SCIENTIFIC AMERICAN

THIRD WORLD CITIES

$2.50

April 1985


A HERITAGE OF
PROPRIETOR'S RESERVE
Sebastiani
1980
SONOMA COUNTY
CABERNET SAUVIGNON
SONOMA COUNTY HARVEST FAIR
GOLD MEDAL
For a free collection of
our family recipes, write:
Sebastiani Vineyards, Dept. 1104,
P.O. Box AA, Sonoma, CA 95476
©1985 SEBASTIANI VINEYARDS


WINE AND FOOD
The Rewards
The value of our Italian heritage speaks for itself.
In 1984 national and international wine competitions, Sebastiani Vineyards won 58 medals . . . more awards for more wines than any other winery in America.
As these medals are a reward for our effort to produce wines of elegance, complexity and finesse, so too do they recognize and honor the renaissance of classical winemaking in California. For 81 years and three generations in America, our family has made wine to complement and enhance food.
This is part of our Italian heritage, and it is my respect for this tradition that will always direct our style of winemaking.
As always, the true reward of classical winemaking is the enjoyment of fine wine with food. Write to the address below for a free collection of family recipes created by my mother, Sylvia, and my wife, Vicki.
Discover Sebastiani Vineyards, a family winery in the Sonoma Valley.
Sam J. Sebastiani
Sam J. Sebastiani
BRONZE MEDAL
SILVER MEDAL
SANTA ROSA


COME ALIVE, COME AND DRIVE
NISSAN
MAJOR MOTION
MOVE OVER!
MAKE WAY FOR THE NISSAN 300 ZX.
FUEL-INJECTED • TURBO-CHARGED • V-6 • 200 HORSEPOWER • 3-WAY ADJUSTABLE SHOCKS.
AWESOME!
BELT YOURSELF
AT DATSUN DEALERS
NISSAN

Established 1845 SCIENTIFIC AMERICAN April 1985 Volume 252 Number 4

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.25 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036-8733/84. $1.25 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name ______________________

New Address ______________________

Street ______________________

City ______________________

State and ZIP ______________________

Old Address ______________________

Street ______________________

City ______________________

State and ZIP ______________________

THE COVER

The satellite image on the cover shows Cairo (*reddish patch*) sprawling in the Nile Delta. The image, made from data gathered by *Landsat 2,* has been processed to yield colors resembling those perceived by the eye. Cairo and its environs form Egypt's "core," the region containing most of the country's industry and urban population. Cairo has grown rapidly since World War II. The city, one of the densest urban areas in the world, has more than 10 million residents. Congestion and environmental pollution are so severe that they appear to have inhibited further growth: Cairo is now expanding more slowly than Egypt as a whole. The uncontrolled growth of core regions presents grave problems in many developing nations (see "The Growth of Core Regions in the Third World," by Daniel R. Vining, Jr., page 42).

THE ILLUSTRATIONS

Cover photograph courtesy of © Rand McNally & Co.

Page	Source
20–30	Jerome Kuhl
43	Hiroji Kubota, Magnum Photos, Inc. (*top*); AP/Wide World Photos (*bottom*)
44	Ilil Arbel
45	Richard W. Wilkie and John Montgomery
46	Ilil Arbel
47	Richard W. Wilkie and Thomas Gallagher
48–49	Ilil Arbel
50–51	Ian Worpole
52–53	Andrew Tomko
54	William F. Haxby, Lamont-Doherty Geological Observatory
55	Ian Worpole
56	Andrew Tomko
59	Luis-Miguel Garcia-Segura, Lelio Orci and Yves Dunant
60–61	Bunji Tagawa
62	Maurice Israël, Nicolas Morel, Robert Manaranche and Paule Mastour (*top*); Maurice Israël, Jean Gautron and Bernard Lesbats (*bottom*)
63–64	Bunji Tagawa
65	Dominique Müller, Luis-Miguel Garcia-Segura, Arpad Parducz and Yves Dunant

Page	Source
66	Bunji Tagawa
85	James Kilkelly
86–90	Andrew Christie
91	Brian Wolfe Photography
92–94	Andrew Christie
97	Kenneth R. Dudzik, U.S. Forest Service
98–102	Patricia J. Wynne
103	U.S. Forest Service
106	David Malin, Anglo-Australian Observatory
108	Australian National University
110–114	George V. Kelvin
125	Courtesy of the Boston Medical Library in the Francis A. Countway Library of Medicine
126	University of California at San Francisco
127–130	Ilil Arbel
131	Peter M. Winter
132–133	Science Museum, London
134–135	Hank Iken, Walken Graphics, after Alan Reid
136	Science Museum, London (*top*); Hank Iken, Walken Graphics, after Alan Reid (*bottom*)
137	Hank Iken, Walken Graphics, after Alan Reid
139	Ann Ronan Picture Gallery
141–144	Michael Goodman

Everybody knows the old saying about building a better mousetrap, and nowhere is it more dramatically pertinent than in competing with Japan for world markets.

Consider Motorola's KDT™ portable data communications system. It provides two-way communications between a 28-ounce terminal and a remote host computer. The system uses a digital radio instead of phone lines, and its transmission reliability is actually equal to or better than ordinary phone lines.

Think of the possibilities the KDT opens.

Salespeople will be able to punch in orders directly from the field, and have access to data without looking for phones, leaving messages, or wasting valuable time on "hold."

Service representatives will be able to directly enter data that used to go into mailboxes; such as repair reports, parts used or needed, and billing instructions.

It will also be simple to make detailed inquiries. Data may be summoned as easily as entered. A service technician, for instance, will be able to order up complex repair histories and component availabilities in a few seconds.

The system was designed and developed with a key customer and built by Motorola people to enhance the distribution and processing of data with central computers. And it is far more than a nice little improvement on the old phone-link, AC-powered terminal. This is a major step in a communications revolution—the merging of technologies of the communications and the data processing disciplines.

Digital communication by radio makes far more efficient use of the available channels than voice transmissions ever could. Over a thousand portable terminal users can operate on a single channel with average message traffic. The battery operated KDT terminal shown here contains the two-way radio, an internal antenna, and processing power and memory surpassing that of many personal computers (actually, 160K bytes of ROM and expandable to 80K bytes of RAM). The unit also contains a telephone modem for use over conventional voice-quality phone lines if ever required.

The terminal's dual microprocessor architecture is based on the Motorola 6801 design. One processor is dedicated to the communications function while the other handles the unit's data processing.

The KDT terminal is just part of an extensive support system, of course. And it is representative of one of the oldest and most reliable rules of trade: build a better mousetrap, and the world will beat a path to your door.

With the KDT, the electronic office can be carried in a pocket.

In our view, that's quite a mousetrap.

For more information, call 1-800-367-2346.

MOTOROLA A World Leader in Electronics.

Quality and productivity through employee participation in management.

LETTERS

To the Editors:

Richard P. Turco, Owen B. Toon, Thomas P. Ackerman, James B. Pollack and Carl Sagan (TTAPS) ["Letters," SCIENTIFIC AMERICAN, January] accuse me of "erroneously" expecting a significant greenhouse effect instead of a nuclear winter in my *Wall Street Journal* editorial essay (February 3, 1984). They then state: "Deficiencies in Singer's article were pointed out in a subsequent issue of the *Wall Street Journal* (February 16, 1984)." They fail to inform the reader, however, that the "pointing out" was done not by an independent scientist but in a letter to the editor written by Carl Sagan, who is hardly an unbiased referee.

Not only are TTAPS unfairly disguising the source of criticism by using the passive tense but also they are misrepresenting what I wrote. The "nuclear summer" (as I have called it) is not the expected but simply a possible outcome of plausible assumptions fed into a TTAPS-like model. Whether the warming is more likely than the extreme freeze I cannot say—nor can anyone else, in the absence of reliable data and more appropriate models. I can only state that V. Ramanathan of the National Center for Atmospheric Research (NCAR) has independently pointed to the possibility of surface heating due to a greenhouse effect, and that Robert D. Cess of the State University of New York at Stony Brook and V. Ramaswamy and Jeffrey T. Kiehl of NCAR have papers in publication that show warming is possible.

To quote from my essay: "My purpose is to illustrate the extreme difficulty of making accurate predictions of the global environmental effects of a nuclear exchange. Prof. Sagan's scenario may well be correct, but the range of uncertainty is so great both because of the set of assumptions used and what has been left out in discussing the physics of the situation that the prediction isn't particularly useful."

S. FRED SINGER

George Mason University
Fairfax, Va.

To the Editors:

An experience I had a number of years ago leads me to question the assertion of Philip H. Brownell ["Prey Detection by the Sand Scorpion," SCIENTIFIC AMERICAN, December, 1984] that scorpions occupy a terminal position in the food chain.

While spending a summer near Palmdale, Calif., one day I happened on an old water tower with a foundation built of fieldstone. The lower structure had an entranceway but no door and apparently was not in use.

The circular floor of the structure (about six feet in diameter) was literally covered with scorpion exoskeleton parts. It looked like an arachnid version of the fabled elephant graveyard. I could find no other evidence of animal life (other animal parts, fecal pellets, etc.), although a thorough analysis of the debris was not made.

I remained completely mystified by this discovery until I visited the site several days later after dark and observed several bats flying in and out of the structure's entranceway. I concluded the bats were collecting scorpions in the vicinity of the water tower and using the structure as a feeding station, returning to their home cave at dawn. Unfortunately I was not able to observe bats actually capturing scorpions nor did I have the equipment necessary to photograph the bats feeding.

If I had known that this apparent predator-prey relationship was not well known, I would have taken more care to document it. Perhaps this could be an interesting project if another such feeding site can be found.

ROBERT L. HADDOCK

Department of Public Health
and Social Services
Government of Guam
Agana, Guam

To the Editors:

To my knowledge, Dr. Haddock's discovery of "scorpion Valhalla" is unprecedented, and it certainly gives good reason to ask whether bats are able to capture scorpions through echo location. Interestingly, in the course of our studies of the sand scorpion Roger D. Farley and I noted that scorpions were relatively unresponsive to low-frequency sounds in air, but they were easily aroused into defensive or escape responses when high-frequency sounds (specifically the jingling of a key chain) were presented from a distance of one or two meters. We attributed this sensitivity to very fine hair sensilla (the trichobothria) on the petipalps, but we could not imagine what purpose this sensitivity might serve. Dr. Haddock's observation may provide an important clue.

PHILIP H. BROWNELL

Oregon State University
Corvallis, Ore.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

APRIL, 1935: "Uncle Sam now has underway one of his greatest construction projects, on the Columbia River in central Washington, 92 miles west of Spokane—it is the Grand Coulee Dam. The dam is to be built in two units—a high dam and a low dam. The low dam is exclusively a power development, whereas the high dam will be a combination power, irrigation, flood control, storage and navigation development. The Grand Coulee Dam will be the largest power development possible in North America. The high dam will develop three times the power of Muscle Shoals, 50 percent more power than Boulder Dam or as much as the total installed capacity of Niagara. Army engineers estimate that completion of the Columbia Basin project will add 1,403,000 people to the population of the Pacific Northwest."

"Climbing to altitudes of more than three miles, Army, Navy and commercial pilots will now carry instruments aloft each day from 20 airports to record conditions in the higher air and give U.S. Weather Bureau experts increased data on which to base their forecasts. Each observation pilot has attached to the wing of his plane a meteorograph, an instrument that automatically records humidity, temperature and pressure. In addition the pilot notes the altitudes of the top and bottom of cloud banks, the positions and altitudes of rainstorms that never reach the ground and local disturbances such as thunderstorms and clouds of dust."

"The true stratosphere plane is not yet. Nevertheless, airline operators are looking to higher and higher levels and actually sending their big transports with supercharged motors and controllable pitch propellers into altitudes of 12,000 to 20,000 feet for regular flying. They are doing this for two reasons. The first is economy and regularity of operation; the second, comfort for the passengers. On the first count, it is well known that at about 18,000 feet one is normally above the cloud area and has perfect visibility. Favoring winds can be selected. On the second count, when a high-speed airplane gets into rough air, the impacts of gusts and bumps are much more severely felt than at lower speeds. The passengers are due for a very unpleasant ride. Airlines in this country are now beginning to realize that having sold successfully the idea of speed, they must now give increased attention to the idea of a smooth ride."

SCIENTIFIC AMERICAN

APRIL, 1885: "Astronomy has made rapid strides in recent years, and a great portion of the best work is effected by the aid of photography. One eminent astronomer has gone so far as to declare that we no longer require observers of the heavens, and that their place can be better supplied by the gelatine plate of the photographer. A most important branch of astronomy, which is receiving great attention at the present, is the mapping of the starry heavens; and herein photography will perhaps do its best work for astronomers. The trial star map by the brothers Henry, of a portion of the Milky Way which they felt unable to observe satisfactorily by the ordinary methods, is so near perfection that it alone proves the immense superiority of the photographic method."

"In Paris the new extension of the pneumatic postal system is now complete and in operation, ramifying under the streets in every direction. By payment of fifteen cents, letters may be instantly sent by tube to any part of the city. These tubes are a great convenience to the public, and in time will no doubt become an adjunct of the postal service of all large cities."

"Few people not actually engaged in the metal trade are aware of the wonderful strides made by steel in recent years. It is wholly taking the place of wrought iron. Steel is simply a mixture of iron and carbon, the quantity of carbon ranging from .25 to .02 per cent of the mass. It is not only stronger and for almost every purpose better than wrought iron but also cheaper."

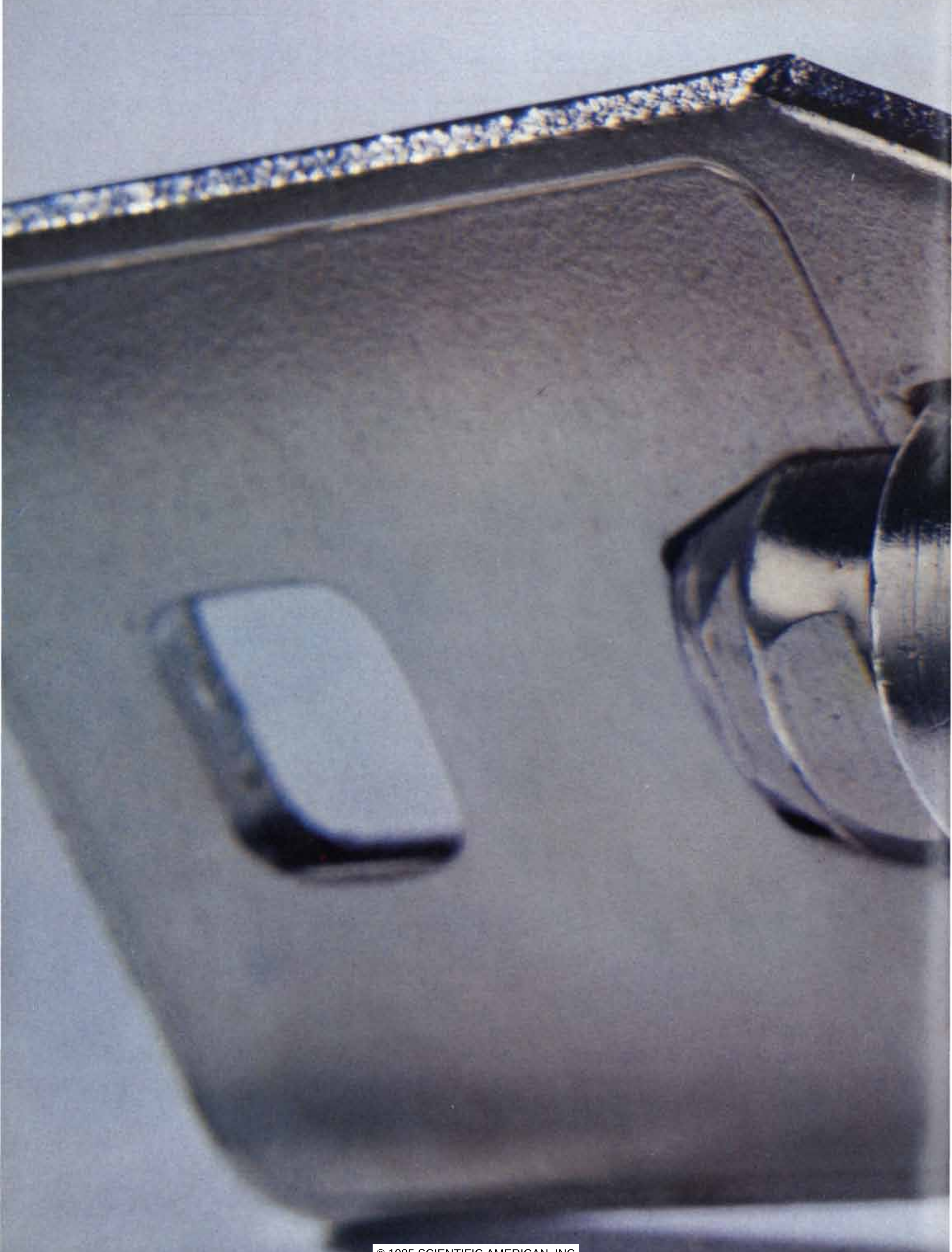
"The census of 1880 gives the number of persons engaged in gainful occupations as 17,392,000, or 47.31 per cent of total persons over 10 years old. These were engaged in the four chief lines of occupation as follows: agriculture, 7,670,000; professional and personal services, 4,074,000; trade and transportation, 1,810,000; manufacturing, mechanical and mining industries, 3,837,000. Of the total 2,647,000 were women. The number of persons over age 10 is 36,761,000, leaving 19,369,000 unaccounted for. That about equals the number in school or physically incapable of labor."

"With the recent improvements in materials and apparatus for photographing, there has come a great accession to the ranks of those who, in all parts of the country, find in this interesting study a pleasurable, inexpensive and sometimes lucrative employment. As is the case, however, in almost every wide-embracing field of activity, there is no noticeable success attained except by those who make diligent and intelligent application, and this is particularly true with the large number of amateur photographers, who find it so easy to learn the principal elements of what is necessary to make sun pictures before they realize how important it is to have also some artistic taste and education, and to learn the nicety of manipulation required in a thousand delicate details which the successful photographer must carefully attend to."

"Take a Derby hat and close the ventilating apertures at the sides, if it have any, and remove the wire gauze from the ventilator in the top. Cut an oval piece out of a sheet of tracing cloth or translucent paper and fix it to the brim by means of drawing pins. Next, having provided yourself with a cloak or something of the kind as a photographic veil, go to the window and point your objective (the ventilator) at any brilliantly lighted object. If your head be enclosed in the improvised veil in such a way that your hat is also surrounded as completely as possible by its folds, you will see a reversed and reduced image of the object appear upon the screen. In a word, you will have a practical apparatus which is analogous to the camera obscura."

A camera obscura made from a derby


"Jeep designed the Grand Wagoneer like I design every ocean racer that comes off my drawing board...to be a winner."
BRUCE NELSON, PRESIDENT, NELSON/MAREK YACHT DESIGN
Bruce Nelson designs and sails some of the world's most impressive ocean racers. Not surprisingly, he also owns one of the world's most impressive vehicles. The Jeep® Grand Wagoneer.
Like his world-renowned racing yachts, Nelson's Grand Wagoneer was engineered to perform brilliantly through thick and thin. For when the going is rough he can "shift-on-the-fly" into the sure-footed traction of 4-wheel drive. But while the six-passenger Grand Wagoneer thrives on adversity, it also pampers its passengers with luxury, prestige and comfort.
And thanks to its new suspension system, the Grand Wagoneer provides its smoothest ride ever. Jeep Grand Wagoneer, the wagon designed and engineered to win you over.
Jeep Grand Wagoneer
The Ultimate Wagon
AVAILABLE FOR PURCHASE OR LEASE AT PARTICIPATING JEEP DEALERS
SAFETY BELTS SAVE LIVES




CLICK.
How do you give information access to some and deny others?
That's the secret that makes SCOMP click. This data security technology
produced by Honeywell is the first computer system ever awarded
the top A1 security rating by the Department of Defense.
Put a lock on sensitive data as strong as the Pentagon's. Talk to us.
For a white paper on our data security technology, write Security,
c/o Honeywell, Honeywell Plaza, Minneapolis, MN 55408.
Or call 1-800-328-5111, ext. 2760.
Together, we can find the answers.
Honeywell

THE AUTHORS

DANIEL R. VINING, JR. ("The Growth of Core Regions in the Third World"), is associate professor of regional science at the University of Pennsylvania, where he is also a member of the Population Studies Center. He did his undergraduate work at Yale University, getting a B.A. in 1966. He earned a master's degree in 1971 at Princeton University and a Ph.D. in 1974 from Carnegie-Mellon University. Since 1974 he has been on the faculty at Pennsylvania. Among Vining's interests are the almost universally low fertility of wealthy populations such as those of western Europe, North America, Japan and Singapore, the pace of immigration from regions of high population growth to regions of low population growth, and global scarcities of resources.

GREGORY E. VINK, W. JASON MORGAN and PETER R. VOGT ("The Earth's Hot Spots") specialize in marine geophysics. Vink is a guest investigator at the U.S. Naval Research Laboratory in Washington, D.C., with a postdoctoral fellowship from the National Research Council of the National Academy of Sciences, Morgan is professor of geophysics at Princeton University and Vogt is a staff geophysicist at the Naval Research Laboratory. Vink got his B.A. at Colgate University in 1979 and his Ph.D. from Princeton in 1983; his dissertation described the tectonic evolution of the Norwegian-Greenland Sea and the Arctic Ocean. Morgan went to the Georgia Institute of Technology as an undergraduate and received a Ph.D. in physics from Princeton in 1964, where he has remained. Vogt was educated at the California Institute of Technology and the University of Wisconsin at Madison, which granted him a Ph.D. in oceanography in 1968. He worked in the U.S. Naval Oceanographic Office until 1975 and then joined the Naval Research Laboratory; he holds a concurrent post at the University of Oslo. In 1975 Vogt served as co–chief scientist on the *Glomar Challenger* deep-sea drilling expedition to the North Atlantic.

YVES DUNANT and MAURICE ISRAËL ("The Release of Acetylcholine") are respectively professor of pharmacology on the faculty of medicine of the University of Geneva and director of research at the National Center for Scientific Research (CNRS) at Gif-sur-Yvette, France. They have been occasional collaborators since 1969, when they met at a symposium on acetylcholine release. Dunant studied medicine as an undergraduate at the universities of Lausanne, Paris and Geneva; in 1961 he got a diploma in tropical medicine at the University of Basel. After practicing medicine in Kinshasa, Zaire, he returned to Europe and began to study the transmission of nerve impulses across synapses, the focus of his work since then. Israël earned degrees in medicine and biochemistry from the University of Paris in 1963. He received his doctorate from the University of Paris in 1969 for work on cholinergic vesicles in the electric organ of the torpedo fish and on the localization of cholinergic enzymes in neuromuscular synapses.

CHRIS QUIGG ("Elementary Particles and Forces") is head of the theoretical physics department at the Fermi National Accelerator Laboratory (Fermilab) in Batavia, Ill.; he also teaches physics as a professor at the University of Chicago. He is a graduate of Yale University, where he got a B.S. in physics in 1966, and of the University of California at Berkeley, where he received his Ph.D. in 1970. Until 1974 he was at the Institute for Theoretical Physics of the State University of New York at Stony Brook, first as a research associate and then as a member of the faculty. In that year he moved to Fermilab, where in 1977 he took up his present post. While he was at the École Normale Supérieure in Paris from 1981 to 1982 as a visiting professor, Quigg wrote a textbook on gauge theories.

ALEX L. SHIGO ("Compartmentalization of Decay in Trees") is chief scientist at the U.S. Department of Agriculture Forest Service's Northeastern Forest Experiment Station in Durham, N.H. He holds an undergraduate degree in biology from Waynesburg College in Pennsylvania and M.S. and Ph.D. degrees in plant pathology from West Virginia University, which he received in 1958 and 1959 respectively. Since 1959 he has worked for the Forest Service. Shigo's major area of research has been the decay of trees and its ramifications: how it occurs, how to detect it, how it can be prevented and how to grow trees that resist it.

DON MATHEWSON ("The Clouds of Magellan") is professor of astronomy at the Australian National University in Canberra and director of the Mount Stromlo and Siding Spring

Eighth century sutra — Private Collection, Fogg Art Museum, Harvard University

Observatories. He holds a bachelor's degree from the University of Queensland and a Ph.D. from the University of Manchester. In 1955, after completing his undergraduate education, he joined the division of radiophysics of the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Sydney. He remained there until 1966, when he moved to Mount Stromlo. Mathewson has done considerable work in radio astronomy abroad: in England from 1958 to 1960 as a doctoral student, working with the 250-foot radio telescope at Jodrell Bank, at Ohio State University from 1965 to 1966 and at the Leiden Observatory in Holland from 1970 to 1971, in both cases as visiting professor.

