind itself, rather than boundaries in the objective world of knowledge.

It is certainly not strange that somewhere on the outskirts of modern science we should find ourselves concerned with the functioning of the human brain. For science is a product of that brain, just as truly as are music and literature and art. Too often we think of science as a structure in the objective world quite apart from our minds, about which we have only to "discover" one fact after another. Science is rather a product of the mind's groping for a way to simplify and organize the outside world, the framework and the methods of the organization stemming from the mind itself.

Physical science impresses us with its vast distances, its immensely long reaches of time, its incredibly small particles; yet all these have meaning only in terms of concepts which the mind derives from its own experience. If the size and the great age of the universe seem at times to reduce us and our affairs to insignificance, we need but reflect that it is our minds which give meaning to distance and time. So far as we know, the human brain is the only device in all the universe which can fathom the depths of space and project itself into remote ages of the past and future.

Suggestions for Further Reading—Part VI

On the sun, stars, and nebulae:

DUNCAN, J. C.: *Astronomy*, Harper & Brothers, New York, 1935. A standard elementary text.

FATH, E. A.: *Elements of Astronomy*, McGraw-Hill Book Company, Inc., New York, 1934. A standard elementary text.

JEANS, J. H.: *The Universe around Us*, The Macmillan Company, New York, 1929. A series of informal essays.

GAMOV, G.: *The Life and Death of the Sun*, Viking Press, Inc., New York, 1940. An entertaining and well-written account of the probable life history of an average star.

On relativity and fundamental concepts:

EINSTEIN, A., and L. INFELD: *The Evolution of Physics*, Simon and Schuster, Inc., New York, 1938. See p. 321.

EDDINGTON, A. S.: *The Nature of the Physical World*, The Macmillan Company, New York, 1932. Nonmathematical, but detailed and fairly difficult reading.

LEWIS, G. N.: *The Anatomy of Science*, Yale University Press, New Haven, 1926. A series of brief, informal essays on relativity and fundamental ideas in all branches of science.

BRIDGMAN, P. W.: *The Logic of Modern Physics*, The Macmillan Company, New York, 1927. A philosophical inquiry into the meaning of fundamental physical concepts.

Appendix

TABLE XXVIII. THE METRIC SYSTEM

Units of length:

	1 meter (m)	= a little over a yard (39.37 in.)
1 centimeter (cm)	= 0.01 meter	= a little less than ½ inch (0.39 in.)
1 millimeter (mm)	= 0.1 centimeter	= about 4 hundredths of an inch (0.039 in.)
1 kilometer (km)	= 1000 meters	= a little over ½ mile (0.62 mi)

Units of area:

10,000 square centimeters (sq cm) = 1 square meter (sq m.)
= almost 11 square feet (10.76 sq ft)

Units of volume:

1,000 cubic centimeters (cc) = 1 liter (l.) = a little over 1 quart (1.06 qt)

(Strictly 1 liter = 1,000.027 cc, but for all ordinary purposes the fraction 0.027 may be disregarded.)

Units of weight:

1 gram (g.) = the weight of 1 cubic centimeter of water.

(Strictly this is not quite true. At 4°C, 1 cc of water weighs 0.999973 g., and the weight is slightly less for higher and lower temperatures.)

1 kilogram (kg) = 1,000 grams = a little over 2 pounds (2.20 lb)

1 milligram (mg) = 0.001 gram

TABLE XXIX. VERY LARGE AND VERY SMALL NUMBERS

Very large numbers are expressed in terms of the following powers of 10:

$10^0 = 1$
$10^1 = 10$
$10^2 = 10 \times 10 = 100$
$10^3 = 10 \times 10 \times 10 = 1{,}000$
$10^4 = 10 \times 10 \times 10 \times 10 = 10{,}000$
$10^5 = 100{,}000$
$10^6 = 1{,}000{,}000$
$10^7 = 10{,}000{,}000$
$10^8 = 100{,}000{,}000$ etc.

Any large number may be expressed as a product involving one of these powers. For example, the number 359,000,000 may be written

$$359 \times 1{,}000{,}000 \quad \text{or} \quad 359 \times 10^6$$

Alternatively it may be expressed:

$35.9 \times 10{,}000{,}000$ or 35.9×10^7
$3.59 \times 100{,}000{,}000$ or 3.59×10^8
$3{,}590 \times 100{,}000$ or $3{,}590 \times 10^5$, etc.

To convert a number like 26×10^{13} into ordinary notation, simply write out the power of 10 and multiply:

$$26 \times 10^{13} = 26 \times 10{,}000{,}000{,}000{,}000 = 260{,}000{,}000{,}000{,}000$$

Note that when 26 is multiplied by 10^{13}, the decimal point is moved 13 places to the right. *In general, the exponent of* 10 *indicates the number of places by which the decimal point should be moved to the right.*

Very small numbers are expressed in terms of the following negative powers of 10:

$$10^{-1} = \frac{1}{10^1} = 0.1 \qquad 10^{-4} = \frac{1}{10^4} = 0.0001$$

$$10^{-2} = \frac{1}{10^2} = 0.01 \qquad 10^{-5} = \frac{1}{10^5} = 0.00001$$

$$10^{-3} = \frac{1}{10^3} = 0.001 \qquad 10^{-6} = \frac{1}{10^6} = 0.000001, \text{ etc.}$$

Any small number may be expressed as a product involving one of these powers. For example, the number 0.000082 may be written

$$82 \times 0.000001 = 82 \times 10^{-6}$$

Alternatively it may be expressed:

8.2×0.00001 or 8.2×10^{-5}
820×0.0000001 or 820×10^{-7}, etc.

To convert a number like 406×10^{-10} into ordinary notation, simply write out the power of 10 and multiply:

$$406 \times 10^{-10} = 406 \times 0.0000000001 = 0.0000000406$$

In general, the negative exponent of 10 *indicates the number of places by which the decimal point should be moved to the left.*

Index

Boldface figures indicate definitions and other important references.