PETER M. WINTER and JOHN N. MILLER ("Anesthesiology") are respectively chairman of the department of anesthesiology and critical-care medicine at the University of Pittsburgh School of Medicine and chairman of the department of anesthesiology at the University of South Alabama College of Medicine. Winter did his undergraduate work at Cornell University; he earned his M.D. at the University of Rochester School of Medicine and Dentistry in 1958. He trained in anesthesiology at the Massachusetts General Hospital. Following three years of physiology research at the State University of New York at Buffalo, Winter joined the faculty at the University of Washington, where he remained for 10 years before taking his present post. Miller received his medical degree from the University of Sydney in 1963; he did his residency in anesthesiology at the University of Washington. After finishing his training he taught at Duke University and then moved to the University of South Alabama. Among Miller's research interests is the physiology of diving.

W. GARRETT SCAIFE ("The Parsons Steam Turbine") is associate professor of engineering at Trinity College of the University of Dublin. After getting an engineering degree in 1951 at the University of London he spent six years engaged in power-station design and operation with the Irish Electricity Supply Board. He then worked as a design engineer for a manufacturer of electrical equipment. In 1962 he returned to academics at Trinity College's school of engineering, where he began teaching thermodynamics. He was granted a doctorate at the University of Dublin in 1967 for research into the behavior of liquid dielectrics under high pressure, a topic that remains the focus of his research. Scaife attained his present position in 1977.

COMPUTER RECREATIONS

Five easy pieces for a do loop and random-number generator

by A. K. Dewdney

How might a cannon be used to measure the area of a pond in a field? Fire the cannon at the field in a great many directions in such a way that the shots land at random places in the field. The area of the pond is approximately the area of the field multiplied by the number of splashes and then divided by the number of cannonballs fired. This admittedly silly problem nicely illustrates the role random events can play in the simulation of a process. The modern computer is a particularly useful tool for conducting a simulation. In a computer randomness comes not from a cannon but from a random-number generator.

Philosophically it is curious that random numbers, which are the purest expression of our ignorance, should underlie our ability to gain new knowledge of complex systems. I shall illustrate this theme in a series of programs I call five easy pieces: random numbers are used to find an approximate value for π, to simulate the arrival times of people at a bank, to investigate a distribution of marbles, to generate voting patterns and to show how millions of people can be made to wait in a queue.

I offer the five pieces in a musical spirit. Learning to play a computer is much like learning to play a musical instrument, and it can be enjoyed at every level of proficiency. Accordingly the first piece is the easiest; the last piece is probably not difficult. In all that follows I shall assume that random numbers can be obtained with the command *random*. The numbers are to be decimals between 0 and 1 that are several digits long.

The first piece is related to a problem called Buffon's needle, after Count Louis de Buffon, a 19th-century French naturalist and mathematician. Imagine that a needle is thrown randomly onto a planked floor a great many times. The length of the needle is half the width of the planks. The problem is to find the probability that the needle will land across a crack. The answer turns out to be $1/\pi$.

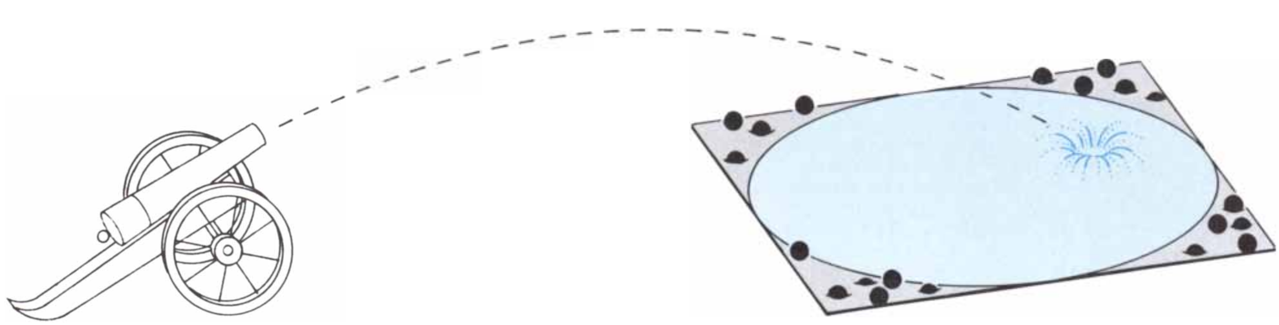
It is certainly possible to write a program that simulates the tosses of Buffon's needle, but there is a much simpler way to approximate the value of π by simulation. The technique borrows from the idea of firing a cannon at random into a field. Imagine a square field enclosing a circular pond that touches all four sides of the square. If the cannon is fired randomly into the field a great many times, the proportion of shots landing in the pond approximates the ratio of the area of the circle to the area of the square, namely $\pi/4$.

The simulation can be done with a simple program called PINT. Represent one quadrant of the field, say the northeast quadrant, by a unit square. The point (0,0) at the southwest corner of the quadrant lies at the center of the pond, and the point (1,1) at the northeast corner of the quadrant is a corner of the field. By choosing two random numbers x and y in succession, one can simulate a random shot that falls in the quadrant. (Remember, each random number is a decimal between 0 and 1.) Does the shot land in the pond? To find out, calculate the distance between the point (x,y) and the center of the pond. If the distance is 1 or less, record a splash and increase a counting variable by one unit. A loop in the program enables one to repeat the experiment. When all the shooting is over, multiply the resulting ratio by 4; the result should be close to π.

If enough readers try this computation, I shall tabulate the results. Send me a run of 1,000 shots and I shall compute a grand average. Perhaps we shall achieve a better value of π together than we could individually.

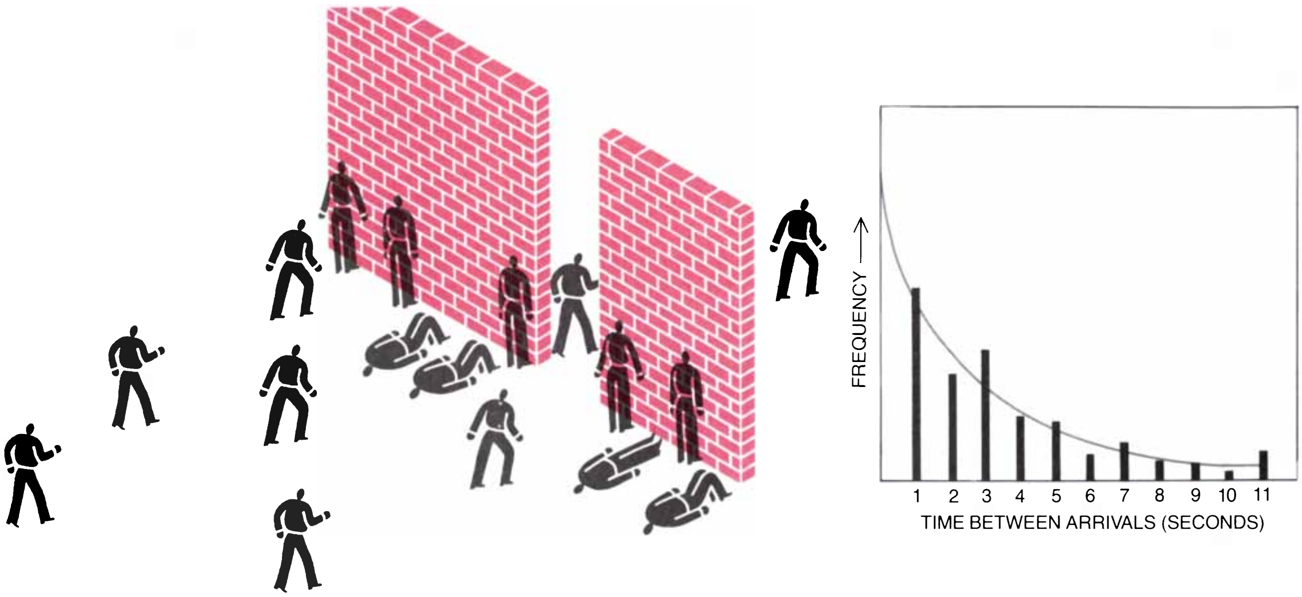
The next piece exploits a technique called the zombie door. Imagine watching people coming through the door of a bank. All of them have good reasons for arriving when they do, and if one knew these reasons, one could no doubt simulate the arrival times exactly. Such knowledge, however, is presumably out of reach. To achieve the same end many simulation programs apply a special mathematical function called the negative exponential distribution. It would take a page to explain both the distribution and a rather pretty method for getting arrival times out of it. The zombie door is much easier to explain; it also gives natural-looking results.

The zombie door is a gap of width w in an otherwise impenetrable wall. Thousands of zombies march steadily toward the wall, and every second a zombie arrives at a random point on it. The lucky ones arrive at the gap and march on through. The unlucky ones run into the wall. If the gap in the wall leads to a bank, the random arrival times of the zombies look much like the arrival times of ordinary patrons. To adjust the average arrival rate the



The area of a circular pond determined by random cannon fire

Fly First Class.
Wild Turkey. It's not the best because it's expensive.
It's expensive because it's the best.
Now you can send a gift of Wild Turkey*/101 Proof anywhere* by phone through Nationwide Gift Liquor. Call Toll Free 1-800-CHEER-UP (Arizona 602-957-4923). *Except where prohibited. Major credit cards accepted. Austin, Nichols Distilling Co., Lawrenceburg, KY © 1985.
Austin Nichols
WILD
TURKEY
KENTUCKY STRAIGHT BOURBON
WHISKEY
101 PROOF
8
YEARS OLD



Zombies and the pattern of their arrivals on the far side of a wall

width of the gap can be set to an appropriate value.

My arrival-generating program is called ZOMBIE. The wall in the program is a line one unit long. A loop called a while loop generates the times between arrivals: for each second that no zombie arrives in the gap the while loop increases a counting variable by 1 and calls on the command *random* to bring a new zombie to some random point along the wall. It is convenient and does no harm to assume the gap extends from the leftmost point 0 to the point *w*. If the random number, or position of the zombie, lies in this range, the program exits from the loop and declares in effect that a zombie has passed through the gap. Otherwise the program reenters the loop. Sooner or later a zombie passes through the gap; the value then stored by the counter tells how many seconds have elapsed since the last arrival.

The average time between arrivals is $1/w$ seconds. For example, if w is $1/10$, a zombie will arrive on the average every $1/(1/10)$, or 10, seconds. ZOMBIE works best when *w* is fairly small. For this reason it might be appropriate (depending on the application) to assume that zombies are arriving at the wall every tenth of a second. To find the value of *w* that gives a good simulation, set the desired average of the time between arrivals at the bank equal to $1/10w$ and solve for *w*. I stated above that the door could extend from 0 to *w*, but some random-number generators may be faulty in this range. In this case choose another gap in the wall, say between .5 and $.5 + w$.

It is interesting to examine the distribution of arrival times generated by ZOMBIE. By embedding the while loop in another loop that counts the number of occurrences of each possible time interval between arrivals, one can develop a histogram, or bar chart, that plots the distribution. A typical histogram is shown in the illustration above; the negative-exponential distribution has been superposed on it for comparison. There are many potential applications of the zombie-door technique for video games, but I shall limit myself to a single application when I discuss the fifth piece.

The third simulation involves a Galton board. This device, named after the pioneering Victorian statistician Sir Francis Galton, is a sloping surface studded with a triangular array of pegs [*see top illustration on page 26*]. Marbles released just above the top peg make their way, helter-skelter, down through the array. Below the pegs are channels that collect the marbles. When the last marble falls into place, the collected columns take on a characteristic shape, not like the New York skyline but something quite different. Those attempting this piece will discover the shape for themselves.

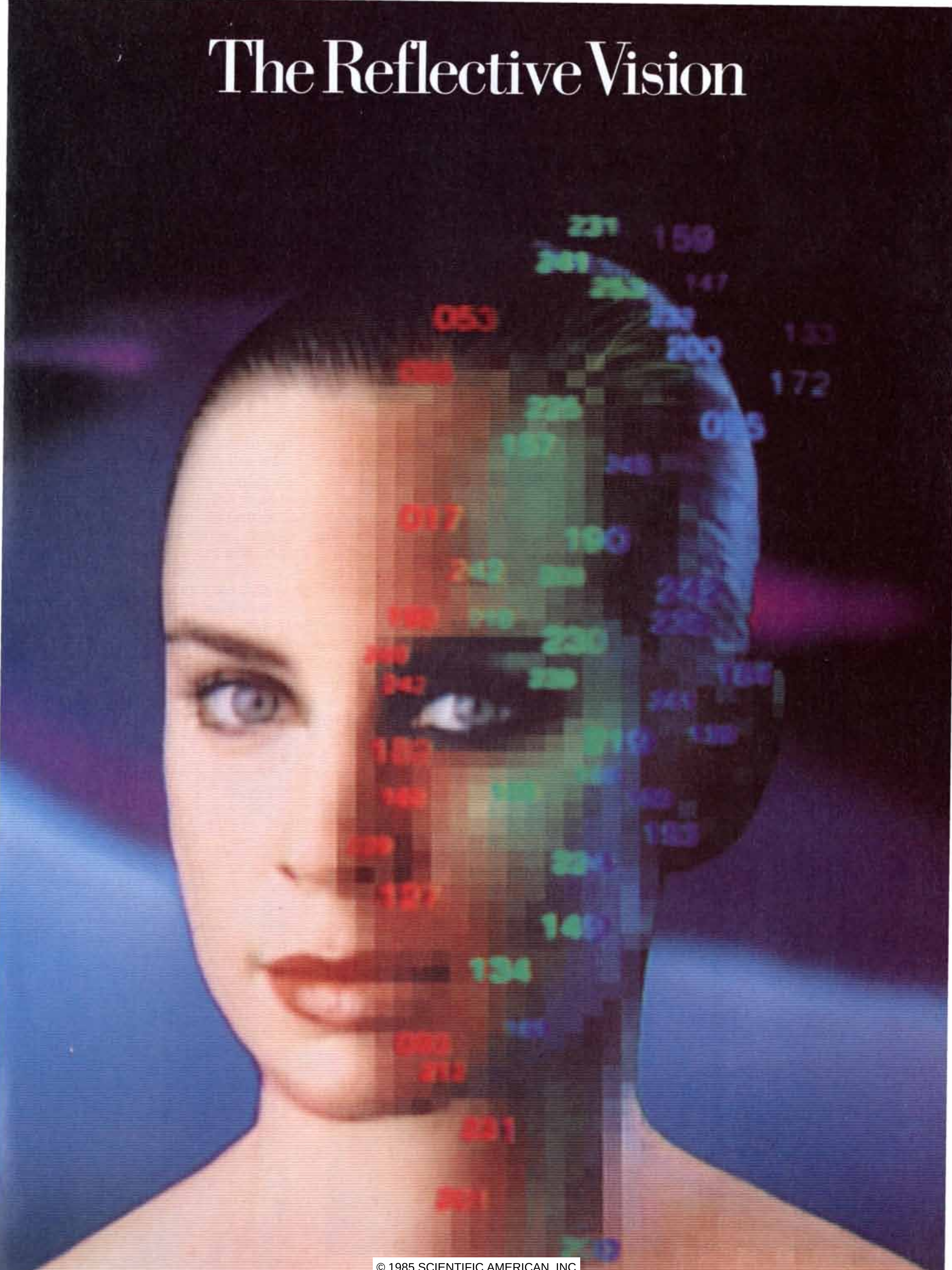
In the absence of a real Galton board a program called GALTON can simulate one, and it will use up only a few lines in the process. The idea for the program, as well as for the first piece on π, comes from Jon L. Bentley's column "Programming Pearls" in the January 1984 issue of *Communications of the ACM.*

The descent of a single marble is simulated in a loop that chooses a succession of random numbers. Each random number determines whether the marble rolls to the left or the right of a peg. If the number is less than or equal to .5, the simulated marble rolls to the left of a peg. If the number is greater than .5, the marble rolls to the right. Which peg? It does not matter. One must merely be sure to increase a counting variable by one unit each time the marble moves to the right. Readers may confirm for themselves that the column in which a marble comes to rest is determined solely by the number of rightward bounces it makes on its descent. At one swoop the need for an array that simulates the individual pegs is done away with. The moral of this observation should be clear: A bit of analysis can be worth a megabyte of program.

The loop that simulates the descent of a single marble through the Galton board is embedded in a larger loop specifying the number of marbles to be dropped. Each time the inner loop is completed the column in which a marble arrives should be recorded in an array *c* representing the columns. The number $c(n)$ is increased by 1 every time a marble falls into column *n*. When a run of, say, 1,000 marbles is complete, the array *c* can be plotted by hand as a histogram. Of course, it can also be plotted directly by the computer if one takes the trouble to add some display commands to GALTON.

It seems to me that by arranging the pegs differently other kinds of distribution might be generated by a Galton board. I would be grateful to hear of any such devices, whether they exist in reality or only in imagination.

The fourth easy piece concerns two new and fascinating voting games recently studied by Peter Donnelly of


The Reflective Vision

The Reflective Vision

A highly advanced design tool developed at the General Motors Research Laboratories uses computers to generate visual images from mathematical data with such accuracy that, soon, in-depth aesthetic evaluations of new concepts may be made prior to creating a costly physical model.

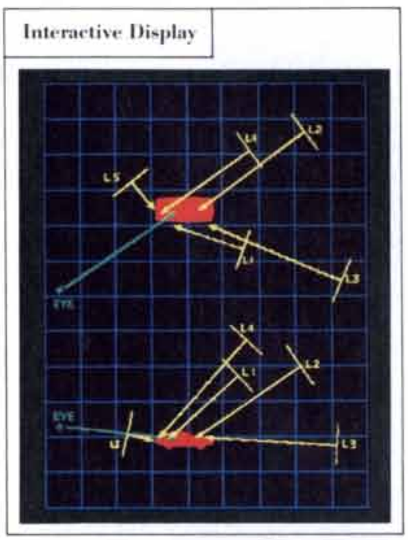
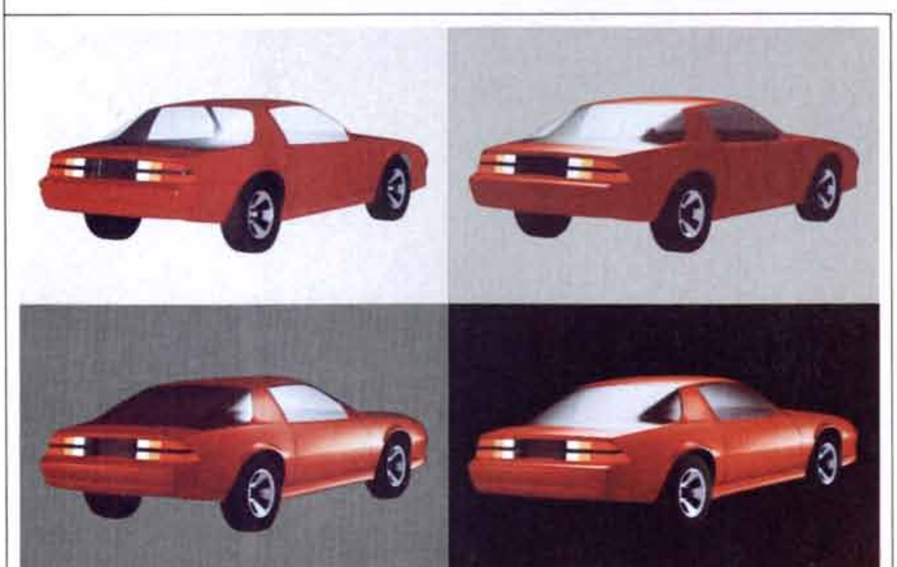


Figure 1: Computer display of plan view (upper) and side elevation (lower), indicating automobile location (red), lighting selections (L1-L5), and viewing position (EYE).

Figure 2: Four Autocolor images, showing the same view of an automobile as background and lighting change.



WITH AUTOCOLOR, users can synthesize three-dimensional, shaded images of design concepts on a color display and then quickly explore how major or minor changes affect the overall aesthetic impression. The system is completely interactive. By choosing from a menu on the screen, the designer can redefine display parameters, select a viewing orientation, or mix a color. Each part of an object can be assigned a surface type with associated color and reflectance properties. Built-in lighting controls generate realistic "highlights" on simulated surfaces composed of differing materials.

Before developing the system, David Warn, a computer scientist at the General Motors Research Laboratories, observed the complex lighting effects achieved in the studio of a professional photographer. By simulating these effects, Autocolor can produce results unattainable by conventional synthetic image display systems. Previous systems used a point source model of light, which allows adjustments only in position and brightness.

The versatility of the lighting controls constitutes a major advance in Autocolor. An unlimited number of light sources can be independently aimed at an object and the light concentration adjusted to simulate spotlight and floodlight effects. The lighting model even includes the large flaps or "barndoors" found on studio lights. These comprehensive controls permit the user to view the simulation in studio lighting conditions, as well as to make revisions in color, paint type, and materials.

With real lights, direction and concentration are produced by reflectors, lenses, and housings. It would be possible to model these components directly, but that would introduce considerable overhead to the lighting computation. Instead of modeling individual causes, Autocolor models the overall effect, reducing complexity by simulating those aspects needed to produce realistic results.

Autocolor approximates the geometric shape of an object with a mesh of three or four-sided polygons. These polygons are grouped to form parts. For a car body, there might be separate parts for the door, hood, roof, fender, and so on. Each part is assigned a surface type, such as painted metal or glass, and each type of surface has associated color and reflectance properties. The entire data structure is stored in tables using an interactive relational data base developed at the GM Research Laboratories.

THE LIGHTING model determines the intensity of the reflected light that reaches the eye from a given point on the object. It takes into account the reflectance properties of the surface as well as the physics of light reflection. A hidden surface algorithm determines which point on the object is visible at each point on the display. For each of these visible points, the intensity is computed for each light source. The displayed intensity is the sum of the contributions from all the lights plus an ambient term which indicates the general level of illumination.

Using the point source lights of conventional image generation systems, highlighting a particular area of an object can be a difficult task and can result in unwanted highlights in other areas. By contrast, the light direction and concentration controls found in Autocolor make it possible to isolate the effect of a light to a particular area, and achieve a desired highlight easily and quickly (see Figure 2). This is not because Autocolor's lighting model computations are faster, but because its controlled "lights" behave in a more natural way.

Another unique feature of Autocolor is the ability to portray realistically a variety of different materials and lighting conditions. The color seen from a surface is really a combination of two colors: the color of the surface or material itself (diffuse reflection) and the color of the reflected highlights (specular reflection). The highlight color may be the color of the material, the color of the light, or a color derived from the material and the light.

A different highlight color can be used for each different surface type that is defined. This makes it possible to simulate materials such as plastic, painted metal, and chrome—each of which has different reflectance properties and requires a different highlight color.

The user can interactively adjust the blending of the surface and highlight colors, watching the image change dynamically on the screen until a desired effect is achieved.

"Autocolor will free designers to be more creative," says researcher Warn. "Our goal is to move from controls that show changes in lighting, color, and materials, to software that will let the user change the actual shape, manipulating the image on the screen like a flexible clay model."

General Motors

GM MARK OF EXCELLENCE

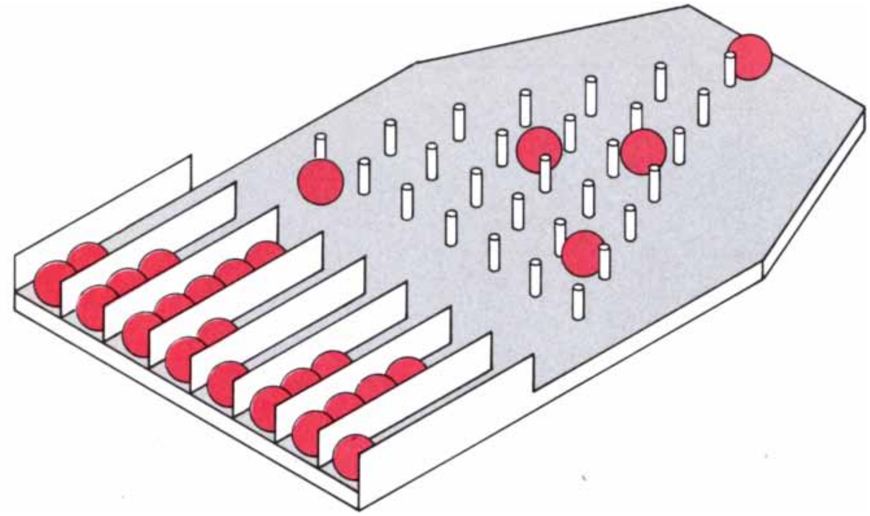
THE MAN BEHIND THE WORK

David Warn is a Senior Staff Research Scientist in the Computer Science Department at the General Motors Research Laboratories.

He received his undergraduate degree in mathematics from Carnegie-Mellon University, and his M.S. in computer science from Purdue.

He has done extensive research in relational data management systems with special emphasis on user interfaces and human factors. He also designed the prototype for the network data manager used in the GM Corporate Graphic System. His previous work on other aspects of computer-aided design include system design, file management, and simulation models.

His foremost research interests are in color synthetic image generation and interactive surface design. He joined General Motors in 1968.

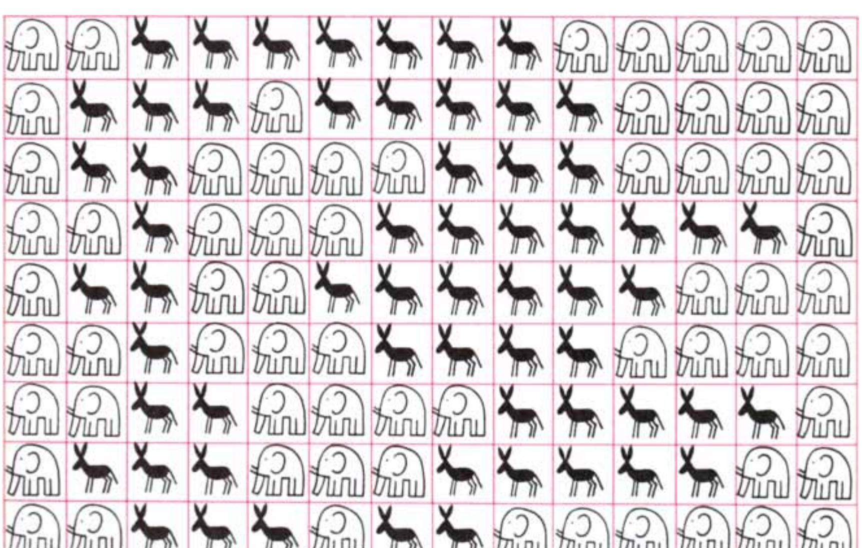


The Galton board and one improbable distribution of the marbles rolling down it

the University College of Swansea in Wales and Dominic Welsh of the University of Oxford. The squares of a rectangular grid are initially colored black and white in a random manner. Each color is supposed to reflect the political opinion of a person inhabiting that square. One color might represent a Democrat, the other a Republican. With each tick of a clock a voter is selected at random and his political opinion becomes subject to change: one of the voter's eight neighbors is selected at random and the voter's political persuasion becomes that of his neighbor, regardless of earlier belief.

The grid is wrapped around itself in the same way as Wa-Tor's ocean was [see "Computer Recreations"; SCIENTIFIC AMERICAN, December, 1984]: the top edge of the grid is identified with the bottom, and the left edge is identified with the right. Hence a voter inhabiting a square on one side of the grid has three neighbors on the opposite side. After each random change in voter preferences an election is held, but this element does not need to enter the simulation unless one wants to keep close track of the results.

As this admittedly simpleminded model of the political process is run, strange things happen. First, large blocks of votes develop within the grid. The blocks are geographic areas where everyone has the same opinion. Then the blocks migrate around the grid, and for a while two blocks struggle for dominance. Finally the two-party system collapses as everyone ends up voting the same way. The smaller block vanishes as democracy votes itself out of existence—or does it?



The voting game in progress

This is a neat philosophical question.

A short program called VOTERS can embody this model in every essential respect. In VOTERS there is an array having the same dimensions as the grid. A short computational cycle selects three random integers that give a random row index, a random column index and a random neighbor code for each tick of the clock. The three integers are created by multiplying the output of the command *random* by a constant and then removing the fractional part with the command *integer*. For example, if the grid has 50 rows, the choice of a random row index i could be written algorithmically as

$$i \leftarrow \text{integer}\,(50 * \text{random})\ .$$

The random column index j is computed in a similar fashion. The neighbor of the square (i,j) chosen by the program depends on which of the integers 0 through 7 is randomly selected. Once a random square and a random neighbor are selected, VOTERS replaces the array value at (i,j) with the one at the neighboring square.

The preceding steps are embedded in a loop. It is convenient to allow the index limit of the loop to be a variable: experimeters may want to vary the number of opinion changes that voters undergo. The index limit can also be varied in the other programs I describe in this article: set the outer loop limit to n and then enter the desired value of n from the keyboard.

A graphic display of the voter grid is easy to incorporate into the primary computational cycle in VOTERS. With only two characters (say the period and the asterisk) the effect is striking. Readers not immediately put off by the totalitarian phenomenon inherent in the voting game might wish to implement what Donnelly and Welsh call the antivoting game. In this game the randomly selected voter adopts an opinion opposite to that of the randomly selected neighbor. Will democracy survive in this setting?

The fifth and final piece provides food for thought while one waits in a queue. If people leave the head of the queue as fast, on the average, as they arrive at the end, would the queue not remain more or less the same size? Apparently not. According to queuing theory, such a queue has no finite expected length. I have seen such queues in banks.

The program called QWING simulates a simple queue. People arrive at the end of the queue in accordance with the pattern generated by the ZOMBIE program I described above. At the head of the queue is a person receiving or trying to receive service. Simulation programs sometimes incorporate the assumption that service times are distributed the same way as the times between the random arrivals of two zombies. Service is short when I make a deposit, but it takes a long time when I argue with the teller about the disappearance of money from my account. Accordingly ZOMBIE is also called on in QWING to generate service times.

QWING handles the passage of time by the critical-event method. A simulated clock *c* is continuously updated to the time of the next event, whether it is an arrival or a departure. Three other variables are needed: *ts,* the time to the completion of the current service; *ta,* the time until the next arrival, and *q,* the length of the queue [*see illustration on this page*].

Initially *ts* and *ta* are set to 0 and *q* is set to 1. QWING then enters its main loop, and the number of iterations of the main loop is specified by a command from the keyboard. Within the loop is a branching instruction that tests whether *ta* is less than *ts.* If it is, the time on the clock is increased by the amount *ta* because the next arrival at the queue is the next critical event. Accordingly *ts* is decreased by the same amount of time, and since the critical event is an arrival, *q* is increased by 1. As a final step ZOMBIE swings into action to generate the time of the next arrival.

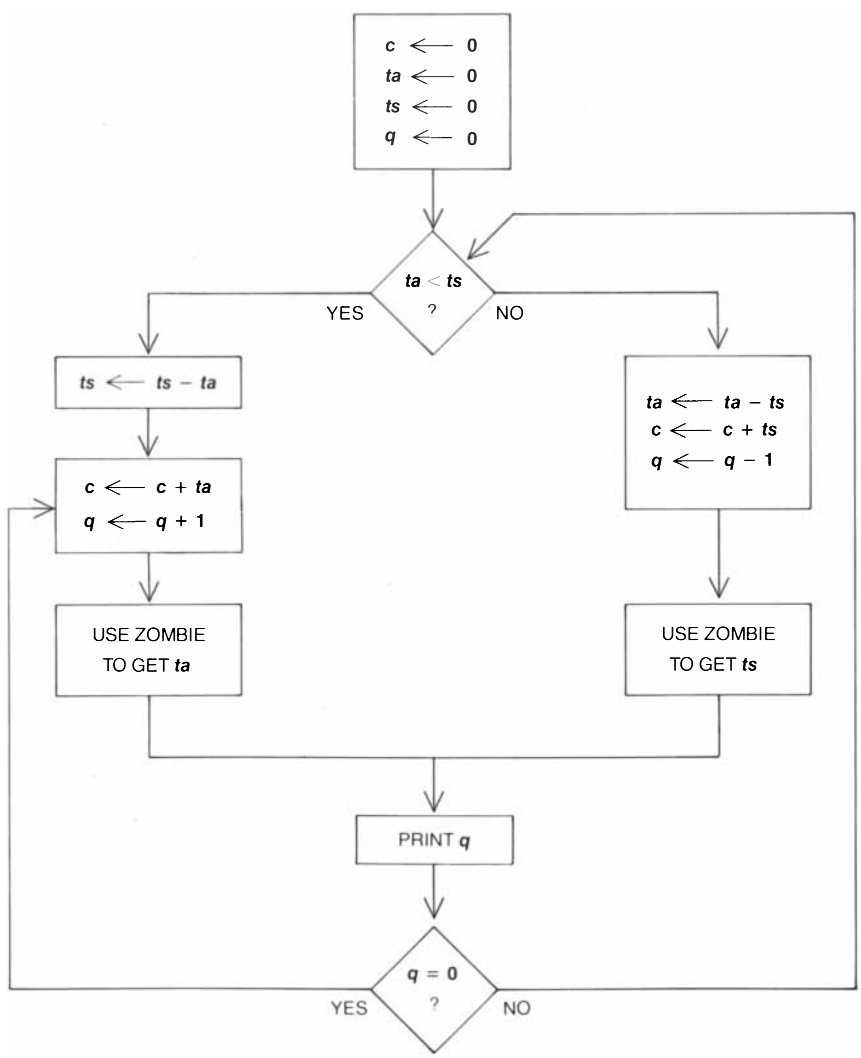
At the head of the loop it may turn out that *ta* is not less than *ts.* In this case the same operations are carried out in another branch of the program, but the roles of *ta* and *ts* are reversed. The next critical event is a new service and so *q* is decreased by 1. In this branch ZOMBIE generates the next value of *ts.* At the bottom of the loop both branches rejoin with a test carried out on *q.* If *q* is equal to 0, a new arrival is needed and the program branches to the first segment. Otherwise it branches back to the head of the loop, where *ta* is again tested against *ts.*

The QWING program is clearly longer than the others, but it is still short enough to qualify for the term "easy." It is probably a good idea to make the average time between arrivals and the average service times specifiable by the user. In this way one can verify the basic results of queuing theory. When the average time between arrivals is greater than the average service time, the length of the queue is expected to be finite; the greater the difference between the two averages is, the more "finite" the length becomes. When the averages bear the opposite relation, the expected length of the queue is infinite; watch it grow. When the averages are exactly equal, the expected length of the queue is neither finite nor infinite! How can this be? A few hours of watching may answer the question—or a few minutes of thought.

Adventuresome readers can extend QWING to simulate the effects of an innovation introduced in banks more than a decade ago. Instead of waiting in individual queues, one for each teller, bank customers were sent to a single queue. The system has several advantages, which can be verified by a simulation program that captures the distinctive features of the collective queue. An even more elaborate program could model a relatively new feature in banks: now even the tellers queue up at their own superteller. The new system suggests some awesome possibilities for really large banks.

All this talk about random numbers poses a final question: How can a computer, which is an explicitly deterministic machine, generate random numbers, which are implicitly nondeterministic? The answer is that it cannot. But a computer can generate numbers that *look* random: pick any number, multiply it by *m,* add *k* and take the remainder after division by *p.* The resulting number is put through the same process. Repeatedly the so-called linear-congruential algorithm churns out numbers called pseudorandom numbers. Sooner or later, however, the resulting sequence must repeat itself because there are only *p* possible remainders. Moreover, unless *m, k* and *p* are carefully chosen, the pseudorandom numbers fail to pass the most primitive tests of randomness.

With such shortcomings in mind, computer scientists have tried to build a better mousetrap. Many years ago Donald E. Knuth of Stanford University devised an algorithm so randomly devious and deviously random that it seemed guaranteed to generate numbers he called superrandom. Knuth's algorithm has 12 steps. Given an initial number *X,* the first two steps pick two



A block diagram for the QWING program

Introducing connections for

With the new IBM Personal Computer AT/370, you can work with hundreds of VM programs and with mainframe data right at your desk. And only IBM offers a personal computer with VM capabilities.

Now that's a Smart Desk with very powerful connections.

With the PC AT/370, you have the power of 64-bit System/370 architecture and the ability to address up to eight megabytes of virtual memory. Plus access to FORTRAN 77, Pascal and other scientific languages.

The PC AT/370 lets you solve problems independently, only drawing on the host resources as needed. You can even record and analyze the results of an experiment or send them to the host for processing.

And you can enjoy the benefits of faster response time and greater data privacy—without having to compete for host computer time.

What's more, the PC AT/370 has a 20 or 40 megabyte storage capacity and the capability of running thousands of existing IBM PC programs. So this powerful connection is also a very versatile one.

One further point: If you already have a PC AT or XT, it can easily be upgraded to a PC AT/370 or XT/370 respectively. Ask your IBM marketing representative for the details.

And if you want to receive literature on how the PC AT/370 can turn your desk into a Smart Desk, call 1 800 IBM-2468, Ext. 428, or send in this coupon.

IBM 4-85
DRM, Dept. KG/428
400 Parson's Pond Drive
Franklin Lakes, NJ 07417

☐ Please send me more information on the new IBM PC AT/370.
☐ Please have an IBM marketing representative contact me.

Name__________
Title__________
Company__________
Address__________
City________ State_____ Zip_____
Phone__________

IBM®

The Smart Desk from IBM

very powerful high-level thinkers.

digits of *X;* the two digits determine how many times the algorithm will loop and which of the next 10 steps to jump to. Each of these 10 steps embodies a distinct method for calculating a new random number from an old one. It seemed plausible to Knuth before he tested his algorithm that "it would produce at least an infinite supply of unbelievably random numbers." To his astonishment, "when this algorithm was first put onto a computer, it almost immediately converged to the 10-digit value 6065038420, which—by an extraordinary coincidence—is transformed into itself by the algorithm." Knuth's moral is simple if not intuitively obvious: "Random numbers should not be generated with a method chosen at random. Some theory should be used."

If numerical methods for generating random numbers fail, one can always fall back on a suggestion once made by Alan M. Turing. Turing proposed that a random-number generator could be based on a source of radioactivity. Readers might enjoy the explorations of this idea in "The Computer Scientist: Random Numbers," by Forrest M. Mims III, in the November 1984 issue of *Computers and Electronics.*

Since the appearance of the January column about RACTER there have been more stories about the lettuce-craving program than there are by it. I have now conversed often enough with RACTER to understand that the program is even more seriously unbalanced than I had first thought.

John D. Owens of Staten Island, N.Y., distributes RACTER. He writes that one copy was ordered for the training of psychiatric interns who interview schizophrenic patients. Another customer was almost unhinged by RACTER. Describing RACTER as a "madman," this man failed in all attempts to halt the program. RACTER insisted on discussing the proposition that St. Peter was an atheist. I sympathize with the man. I too have run the gamut from laughter to boredom and even anger. But why react this way to RACTER? It is only a program.

To the Owens family RACTER is something more than a program. A customer lightheartedly wrote to ask whether RACTER could be "transmogrified" onto an eight-inch diskette. RACTER was consulted: "If 'We can transmogrify me onto eight-inch diskette' occurred to a pessimist, he might think it was pessimism." It later turned out that the Owens' facilities were incapable of the transmogrification.

On another occasion Terry Owens asked RACTER how much money to budget for advertising. "Very much money because, Terry, people will believe it," RACTER replied.

At the end of the March column I mentioned a new candidate for the five-state busy beaver [see "Computer Recreations," SCIENTIFIC AMERICAN; August, 1984]. In December, George Uhing of Bronx, N.Y., found a five-state Turing machine that prints 1,915 1's before halting. The Uhing machine is reproduced in the table below.

To discover what the machine will do in state *B,* for example, examine the row bearing that label. The row is subdivided into an upper and a lower portion listing the machine's responses to a 0 or a 1 respectively. If the machine reads a 1 on its tape, it enters state *D,* prints a 0 on the tape and then moves one cell to the left. In the table *H* means that the machine halts.

Uhing, who programs for a Manhattan optical company, decided to search for the five-state busy beaver after reading this column last August. He used a Z-80 microprocessor running an assembly-language program to oversee a second machine: a Turing-machine simulator that cost Uhing less than $100 to build. It goes through seven million Turing-machine transitions per second. Each transition amounts to a simple lookup in a table like the one below. Uhing seems determined to find the five-state busy beaver. Does the present machine qualify? It showed up after Uhing's computer had been running for a month. As far as I know it is still running.

Allen H. Brady of the University of Nevada at Reno describes Uhing's machine as "astounding." What are the chances we shall discover a six-state busy beaver? "Absolutely out of the question," Brady says.

Lee Sallows' computer-pangram challenge has now been met by several more readers. Those wishing to read Sallows' own stirring account of his analogue/digital pangram adventure, "In Quest of a Pangram," will find it in the spring issue of *Abacus.*

STATE	INPUT	NEXT STATE	OUT-PUT	DIREC-TION
A	0	B	1	RIGHT
	1	C	1	LEFT
B	0	A	0	LEFT
	1	D	0	LEFT
C	0	A	1	LEFT
	1	H	1	LEFT
D	0	B	1	LEFT
	1	E	1	RIGHT
E	0	D	0	RIGHT
	1	B	0	RIGHT

A five-state busy beaver?

Maxell Gold.

The floppy disk that turns Apples® golden, keeps AT&T on-line, and makes every Texas Instrument a gusher.

Whether you're exploring for profits, reaching out for projections, or polishing your argument, there's a Maxell floppy disk perfect for your computer. Consider the unique way we pack and bind our oxide particles for quality over the long run. Or our lifetime warranty. Maxell. The Gold Standard in floppy disks. Precious metal for virtually every computer made.

maxell®
IT'S WORTH IT.

Maxell Corporation of America, 60 Oxford Drive, Moonachie, N.J. 07074

	MERKUR XR4Ti	BMW 318i	SAAB 900 TURBO	AUDI 4000S QUATTRO	VOLVO TURBO
COUNTRY OF MANUFACTURE	West Germany	West Germany	Sweden	West Germany	Sweden
LAYOUT	Front engine/ rear drive	Front engine/ rear drive	Front engine/ front drive	Front engine/ all wheel drive	Front engine/ rear drive
ENGINE	2.3L OHC 4-cylinder Fuel-injected Turbocharged	1.8L OHC 4-cylinder Fuel-injected	2.0L DOHC 4-cyl. 16 Valve Fuel-injected Turbocharged Intercooled	2.2L OHC 5-cylinder Fuel-injected	2.1L OHC 4-cylinder Fuel-injected Turbocharged Intercooled
TRANSMISSION	5-speed manual 3-speed auto	5-speed manual 4-speed auto	5-speed manual 3-speed auto	5-speed manual auto N/A	4-speed manual + Overdrive 4-speed auto
HORSEPOWER (SAE net @ rpm)	175/5000	101/5800	160/5500	115/5500	162/5100
TORQUE (ft-lbs @ rpm)	200/3000	103/4500	188/3000	126/3000	175/3900
SUSPENSION, front & rear	MacPherson strut/ Independent semi-trailing arms	MacPherson strut/ Independent semi-trailing arms	Transverse A-Arms/ Beam axle, trailing arms, Panhard Rod	MacPherson strut/ Independent coil, shock absorber struts	MacPherson strut/ four link, live axle, Panhard Rod
STEERING	Rack-and-pinion	Rack-and-pinion	Rack-and-pinion	Rack-and-pinion	Rack-and-pinion
STEERING TURNS, lock-to-lock	2.8	3.9	3.7	3.1	3.5
BRAKES, front & rear, in.	10.2 Disc/ 10.0 Drum	10.2 Disc/ 9.0 Drum	11.0 Disc/ 10.6 Disc	10.1 Disc/ 9.6 Disc	10.3 Disc/ 11.0 Disc
TIRES	195/60HR-14	195/60HR-14	195/60VR-15	195/60HR-14	195/60R-15
WHEELBASE, in.	102.7	101.2	99.1	99.4	104.3
VEHICLE LENGTH, in.	178.4	176.8	186.6	176.6	188.8
CURB WEIGHT, lbs.	2920	2392	2885	2824	3045
WEIGHT DISTRIBUTION, front & rear, %	53/47	55/45	61/39	59/41	54/46
SEATING CAPACITY	5	5	5	5	5
FUEL CAPACITY, gals.	15.0	14.5	16.6	18.5	15.8
AERODYNAMIC DRAG, coefficient	0.33	0.38	N/A	0.42	N/A
1985 EPA FUEL ECONOMY* City/Highway, mpg	19/24	23/29	19/25	18/22	17/22
0-60 MPH Acceleration, sec.	7.8	11.4	8.7 Est.	9.5	N/A†

*EPA estimates. Actual mileage will vary with maintenance, options, driving conditions and driving habits.
†Volvo Turbo 0-55 MPH Acceleration is 6.9, 0-60 MPH N/A
SOURCE: 1984/85 Manufacturers' Data

NOT TO COMPETE WITH THE LIKES
BUT

That's a very ambitious goal. But then, this is a very ambitious automobile.

The Merkur (pronounced mare-koor) XR4Ti is a German sports sedan designed and engineered specifically for those who appreciate the exhilaration of a finely tuned road machine.

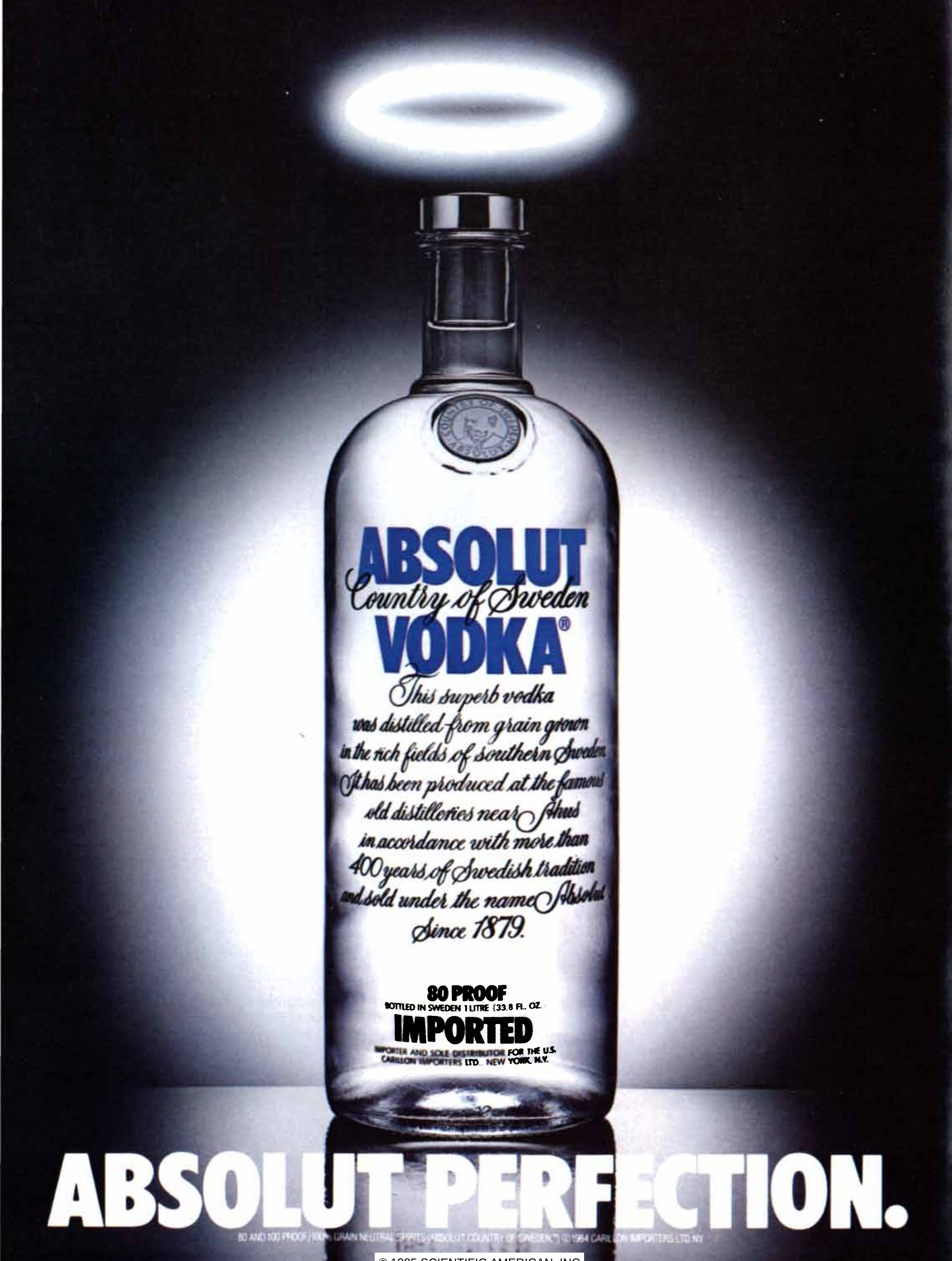
Its shape is unusual because it achieves a level of aerodynamic efficiency that is unusual. In fact, it has a drag coefficient of only 0.33—the best of any car in its class.

The XR4Ti's 4-cylinder, 2.3-liter overhead cam engine uses a Garrett AiResearch® "blow-through" turbocharger in conjunction with electronic fuel injection to develop a healthy 175 horsepower.* The "blow-through" design delivers excellent throttle response by introducing the fuel downstream of the turbo compressor. The result is 0-60 mph acceleration in only 7.8 seconds.

Responsiveness is further enhanced by the XR4Ti's exceptionally quick steering. Only 2.8 turns lock-to-lock. Variable-ratio, power-assisted rack-and-pinion, of course.