A

Absolute temperature, **122**, 131
Absolute zero, **122**, 131
Absorption spectra, 266, 307–308
  stellar, 607, 626–628
Acceleration, **41**, 53–55
  of gravity, 55, 58
Acetate ion, 355–356
Acetic acid, **410**
  equilibrium in, 383
  ionization of, 351
  structure of, 404
Acetylene, 407
Acidic oxides, 359–360
Acidic rocks, 481
Acids, **349–351**, 353–357
  names of, 189
  relation to nonmetals, 196, 359–360
Activated molecule, 369, 380
Activation energies, 368–370, 380
Adams, J. C. (1819–1892), 84, 85
Adams, W. S. (1876–), 622
Air, composition of, 443
  resistance, 44, 62–63
Air masses, **460–462**, 464–467
Alchemy, 149–150
Alcohols, 409–410
Alkali, **352**
Alkali metals, **200**, 323–324, 328–329
Alkaline solutions, 352
Alloy, **398**
Alluvial fan, 510
Alpha particles, **277–280**, 282, 285–287, 292
Alpha rays, 277
Alternating current, 249
Aluminum, metallurgy of, 399
Amino acids, 413–414
Ammeter, 243
Ammonia, 357–359, 379, 386–387
  equilibrium in solution, 384–385
Ammonites, 586
Ammonium chloride, 358
Ammonium salts, 358
Amorphous solids, **137**, 139–140
Ampère, A. M. (1775–1836), 245, 248
Ampere (unit), **245**
Amphibians, 580
Anderson, C. D. (1905–), 290
Andesite, 480
Anesthetics, 415
Anode, 228
Antares, 624–625, 630
Anticline, 535, 538
Anticyclones, 462, 464–467
Appalachian geosyncline, 558, 573–574
Appalachian Mountains, 498, 535, 584, 592
Appalachian revolution, 567, 577–578
Apparent brightness of stars, 622
Argon, 200, 201
Aristotle (384–322 B.C.), 11, 17, 21, 44, 45, 148–149, 156, 620
Arrhenius, S. A. (1859–1927), **342–344**; 433
Asteroids, 18
Aston, F. W. (1877–1947), 293–294
Atmosphere, circulation of, 454–458
  composition of, 443–444
  moisture in, 452–454
  pressure of, 119

Atmosphere, structure of, 442–444
temperatures of, 442–443, 451–452
unit of pressure, **119–120**
Atmospheres of planets, 30–33, 36, 134–135
Atomic bomb, 281–282, 298–303
construction of, 302–303
possibilities of, 303
Atomic bombardment experiments, 282–287, 288–291, 616
Atomic collision, 285
Atomic energy, 295–303, 375–376
Atomic number, **287–288**, 289–291, 293–295
Atomic theory, 167–179
Atomic weights, **173–174**, 291
of isotopes, 291–295
table, 175
Atoms, in Dalton's atomic theory, 167–178
electron clouds of, 287, 322–324
nucleus of, **287**
relation to molecules, 171–172
in stars, 616
structure of, 282–295, 309–310, 322–324
Aurora borealis, 443, 614
Avogadro, A. (1776–1856), 171–173
Avogadro's law, 171–173

B

Baking, 363–364
Balancing equations, 191–192
Barometer, 119
Basalt, 480, 522
Bases, **351–354**, 355–359
names of, 190
relation to metals, 196
Basic oxides, 359–360
Basic rocks, 481
Batholiths, **526–528**
Battery, 237–238, 246, 247
Becquerel, A. H. (1852–1908), 272–273, 275, 276
Benzene hydrocarbons, 407–408, 416, 582
Berthollet, C. L. (1748–1822), 163
Berzelius, J. J. (1779–1848), 176, 190
Bessel, F. W. (1784–1846), 621
Beta rays, 277
Birds, 589
Blast furnace, 397–398
Bohr, N. (1885–), 309–312
Boiling point, 123–124, **138**
Boyle, R. (1627–1691), 121, 150
Boyle's law, 120–121, 130–131
Brachiopods, 563, 576
Brahe, Tycho (1546–1601), 12
Bricks, 427–428
Bright-line spectra, 266, 306–307
Bromine, 198–199, 392
Bronze, 397
Brownian movement, **129**, 135, 431
Burning, 161
Butane, 406

C

Calcite, **476**
Calcium carbonate, 346, 361–363
Calendar, 5
Caloric, 110, 163
Calories, **110**
in food, 374
Canals on Mars, 31
Carbohydrates, **411–413**
cellulose, 412, 413
energy from, 374
formed by photosynthesis, 371–372
starches, 412, 413
sugars, 412
Carbon, **183–185**
compounds of, 401–418
electron structure of, 401–402
structural formulas of, 402–405
Carbon dioxide, 184, 361–365
structure, 423
on Venus, 30
Carbon monoxide, 184
Carbonates, 362–363, 472
Carbonic acid, **361–365**
equilibrium in, 382–384
ionization of, 351
in weathering, 489
Carboniferous period, 577
Carborundum, 419
Cascade range, 524, 592
Catalysts, 381–382, 386–387
Catastrophic hypothesis, 546–547
Cathode, 228
Cathode-ray tube, 227, 228, 241
Cathode rays, **227–230**, 272, 273
Causality, 319