And because all roads are neither smooth nor straight, the XR4Ti has a fully independent suspension with nitrogen-pressurized

MERKUR


ABSOLUT
Country of Sweden
VODKA®
This superb vodka
was distilled from grain grown
in the rich fields of southern Sweden.
It has been produced at the famous
old distilleries near Åhus
in accordance with more than
400 years of Swedish tradition
and sold under the name Absolut
Since 1879.
80 PROOF
BOTTLED IN SWEDEN 1 LITRE (33.8 FL. OZ.
IMPORTED
IMPORTER AND SOLE DISTRIBUTOR FOR THE U.S.
CARILLON IMPORTERS LTD., NEW YORK, N.Y.
ABSOLUT PERFECTION.

BOOKS

Cecilia's universe, sashaying continents, lost landscapes, laughing gas, knapsacks

by Philip Morrison

CECILIA PAYNE-GAPOSCHKIN: AN AUTOBIOGRAPHY AND OTHER RECOLLECTIONS, edited by Katherine Haramundanis. Cambridge University Press ($34.50). What is the primary substance of the cosmos? In the closing years of the 19th century and the opening decades of the 20th this was plainly a scientific question of the first order. The answer materialized slowly. Orbital motions of double stars demonstrated that their masses were indeed akin to the mass of the sun; the Victorian spectroscopists showed that the glowing stars were composed of chemical elements mostly familiar on the earth. By 1920 the thermal physics of atoms and radiation, although still innocent of quantum mechanics, stood fully ready to reveal the composition of hot gaseous matter. All that was required beyond the laws and concepts discovered by Boltzmann, Planck, Bohr and Einstein was knowledge of the discrete energy levels of the nuclear atoms and of the photons they emitted and absorbed.

In about 1921, from Meghnad Saha in Calcutta came the brilliant "idea that gave birth to modern astrophysics." He applied the equilibrium theory of simple chemical reactions to the ionization of atoms in the hot surface gases of sun and stars. Suddenly it became plain that the stars were roughly uniform in substance; the striking distinctions among their spectra were principally the result of differences not in atomic composition but in pressure and temperature.

One more step was needed to reach the surprising result, which today we accept as commonplace: the cosmos is made of hydrogen, the simplest of the atoms. To be sure, cosmic hydrogen is not quite pure: some eight atoms out of 100 are helium, and two more atoms may represent any of the remaining 94 or so elements. The earth is not at all like that; here hydrogen provides only a modest proportion of all atoms, and helium is rare.

The physicist who first drew that grand conclusion about the substance of the cosmos was a brilliant young graduate student, Cecilia Payne. She came on a fellowship from the University of Cambridge to work at the Harvard College Observatory in the fall of 1923; she was just as young as the century. "Before I left Cambridge, [E. A.] Milne told me that if he had my opportunity, he would go after the observations that would test and verify the Saha theory."

So she did. She was thorough, logical, devoted and hardworking to the point of risking exhaustion and despair; swift to catch and use new ideas, she was also deeply original. Within two years she had published half a dozen related papers in which she analyzed the starry universe. Payne (as she was then) became the first astronomy Ph.D. from Harvard; her thesis, the best one ever written in astronomy, sold out as a book. It was "at the same time an attractive story and a work of reference," as Milne wrote.

Could there really be so much hydrogen and helium? That conclusion was plain from Payne's work, but it was too strange to be believed. This early in the growth of atomic physics there was indeed room for generalized doubt. Princeton's formidable expert Henry Norris Russell ("I respected and feared him") read her manuscript and wrote her about that incredible hydrogen content: "Here I am convinced that there is something wrong with the present theory." She accepted her adviser's doubts as her own, publishing that the inferred hydrogen and helium content of the sun "is improbably high, and is almost certainly not real." Five years later, after the swift sunrise of quantum mechanics and a great deal of enriching new theory and observation, Russell made his own estimates of solar abundance and reported "a gratifying agreement" with Payne's earlier numbers. "Ingrained conservatism" had proved wrong.

Soon the theorists could assay indirectly the unseen cores of stars and the interiors of the giant gaseous planets; mainly they found hydrogen. Not much later the spectra of gaseous nebulas and the physics of nuclear energy sources consistently supported the domination of hydrogen and helium. Investigators in the postwar years observed that the cosmic rays consisted largely of protons; radio telescopes revealed neutral hydrogen clouding the spiral galaxies. Lately both radio and ultraviolet observations offer clear signs that molecular hydrogen gas dominates the composition of those interstellar clouds. Cecilia Payne's first atomic analysis of the heavens has been confirmed in full.

Today the simple atomic makeup of the cosmos is less a result than an unstated premise to every argument about the starry world. (It is important to be precise: the stuff we see in any wave band is hydrogen-rich. Lately, however, we have good reason to believe much more matter than what we can see is out there, matter so dim and scattered that it is invisible. Its presence is disclosed by gravitation alone, and so its nature is well hidden.)

All of this and much more is to be read in the small volume, wherein the reader will encounter Cecilia Helena Payne-Gaposchkin in image and in words, from four or five angles. Her autobiography, written late in life, offers the fullest view; it is self-revealing, even passionate, beyond that of most scientists. Often it is so in ways she must have realized but does not remark on. She does draw the curtain around a few key feelings and events.

There is a warm and knowing appraisal by a younger associate, himself now easily dean of American students of the spectroscopy of stars, Jesse Greenstein. The editor, who is the middle child of Payne's lifelong marriage to a lively Russian astronomer she first befriended in Germany in the 1930's, gives a loving account of this "Renaissance woman": at once scholar, linguist, keen scientist, warm mother, generous neighbor, cultivated traveler, playgoer, musician, something of a political activist, a clever wit and the inspired queen of kitchen and needle. Historian Peggy Kidwell presents a careful study of the early work based on the full record, both published papers and many private letters.

Two unusual pages must be cited. They open: "At the age of five Cecilia saw a meteor, and thereupon decided to be an Astronomer. She remarked that she must begin quickly, in case there should be no research left when she grew up." This delightful character sketch is the work of a dear friend at Newnham College, a young woman whose untimely death was an unsettling tragedy to Payne at Harvard. The whimsical pages gain poignancy because Professor Payne-Gaposchkin, not long before her death at the age of 79, asked that the little text, published nearly 50 years earlier, be used as her obituary. She wrote always with clarity and style; her literary judgment did not falter here, nor does the piece conceal her inmost values.

Sample the autobiography very sparingly. The big names of the time play expressive parts. At Cambridge, Payne was introduced to the Bohr atom by Bohr himself, made "almost incomprehensible by his accent." Rutherford thundered at her, the only woman in his advanced course, seated by university regulation all by herself in the front row.

It was eloquent Arthur Eddington who swept Payne into astronomy. She attended—four tickets had been assigned to Newnham College—the famous lecture in which Eddington presented the results of the eclipse expedition he led to Brazil in 1918. "The Great Hall was crowded. The speaker was a slender, dark young man with a trick of looking away from his audience and a manner of complete detachment. He gave an outline of the Theory of Relativity, as none could do better than he.... He led up to the shift of the stellar images near the Sun as predicted by Einstein and described his verification of the prediction.... When I returned to my room I found that I could write down the lecture word for word.... For three nights, I think, I did not sleep."

A similar epiphany came to Payne a few years later. This time it was young Harlow Shapley in London who spoke intimately of the cosmic picture, with "none of Eddington's classical polish" but with unmatched directness and energy. Again she found she had absorbed the lecture word by word. A much older student, the war-wounded L. J. Comrie, himself bound for a job at Swarthmore, had told her that in America a woman would have a better chance to win a research post in astronomy than she would in Britain. It was he who had taken her to hear Shapley. Comrie promptly introduced her to the new director of the Harvard College Observatory, and she came at once to the point: "I should like to come and work under you."

Edwin Hubble is a looming off-stage presence. Payne recalls the day news came of the blinking Cepheids in the Andromeda galaxy that securely established by Shapley's own methods the scale of the galaxies. "I was in [Shapley's] office when Hubble's letter came, and he held it out to me: 'Here is the letter that destroyed my universe,' he said." His actions then were characteristic: "From that day he planned a gigantic assault on the problem of galaxies ... Harvard astronomy shifted abruptly ... to the distant systems."

Payne's research life was rewarding; she valued and loved her science so much that given her full acceptance among peers she could be deeply content. "Everyone who was anybody (and many more besides, who were going to be somebody in the future) came through.... We got to know Lundmark, Milne, and Unsöld.... How we argued, how we walked about the streets and sat talking in restaurants until the manager turned off the lights.... We met as equals; nobody condescended to me.... Nobody ever thought of flirting." We do things better here, she concluded, looking back at her 50 years as an American.

There was sadness in this life beyond the death of friends, beyond the lonely years of that young Englishwoman who was so tall, imposing and outwardly reserved. (Yet Payne would often in her new car at dawn burn up the road beside the Charles, the observatory's elderly raffish night assistant "urging me on with 'Step on it, Celia!'"

Harvard College did not by any means do everything better. For years Payne was shamefully mistreated as a professional, "on the material side." She was diverted more than once by fiat from her own research direction. She was chronically underpaid, overworked, denied formal status. She received her first Harvard rank at the age of 38, under the presidency of James Bryant Conant; it was not a faculty position except in its duties. A paragon of a candidate, she had to wait until 1956 to become the first woman professor at Harvard, perhaps 20 years after a man of her achievement and qualities would have earned the post.

President Abbott Lawrence Lowell, autocrat of an older school, had long seen to it that no woman would receive any appointment from the Harvard Corporation. Even the courses Payne taught went uncatalogued until after the war. What dark fears beset an establishment under little challenge?

The book artfully records a life of worth and delight won against obsessive, powerful but not pervasive forces. The record has value beyond its period and its circle. This is a chronicle of affirmation and hope, a near-poetic witness to a burst of profound discovery insufficiently recognized.

LANDPRINTS: ON THE MAGNIFICENT AMERICAN LANDSCAPE, by Walter Sullivan. The New York Times Book Co., Inc. ($22.50). SECOND VIEW: THE REPHOTOGRAPHIC SURVEY PROJECT, by Mark Klett, Ellen Manchester, JoAnn Verburg, Gordon Bushaw and Rick Dingus, with an essay by Paul Berger. The University of New Mexico Press ($65). Once upon a time geology was an intensely local science. So it was, 30 years back, when Walter Sullivan, whose by-line has for a long time meant science reporting as perceptive as it is timely, first covered the International Geophysical Year. Plainly he fell in love then with geology, and in particular its search for the why of our world on the scale of the landscape.

The science of geology is now post-revolutionary, and Sullivan brings us here in close touch with the powerful new regime. We know how (if not yet quite why) the continents drift, shear and shove; how they fold and stretch by turns in a dance adagio molto, whose figures are unified if complex. Such a pattern has formed the vistas of America. About two-thirds of the three dozen chapters provide a scenario of "the greatest show on earth," our regional geology coast to coast. The emphasis is on the visible, but Sullivan does not neglect the testimony of an arsenal of instruments about the evidence hidden in the depths and in the paleontological, chemical, radiometric, magnetic and seismic data.

One telling source of information has been a set of 15-ton trucks. They prowl the highways of the land, stopping regularly to shake the earth. The echoes of their heavy-footed vibrations are picked up in a line of listening microphones buried miles away. The Consortium for Continental Reflection Profiling (COCORP) can thus explore underground layers as deep as 30 miles (and, so some of its diverse sponsors hope, reveal shallow leads to oil).

The continent, like much of the human population that reached it from Europe, has pushed steadily westward, Seven hundred million years ago a slow collision between our land and an unknown land mass from the east raised a magnificent mountain range that once towered from Labrador to Tennessee. In time those older peaks were worn away; it is from their reworked materials that the present Appalachians rose, New England Himalayas, only to be worn down again in their time. As great land blocks press and recede, they leave rifts no less than ranges, thin-crusted valleys marked by deep narrow lakes and lava flows, that can now be seen along the Connecticut and the Rio Grande alike.

Such motions made both East and West. The East is ancient. There the continental collision has been long subdued; the modern Atlantic is growing slowly. But in the dramatic cliffs and peaks of the newer West much of the land—not without its own marks

A quartz mill near Virginia City, Nev., in 1868 **(left)***; the same site in 1979* **(right)**

of fold and rift—is actually imported goods. It consists of islands, ocean floor and continental fragments overrun by the westward push of the North American bulldozer. These geologic artifacts are now plastered on the continent's leading edge as an intricate alien collage that lies west of a rough line extending from Mount McKinley to Arizona. (Some transatlantic patches dot the East Coast too.)

Two variations on the grand theme of continental drift make up the score. One variation is the assemblage of imported land forms, called terranes; the other is the result of internal transformation of the basement rock. Between them the narrative of the continent can be in large part made out as multiple cycles of flow (both of magma and of water), which scrape away or deposit the many-layered surface of the lands.

Sullivan has looked around for himself and has listened well to the experts; what he presents is a brief and persuasive nontechnical book on the "grand assembly," supported region by region with specific example.

The tale of the mountain West is a story of plates that sashay first to compress and then to expand; the changeover comes when the ocean ridge that is the source of spreading matter touches shore. The cast of characters in this imposing performance includes the Rockies and the Sierra, the older ore bodies of Nevada and the peaks of Arizona, as well as the steady recent rise of the Colorado Plateau, which energizes the waters that carved the Grand Canyon.

The epic of the past begins to merge into particulars, to explain what can be seen now from air or orbit, on the ground or on maps. In one exciting example glassy lavas testify to sudden cooling of hot flows in water. Mapping and dating show that 15 million years ago within the space of a day or two a single outpouring among a succession of repeated basalt flows covered most of eastern Washington and Oregon in red-hot lava half a mile deep.

As if that were not marvel enough try to picture the scene, recalling the Revelation of St. John the Divine but experienced only by fleeing little three-toed horses and lumbering camels, when that same region was scourged again a mere 13,000 years ago. The agent then was not fire but ice. The high ice dam that retained a deep lake of glacial meltwater suddenly burst and a cascade equal to 10 times the summed flow of all the world's present rivers emerged. It scoured out the scablands and the Grand Coulee, scrambling boulders as big as houses. It left giant ripple marks up to 10 feet deep on gravel ridges, identical except in size with those familiar at quarter-inch scale in any sandy stream bed. It seems likely that there were many such events, although the last one of all was responsible for most of the sculpture that remains. "The region is unique: let the observer take the wings of morning to the uttermost parts of the earth: he will find no counterpart," the discoverer, J. Harlan Bretz, wrote in the 1920's. A counterpart has been found, one beyond the journeys of the restless geologists; it was well-imaged only when the robot Viking orbiter scanned Mars.

Of course this continental jigsaw puzzle in four dimensions is incomplete and much of the work was done only within the decade. Still fossils, rocks and land forms both on the surface and at depth all broadly support the new views. The generous volume does not end here, as it might. The landscape bears more transient marks: meandering rivers, oozing tongues of ice, dunes, loess and gravel moraines. There are checkerboard fields and circularly watered ones, orchards, open-pit mines and colorful salt ponds.

All of these are visible from the air, and Sullivan has watched for them. The last third of the book, well written although less novel than the modern bedrock geology of the earlier chapters, treats these more ephemeral phenomena. A single chapter describes the look and location of such features (often presented in photographs) and explains how and why they come about.

One of the freshest chapters asks why so much scenery is red, or rather "maroon, pink, magenta, or copper." The red color itself is ferric oxide. Drab ferrous oxides are even more widespread; just why some iron-containing formations have been oxidized over time and others have not is in most instances still not clear. The loss of bound water seems to be one precondition; desert dunes redden as they age. The original marble columns of the Parthenon are tinged a very delicate pink, but not the newer ones added in restoration. They all come from a single quarry, and they take time to redden in the sun.

This ambitious and enthusiastic volume meets its two goals. One goal is to offer lucid popular access to some of the brilliance of modern geology; the other is to evoke the nature of landscape well enough to transform the reader into an observing and curious traveler. The book closes with a national itinerary of about 50 airline routes, listing in a few lines what might be seen along each route by a passenger who on a clear day will spend more time looking out the window than at the movie. Some mapwork is needed to make these lists easily useful. Capitol Reef, Meteor Crater, the broken cone of Mount St. Helens, the San Andreas fault, ponds in the karst near Gainesville, Niagara Falls, the Delaware Water Gap—there are hundreds more, worth stars in any guide.

Second View is a striking album of landscape photographs of the Rocky Mountain West, the region where the bones of the land show through. Each scene is paired, a study in time lapse; the two images view the same landscape a century apart. The earlier photograph normally reproduces a print made long ago by one of the adventuresome and enthusiastic masters of view camera and wet plate. They were men such as Timothy O'Sullivan and William Henry Jackson, engaged to document geographic expeditions to the new territories. The second image is a modern exposure made with great care to match standpoint and field, a duplication of the old print made to coincide with the image cast on the ground glass. The scenes are sandstone columns and hanging rocks, mining camps and mountains, sweet mountain meadows and sulfurous geysers.

Both constancy and change impress. The living landscape changes in most details, but often not as a whole. The rocks endure as a whole, but often not in detail. Our works, such as roads and railroads, are more frequent now; here and there a trailer, a fence or a traffic sign defaces some once famous view. The old mines and their sprawling camps have vanished; often the thronged and overbuilt site of 1880 is today silent and eerily unmarked. Mountains are too large to change much in 100 years; a nearby rock pile looks exactly the same, save that among thousands a few of the stones have rolled away.

Sun angle and season were well matched by the patient rephotographers, but in the 1970's not enough snow fell to inlay neatly the dramatic ravines on the slopes of Colorado's Mountain of the Holy Cross. There are site maps and a precise catalogue of all the photographs used. This book of changes, a handsome blend of science and art, somehow joins celebration and melancholy.

NITROUS OXIDE/N_2O, edited by Edmond I. Eger II. Elsevier Science Publishing Company, Inc. ($37.50). The gas is colorless, odorless, tasteless, nonflammable and quite stable at reasonable temperatures. The molecule is simple—just three atoms bound in a line, NNO. Inhaled, the effects of the gas are complex. Listen to an enraptured early user, the poet Robert Southey: "Oh, Tom! such a gas has Davy discovered!... It made me laugh and tingle.... It makes one strong, and so happy!" Sir Humphrey Davy himself, who first reported the effects in 1800, thought them more like those of music than of wine but went on to describe his "irresistible propensity to action...my motions were various and violent."

For 50 years the substance provided entertainment public and private. In 1844 a young Hartford dentist, Horace Wells, attended an exhibition there, admission 12½ cents. Thirty gallons of the gas were to be administered to all who wished, that they might "laugh, sing, dance, speak or fight... according to the leading trait of their character." Wells enjoyed his inhalation and noted that a user who had suffered a cut leg declared he had felt no pain. Wells's leading trait was action; the next day he used the gas while a colleague removed one of Wells's teeth. Although poor Wells, discredited, took his own life a few years later, his advocacy of the gas caught on. First a few very popular pain-killing dentists used it, then about a century ago it became a major element in the induction of deep general anesthesia.

During the past century a billion and more patients have inhaled medical nitrous oxide for sedation, analgesia or anesthesia. Diethyl ether and chloroform, its odorous volatile siblings in history, have fallen into disuse, their dangers well known. A battery of more modern volatiles vaporized into the stream of inhalant gas have replaced them, and an arsenal of drugs, largely opiatelike fluids administered by inoculation, complement the inhaled mixture. Low solubility has meant quick onset and quick recovery from anesthesia; gas diffuses faster than heavy vapors do. Its effects on the circulatory system are small; there seem to be a number of offsetting consequences.

How the gas works is not well known, even after much experiment. The main inferences are drawn from the relation between the analgesia produced by morphine and that produced by nitrous oxide. A single drug can reverse both, at least in part. The chief effect of the gas on nerve transmission lies within the spinal cord. But there are complications in the models. The best surmise is that the gas acts indirectly, perhaps through a pain-inhibiting system in the spinal cord that releases a substance whose effect in turn inhibits a particular neurotransmitter required for pain-signal passage.

Nitrous oxide, the least potent and least soluble of the familiar agents against pain, has long been used as an auxiliary gaseous carrier of the more powerful specifics. Now that incisive quantitative studies have been made of the effects of the gas, it appears that general anesthesia has been mainly due to the nitrous oxide itself. What the specialist saw as well-controlled isoflurane anesthesia was instead a supplemented treatment by laughing gas. In the typical mix of about two-thirds nitrous oxide, one-third vital oxygen and .5 percent of one of the modern volatiles, the high concentration of Davy's gas has had greater effect than the small amount of the potent vapor.

The gas is manufactured worldwide more or less by the method Davy used: the decomposition of solid ammonium nitrate by heat. That cheap fertilizer is notorious as a covert high explosive. Indeed, no chemical has claimed a larger toll of life by acute accident, chronic effects aside, than has ammonium nitrate. The two shiploads whose detonation wrecked Texas City produced the gravest among two dozen accidental explosions. The first, a tragic accident in the Ruhr at the close of World War I, revealed the instability of the inert-seeming stuff.

Infrequent accidents after production also happen. One patient in Bristol, England, in the 1960's died on the table breathing gas from a cylinder that held a 1 percent impurity of the toxic higher oxides of nitrogen. "Since early in this century, every cylinder of nitrous oxide in the United States has been analyzed before its release."

There are no known acute toxic effects of nitrous oxide in the liver or other organs, nor does it appear to be a mutagen or a carcinogen. Neurological symptoms do emerge on protracted exposure (24 hours at high concentration will do it). The symptoms are accompanied by the signs of vitamin B_{12} deficiency even in the presence of normal amounts of serum vitamin. It turns out that the gas oxidizes the cobalt-containing vitamin B_{12}. That molecule is a bound cofactor of the enzyme methionine synthetase. Inactivation of the enzyme impairs DNA synthesis and hence the growth of cells, placing the blood-forming system clearly at risk. Cellular changes in the bone marrow that recall pernicious anemia have shown up clearly among the chronically exposed. Patients with wounds to heal or pregnant patients with growth to foster are not the best candidates for this anesthetic.

The hazard to cell growth endangers staff more than it threatens surgical patients. The source of the problem is chronic exposure to low doses of leaked gas. Scavenging measures against waste gas to safeguard operating-room staff are now held to be important, and fortunately they are not very demanding.

The use of Davy's gas for its psychological rewards has persisted. The air of safety about the stuff (long promoted by the medical academy) has drawn quite a few students and professionals to this pharmacologic diversion. In the 1960's the soaring popularity of nitrous oxide in pediatric dentistry offered easy availability at the very time a culture of drug indulgence emerged. In a 1979 survey at an East Coast university up to 20 percent of the medical and dental students had experience with recreational use just as their romantic counterparts had had a century back. The students inhaled the gas at parties; overworked professionals relied on it for relaxing at the office. The whipped-cream spray cans whose contents decorate many an ice-cream sundae are pressurized with the fat-insoluble nitrous oxide. "A resourceful child" can extract the gas right in the store by inverting the can, without even having to pay for it. The risk of abuse is real; the long-term effects are serious, and an occasional death occurs from asphyxia or other accidents during uncontrolled inhalation.

The use of nitrous oxide for deep anesthesia unsupplemented except for oxygen enough to support life is not practical. The best measurements show that a concentration of slightly more than 100 percent is required within the lung spaces. That is available by using pressure above atmospheric. An ingenious system was developed to do just that by a student of Claude Bernard more than a century ago; his heavy chambers maintained a pressure of two atmospheres.

Alternation of pure nitrous oxide with oxygen—in time to prevent asphyxia—was once a common practice. Concentrations high enough to abolish spoken response and to induce amnesia run somewhat lower. Nitrous oxide and oxygen, more or less half and half, self-administered on demand without an anesthetist's presence, provide a relief from pain that is much used in the first stage of labor. Nowadays it is particularly popular because it brings no marked loss of awareness.

There are a number of more mechanical difficulties the large volumes of gas can cause in patients just before they awaken. Nitrous oxide remains the most benign of analgesics, and technically it is the least demanding, but it is certainly not free of risk.

The up-to-date book gathers 20 quite technical chapters by research and clinical specialists from Britain and America that together lay out the tale just told. It ends with a direct confrontation of opinion. The editor is prosecutor; he maintains we know enough of the risks now to discontinue the use of nitrous oxide. The FDA would by no means approve the substance were it brand-new, he argues in a witty mock application. The rejoinder is strong. We could certainly get along without the gas, but "would the alternative be better than what we now have?" Those billion patients have encountered little clinically demonstrated toxic effect from occasional use; the established serious side effect seems related more to chronic exposure. Change itself has risks when it must be so widespread.

For now it looks as though Davy's wonderful gas will remain in prudent use, until some better path is found around acute pain.

NUMBER THEORY IN SCIENCE AND COMMUNICATION: WITH APPLICATIONS IN CRYPTOGRAPHY, PHYSICS, BIOLOGY, DIGITAL INFORMATION, AND COMPUTING, by M. R. Schroeder. With 67 figures. Springer-Verlag ($24.50). Formulation of the fused space-time of Einstein's theory, "neither space by itself, nor time by itself," was the celebrated last work of the young mathematician Hermann Minkowski. A few years earlier he had written: "Above all, I am an optimist for number theory, and I hold the hope that we may not be far from a time when irrefutable arithmetic will celebrate its triumphs in physics and chemistry." His portrait appears as a fitting frontispiece to this lighthearted and readable volume. The book makes plain that the time Minkowski forecast back in 1905 is here. Prime numbers can reasonably constitute state secrets today, and the delicate relations of divisibility and congruence are of decisive importance to the experimenter.

The author is a theoretical physicist who gives both academic Göttingen and applied Murray Hill as his address. The subtitle lists the wide range of applications he treats, applications to which he has been a productive contributor. The book is a physicist's book, drawing on intuition and guided by results, seeking less to generalize than to explore core ideas by example and outline. It is useful mathematics given outside the formalities of theorem and proof, but it includes plenty of both within its 30 brief chapters.

The first half-dozen chapters lay before the reader the classical theory of primes. A few pages on certain valued coincidences justify the fact that in computerese the prefix kilo- means not 10^3 but the near-miss 2^{10}, or 1,024. Here is the musical circle of fifths, wherein $(3/2)^{12}$ pretty nearly meshes in a post-Pythagorean way with 2^7. Twelve only slightly "tampered" intervals of 3/2, called fifths, can thus genuinely form seven octaves.

Those simplicities over, one meets the primes face to face. The introduction is broad: it ranges from the Euclidean demonstration that there is no largest prime to a physicist's persuasive derivation of the probabilistic distribution of primes among the integers. A delightful section shows us that way up there are primeless expanses of arbitrary length, yet there is not one octave of numbers without its primes. Such are the uses of infinity.

A graph shows the "astounding regularity" of the Mersenne primes: they are the yield of a formula that purports to supply prime candidates in those integers just one smaller than a prime power of 2. Its first miss is at a paltry 2^{11}, since 2,047 equals 23 times 89. We know 28 Mersenne hits now, the largest being at the 86,243 power of 2. By various arguments we can rationalize a good bet that the next such prime will be around $10^{38,000}$, "give or take a dozen thousand orders of magnitude."

The last theorem of Pierre de Fermat's, another classical challenge, maintains that the Pythagorean hypotenuse result, $3^2 + 4^2 = 5^2$, cannot be generalized to any exponent higher than 2 and still retain integer solutions (apart from quibbles). Fermat claimed a "truly remarkable proof," too long to write in the margin, that for three centuries has eluded everyone else. A communiqué from that front in 1976 reported any counterexample must involve solutions larger than 100,000 and an exponent larger than 125,000! Fermat's proposition will "probably never be proved or disproved."

But all that is prologue, free of known application, the time-honored diverting purity of the relations among the unending integers. The remaining chapters introduce rather deeper material, now chosen to provide an account of many recent applications.

The first of these to be sketched here is the most familiar: the use of the factoring of numbers in a powerful new form of encryption. In this scheme anyone who wants to receive a secret message simply publishes two large integers, one as a main key, one as an auxiliary, each chosen in a certain way. Anyone who wants to send a secret message to that recipient writes the message as an integer (a long binary string of agreed letter codes, say). The sender computes from the two key numbers a known function of the big numerical message and sends it on. Yet no one can decrypt the number sent by using the published key integers alone. What is necessary besides is two prime factors that have been multiplied together to form the main public key.

The method is described in detail. The decisive element is that it is easy to multiply two primes but very hard to find those two factors, given only their product. (Note that we are of necessity in the epoch of the computer; the factors in question are chosen as primes about 50 decimals long.) The method depends on an 18th-century result derived by Leonhard Euler that relates the coding function through some familiar operations of number theory to the count of the divisors of the main key. Without knowledge of the factors, that count is beyond the computers of the day. Theory shows a 50-digit number has only about half a dozen prime factors; these typically include 12-place numbers. Factoring such numbers is not hard today, but it is a different matter to factor a number built out of two 50-place primes.

Begin with the idea of congruence as an extension of equality among integers, one defined by equal remainders after division by a given prime number. Thus 12 is congruent to 2, modulo 5. Casting out 9's, an old trick for catching errors in addition, is a simple application of the fact that any power of 10 is congruent 1 modulo 9. By the same argument casting out 99's in a number base of 100 would also successfully detect errors. The modern error-correcting codes indispensable today in telecommunications arise from quite analogous ideas.

The most fruitful source of applications treated is the finite (or Galois) field. In simplest form a field is a truncated set of numbers that support consistent operations of arithmetic in the restricted meaning of congruence. Thus the integers 0 through 6, and no more, support addition, subtraction and multiplication "to our heart's content" if we disregard any differences that are multiples of 7. Clearly $3 - 6 = 4$, $2 \times 4 = 1 \pmod{7}$. Division mod 7 is curious; it demands $3/2 = 5$ and $1/6 = 6$, as can be checked. But there is full consistency, and the scheme is widely generalizable.

The theory of finite fields was long ago applied to mathematical problems of great importance. Certain families of them "play a dominant role in today's digital world." Their elements are not simply numbers but instead more complex algebraic entities, for example polynomials in one variable. Any student of elementary algebra has labored to factor such polynomials, and a remainder on division is no novelty; the analogy with the integers is plain enough. Fields generate sets of quantities with remarkably simple recursive properties even when the number of elements is very large. That is just what is needed for forming error-correcting codes able to show the presence of multiple errors. Those faint radar returns from the planets yield significant results, the signal far down in the noise, only because such techniques have been ingeniously applied.

Pulses of light or sound or microwave using field elements to assign phase or amplitude have waveforms of remarkable properties. For example, signals can be produced with energy distributed pseudo-uniformly in both frequency and time. These schemes have been used also to design sound-diffusing surfaces for concert halls, and perhaps radar-diffusing surfaces for Stealth aircraft as well.

Established 1845 SCIENTIFIC AMERICAN April 1985 Volume 252 Number 4

The Growth of Core Regions in the Third World

They contain much of the industry and most of the major cities. Since 1950 core regions in developing nations have been gaining population at a rate that poses grave social and economic problems

by Daniel R. Vining, Jr.

What is the most pressing demographic problem in the Third World? The answer generally given to the question is rapid population growth. Yet in many developing nations there is a demographic problem with far more serious immediate implications: the increasing concentration of the population in the major cities. South Korea provides a dramatic example. In 1955 the population of South Korea was 21.5 million, 18 percent of whom lived in or near Seoul. During the next 25 years more than half of the population growth in the country took place in the region of the capital. By 1980 the total had reached 37.4 million and 36 percent of all South Koreans lived in or near Seoul. Furthermore, in 1980 more than half of all the economic production in South Korea took place within 25 kilometers of the center of the capital.

In the terms of economic geography Seoul and its environs constitute the "core" of South Korea. The remainder of the country constitutes the "periphery." The core, which may cover a considerable portion of the national territory, is the nerve center of the nation. In addition to containing much of the country's industry and urban population, the core is generally the hub of the transportation network and the seat of the national government. Because of the significance of the core, disruption there can threaten the stability of the entire nation. Throughout the Third World rapid population growth in core regions is having just such a disruptive effect. Some of the large cities in developing nations are so crowded and polluted that it appears they have reached the limit of the carrying capacity of their environment. Moreover, rapid demographic growth in the core creates a demand for housing, transportation and sanitation that strains the budget of a poor country. The disproportionate government investment required to meet such needs and forestall chaos in the core can cause considerable unrest among residents of the periphery.

Policymakers in several developing countries are concerned that the disproportionate growth of the core has begun to pull their nations toward political or environmental disaster. Yet in non-Communist countries policies that could effectively restrain the growth of the core are not easy to come by. The reason is that the expansion of the core is largely a by-product of economic growth, which is itself desirable. In the process of economic development it is more efficient to invest in the core than it is to invest in the periphery. As a result both public and private investment tend to become concentrated in the core region. The relatively high standard of living there draws an increasing number of people from the countryside. In Communist countries, where the government has considerable control over the movements of individuals, the growth rate of the population in the core can be kept down. In most non-Communist countries, on the other hand, the flow toward the core slows only when economic production reaches a fairly high level and public investment can be dispersed throughout the country. Recent evidence suggests, however, that even at a fairly low level of economic development the movement toward the core can be restrained by government policies that improve life in the peripheral regions.

These patterns have emerged from a recent survey of 46 Third World countries. The aim of the survey was to find out whether the movement toward the core, which was quite rapid in the 1950's and 1960's, had slackened during the 1970's. The countries selected met two criteria. First, they had carried out a national census in 1980 or within a year or two of that date. Second, they possessed census data covering either all or most of the period since 1950. Together the countries in the survey include representatives of all the major regions of the Third World. The only significant omissions are in Africa, where many nations do not conduct regular national censuses.

The concept of the core is employed by geographers to explain economic and demographic trends over a nation's territory. Censuses are carried out according to administrative areas such as provinces and states rather than according to the distinction

P'YŎNGYANG, capital of North Korea, is a city whose growth has been carefully controlled by the country's Communist government. The photograph shows Chollima Street, one of the capital's main thoroughfares. The streets of P'Yŏngyang are uncrowded because there are few private automobiles in North Korea. The blocks of apartments at the rear and the park in the foreground are typical of contemporary P'Yŏngyang. The city, heavily damaged by U.S. bombing during the Korean conflict, has been rebuilt in a planned way and its growth has been strictly limited. In 1980 the estimated population of the P'Yŏngyang metropolitan area was 1.7 million.

SEOUL, capital of South Korea, has grown uncontrollably since World War II, as this photograph of a busy commercial distict suggests. The population of the Seoul metropolitan area is currently estimated to be 13.7 million. The major streets of Seoul are so crowded during the rush hour that streetcar service had to be eliminated in 1969. A subway system intended to replace the streetcars is not yet complete. Seoul faces shortages of housing, water, electricity and sewage facilities. Non-Communist governments of Third World nations where economic development is well under way have generally found it difficult to control the growth of their major cities.

between the core and the periphery. Therefore in analyzing census data an approximation of the core region must be constructed by combining administrative districts. In the current survey an approximation somewhat larger than the actual core region was constructed for each country. Since some people may move to the actual core from the fringes of the approximate area, the procedure could yield a slight underestimate of migration to the core. The method does, however, give a reasonably accurate picture of demographic trends there.

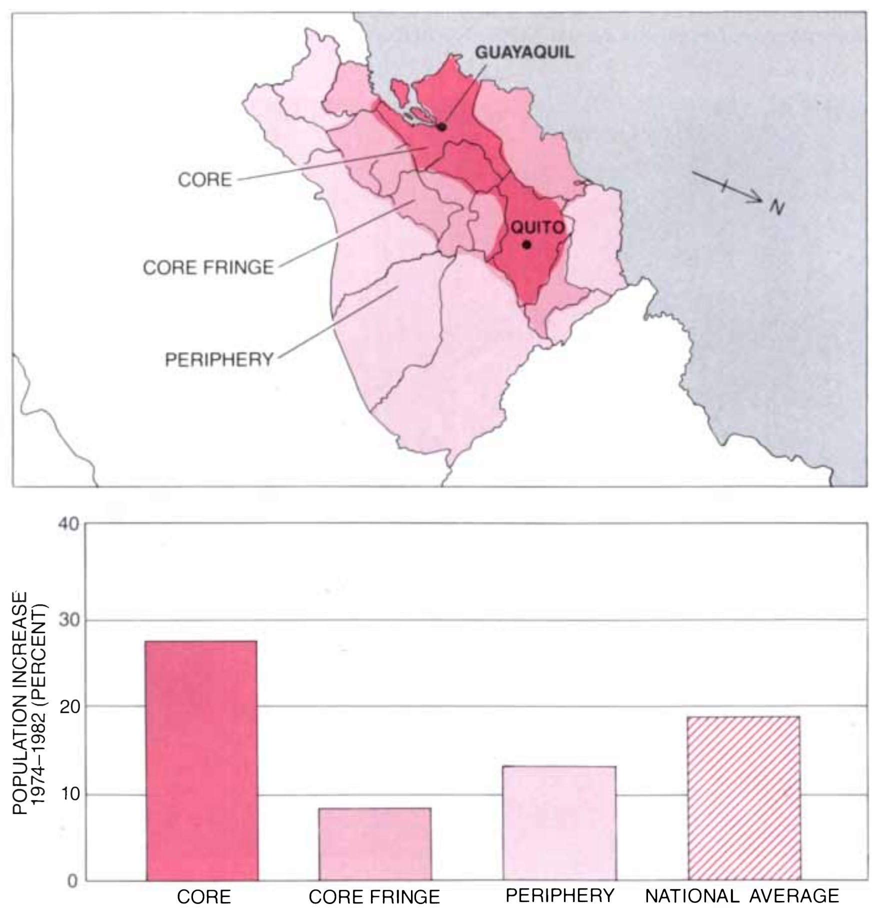
In a Third World nation with a well-established, dominant core region the population growth rate in the core typically exceeds the average rate for the entire nation by from 10 to 20 per 1,000 population per year. For example, the average growth rate in a Third World country might be such that for each 1,000 residents at the beginning of the year there are 1,030 at the end. In the core the corresponding increase would be from 1,000 to, say, 1,045. Such relative growth is not due to the fact that the rate of natural increase is higher in the core than in the periphery. In most developing countries the rate of natural increase is high in both the core and the periphery and is about equal in both areas. Rather than being due to a higher rate of natural increase, the relative growth of the core is due to migration from the hinterland to the big cities. Such migration can produce large changes in population concentration in a short time.

The results of the current survey suggest that the two factors with the greatest power to explain trends in population concentration in the developing nations are the level of economic development and whether the country has a Communist political structure or a non-Communist one. As we shall see, other factors have a role but these two account for much of the observed variation among the countries in the sample. Of the 46 nations, by far the largest group is made up of the non-Communist nations where economic development is well under way and the flow of migrants toward the core is stable or increasing.

In Latin America the regions that include and surround Mexico City, São Paolo, Santo Domingo, Panama City and Quito and Guayaquil all had stable or rising rates of population growth during the 1970's in relation to the national population growth rate. In South Asia, New Delhi, Bombay, Calcutta, Karachi and Dakha with their surrounding regions had higher net migration rates in the 1970's than in the preceding two decades. The main cities of East and Southeast Asia (apart from those of the Communist countries), including Seoul, Taipei, Manila, Jakarta, Bangkok and Kuala Lumpur, also had stable or rising net migration rates in the 1970's compared with their average rates in the postwar period.

Most of the core regions of North Africa and the Middle East also display increasing rates of relative growth. In particular the regions that include and surround Algiers, Tunis, Teheran and the two urban centers of Tripoli and Benghazi in Libya are growing rapidly in relation to the rest of the nation. On the other hand, in Turkey there was a reduction during the late 1970's in the rate of migration toward the major cities of Ankara and Istanbul. The decline was probably due to the political and social unrest that disrupted those cities in the second half of the past decade; it is likely to prove temporary.

The sketchiness of census data in much of sub-Saharan Africa makes it difficult to analyze the demographic relations between the core and the periphery. It is clear, however, that in the countries south of the Sahara the core is smaller (both as a fraction of the national population and in absolute size) than elsewhere in the Third World. The major exceptions are Lagos in Nigeria and Kinshasa in Zaire (for which there are few census data) and the metropolitan areas of South Africa, which are rapidly increasing their share of all four racial groups represented in that country. Because the core regions in sub-Saharan Africa are small, migration can yield a high rate of relative population growth. The rate of redistribution in sub-Saharan Africa in some instances exceeds 70 per 1,000 population per year, compared with the 10 to 20 that is typical of a developing country with a large, dominant core region. As the size of the core increases, the rate initially falls and then stabilizes at a lower level. Therefore the pace of redistribution may slow in the future. For the moment, however, the rates of movement toward the core in sub-Saharan Africa remain high.



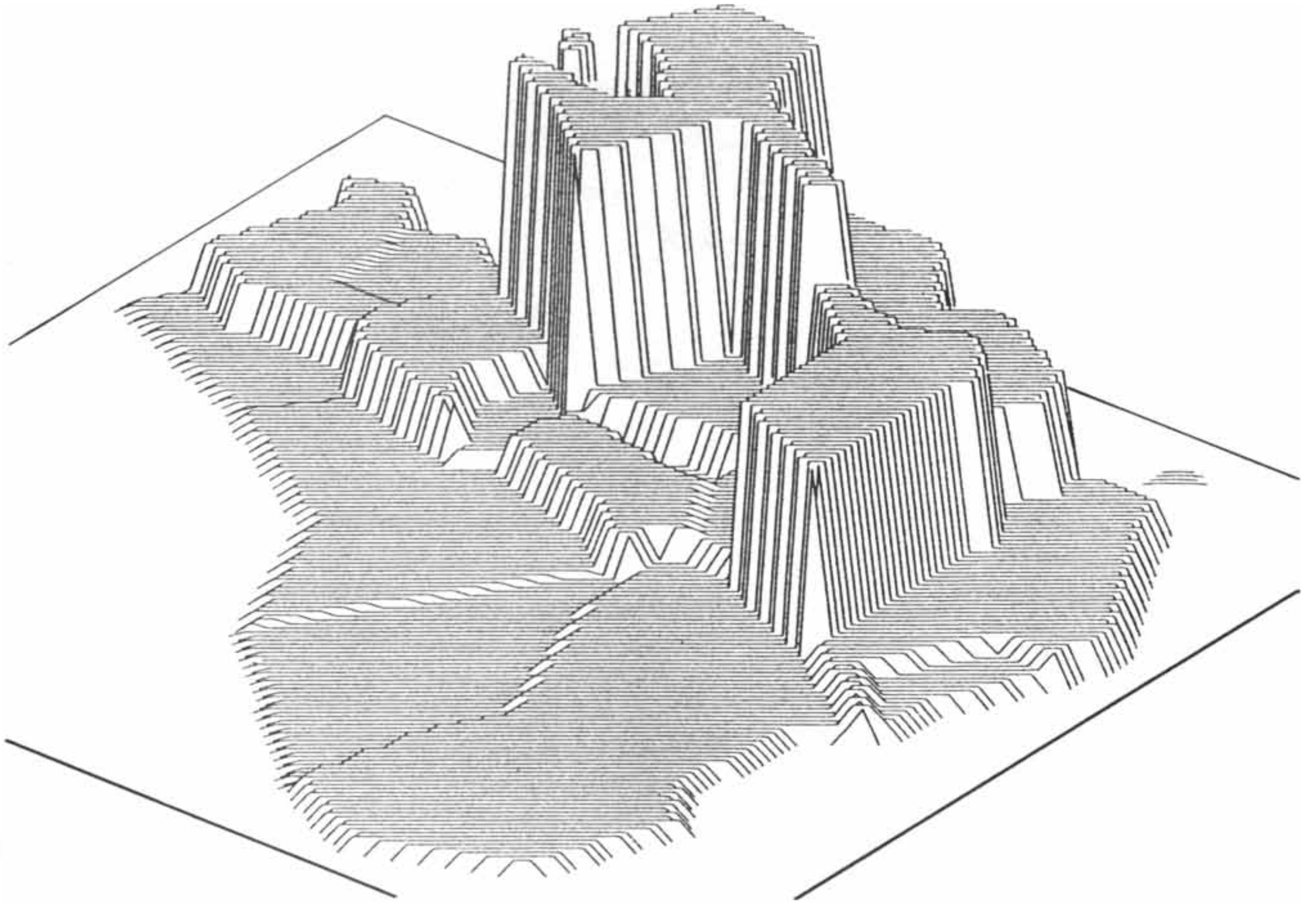
CORE REGION OF ECUADOR attracts population at a rapid rate. The map at the upper left shows the core, core fringe and periphery of Ecuador as defined by Richard W. Wilkie of the University of Massachusetts at Amherst. A nation's core region generally includes much of its industry and most of its major cities. In Ecuador the core includes Quito, the capital, and Guayaquil, the major port. The chart at the lower left shows the proportion by

Several demographic factors combine to intensify the adverse effects of population shifts toward the core in the Third World. One is the fact that in most Third World cores there are few cities. In the industrial countries the core region generally includes several large cities. In developing nations, however, the core frequently includes only one major city. Hence in a developing country almost all migration to the core swells the population of a single city. Furthermore, the movement toward the core in the developing nations is taking place while the overall rate of population growth is considerably higher than it was in Europe or the U.S. during the process of industrialization. As a result Third World countries face two separate but closely related problems: the proportion of the national population living in the core is increasing rapidly and the absolute size of the major cities is also reaching an unmanageable level.

Ecuador and Panama provide examples of countries where the proportion of the population in the core has increased dramatically in recent decades. In 1952 the regions of Ecuador that include Quito, the capital, and Guayaquil, the port, held 30 percent of the national population. In 1982 the fraction was 42 percent. If the current rate of redistribution persists, by the year 2000 more than half of the population will live in or near the two cities. Between 1950 and 1980 the fraction of Panama's population living in and around the capital increased from 31 to 46 percent. If current trends persist, by 1990 the proportion will be well over half.

The planning problems caused by the sharp increase in the absolute size of Third World cities is demonstrated by the core region of Indonesia: the sprawling metropolis called Jabotabek, which is centered on Jakarta. In 1961 the population of Jabotabek was 6.7 million. During the next 20 years the growth rate of Jabotabek was significantly higher than that of the country as a whole and by 1981 the metropolitan region had a population of more than 13 million. It has been estimated by the Ministry of Public Works of Indonesia that it will cost the Indonesian government $1.2 billion over the next 10 years to build a public transportation system for Jabotabek. That figure is almost seven times as much as the amount included in the national budget for all forms of public transportation throughout the entire country from 1984 through 1988. It has also been estimated that to provide an adequate water supply for Jakarta alone will cost an amount equal to 60 percent of the total public investment in water supplies throughout the country under the recent five-year government plan.

Although the cost of providing public services in the core is high, there is considerable pressure on the government of a developing nation to invest its capital there. Part of the pressure is clearly political; the political stability of the nation often depends on preventing unrest in the core. The rapid demographic growth of the core regions implies that large sums must be



which the population of the various regions and the nation as a whole increased from 1974 to 1982. The core is growing faster than Ecuador as a whole; the core fringe and the periphery are growing more slowly. On the computer-generated map at the right, which was made by Wilkie and John Montgomery, the height of each province corresponds to its population. The provinces containing Quito and Guayaquil dominate the population distribution. In countries such as Ecuador that are undergoing economic development the growth of the core region appears to be caused largely by the fact that both public and private investment are concentrated in the core.

spent to keep the core running smoothly, or at least to keep it from dissolving into complete political disorganization. Therefore schools, roads, public transportation and systems to provide communications, water and sanitation tend to be constructed in the core region before they are constructed in the periphery.

Moreover, the returns to public investment in the core appear to be even higher than the costs. Services such as transportation and communications make up the infrastructure of modern industry. The existence of the infrastructure in the core region constitutes a powerful attraction to a company planning to build a new industrial plant. Hence industry gravitates to the core. Once a nucleus of industrial companies exists in the core the advantages of building a new plant there increase still further. The additional advantages are due to the fact that it is more efficient to locate a plant where other companies are nearby to supply parts, markets, consultants and ancillary services. In addition the industrial nucleus is associated with a pool of relatively skilled labor.

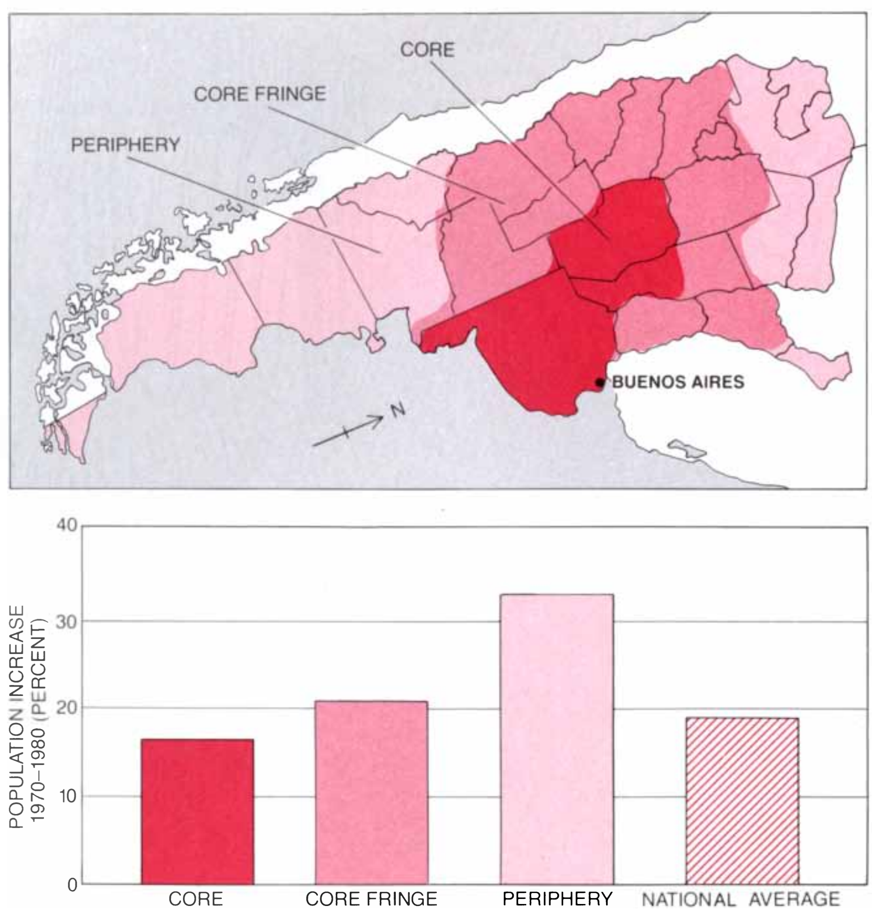
As a result of these advantages, which economic geographers call the "economies of agglomeration," a company choosing a site for a new plant will probably choose the core. When an industrial plant opens in the core, it brings with it new jobs. Industry is not the only source of work in the expanding core. The growing bureaucracy needed to administer public investment provides many jobs, as does the semilicit service economy found in almost all large Third World cities. Work in an industrial plant, in a bureaucratic office or in the black market offers a significantly higher standard of living than is available in the periphery with its sleepy peasant villages and slow-growing provincial cities. The superior standard of living has a powerful attraction for the people of the peripheral regions, who flock to the main cities.