Cavendish, Henry (1731–1810), 80
Caves, 363–364, 508
Cellulose, 373, **413**, 416–417
Celsius, A. (1701–1749), 109
Cement, 428
Cenozoic era, 566, 567, 590–597
Center of gravity, 77
Centigrade scale, 109–110
Centrifugal force, **70**, 71, 76
Centripetal force, 71
Cephalopods, 577, 586
Cepheid variables, **638–639**, 643
Ceramic products, 427–428
Cesium, 200
Chain reaction, 298–302
Chalcedony, 483
Chalk, 482
Chamberlin, T. C. (1843–1929), 96, 97
Changes of state, 123–125, 137–140
Charles' law, 122–123, 131
Chemical change, 147, 154–155
  early ideas of, 148–150
  rate of, 379–382
Chemical combination, 324–330
  by electron transfer, 325–326
  by sharing electron pairs, 326–327, 329
Chemical energy, 107, 197–198, **366–368**
Chemical equilibrium, **382–386**
Chemical names, 187–190
Chemical properties, 154
Chemical symbols, 176
Chert, 482–483
Chile saltpeter, 378
Chloride ion, 325, 331–332, **344–345**
Chlorides, 186
Chlorine, **186**, 198–199
  activity of, 392
  isotopes of, 293–294
Chlorophyll, 372
Circuits, closed and open, 238
Circulation of atmosphere, 454–458
Cirques, 503, 504
Classification, of elements, 202–207
  of igneous rocks, 480
  of matter, 152–153
  of metamorphic rocks, 485
  of rocks, 476–479
  of sedimentary rocks, 482
Cleavage, 474–475
Cleopatra's Needle, 492
Climates, **450**, 458–460
Cloud chamber, 284–285
Coagulation of colloids, 436
Coal, as a fuel, 373
  origin of, 577, 582
Coal tar, 416–417
Cocaine, 416
Coils, 241–244
Coke, 373
Cold front, 460–461
Collision hypothesis, 97
Colloid, 432
Colloid chemistry, 431
Colloidal solutions, 430–438
  definition of, **431**
  kinds of, 431–432
  precipitation of, 436
  preparation of, 434–436
  properties of, 432–434
Color, 265–267
Combustion, 156–161
Comets, 18, 19, **35–36**
Commutator, 244, 249
Compounds, **151**
  formulas of, 188–190, 202
  naming of, 187–190
  stability of, 197–198
Concrete, 428
Conductors, 215–217, 230
Conglomerate, 482
Conservation of energy, law of, **111–112**, 297
Conservation of mass, law of, **162–163**, 167, 297
Constellations, **4**
Contact metamorphism, 527
Continuous spectra, 266, 267
Convection currents, 454–455
Copernican hypothesis, 8–13, 16, 87
Copernicus, Nikolaus (1473–1543), 8–11
Copper, activity of, 197, 391
  metallurgy of, 396–397
Copper ion, 344
Cordilleran geosyncline, 558, 573–574, 584, 589
Corona, 29, 615
Corpuscular theory, 270, 314
Correlation of geologic events, 561–566
Cosmic rays, 290, 315–316
Coulomb (unit), **244–245**
Coulomb's law, 217–219
Covalent substances, **327**, 329

Covalent substances, solubility, 339–340
Craters, of the moon, 24–26
  of volcanoes, 518, 519
Cross-bedding, 509
Crust of the earth, 445
Crystal form in minerals, 474
Crystalline solids, **137**, 139–140
Crystals, 117
  analysis by X rays, 275
  ions in, 326–327
  structure of, 137
Curie, Marie (1867–1934), 275–276
Currents, alternating, 249
  deposition by, 513–514
  direct, 249
  electric, 237–254
  induced, 248
  ocean, 467–469
  produced by waves, 505–506
Cuvier, Georges (1769–1832), 547
Cyclones, 461–467
Cyclotron, 283

D

Dalton, J. (1766–1844), **166–167**, 170, 172–173, 176
Darwin, C. (1809–1882), 550, 565
Decay, 395–396
Definite proportions, law of, 163–164
Delta, 510
Density, **146**
  of earth, 446–447
Democritus (460?–362? B.C.), 127, 167
Derivatives of hydrocarbons, 408–411
Deserts, 459
Deuterium, 294–295
Deuteron, 295
Diamond, 183
Diastrophism, 529–543
  causes of, 542–543
  definition, 518
  evidences of, 529–532
  kinds, 532–539
Diffusion, of colloids, **432**
  of gases, 114
  of liquids, 115
  of solids, 117
Digestion, 411–412
Dikes, **525–526**, 528
Dinosaurs, 586–587
Diorite, 480
Direct current, 249
Dispersed substances, 432
Dispersing medium, 432
Displacement reactions, 391–393
Dissipation of energy, law of, **143**
Divide, 494
Doldrums, 455, 456
Double stars, **618–619**, 623–624
Drowned valley, 534
Dry ice, 184
Dunes, 513
Dwarf stars, 628–630
Dynamite, 375
Dynamo, **249–250**
Dynamo effect, 248–249
Dynamothermal metamorphism, 540–541
Dyne, **58**

E

Earth, age of, 562
  atmosphere of, 442–444
  composition of core, 447
  composition of crust, 445
  density of, 446–447
  history of, 569–597
  as a magnet, 222–223
  mass of, 82–83
  origin of, 95–98, 569–570
  as a planet, 20–23, 30, 441–442
  revolution of, 10, 21–22
  rigidity of, 446
  rotation of, 7, 10, 20–21, 94
  shape of, 80–81
Earthquake waves, 447–448
Earthquakes, 531, **539–540**
Eclipses, **24–25**, 28–29, 94
Einstein, A. (1879–), 281, 295–297, 313, 649–655
Electric charges, 228, **213–214**, 230
  compared with magnetic poles, 221
  units of, 218, 244
Electric currents, **237–253**, 331–332
  alternating, 249
  direct, 249
  heating effect of, 238–239
  magnetic effect of, 239–240
  units of, 244–247
Electric energy, 107, 233, 246–247
Electric eye, 313