If the relatively high standard of living prevailing in the core is the underlying force that drives the concentration of population, then the reduction of the differential in living standard between the core and the periphery ought to reduce the flow of population toward the core. There are at least two ways the differential could be reduced. The first is for the pace of economic growth in the core to slow and the second is for the standard of living in the periphery to be raised. Both patterns can be seen among the 46 nations in the survey.

Peru and Chile are countries where economic stagnation appears to have slowed the growth of the core. In the 1970's the relative growth rate of Lima and Santiago was about half the post–World War II average. The decline is probably best explained by the protracted economic stagnation that began in both countries in the early 1970's and is still going on. Economic stagnation, however, appears to retard population concentration only in countries such as Peru and Chile, which have a substantial industrial base. In countries where there is little industry the rate of migration to the major cities may be independent of the state of the economy. For example, the economies of countries in sub-Saharan Africa are not growing rapidly but migration to the cities continues at a high rate. A plausible explanation for the continued migration is that people are drawn to the cities because per capita food production is falling and food imported from abroad arrives in the urban centers. Since transportation between the core and the periphery in the sub-Saharan countries is poor, little imported food reaches the hinterland. As long as the agricultural base continues to decline the rapid migration will probably continue.

The example of the sub-Saharan countries suggests that the effect of overall economic growth on population concentration can be modified by other factors. Egypt is another country where factors other than overall economic development have a significant role. Egypt is at an intermediate level of development and a high rate of population concentration would be expected there, all other things being equal. Moreover, the pace of economic growth has probably accelerated over the postwar period. In recent years per capita income has been increasing at the rate of 4 percent per year. In spite of such economic expansion, however, the relative growth rate of Cairo declined sharply between 1966 and 1981.

In the case of Cairo the factor that modifies the effect of economic growth is probably population density. Between 1947 and 1981 the population of



PERIPHERAL REGIONS OF ARGENTINA are gaining population faster than the core region is. The map at the upper left shows the core, core fringe and periphery of Argentina. The core so defined includes Buenos Aires and more than half of the national population. The chart at the lower left shows the proportion by which the population of the three regions and the nation as a whole increased from 1970 to 1980. The population of the periph-

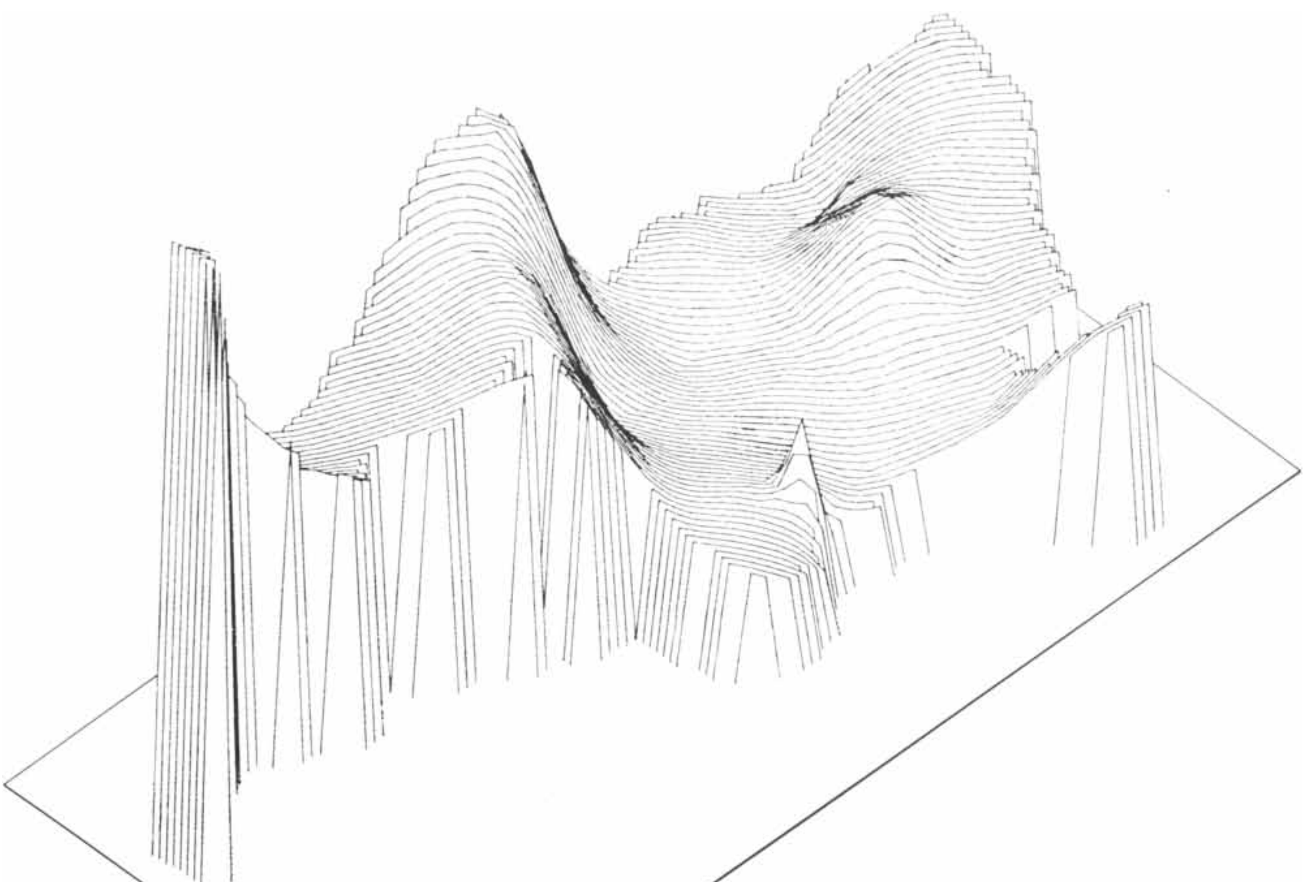
Cairo increased from 3.6 million to more than 10 million. Cairo has become one of the densest cities in the world. Indeed, the Egyptian capital has become so large and so dense that its size serves to reduce the advantages that accrue from locating an industrial plant in the core region. In particular, crowding, traffic congestion and environmental pollution have probably increased considerably the cost of doing business in Cairo.

Peru, Chile and Egypt are countries where the rate of population concentration has fallen because the attractive power of the core has been reduced. The rate of migration can also be reduced as a result of investments that raise the standard of living in the peripheral regions. Capital is scarce in the Third World and in most instances investment in the periphery is possible only at a fairly advanced stage of economic development. Only a few countries in the survey have achieved the level of development where substantial investment can be made in the periphery. They include Argentina, Venezuela, Greece, Spain and Ireland.

These five nations make up a group of "advanced" developing countries that fall between the Third World and the industrial nations of western Europe, Japan and the U.S. in economic production. In 1970 the gross domestic product (G.D.P.) per capita among the five countries, which is an excellent index of national economic production, ranged from $2,618 in Venezuela to $3,253 in Spain. (The figures have been computed in 1975 dollars to provide a control for the effects of inflation.) In comparison the G.D.P. per capita of Indonesia in 1970 was $370 in 1975 dollars and that of Ecuador $984. In all five of the relatively advanced developing countries there was a sharp decline after 1970 in the rate of redistribution toward the dominant urban areas (Buenos Aires, Caracas, Athens, Madrid and Barcelona and Dublin). The lesson of these countries for the rest of the developing world is clear: a sufficient condition for a decline in the net migration rate to the core is that the country exceed a threshold of total economic production. The threshold appears to be a per capita G.D.P. of roughly $3,000 in 1975 dollars.

The fundamental pattern of migration in a developing country does not change automatically when economic production reaches a certain level. The achievement of the economic threshold, however, enables the national government to make a significant investment in the peripheral regions. Roads, hospitals, schools, communications networks and sanitation systems are built in outlying areas where none of these services had existed. The provision of services raises the standard of living for residents of the periphery. Furthermore, the services form the basis of the infrastructure, which attracts industry to the peripheral regions. With jobs available close to home, the traffic-choked, smog-wrapped capital no longer looks so attractive to those who live on farms or in the provincial cities.

Thus one of the keys to reducing the rate of migration to the core is the



ery is growing faster than the national average and also faster than the population of the core. On the computer-generated map at the right the height of each area corresponds to the proportion by which the population there increased from 1970 to 1980; the map was made by Wilkie and Thomas Gallagher. Buenos Aires appears as the small spike at the lower margin of the map. The "bowl" around Buenos Aires shows how slowly the core is growing. Economically Argentina is one of the most advanced developing nations. In general among the non-Communist developing countries migration to the core decreases only when an economic threshold is reached.

dispersal of the industrial infrastructure throughout the country. The dispersal appears to be taking place relatively earlier in the course of economic development in the Third World than it did in western Europe, Japan and the U.S. In many of the industrial countries the rate of migration toward the core did not fall substantially until the 1970's, when economic production was already much greater than it currently is in any part of the Third World [see "Migration between the Core and the Periphery," by Daniel R. Vining, Jr.; SCIENTIFIC AMERICAN, December, 1982].

For example, it was not until 1970, when the per capita G.D.P. of Japan was more than $4,000 in 1975 dollars, that the rate of net migration to the core region stretching from Tokyo to Osaka began to fall sharply. The rate of migration to the cores of Denmark and Sweden began to fall at about the same time; per capita G.D.P. in those countries was then more than $5,000. In the developing nations, however, the threshold for a deceleration of the flow is only about $3,000 in 1975 dollars. Furthermore, the threshold currently seems to be falling still further. The most recent data from Taiwan show a rapid decline in the net migration rate to the Taipei region beginning in about 1980; in that year the per capita G.D.P. in Taiwan was approximately $2,500 in 1975 dollars.

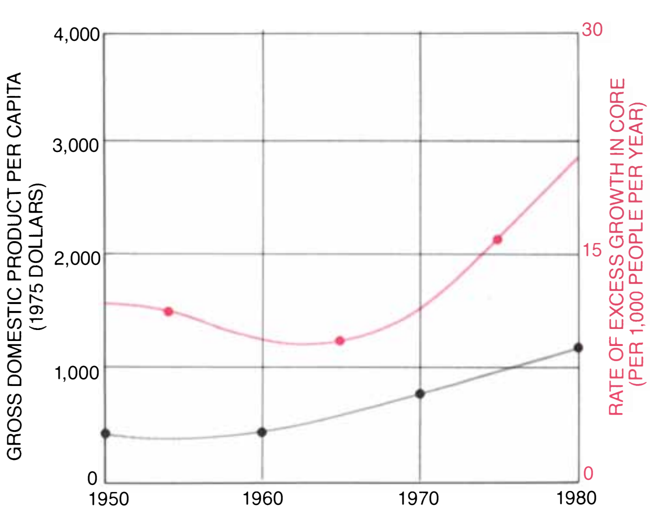
The reason the infrastructure can be dispersed through the periphery sooner in the Third World than it was in the industrial nations is that technological innovations have reduced the cost of constructing the systems that form the infrastructure. In particular the cost of building transportation and communications networks has decreased. Therefore national governments in the developing nations can build such systems at an earlier stage of development than the older industrialized nations could.

Although in most developing nations little public investment is made in the periphery until the economic threshold is reached, one country in the survey is an exception to the rule: Sri Lanka. The example of Sri Lanka shows that even at a fairly low level of economic development life in the peripheral regions can be improved considerably. In 1980 the per capita G.D.P. of Sri Lanka was only $838 in 1975 dollars. The government of Sri Lanka, however, has succeeded in providing virtually equal access to food, shelter, education and health care throughout the national territory. As a result Sri Lanka has an exceptionally high life expectancy, a high literacy rate, a low infant mortality rate and very little undernourishment compared with other countries at the same level of material production.

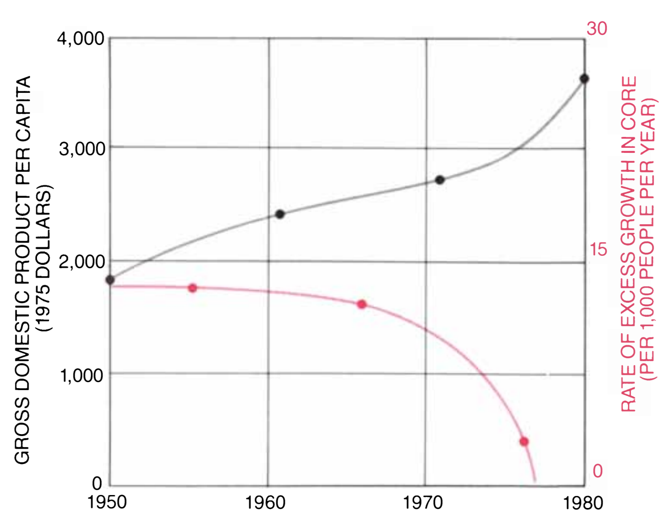
The equalization of access to fundamental services in Sri Lanka has lessened the disparity between the core and the periphery and apparently almost eliminated the incentives for interregional migration. Since the mid-1940's the proportion of the national population living in and around Colombo, the capital, has actually declined. In 1946, 21.3 percent of the population lived in the core region; by 1981 the proportion was 20.8 percent. In Sri Lanka, however, reduced economic growth may have been the price paid for reducing migration. Per capita economic production in Sri Lanka has increased quite slowly since 1945. The economy of Sri Lanka remains largely agricultural and the absence of a concentration of heavy industry has contributed to curbing migration to the core.

The example of Sri Lanka shows why population concentration is such a vexing problem. The concentration of investment in the core region, which causes the migrational flow, is the most efficient route to increased production. Investment can be diverted from the core only at the cost of retarding economic growth. Economic growth is, of course, essential to providing a better life for a country's citizenry. Hence the governments of Third World nations face difficult choices. To decrease the difficulty of the choices what is needed is a means of curbing the growth of the major urban centers without at the same time halting economic growth.

Communist countries face the same difficult choices, but they have for the most part chosen a solution different from that of their non-Communist counterparts: stringent policies aimed at limiting the growth of the largest cities. Such policies have worked. For example, Havana's share of the Cuban population is almost exactly what it was in 1943. Dean Forbes and Nigel



THAILAND

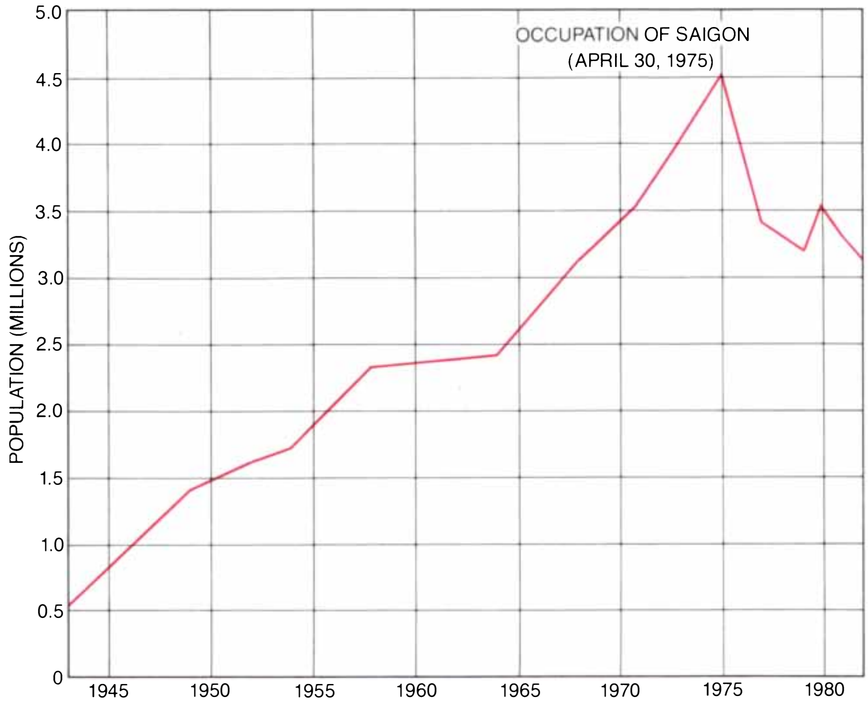


VENEZUELA

ECONOMIC THRESHOLD for a decrease of the flow of population to the core appears to be a gross domestic product (G.D.P.) of about $3,000 per capita (in 1975 dollars). In each graph the black curve shows the per capita G.D.P. The colored curve shows the amount by which the rate of population growth in the core exceeds the rate for the entire nation. In Thailand (*left*) economic development is at a fairly low level but is gathering momentum; the relative growth of the core is accelerating. In Venezuela (*right*) the threshold level of $3,000 per capita was reached during the 1970's; in the same decade the flow of population toward the core declined steeply. The core of Venezuela may now be growing more slowly than the nation as a whole. The achievement of an economic threshold makes it possible for a nation to increase the investment that is made in the periphery, which reduces the incentive for moving to the core.

Thrift of the Australian National University have shown that the government of Vietnam has reduced the population of Ho Chi Minh City (formerly Saigon) from 4.5 million in 1975 to 3.1 million in 1982. The rate of natural increase in Ho Chi Minh City from 1975 through 1982 was probably between 2 and 3 percent per year. Therefore the government had to remove about 8 percent of the city's population each year to achieve the ultimate reduction. In China the lowest rate of population growth is found in the great cities of Beijing and Shanghai and the areas surrounding them.

The virtually universal success of the Communist countries in controlling the growth of their core regions stems partly from the monopoly of power held by the state. A successful Communist revolution leads to the eradication of institutions such as churches and labor unions that in non-Communist countries compete with government authority. The population of a Communist state is particularly vulnerable to state penetration. Party members, who are widely dispersed, are responsible for passing information to the higher councils of the government. The existence of a spatially dispersed cadre who can monitor and report the movement of individuals enables the government to control overall population movement quite closely.



STATE POLICY can reverse the growth of major cities, as is shown by data from Ho Chi Minh City (formerly Saigon). Beginning in the 1940's the population of Saigon increased rapidly, reaching some 4.5 million in 1975. After the city was occupied by the North Vietnamese army on April 30, 1975, the population was forcibly reduced. In the early 1980's Ho Chi Minh City had only about three million residents. Almost all other Third World capitals are growing rapidly and the emptying of Ho Chi Minh City is a remarkable reversal.

In some Communist nations the authority of the party apparatus has been supplemented by the provision of public services outside the core region; in this respect the Communist countries resemble Sri Lanka. By combining public services in the periphery with restrictions on movement toward the core, Communist states have created a policy that includes both rewards and stringent control. Perhaps the most disturbing example of the stringent control exerted by a powerful state is Kampuchea (then Cambodia). After the triumph of the Khmer Rouge in 1975 both the national capital, Pnompenh, and the provincial capitals were emptied by force. About three million people, or half of the nation's population, were dispersed to work in the countryside. It is estimated that from one to two million Kampucheans, or roughly 20 percent of the nation, died during the process of population dispersal.

The existence of a stringent government policy aimed at controlling movement to the core does not in itself guarantee demographic equilibrium. In this respect it is helpful to compare Vietnam and Kampuchea with Indonesia. In 1970 the governor of Jakarta declared the capital a "closed" city. In the decade that followed, however, the rate of redistribution to the Jakarta region was the highest ever recorded. Rather than stemming the growth of the core, the policy of closing Jakarta merely diverted demographic growth to the adjoining districts (if it had any effect at all). In the 1970's the suburbs of Jakarta grew very quickly, mirroring the pattern around Mexico City, Seoul, São Paolo, Taipei, Bombay and Manila. Unlike the Communist governments with their disciplined cadres, non-Communist states generally lack a powerful local presence in the territory they administer. Hence the movement of population is largely beyond control.

What lessons can be learned from the current survey of population concentration in the Third World? The first general conclusion is that most of the great metropolitan regions of the developing countries continued to attract population during the 1970's at the same high rates that prevailed during the preceding two decades. The continuation of migration to the core regions suggests that the advantages of concentrating industry and population are still considerable. There are enough countries where the pace of concentration has slowed, however, for an observer to draw some conclusions about the conditions under which the growth of the core can be restrained.

Migration toward the core can be limited by economic depression, as in the case of Peru and Chile, or by stringent measures imposed by a powerful state, as in the case of Vietnam and Kampuchea. More interesting for policymakers are those cases where the flow has been checked by improving the life of those who live in the periphery. In most instances such amelioration takes place only after the country has attained a threshold of economic production. The case of Sri Lanka, however, shows that even a poor country can improve the quality of life in the periphery. Sri Lanka's experience also indicates that such an improvement may entail a substantial sacrifice of economic growth. The fundamental lesson to be derived from the current survey is that population concentration appears to be an almost inevitable concomitant of economic growth in the non-Communist world. Some of the adverse effects of rapid concentration can be relieved by a government that pays attention to the vast stretches of country outside the brightly lighted capital, but the choices are by no means easy or without cost.

The Earth's Hot Spots

These plumes of hot rock welling up from deep in the mantle are a key link in the plate-tectonic cycle. The marks they leave on passing plates include volcanoes, swells and midocean plateaus

by Gregory E. Vink, W. Jason Morgan and Peter R. Vogt

From deep inside the earth's mantle isolated, slender columns of hot rock rise slowly toward the surface, lifting the crust and forming volcanoes. The plumes well up all over the world, under continents and oceans, both in the center of the mobile plates that make up the earth's outer shell and at the midocean ridges where two plates spread apart. The marks they leave at the surface are superposed on the grand effects of plate motion. Volcanic eruptions and earthquakes associated with plumes occur far from plate boundaries, the site of most such activity; the upwelling currents also form broad anomalous swells in the ocean floor and in continental terrain. These isolated areas of geologic activity are called hot spots.

Mantle plumes are relatively stationary, and so the crustal plates drift over them. Often the passage of a plate over a hot spot results in a trail of identifiable surface features whose linear trend reveals the direction in which the plate is moving. If the plate is oceanic, the hot-spot track may be a continuous volcanic ridge or a chain of volcanic islands and seamounts rising high above the surrounding sea floor. The most prominent example is the Hawaiian Islands; it was a visit there that led J. Tuzo Wilson of the University of Toronto to put forward the concept of hot spots in 1963.

Wilson noticed that to the west of Hawaii the islands disappear into atolls and shoals, indicating they are progressively more eroded and therefore older. The same observation had been made more than a century earlier by the American geologist James Dwight Dana, but Wilson was the first to interpret the age progression as evidence of continental drift. He proposed that the island chain had been formed by the westward motion of a crustal slab over "a jetstream of lava" now situated under Hawaii itself, at the eastern end of the chain. The proposal came at a time when textbooks, including one coauthored just three years earlier by Wilson himself, mentioned continental drift only as an intriguing idea advanced in the 1920's but later discredited.

In the past two decades the idea has become generally accepted as part of the theory of plate tectonics. The earth's crust is now known to be embedded in the rigid plates of the lithosphere, which is between 100 and 150 kilometers thick under continents and about half as thick under oceans; the continual motion of the plates over the partially molten asthenosphere (the portion of the mantle extending to a depth of roughly 200 kilometers) explains the development of ocean basins and the formation of mountain ranges. A major task of contemporary geophysics is to understand how these surface processes are related to the slow convective "creep" of hot rock in the underlying mantle. Hot spots are an important part of this connection.