Electric fields, 224–226
Electric motor, **243–244**
Electric power, 246–247
Electric spark, 216, 234
Electrical units, 244–247
Electricity, 213–226, 234–236
Electrodes, 228
Electrolysis, 331–333
  in metallurgy, 398–399
  of water, 182
Electrolytes, **342–344**
Electromagnet, 242–243
Electromagnetic waves, **256–258,** 261–262, 317
  table, 305
Electron cloud, **287,** 288, 308–311
  structure of, 321–324
Electron orbits, 308–311
Electron shells, 310, 321–324
Electrons, 217, **229–231,** 234–236
  in atoms, 285, 291, 308–310
  in chemical combination, 324–327, 369
  deflection by magnets, 241, 248
  in electric currents, 238
  in electrolysis, 331–333
  emitted from metals, 312–313
  in oxidation and reduction reactions, 389–391
  from radium, 277
  wave nature of, 317–320
  in X-ray tubes, 273
Electrophorus, 233
Electroscope, **215,** 274, 312, 315
Electrostatic machine, 233, 234, 247–248
Electrostatics, 237
Elements, **150–151**
  active and inactive, 197–198
  Aristotle's, 148–149
  families of, 198–201
  modern definition of, **295**
  periodic classification of, 202–207
  table, 175
Ellipse, 12
Emission spectrum, 306–307
Emulsifying agent, 436
Emulsion, 432, 435–436
Endothermic reactions, **366–368,** 385
Energy, 101–112
  of activation, 368–370
  atomic, 295–298
  conservation of, 111–112, 297
Energy, definition of, **103**
  dissipation of, 143
  of food, 374
  forms of, 107–108
  of photons, 312–314
  of quanta, 311–315
  of radiation, 309–312
  relation to mass, 281–282, 295–297
  transformations of, 107–108, 371–372
  units of, 102, 105, 110
  in wave motion, 259
Enzymes, 411–413
Equations, algebraic, 58–60
  chemical, 190–192, 345–347
  ionic, 345–347
Equatorial belt, 456, 458
Equatorial current, 467
Equilibrium, chemical, **382–386**
  of forces, 67–68
Eras, geologic, **566–568**
Erg, **102**
Erosion, 488–508
  agents of, 492
  definition, **492**
  by glaciers, 500–504
  by groundwater, 508
  by streams, 492–500
  by waves, 505–506
  weathering, 489–492
  by wind, 504–505
Escape velocity, 134
Esters, 409–411
Ether, 223, 262
Ethyl alcohol, 404, 409, 412
Ethylene, 407
Evaporation, 137–139
Evolution, theory of, 550, 565
  of horse, 592, 593
Exothermic reactions, **366–368,** 385
Explosives, 374–375, 379
Extrusive rocks, 480

F

Fahrenheit, D. G. (1686–1736), 109
Fahrenheit scale, 109–110
Falling bodies, 44, 45, **52–55**
Faraday, M. (1791–1867), 248, **255–256.** 333, 342–343
Fats, 411, **413**
Fault scarps, 534–536

Faults, 530, **532–535**
Feldspar, **475**
Ferric compounds, 393–395
Ferromagnesian minerals, 475
Ferrous compounds, 393–395
Fertilizers, 378–379
First law of motion, **38–40**, 63, 68
Fishes, 577–578
Fission, uranium, 298–301
Fixed nitrogen, 377–379
Fixed stars, 4
Flint, 483
Flood plain, 497
Fluids, 113
Fluorescence, 228
Fluorine, 198–199, 392
Flux, 397
Folds, **535**
Foliation, 485, 540–541
Foods, 374, 396, **411–414**
Forces, centrifugal, 70
  centripetal, 71
  definition, **40**, **43**
  in equilibrium, 67–69
  fields of, **223–226**
  of gravity, 43, 58, 75–77, 80
  units of, 57, 58
  as vectors, 65–67
Formulas, chemical, 176–178, 187–190, 202
  of ions, 341
  molecular, 403
  structural, 403
Fossils, 482, 547, 549, **562–566**
Foucault pendulum, 20
Fourth dimension, 653–654
Fragmental rocks, 482
Franklin, B. (1706–1790), 213, 214, 217
Frequency of waves, 260–261
Friction, 40, 62–63, 110–111, 133–134
  tidal, 92–95
Front, **460**, 461–467
Frontal surface, 460–461
Fuels, 372–374, 376
Fundamental particles, 291

G

Gabbro, 480
Galactic nebulae, **641**
Galaxies, 634–637
  definition, 635
Galaxies, island universes, 643–645
  shape of, 636
Galilei, Galileo, (1564–1642), 12, 13, **16–18**, 28, 37, **45**, 76–77, 87, 99
Galvanometer, 243
Gamma rays, 261, 277, 282, 297
Gangue, 396
Gas, formulas of, 177
  general properties of, 113–114
Gas, molecules of, 129–135
Gas fuels, 373–374
Gasoline, 373
Gay-Lussac, J.-L. (1778–1850), 122, 171–172
Geiger-Müller counter, 285
Gels, 433, **436–437**
Generator, 248–249
Geologic time, 561–562, 566–567
  table, 567
Geosyncline, 557–558
Giant stars, 628–630
Gilbert, Sir William (1540–1603), 213
Glaciers, 500–504
  deposition by, 511–513
  Pleistocene, 594–597
Glass, 425–427
Glazing, 428
Globular clusters, **637–639**, 644
Glucose, 412
Glycerin, 410
Glycogen, 412, 413
Gneiss, 485, 486, 540–541
Gradient, of a stream, 492
Grand Canyon of Arizona, 495, 496, 498
  geologic history of, 559–561
Granite, 480, 525
Graphite, 183–184, 301
Graphs, 48, 50, 59
Gravitation, 74–85
  constant of, 80
  law of, **77**
Gravitational fields, 224–225
Gravity, acceleration of, 55, 58
  center of, **77**
  force of, 43–45, 58, 75–77
  surface gravity of moon, 83, 84
  in theory of relativity, 654–655
  variations of, 81, 82
Grounding a charge, 216
Groundwater, **506–508**
  deposition by, 514