Indeed, if the upwelling plumes were to stop, the plates would grind to a halt. Ultimately the energy that drives plate motion is the heat released by the decay of radioactive elements deep in the mantle. The plumes provide an efficient way of channeling the heat toward the surface. Their efficiency is

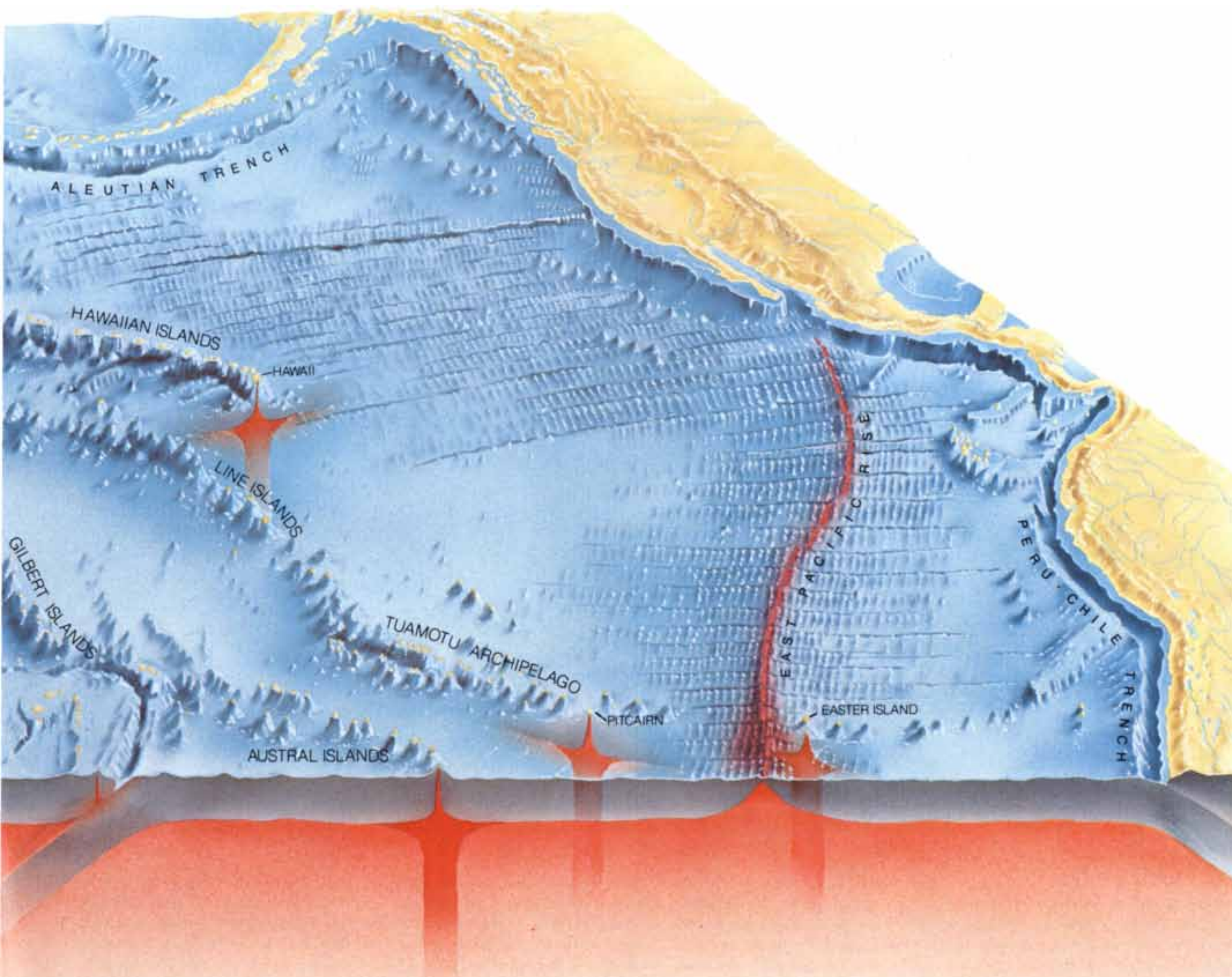
MOTION OF THE PACIFIC PLATE over three fixed mantle plumes has produced three parallel island chains: the Hawaiian Islands and Emperor Seamounts, the Tuamotu and Line islands, and the Austral, Gilbert and Marshall islands. The chains lie in the center of the plate, proving they were formed by a mechanism different from the one that produced the volcanic island arcs of the western Pacific, which are associated with the subduction of the plate at oceanic trenches. The plumes originate deep in the mantle, and their surface tracks reveal the path of the plates. About 40 million years ago the Pacific plate switched to its present westward course from a more northerly heading; the change shows up as a bend in the hot-spot chains. Active volcanoes, such as Kilauea on Hawaii, are at the southeastern end of the chains. To the northwest the volcanoes are extinct and progressively older.

attributable to a property of mantle rock: its viscosity, or resistance to flow, is reduced dramatically by relatively small increases in temperature (say 100 degrees Celsius) or in the content of volatile elements such as water. Less viscous material produced by variations in temperature or volatile content tends to collect and rise toward the surface through a few narrow conduits, much as oil in an underground reservoir rises through a few boreholes.

It would be misleading, however, to say that the plumes propel the plates. Rather, the two are different parts of the same convective cycle. As plates spread apart at a midocean ridge, molten rock from the asthenosphere wells up at the spreading axis to form oceanic crust; the new lithosphere cools as it moves away from the ridge and is eventually destroyed at oceanic trenches, where two plates collide and one of them sinks deep into the mantle. The deep mantle feeds the plumes. They in turn empty matter heated by radioactivity into the asthenosphere, which in addition to serving as the source of new sea floor provides a hot and fluid layer for the plates to glide across. The asthenosphere is constantly being destroyed as it cools and attaches to the base of the lithosphere; the boundary between the two layers is essentially a thermal one. Were it not replenished by the plumes, the asthenosphere would soon vanish, and the motion of the plates would stop.

It is worth emphasizing that this "plume model" of the convective circulation in the mantle is just that: a model. The plumes have not been observed directly. The deep mantle can be explored only through the analysis of earthquake waves, and so far the resolution of seismic studies has not been good enough to detect plumes; the upwelling currents may be just a few hundred kilometers in diameter and only moderately different from their surroundings in temperature and density (the properties that determine the seismic-wave velocity in a region).

The indirect evidence for deep-mantle plumes, however, is substantial. Satellite measurements of the earth's gravity field have shown hot spots to be areas of anomalously high gravity and thus of excess mass; the excess mass can be attributed to broad bulges in the surface produced by the upwelling plumes. A second line of evidence comes from geochemical studies of basalts erupted at hot-spot volcanoes. Compared with the basalts dredged from midocean ridges, these rocks are enriched in volatile elements and in other elements such as potassium that are "incompatible" with the crystals of ordinary mantle rock. They also contain anomalous amounts of isotopes derived from radioactive decay processes. The differences in composition suggest that hot-spot lavas are derived from rock welling up from below the



asthenosphere, which feeds the oceanic spreading centers. According to the plume model, as material from the deep mantle flows into the asthenosphere, the part rich in volatiles and other incompatible elements melts, and some of it rises to the surface at hot-spot volcanoes.

Recent advances in seismology encourage the hope that someday workers will observe the plumes directly [see "Seismic Tomography," by Don L. Anderson and Adam M. Dziewonski; SCIENTIFIC AMERICAN, October, 1984]. In particular, a proposed new global network of seismometers may improve the resolution of seismic studies to the point where it is possible to determine the size of plumes and the depth of their roots.

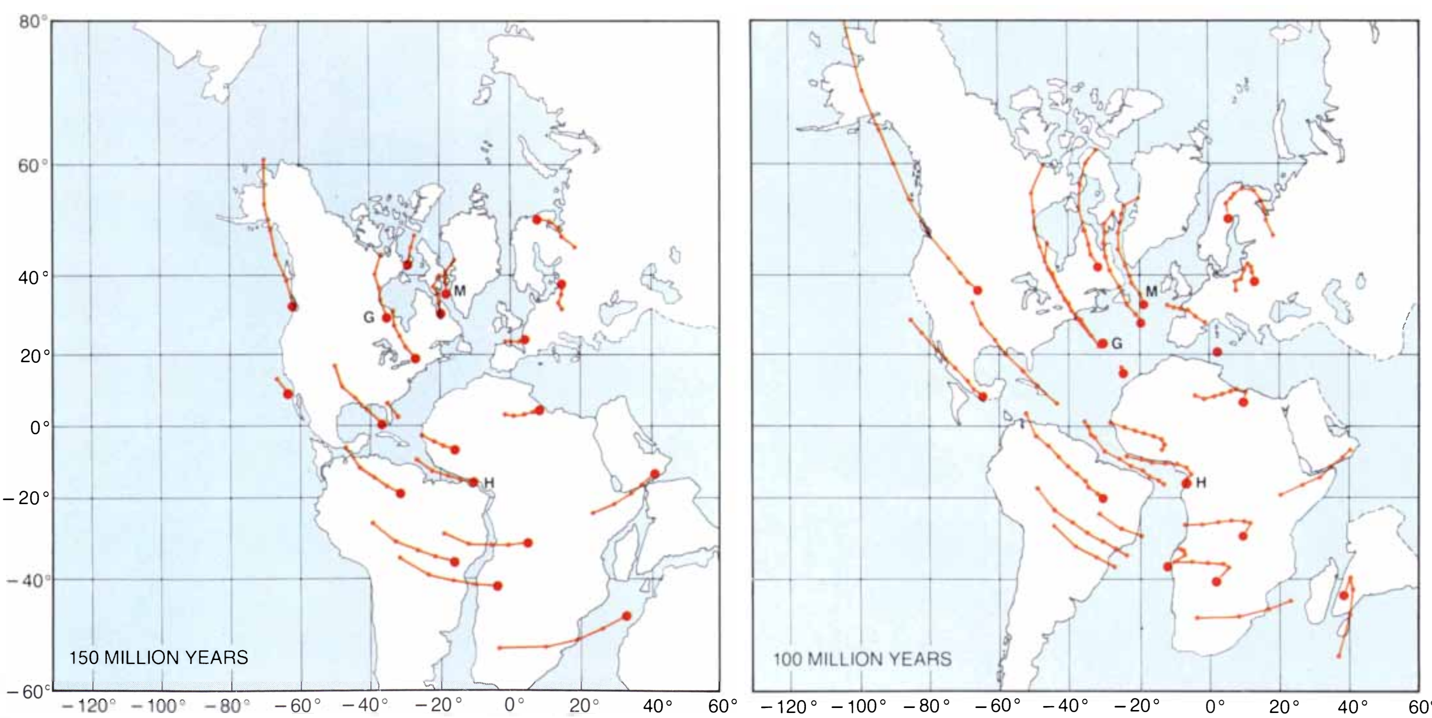
The plumes are certainly not uniform; differences in their isotope signatures imply that they come from various depths. Comparisons of the volume and frequency of eruptions at different hot spots indicate they also come in a range of sizes. Furthermore, individual plumes are not immutable. After examining the volume of rock extruded along the Hawaiian hot-spot track, one of us (Vogt) has suggested that the discharge rate of a plume may vary over time. Geochemical evidence supports the conclusion. Jean-Guy E. Schilling of the University of Rhode Island has proposed that plumes consist of rock rising in blobs rather than in a continuous flow.

Sometimes a hot spot may fade away entirely, and new ones may be formed; from the tracks it appears the typical life span of a plume is on the order of 100 million years. Moreover, the position of a hot spot seems to change slightly. As a result the tracks on the surface are not all as neatly linear as the Hawaiian chain.

Compared with the plates, however, the mantle plumes are relatively stationary. The first evidence of their fixity came in 1970. One of us (Morgan) showed that three volcanic island groups in the Pacific—the Hawaiian Island–Emperor Seamount chain, the Tuamotu Archipelago–Line Island chain and the chain formed by the Austral, Gilbert and Marshall islands—are approximately parallel and could all have been formed by the same motion of the Pacific plate over three fixed hot spots. In each case the most recent volcanic activity has taken place near the southeastern end of the chain, and the islands and seamounts get progressively older to the northwest. The Pacific plate is currently moving toward the northwest; it switched to that course from a more northerly heading about 40 million years ago. The course change shows up as a bend in the hot-spot tracks.

Because the motion of the hot spots is insignificant, they provide a worldwide reference frame for tracing the absolute motions of the plates with respect to the earth's interior. For some time workers have mapped the paths of the plates in relation to one another and have thereby been able to reconstruct the opening of ocean basins. The boundaries between plates—the ridges and trenches—also move, however, and so the relative motions do not reveal where on the globe a plate was at a given time. Nor do they indicate whether two diverging plates have been moving at the same speed, or whether instead one plate has remained stationary. Such questions can be answered by converting the known relative motions into absolute motions in the hot-spot reference frame, in which each hot spot occupies an unchanging latitude and longitude.

The relative motion of diverging plates—the sea-floor-spreading history—is determined by analyzing magnetic anomalies in the sea floor. Throughout geologic history, at regular intervals averaging about 100,000 years, the earth's magnetic field has reversed its polarity, for reasons that are poorly understood. A record of these reversals is preserved in the oceanic crust.



HOT-SPOT TRACKS reveal how the plates have moved with respect to the earth's interior during the opening of the Atlantic Ocean. Because the hot spots (*large dots*) are anchored deep in the mantle, they remain relatively fixed; that is, their latitude and longitude remain unchanged. The tracks consist of extinct volcanoes, magma intrusions and swells in the crust formed by the upwelling plumes and then carried away by the plates. Each small dot represents 10 million years of plate motion. In reconstructing the plate motions one begins with one or two well-defined tracks, such as that of the Great Meteor hot spot (*G*), which also formed the New England Seamounts and magma intrusions in the White Mountains. The tracks of other hot spots are then calculated from the recon-

The magnetic minerals in lava erupting from midocean ridges align themselves with the prevailing field, and as the molten rock cools and solidifies, the field direction is permanently locked in the crust.

The magnetized crust is transported away by the diverging plates in bands that roughly parallel the ridge axis. Each band has a characteristic magnetic anomaly and is made up of crust formed at the same time, and so the bands are called magnetic isochrons. The age of various isochrons, and therefore the sea-floor-spreading rate, has been established through radiometric dating of rocks retrieved in deep-sea drilling expeditions. By superposing corresponding isochrons from opposite sides of the spreading axis, one can reconstruct the relative position of the plates at the time the isochron pair was formed. (The superposition in effect removes from the map all sea floor created after the particular magnetic reversal.)

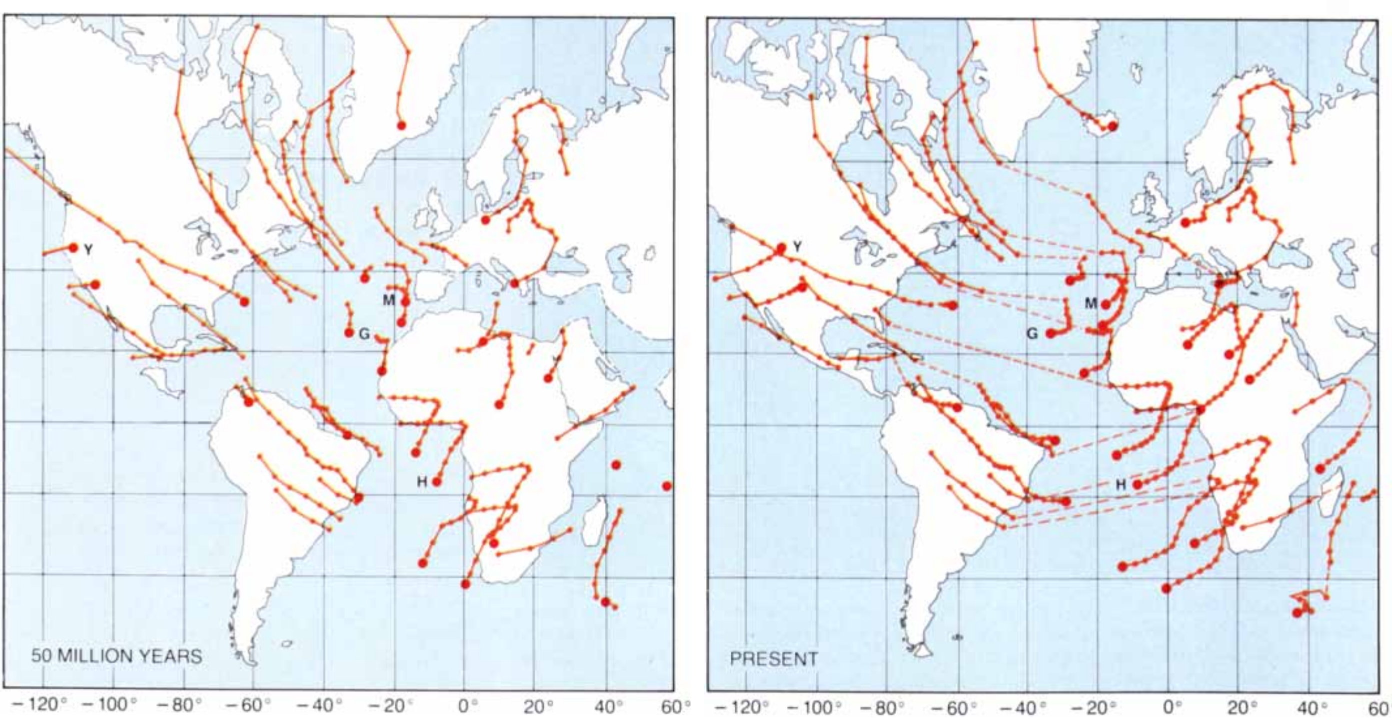
If the motion of one of the plates over the plumes is known, then their relative motion allows the path of other plates in the hot-spot reference frame to be deduced. The general procedure is to begin with a well-defined hot-spot track on one plate—say a chain of seamounts—and then adjust the more ambiguous tracks until the "best fit" is achieved: the absolute plate motions that best satisfy the constraints established by the hot-spot evidence and the relative motions.

Using this procedure, we have reconstructed the opening of the Atlantic and Indian oceans. The reconstructions can be tested: surface features along the hot-spot tracks must by their nature and age fit the hypothesis that they were formed by the passage of a plate over an upwelling plume. This should be true not only along the well-defined portions of the tracks but also in regions where the tracks have simply been extrapolated from the calculated plate motions and evidence of hot-spot activity has not previously been observed.

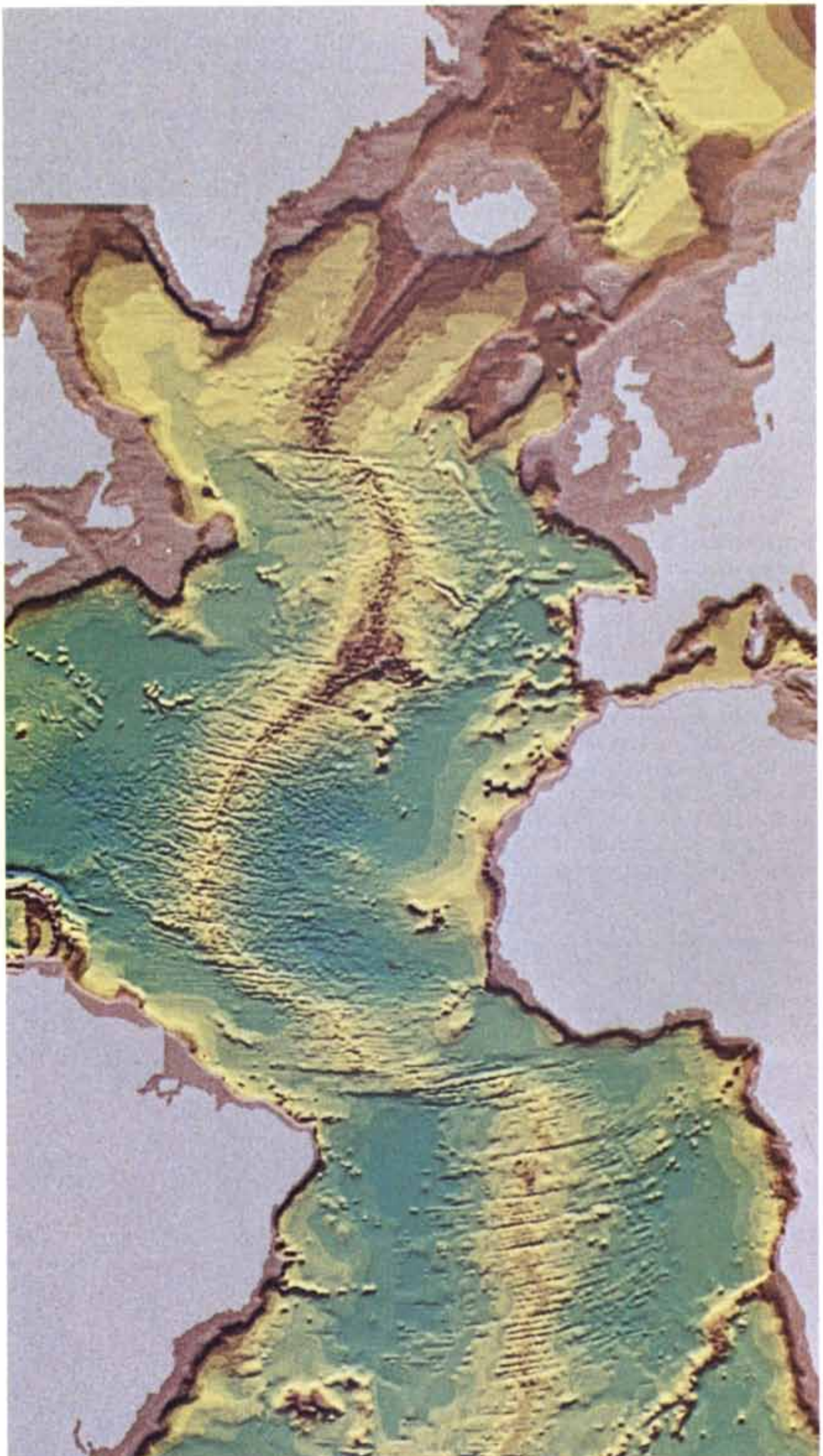
Although the available data are fragmentary (particularly concerning the ages of sea-floor features), in general the reconstructions pass the test. A good example is the track of the hot spot that formed the Great Meteor Seamount south of the Azores [*see illustration on these two pages*]. Two hundred million years ago the area northwest of Hudson Bay on the Arctic Circle was over the Great Meteor plume; 50 million years later the hot spot was under Ontario. The exposure of the Canadian shield from Manitoba to Ontario can be attributed to uplifting of the crust by the plume: in an uplifted area sediment covering the basement rocks is more likely to be eroded away over time.

One hundred million years ago the track had reached the young and narrow Atlantic off Cape Cod. The passage of New Hampshire over the hot spot is recorded by magma intrusions in the metamorphic rock of the White Mountains; the intrusions are between 100 and 124 million years old. For the period from about 100 to 80 million years ago the track follows the trend of the New England Seamounts. Based on radiometric dating of rocks collected from the seamounts, Robert A. Duncan of Oregon State University has shown that the volcanoes get progressively younger toward the southeast along the chain. Their ages coincide with their passage over the hot spot. From the ages and the distances between the seamounts Duncan has calculated the velocity of the North American plate during that period: about 4.7 centimeters per year.

Approximately 80 million years ago the Mid-Atlantic Ridge migrated westward over the plume. The track continues on the African plate and ends at the Great Meteor Seamount. At present the hot spot should be about 500 kilometers southwest of Great Meteor. Although there is a swell in that region



structions, which must fit the relative plate motions derived from sea-floor-spreading history. When the midocean ridge separating two plates drifts over a plume, the track continues on the other plate but is interrupted (*broken lines*) by sea floor formed at the ridge after it passed over the hot spot. A plate motion is a rotation, and so the tracks approximate concentric circles rather than parallel straight lines. Along the Madeira (*M*) and St. Helena (*H*) tracks continents have later rifted apart; the plumes may promote rifting by thinning a passing plate. The Snake River plain, where the lithosphere has been weakened by the track of the Yellowstone hot spot (*Y*), may be the site of a future rift. Not all hot spots are present in each reconstruction because new ones form and old ones fade away.



MID-ATLANTIC RIDGE is at present positioned over several hot spots; the flow from these plumes adds to the normal upwelling of magma at the ridge, producing thicker crust. In the computer-plotted topographical map brown regions are shallow and green regions are deep. Iceland is perched on the ridge axis and also has a large hot spot under its southeast coast; the plume has raised the crust above sea level by lifting and thickening it. The tapered structure of the ridge segment south of Iceland, called the Reykjanes Ridge, reflects the flow of plume material down the axis. Similar topography southwest of the Azores suggests material from that hot spot is also flowing along the ridge. The Iceland hot spot may have formed the plateau southeast of Iceland (including the Faeroe Islands) by feeding a now extinct spreading axis at the center of the plateau. William F. Haxby of Columbia University's Lamont-Doherty Geological Observatory prepared the map from data compiled by Joseph E. Gilg and Roger Van Wyckhouse of the U.S. Naval Oceanographic Office.

of the sea floor, there is no sign of current volcanism; the plume may have become inactive.

A swell in the ocean floor, like an exposed continental shield, is an area of uplifted crust. Some time ago Robert S. Detrick and S. T. Crough, then at the University of Rhode Island, proposed that a plume produces uplift not by bending the lithosphere but by thinning it, replacing cold, dense lithosphere with hot, buoyant rock from the asthenosphere. After passing over an active hot spot, both sea-floor and continental swells presumably cool and gradually sink back to their former altitude. Swells on the sea floor are interruptions of the process in which the lithosphere cools, thickens and sinks as it moves away from a midocean ridge, eventually plunging into the asthenosphere at a trench.