Gulf stream, 469
Gunpowder, 375

H

Haber, F. (1868–1934), 379
Haber process, 386–387
Half life, 279
Hall, C. M. (1863–1914), 399
Halogen derivatives, 409
Halogens, **198–199,** 323–324, 392
Hanging valleys, 503
Heat, of fusion, 124, 140
  of vaporization, 124, 139
Heat energy, 107–111
  molecular explanation of, 132–134, 140–143
  unit of, 110
Heat engines, 140–142
Heavy water, 294–295
Helium, 201, 277–278, 295–297, 616–617
Hematite, 397
Hemoglobin, 396
Herschel, F. W. (1738-1822), 33, 604–605, 643
Heterogeneous substances, 147, 151–153
High-frequency radiation, 305–306
Hipparchus, (555? -514 B.C.), 7
Homogeneous substances, 147, 151–153
Hornfels, 527
Horse, evolution of, 592, 593
Horse latitudes, 455, 456, 458
Hubble, E. (1889–), 643
Humidity, 452–454
Hurricanes, 464
Hutton, J. (1726–1797), 548–549
Hydrocarbons, **405–411,** 582–583
  benzene, 407–408
  derivatives of, 408–411
  saturated, 407
  unsaturated, 407
Hydrochloric acid, 186, 349–351
Hydrogen, **181–183,** 196
  displacement reactions, 392
  electron orbits, 309–310
  isotopes, 294–295
  reaction, with nitrogen, 379
    with oxygen, 366–370, 380–381
  structure of atom, 289
  on sun, 616–617
  in water gas, 374
Hydrogen chloride, 186
Hydrogen ion, 349–350
Hydrogen peroxide, 182, 381
Hydrogen sulfide, 185
Hydronium ion, **350**
Hydrosphere, 442, 444
Hydroxide ion, **351,** 352
Hydroxides, 190, 351–352
  insoluble, 352
Hypothesis, 87

I

Ice age, 593–597
Icecaps, 501–504
  Pleistocene, 594
Ichthyosaur, 563, 587, 588
Igneous rocks, **478,** 479–481
Imponderable, 110, 157, 163
Indicators, 354
Indigo, 415
Induced charges, 231–234
Induced current, 248–253
Induced magnetic poles, 232–233
Induction, **233**
Industrial organic chemistry, 414–418
Inert gases, 200–201, 323–324
Inertia, **38,** 39, 70
Infrared radiation, 261, 305, 307
Inorganic chemistry, 401
Insulators, 215–216, 230
Interference, 268–270
  of electrons, 317
  of light, 269–270
  of water waves, 268–270
  of X rays, 274–275
Intrinsic brightness of stars, 622–623
Intrusive rocks, 480–481, 524–528
Inverse-square laws, **224**
Inverse-square relationship, 77–80
Invertebrates, 575
Iodine, 198–199, 392
Ionic compounds, **326,** 329–331, 340
Ionic theory of solution, 341–344
Ionization, of acids, 349–351
  of air, by cosmic rays, 315
    by X rays, 273–274
  in solution, 331–332, 341–344, 349–351
Ionization chamber, 285
Ions, **231,** 271, **330–333**
  in crystals, 326

Ions, in equations, 345–347
formulas of, 341–342
properties of, 344–345
in solution, 331–332, 341–344
Iron, activity of, 391–393
compounds of, 393–395
metallurgy, 397–398
Island universes, 643–644
Isomers, **403**
Isostasy, **542–543**, 558–559
Isotopes, 292–295, 299–300

J

Japan current, 469
Jasper, 483
Jeans, J. (1877–1947), 97
Jeffreys, H. (1891–), 97
Joule, J. P. (1818–1889), 111
Joule (unit), 102
Jupiter, 18, 19, 30, **32**, 81

K

Kant, I. (1724–1804), 95
Kaolin, 427, **475–476**
Kepler, J. (1571–1630), 12, 13
Kepler's laws, **13**, 75–78
Killarney revolution, 567
Kilowatt, 247
Kilowatt-hour, 247
Kinetic energy, **101–102**, 105, 107, 297
of molecules, 131–143
Kinetic theory, 127–143
Krypton, 201

L

Labrador current, 469
Lake deposits, 514
Landscapes, 497
Laplace, P. S. (1749–1827), 95
Laramide revolution, 567, 590
Laue, M. (1879–), 274–275
Lava, 518
Lavoisier, A. L. (1743–1794), **158–160**, 162–163
Law, 87
Avogadro's, 171–174
Boyle's, 120–121
Charles', 122–123
of conservation of energy, **111–112**, 124, 162–163, 297
Law, of conservation of mass, **162**, 167, 296
Coulomb's, 217–219
of definite proportions, 163–164, 167
of dissipation of energy, **143**
of gravitation, **77**
inverse square, 224
periodic, **202–207**
of uniform change, 548–**552**
Lawrence, E. O., 283
Laws, inverse square, 224
of motion, 38–45, 68–69, 74
of planetary motion, 13, 75–78
Leverrier, U. J. J. (1811–1877), 84, 85
Life, on Mars, 31, 36
on planets, 36
Light, 262–270
color, 265–267
corpuscular theory of, 270
electronic explanation of, 308–311
interference, 268–270
photons of, 312
quanta of, 314
ray, 263
reflection, 263–264
refraction, 264
speed, 261, 650–652
Light year, 622
Lightning, 234
Limestone, 363, 482, 489
Limestone caverns, 363–364
Limewater, 362
Limonite, 397
Line series in spectra, 308–310
Lines of force, **224–226**, 239, 255–257
Lithium, 200
Lithosphere, 442, 444–446
Liquid, formulas of, 177–178
general properties of, 113–116
structure of, 135–136
Loess, 513
Long-shore currents, 505
Longitudinal waves, 260
Lowell, P. (1855–1916), 31, 85
Lucretius (96?–55 B.C.), 127, 167
Lyell, C. (1797–1875), **549–551**

M

Magma, **518**, 520–521
Magnetic compass, 222

Magnetic energy, 107
Magnetic fields, 225–226, 239
  around coils, 241–243
  effect on currents, 240–241
  produced by electric currents, 239–242, 250–251
Magnetic north pole, 222
Magnetic storms, 223, 614
Magnetism, 219–226
  terrestrial, 221–223
Magnetite, 222
Magnets, **219–221**
  north pole, 219
  south pole, 219
Man, geologic history of, 596–597
Mantle rock, 491, 492
Mammals, 589, 592–593, 596
Marine sediments, 513–514
Mars, 10, 18, 19, **30–31**
Mass, conservation of, 162–163, 297
  definition, **39**
  at high velocities, 653
  of planets, 30
  relation to energy, 281–282, 296–297
Mature landscapes, 497–499
Maxwell, J. C. (1831–1879), **255–256,** 258, 262
Meandering streams, 497
Mechanical energy, 107, 247
Mechanical equivalent of heat, 111
Melting point, 123–124, 139–140
Mendelyeev, D. I. (1834–1907), 195, **202–207**
Mercury, 18, 19, **30**
Meson, 291
Mesozoic era, 566–567, 584–590
Metallurgy, 396–399
Metals, 187, **195–196**
  activity of, 327–329, 392
  electron structure of, 324–325, 328–329
  valence of, 201, **329**
Metamorphic rocks, **478,** 482–486
Metamorphism, contact, 527
  dynamothermal, 540–541
  thermal, 527
Meteorites, 34
Meteorology, 450
Meteors, 18, 26, **34,** 36, 82
Methane, 374, 406
Methane series, 405–407
Methyl alcohol, 409
Metric system, 661
Mica, 475, 540
Milky Way, 634–637
Mineral salts, 411
Minerals, **471–476**
Miscibility, of gases, 114
  of liquids, 115
Molecular energy, 132–134
Molecular formulas, 403
Molecular weight, 174, 176
Molecule, **128–129**
  modern ideas of, 177–178
  structure of, 171–173
Momentum, 101
Monsoons, 457–458
Moon, **23–27,** 78
  atmosphere of, 134
  craters of, 24, 26
  eclipse of, 24
  as falling body, 78–80
  motion of, 5–13
  phases of, 23
  surface gravity, 83–84
  tides, 92–95
Moraine, 511, 512
Motion, laws of, 38–40, 42–44, 56–58, 63, 68
Motor effect, 240–241
Moulton, F. R. (1872–), 96, 97
Mountains, formation of, 556–559

N

Natural gas, 373–374, 406, 583
Neap tides, 93
Nebulae, dark, 640–641
  galactic, 640–641
  spectra of, 642, 644–646
  spiral, **641–643**
Nebular hypothesis, 95–96
Negative charge, **214**
Neon, 201
Neptune, **33,** 84, 85
Neutralization, of acids and bases, **354–355,** 355–357, 367
  in electrolysis, 331
Neutron, **290,** 291–294
  in fusion reactions, 298–301
  in nucleus, 291–292
Newton, I. (1642–1727), 12, 37, 45, **74–75,** 76–80, 88, 99, 270
  laws of motion, 38–45, 68–69

Nitrocellulose, 375, 413
Nitrogen, 375, 377–379
  fixed, 377
Nitrogen cycle, 378
Nitroglycerin, 375, 411
Nonelectrolytes, 342
Nonmetals, **195–197**
  activity of, 327–329, 392
  electron structure of, 324–325, 329
  valence of, 201, **329**
Normal fault, 533, 535
North star, 3, 4
Northern lights, 443, 614
Novocaine, 416
Nucleus of atom, **287**, 288–290
  fission of, 298–301
  reactions of, 295–297
  structure of, 291–295

O

Obsidian, 479, 522
Ocean, 444
  deposits in, 513–514
Ocean currents, 467–469
Oersted, H. C. (1777–1851), 238–240, 248
Oil (petroleum), 582-584
Old landscapes, 497–499
Orbits, of comets, 19, 35, 36
  of electrons, 308–311
  of the moon, 24, 79, 94
  planetary, 7–13, 19, 30
  of planetoids, 19
  of satellites, 19
Ore, 396
Organic acids, 410
Organic chemistry, 401–418
  in industry, 414–418
Origin of graphs, 50
Oscillating charge, 257–258
Oxidation, **161, 389–390**
  of food, 396
  of metals, 395
Oxidation-reduction reactions, 389–399
Oxides, 161–162
  acidic and basic, 359–360
  as minerals, 472
Oxygen, **160–162**
  in atmosphere, 372, 395–396
  in atmosphere of Mars, 31

P

Paleozoic era, 567, 573–582
Paracelsus (1493–1541), 150
Parallax, stellar, 21, 621
Pendulum, 106
Periodic law, 204, 288, 323–324
Periodic table, **204–207**, 327–328
  columns, 204
  periods, 204
Periods, geologic, 567
Petrified wood, 564
Petroleum, 373, 405–406, 417
  origin of, 582–584
Phases, of moon, 23, 24
  of Venus, 16, 17
Phenol, 410
Phlogiston, 156–160
Photochemical reaction, 186, 372
Photoelectric effect, 312–314
Photon, **312**, 313–314
Photosynthesis, **372**, 374
Physical properties, 154
Planck's constant, 314
Planetesimal hypothesis, 96–97
Planetoids, 18, 19, 82
Planets, **5**, 16–23, **30–33**
  apparent motion of, 5–6
  atmospheres of, 31, 32, 33, 134–135
  diameters of, 18, 30
  laws of motion of, 13, 75–76
  masses of, 30
  orbits of, 7–13, 19, 106
  revolution of, 19, 30
  rotation of, 19, 30–32
  temperatures of, 31, 32
  shapes of, 81, 82
Plant fossils, 564, 577, 578, 588
Plastics, 416
Playfair, J. (1748–1819), 548–549
Pleistocene epoch, 594–597
Plesiosaurs, 587, 588
Pluto, 18, 19, 30, **33**, 85
Plutonium, 300, 375–376
Polar liquid, 339, 350
Polar molecule, 327, 339–340
Polar regions, 455–456, 459
Polaris, 3, 4
Polymerization, 416
Porcelain, 427–428
Portland cement, 428

Positive charge, **214**
Positron, 290, 291
Potash deposits, 578
Potassium, 200
Potential difference, 245–246
Potential energy, 104–107, 245
  of electrons, 368
  of molecules, 139–140
Pottery, 427–428
Pre-Cambrian, 567, 569–572
Precipitate, 343
Pressure, 117–121
  atmospheric, 119–120
  definition, **118**
  in earth's crust, 81
  in earth's interior, 446
  in fluids, 118
  in gases, 130–131
  units of, 118–120
Priestley, J. (1733–1804), 159–160
Prism, 265–266
Prominences, 28, 29, 614–615
Properties of substances, **145**
Proportionality, 47–60
  constants of, **50**, 56–59
  direct, 47, **49–51**
  inverse, 51, 52
  rules of, 55–56
Protective colloid, 435
Proteins, 377, 411, **413–414**
Proton, **288–289**, 291–293
  in nucleus, 291–292
  in acid-base reactions, 350–353
Proust, J. L. (1755–1826), 163, 164, 167
Pterosaurs, 588, 589
Ptolemaic hypothesis, 6–13, 87
Ptolemy, C. (A.D. 100–170), 7
Pumice, 522
Pyrex glass, 426
Pyroxylin, 413, 417

Q

Quanta, 314
Quantum theory, 311–315
Quartz, 421, **475**
Quartzite, 485, 486
Quaternary period, 590, 593–597

R

Radar, 306
Radiant energy, 107
Radiant energy, photons of, 312
  quanta of, 314
Radiation, 305–316
Radio transmission, 258
Radio waves, 261–262, 305–306
Radioactivity, 272–273, **275–280**
  age of rocks by, 561–562
Radium, **276–277**, 278, 282
Radon, 200, 201, 278–279
Rainbow, 266
Rainfall, causes of, 453–458, 460–463
Raised beaches, 532–533
Ray of light, 263
Reaction rates, 379–382
Red giants, 630
Reduction, 183, **389–390**
  from ores, 397–399
Reflection of light, 263–265
Refraction of light, 264–265
Refrigerator, 140–141
Regional uplift and subsidence, **538–539**
Relative humidity, 453
Relativity, 649–655
  fourth dimension, 653–654
  general theory of, 654
  gravitation, 654–655
  special theory of, 651
  variability of mass and length, 652–653
Reptiles, 581, 586–589
Revolutions, geologic, **566–567**
Rhyolite, 480, 522
Rings of Saturn, 32–33
River valleys, 492–500, 595
Rock crystal, 475
Rocks, age of, 561–562, 567
  chemical analysis of, 481
  classification of, 476–487
  igneous, **478**, 479–481
  metamorphic, **478**, 483–486
  sedimentary, **478**, 481–483
Rocky Mountain revolution, 590, 596
Roentgen, W. K. (1845–1922), 272
Rubber, 415–416
Rubidium, 200
Rumford, Count (Benjamin Thompson), (1753–1814), 110, 133
Russell, H. N. (1877–), 628
Russell's diagram, 628–630
Rutherford, E. (1871–1937), **285–289**

S

Salt, 186
Salts, **354–355**
  solubility, 354
San Andreas fault, 534–535, 537
San Francisco earthquake, 531, 533
Sandstone, 482
Satellites, **18,** 19, 30
Saturated air, 452
Saturated hydrocarbons, 407
Saturated solutions, **338**
Saturn, 18, 19, 30, **32–33,** 81
Schist, 485, 486, 540–541
Scientific method, 85–88
Seasons, 21–23
  on Mars, 31
Second law of motion, **42–44,** 56–58
Sedimentary rocks, **478,** 481–483, 515–516
Sedimentation, 508–516
  by currents, 513–514
  by glaciers, 511–513
  by groundwater, 514–515
  by streams, 510–511
  by wind, 513
Shale, 482, 540
Shells of electrons, 310, 322–324
Shooting stars, 34
Short-wave radio waves, 305, 306
Sierra Nevada, 527, 585, 592
Silica (silicon dioxide), **421,** 423
Silica glass, 421
Silicate minerals, 472, 475–476
Silicates, 420, **421–423**
  structures of, 423–425
Siliceous rocks, 481
Silicic acid, 421, 437
Silicon, **419–421**
Silicon compounds, 419–429
Silver ion, 343, 344, 391
Silver nitrate, 346, 391
Slag, 397–398
Slate, 485, 486, 540–541
Smith, William (1769–1838), 549
Soap, 352, 436
Sodium, **186–187,** 200, 331
Sodium acetate, 352, 356
Sodium bicarbonate, 364–365
Sodium chloride, 186, 346, 354, 367
Sodium hydroxide, 187, 352, 355
Sodium ion, 325, 331–332
Sodium nitrate, 346, 378
Sodium silicate, 421–422
Soil, 492
  color of, 394
Solar system, **17–19,** 90–91
  origin of, 95–98
Solids, amorphous, **137,** 140
  crystalline, **137,** 140
  formulas of, 178
  general properties of, 113, 116
  structure of, 135–137
Solubility, **338–341**
Solute, 337
Solutions, **151–152,** 337
  of electrolytes, 341–347
Solvent, 337
Sorting of sediments, 509
Sound waves, 260
Specific gravity, 146
Spectra, **266–267,** 306–308
  absorption, 266, 307
  bright line, 266
  continuous, 266, 267
  discontinuous, 266, 267
  electronic interpretation of, 308–311
  emission, 306–307
  stellar, 606–611, 626–630, 645–646
  X-ray, 307
Spectroscope, 266, 606
Speed, 40, 48, 49
Spiral nebulae, **641–646**
Spontaneous combustion, 395
Spring tides, 93
Springs, 507
Stahl, G. E. (1660–1734), 150, 156
Stalactites, 364
Star clusters, 637–639
Starches, 413
Stars, 618–632
  apparent brightness of, 622
  apparent motion of, 3–4
  composition of, 608
  densities of, 625
  diameters of, 625
  distances among, 603, 620–623
  energy of, 630–632
  evolution of, 630–632
  giants and dwarfs, 628–630
  intrinsic brightness of, 622

Stars, masses of, 623–624
motions of, 610–611, 625–626
parallax of, 21, 621
spectra of, 606–611, 626–630
in spiral nebulae, 642
structure of, 607
temperatures of, 608, 624
Steam engine, 140–141
Steel, 398
Storms, 461–464
Stratification, 509
Stratosphere, **443**
Stratosphere ballons, 442
Streams, erosion by, 492–496
landscapes produced by, 496–500
meanders of, 497
sedimentation by, 510–511
valleys, 493–496
Striated boulders, 502
Strike-slip fault, 533–534, 536
Strong acids, **350–351**, 353
Strong bases, **352**, 353
Structural formulas, 402–405
Sublimation, 125
Sucrose, 412
Sugars, 412
Sulfates, 189
Sulfides, 185
as minerals, 472
Sulfur, 147, **185–186**
Sulfur dioxide, 185
Sulfuric acid, 185–186
Sun, 18, **27–29, 611–617**
apparent motion of, 5–13
eclipses of, 24, 25, 28, 29
energy from, 108
motion of, 625–626
radiation from, 29
rotation of, 28
size of, 18
source of energy, 615–617
spectrum of, 612
temperature of, 28
Sunspots, 27–28, 612–614
Supersaturated solutions, 339
Surface tension, 116
Symbols, chemical, 176
Syncline, 535, 538
Synthetic ammonia, 386–388
Synthetic rubber, 415–416

T

Telescopes, 16, 604–606
Temperature, **109–110**, 121–122
of atmosphere, 442, 451
effect on reaction rates, 380–381, 385–386
in interior of earth, 446
relation to molecular energies, 131–132
Tertiary period, 590–593
Thales (c. 600 B.C.), 213
Theory, 87
Thermal metamorphism, **527**, 540
Thermocouples, 247
Thermometers, 109–110
Third law of motion, 68, **69**, 71
Thomson, J. J. (1856–1937), 228, 284, 293
Thrust fault, 533, 535
Thunderstorms, 234, 462–463
Tidal friction, 92–95
Tidal hypothesis, 97
Tides, 92–93
Till, 512
Time divisions, geologic, 566–568
table, 567
TNT (trinitrotoluene), 375, 416
Toluene, 407–408
Tornadoes, 461, 464
Trade winds, 455, 456, 459
Transformers, **250–253**
Transmutation, 149, 289
Transverse waves, 260
Trilobites, 576, 578
Troposphere, 443
Tuff, 522
Tyndall effect, 432
Typhoons, 464

U

Ultraviolet radiation, 261, 305, 307
Uncertainty principle, 318–320
Unconformities, **554–556**
Undertow, 505
Uniform change, law of, **550–552**
Unsaturated air, 452
Unsaturated hydrocarbons, 407
Uranium, 272, 275–276, 278–279, 562
fission of, 298–301
isotopes of, 299–300
Uranium-graphite piles, 300–301, 303

Uranus, 18, 19, 30, **33**, 84
Urea, 413

V

Valence, **201–202, 329–330,** 389
  changes of, 389–395
Valence electrons, 330
Valley glaciers, 501, 502–503
Valleys, development of, 493–496
Vapor pressure, **138**
Vaporization, 123, 137–139
Variable stars, **619–620,** 638–639
Vectors, **65,** 66–70, 78–80
Veins, **514–515,** 528–529
Velocity, **40,** 48, 49
  escape, 134
  vector representation of, 67, 70, 79
Venus, 16–20, **30,** 36, 81
Vertebrates, 578, 580
Viscosity, 115
Vitamins, 411, **414**
Volatile liquids, 138
Volcanic bombs, 522
Volcanic breccia, 522
Volcanic rocks, 480, 522
Volcanoes, 518–524
Volt, **246**
Voltmeter, 243
Vulcanism, 518–529
  definition, **518**
  problems of, 528–529

W

Warm front, 460–461
Water gas, 374
Water table, 507
Water witch, 507
Watt, **246–247**
Wave front, 263
Wave length, **261–262**
Waves, 258–262
  electromagnetic, **256–258,** 261–262, 305–306
  erosion by, 505–506
  gamma, 305
  infrared, 305
Waves, interference, 268–270
  light, 261–270, 305
  longitudinal, 260
  radio, 258, 262, 305–306
  relation to particles, 317–320
  sedimentation by, 513–514
  sound, 260
  transverse, 260
  ultraviolet, 305
  water, 259
  X-ray, 262, 273, 275, 305
Weak acids, **350–351,** 355–357
Weak bases, **352,** 355–357
Weather, **450–457,** 461–464
Weather forecasting, 464–467
Weather maps, 465, 466
Weathering, 489–492
  chemical, 489–490
  mechanical, 492
Weight, **38–39,** 43, 58
  on moon, 83, 84
  variations in, 82
Westerlies, 455, 456, 459
Werner, A. G. (1749–1817), 546–547
White dwarfs, 630
Wilson, C. T. R. (1869–), 284–285
Wind, 454–458
  erosion by, 504–505
  sedimentation by, 513
Wöhler, F. (1800–1882), 401
Wood, 373
Work, **103–104**

X

X-ray tube, 273
X rays, 261, **272–275,** 282–283, 305
Xenon, 201

Y

Year, length of, 5
Yeast, 412
Young landscapes, 497–499

Z

Zinc, 147
Zinc sulfide, 147