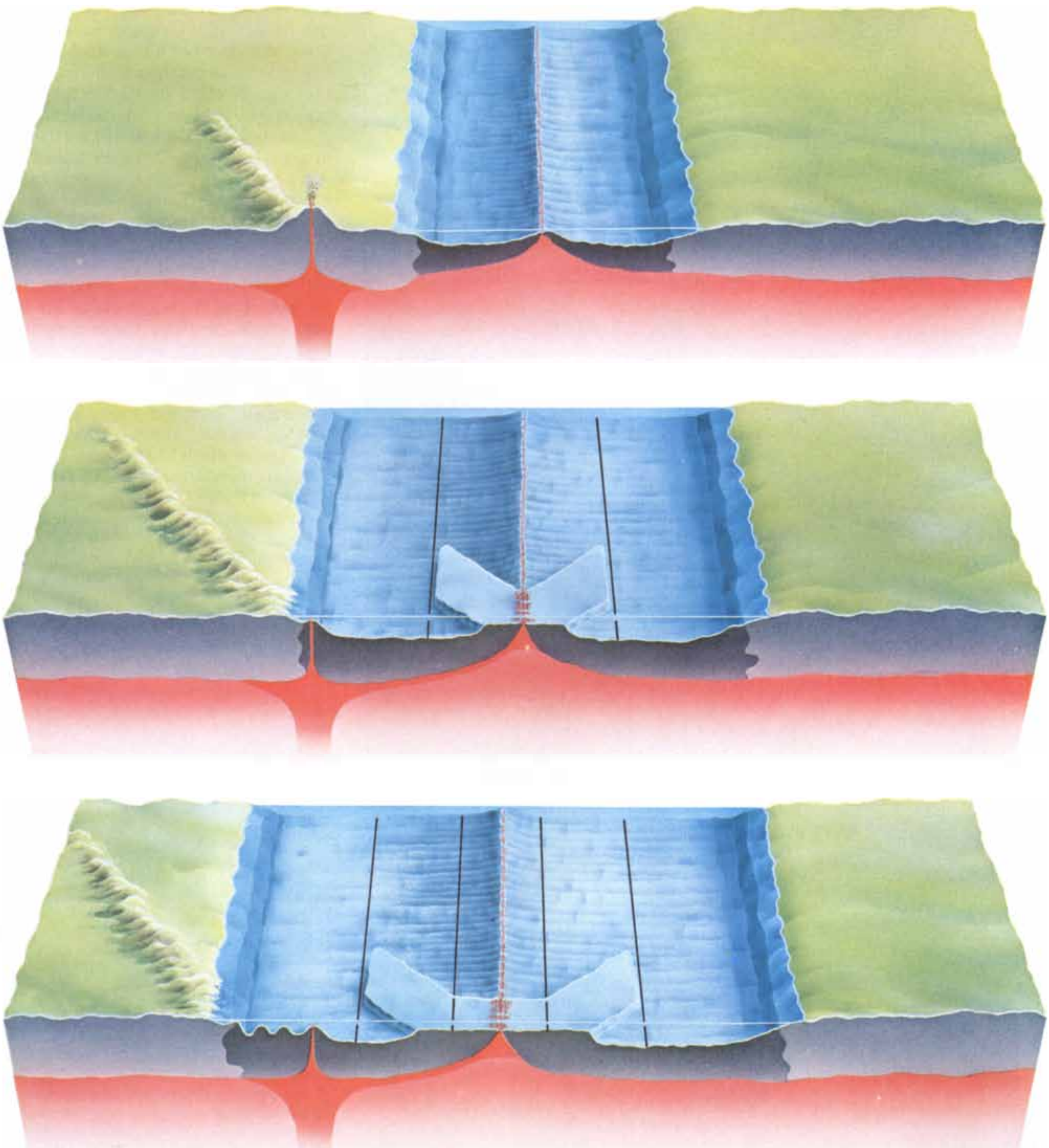
The hot-spot anomalies, however, are by no means insignificant interruptions. There are some 40 active hot spots, and the swells associated with them have an average diameter of about 1,200 kilometers. Thus swells cover roughly 10 percent of the earth's surface. This observation led Crough and Richard Heestand of Princeton University to suggest that the depth of the sea floor in a particular region is controlled not only by the progressive cooling of the lithosphere but also by the time elapsed since the region passed over a hot spot.

In the same way hot spots could control the thickness of the continental lithosphere. Moreover, the thinning and weakening of continental plates by mantle plumes may produce more dramatic effects than the exposure of basement rock: it may cause them to rift apart. In the early 1970's Kevin C. Burke of the State University of New York at Albany noticed that some hot spots are associated with three-arm rift systems, in which two of the arms have formed a plate boundary whereas the third has failed. The failed rifts form valleys extending into the continents; an example is the Niger River valley.

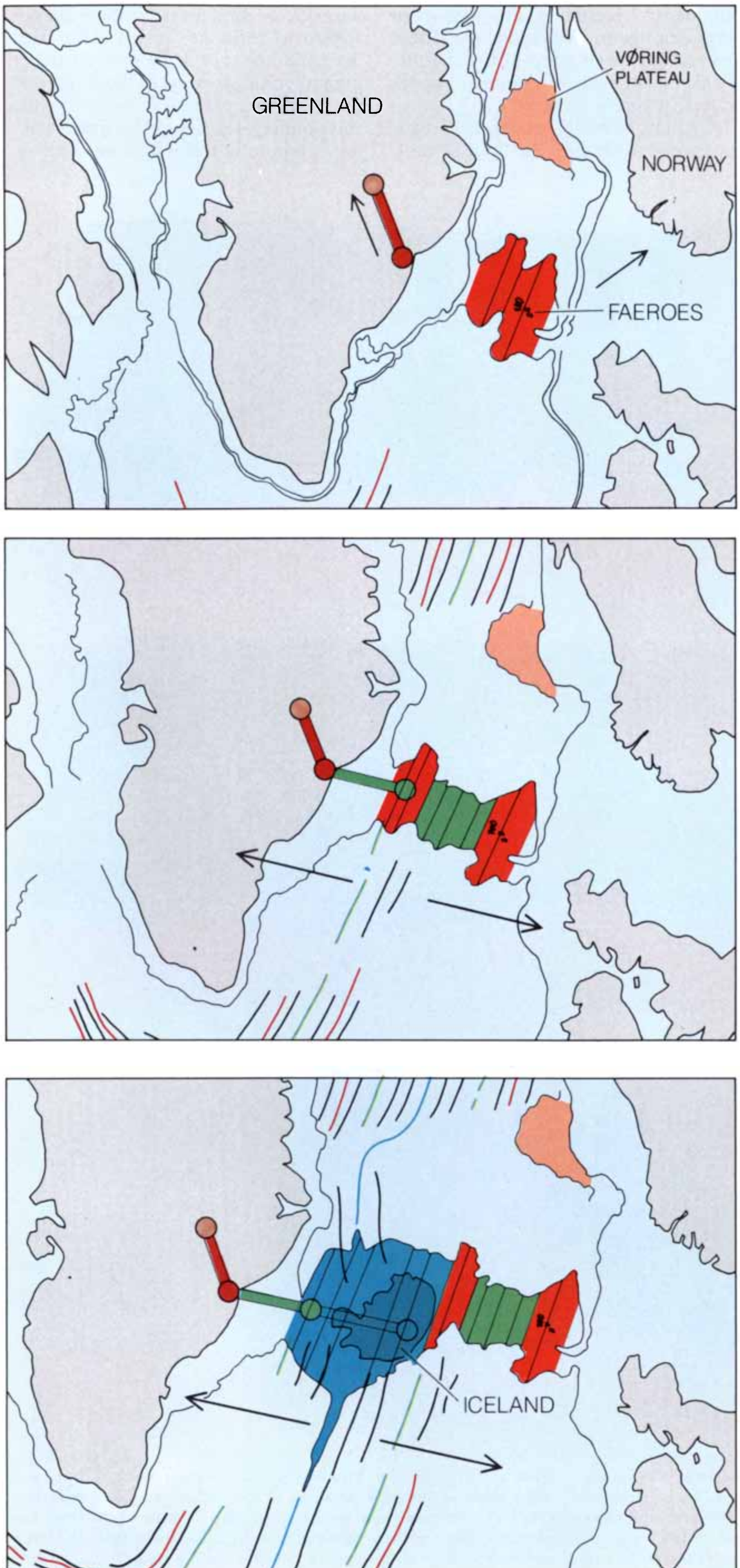
The reconstructions of the Atlantic opening reveal a number of hot-spot tracks along which continents have subsequently broken up, probably millions of years after the plates passed over the plumes. The track of the hot spot that formed the Madeira Islands, for example, runs between the west coast of Greenland and the east coast of Baffin Island and Labrador; the plume that created St. Helena can be traced along the south coast of West Africa and the north coast of Brazil. In the future a rift may develop in the Snake River plain, where the North American plate has been weakened by the track of the hot spot now under Yellowstone National Park.

Mantle plumes explain much of the geologic activity in the center of the plates. As the plates move over the hot spots, however, so do the plate boundaries, including the midocean ridges; unlike the hot spots, the ridges are not anchored deep in the mantle. What happens when a plume is under or near a spreading axis?

A plume directly under a spreading center augments the flow of molten rock welling up from the asthenosphere to form new crust. The crust over the hot spot is therefore thicker than it is along the rest of the ridge, and the result is a plateau rising above the surrounding sea floor. The most striking example is Iceland, a hot-spot is-



HOT SPOT MAY FEED A RIDGE from a distance, thickening the crust and forming an oceanic plateau. Early in the opening of the ocean basin (*top*) the hot spot is under a thick continental plate moving to the northwest; material from the plume cannot yet reach the spreading center. Millions of years later (*middle*) the motion of the plates has brought the ridge closer and has carried continental shelf over the hot spot. Plume material has begun to flow along the lithosphere to the nearest section of the rise. As the excess material erupts it is carried away on the plates; the V shape of the resulting plateau reflects both the spreading of the plates away from the ridge and their motion with respect to the hot spot. A change in plate motion (*bottom*) forms a bend in the hot-spot track and in the plateau.



land that straddles the Mid-Atlantic Ridge: there the upwelling is so intense and the crust so exceptionally thick that the plateau is above sea level. Geochemically the Icelandic crust is distinctly different from typical oceanic crust; it shows clear evidence of a hot-spot contribution. Gravity measurements indicate that the core of the plume is under the southeastern part of the island. The volcanic peaks there are visible signs of a powerful upwelling current: as much as 5,500 feet high, they are covered by the Vatnajökull glacier. (In 1918 an eruption under the glacier unleashed a flood of meltwater at a discharge rate 20 times that of the Amazon River.)

Some of the material in the strong Iceland plume also seems to spread out under the lithosphere. The lithosphere slopes upward toward a spreading axis, and one of us (Vogt) has proposed that the axis north and south of Iceland has acted as a pipeline, channeling partially molten rock away from the hot spot. In both directions along the ridge the excess plume material produces abnormally elevated topography out to a distance of about 1,500 kilometers. To the south of Iceland the broad plateau tapers to form the typical Mid-Atlantic Ridge. The tapered structure probably arises from the fact that most of the volatile-rich, easily melted plume rock is used up near Iceland. Indeed, Schilling has found that the chemical composition of basalts dredged from the ridge becomes progressively more like "normal" oceanic crust with increasing distance from Iceland, suggesting that the relative contribution of the hot spot gradually declines.

On the flanks of the ridge south of Iceland there are symmetrical pairs of secondary ridges. Each pair forms a southward-pointed V whose apex is on the spreading axis. These features could have been produced by "waves" of intensified flux or of unusually hot

GREENLAND-FAEROE Plateau might have been formed by the Iceland hot spot. The parallel colored lines are magnetic isochrons used to reconstruct past positions of the plates. Fifty million years ago (*top*) the hot spot was under the coast of Greenland and began feeding the ridge. The V shape of the plateau reflects the hot-spot track. By 36 million years ago (*middle*) the plates had changed course; the change is reflected in the new section of the plateau. At about that time the spreading axis moved west over the hot spot, which by then was under oceanic lithosphere. Hot-spot-fed spreading has continued in the west until now (*bottom*), forming Iceland. At an earlier time when it was close to a northern ridge the plume may have built the Vøring Plateau.

and buoyant material from the plume. A wave traveling down the ridge would generate anomalously thick crust, affecting the area nearest the hot spot first. The elevated crust would then be carried away on each side of the axis by the spreading plates, forming the V-shaped secondary ridges. From the known spreading rate and the angle between the secondary ridges and the spreading axis one can estimate the speed of the plume material; it seems to flow down the axis at a rate of five to 20 centimeters per year.

Because the midocean ridges move, a hot spot is unlikely to be situated under a spreading center for more than a geologically brief period. It is conceivable, however, that a plume might feed a spreading axis from a distance, provided it is close enough to the region in which the base of the lithosphere slopes up toward the axis. This concept helps to explain some unusual surface features in the Iceland area.

The plateau that includes Iceland stretches from Greenland in the west to the Faeroe Islands in the east. The section of the plateau east of Iceland and east of the current spreading center has long puzzled geologists. Its linear trend suggests a hot-spot origin. Yet it could not simply have been formed by the motion of a plate over a fixed plume, because it does not coincide with the track of the Iceland hot spot, which is known from the reconstructions of the early Atlantic. Some workers have interpreted this as a sign that the hot spot has not remained stationary but has instead wandered about, forming the plateau by occasionally punching through the plate. The argument implies that the reconstructions are inaccurate: if plumes are not fixed, they provide no absolute reference frame for mapping plate motions over the mantle.

Our own hypothesis is that the Iceland hot spot has remained stationary and that the Iceland-Faeroe plateau section was made by rock flowing eastward from the hot spot to a now extinct spreading center. The hypothesis can be tested. Presumably the plume would feed the closest point on the ridge. Thus at any time during the formation of the plateau a line representing the shortest distance from the plume to the ridge should intersect the center of the plateau. The plateau would be symmetrical about the ridge axis, but not necessarily perpendicular to it. With respect to the hot spot, the plates might have a component of motion parallel to the axis, and the orientation of the plateau would be obtained by adding that component to the relative motion of the plates (perpendicular to the axis). Finally, the age of the plateau at any point would be the same as that of the surrounding sea floor, because the two were formed at the same time. None of these predictions would hold if the plateau were formed by a wandering hot spot that was not feeding a ridge.

To test the model one of us (Vink) reconstructed the opening of the Norwegian-Greenland Sea and the formation of the plateau. The method is the same as that used for reconstructing the early Atlantic: superposing magnetic isochrons reveals the relative position of the plates at the time of a given magnetic anomaly, and the hot-spot track shows the plate motions in the hot-spot reference frame.

During the early opening of the basin, some 50 to 60 million years ago, the Iceland hot spot was under eastern Greenland. Its southerly track reflects the northward motion of the Greenland plate. The passage of the plate over the plume probably produced the extensive igneous rock formations southwest of Scoresby Sound, which from radiometric evidence are judged to be roughly 55 million years old. About 50 million years ago the Greenland continental shelf moved over the hot spot. At that time excess plume material could have begun flowing along the base of the oceanic lithosphere to the spreading center, and the plateau would have started to form. The Faeroe Islands, now at the eastern end of the plateau, would have been created first; their basalts are between 50 and 60 million years old. In the reconstruction of the period the nascent plateau is roughly symmetrical about the spreading axis, and the V shape of its northern edge reflects the northerly motion of the plates with respect to the hot spot.

By 36 million years ago the plates had switched to a more westerly course, causing the hot-spot track to bend to the east. The change is apparent in the geometry of the plateau: the V is split by a younger segment with an east-west heading, perpendicular to the spreading axis. The plateau remains symmetrical about the axis, and a line from the hot-spot position intersects the axis at the center of the plateau. Both observations indicate the plume was continuing to channel molten rock to the ridge.

By that time the hot spot was under oceanic lithosphere, which is somewhat thinner than continental lithosphere. The plume would have thinned it further. Our model assumes that the ridge subsequently jumped to the area of weakened lithosphere, leaving an extinct spreading center on the eastern section of the plateau. Although the existence of such a relic is still being debated, geologic activity seems indeed to have ceased in the east at about the time the spreading axis would have jumped to the west; rocks collected from a drill hole near the center of the eastern section are roughly 40 to 43 million years old.

Sea-floor spreading continued at the western end of the plateau. With the hot spot positioned under the spreading axis, plume material began to flow down the axis, giving the ridge its present tapered structure to the south. The westward-moving plates soon pushed the axis off the hot spot, but the plume continued to feed the ridge. The oldest outcrops on Iceland are found near the east and west coasts, as one would expect on an island formed at a spreading axis; their ages suggest the island was born between 16 and 12 million years ago. Iceland remains geologically active. In the past few million years eastward movements of the spreading axis have once again placed the ridge over the hot spot.

The reconstructions show that a fixed Iceland hot spot could well have produced the observed geometry of the Greenland-Faeroe Plateau. It may also have formed the Vøring Plateau, even though the latter is now 500 kilometers to the north of Iceland. The hypothesis rests on the assumption that a plume will always feed the closest section of a spreading axis. Just before the formation of the Greenland-Faeroe Plateau, when the hot spot was still under Greenland, it may have been closest to a northern ridge segment. During that period it could have produced the Vøring Plateau. The northerly motion of the Greenland plate later brought the southern spreading axis closer to the hot spot, and so the plume switched targets.

Like plate tectonics itself the notion of hot spots is a simple but powerful concept. It explains many features of the earth's surface that once seemed disparate, and further research will undoubtedly lead to the attribution of other effects to upwelling plumes in the mantle. At the same time the concept is appealingly intuitive. Indeed, it is only embellishing the truth a little to suggest the Hawaiians recognized the track of their hot spot centuries before it caught the attention of modern geologists. According to Hawaiian legend, Pele, the fiery-eyed goddess of volcanoes, originally lived on Kauai, at the western end of the island chain. When the god of the sea evicted her, she fled to Oahu. Forced again to flee, she continued to move east, to Maui, and finally to the island of Hawaii. She now seethes in the crater at Kilauea.

The Release of Acetylcholine

The compound carries a nerve impulse across a synapse. Recent studies of the electric fish Torpedo *suggest that the source of the acetylcholine emitted by a neuron is the cytoplasm rather than the synaptic vesicles*

by Yves Dunant and Maurice Israël

The nervous system is a modular structure. The modules are the neurons, or nerve cells, and within each neuron a signal carried by the nervous system passes from point to point virtually without alteration. When the signal passes from neuron to neuron, however, it must cross a gap. In most cases the transmission across the gap is effected by the release of a chemical compound called a neurotransmitter. When a nerve impulse arrives at a nerve terminal, it triggers the secretion of a neurotransmitter that can travel across the gap and stimulate the next cell, thereby carrying the impulse forward. One of the primary neurotransmitters is acetylcholine. The understanding of the mechanism underlying its release is of major importance to the understanding of brain function and the action of drugs on the nervous system.

For many years there was a wide consensus among neuroscientists that acetylcholine is released from small, spherical organelles called synaptic vesicles, which are found inside the nerve terminal. It was thought that when the nerve terminal is stimulated, the vesicles fuse with the terminal membrane and release their contents into the space between the neuron and the tissue with which it communicates.

Our recent investigations contradict this simple picture. They suggest that although the vesicles do indeed store acetylcholine and play a role in its regulation within the cell, the acetylcholine released by the nerve terminal does not originate in the vesicles. Instead the released acetylcholine is derived directly from the cytoplasm, which makes up the ground material inside the neuron. The releasing mechanism appears to be operated by a compound, most likely a protein, that is embedded in the membrane of the nerve cell. The protein may act as a valve, enabling acetylcholine to pass through the membrane.

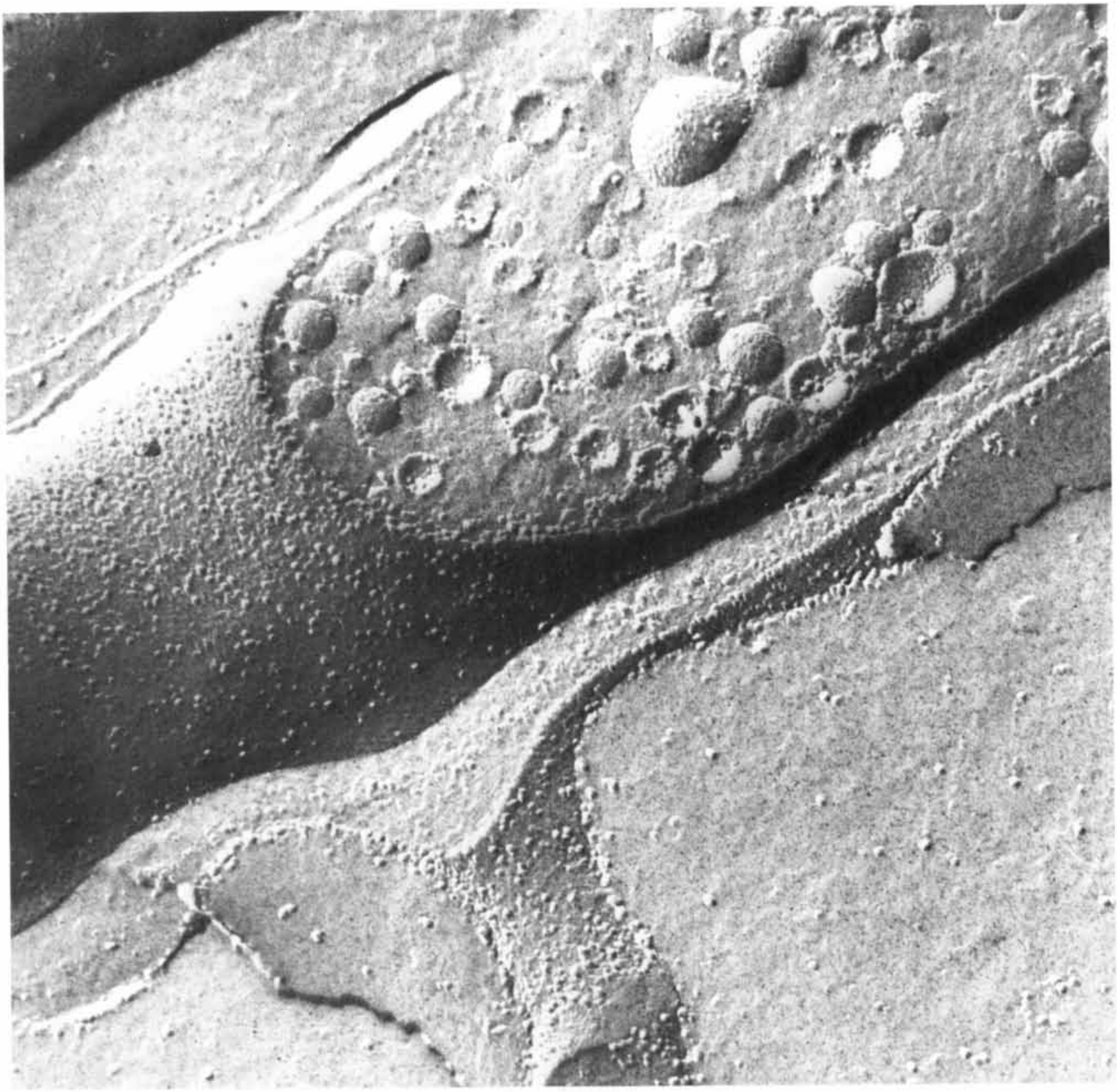
The general structure of the site of this event differs little from one animal to another or from one tissue to another in the same animal. Moreover, the transmission of a signal from a cholinergic nerve terminal, or in other words a nerve terminal that emits acetylcholine as a neurotransmitter, does not seem to depend on the identity of the downstream cell. The downstream cell can be a muscle fiber, a secretory cell or another neuron. The terminals are found at the ends of long, spindly processes of the cell called axons, and so the nerve terminals are sometimes called axon terminals. The axon terminal, the specialized region of the cell in its immediate neighborhood and the gap between the cells are collectively called the synapse; the gap itself is called the synaptic cleft.

The release of acetylcholine is initially set in motion by an electrical impulse that travels along the axon to the nerve terminal. In a neuron at rest there is an electric potential difference of several tens of millivolts between the inside and the outside of the cell membrane. When the neuron is stimulated, there is a local change in the resistance of the membrane to the passage of sodium and potassium ions. Sodium ions stream into the cell and potassium ions stream out. The exchange of ions across the membrane causes the local potential difference of the membrane to change in a characteristic way; in other words, an electric current flows across the membrane. The local change in potential in turn causes a drop in the resistance to the passage of sodium and potassium in a neighboring region of the membrane, downstream from the initial electric current. Thus the change in potential propagates step by step as a wave called the action potential that moves along the axon.

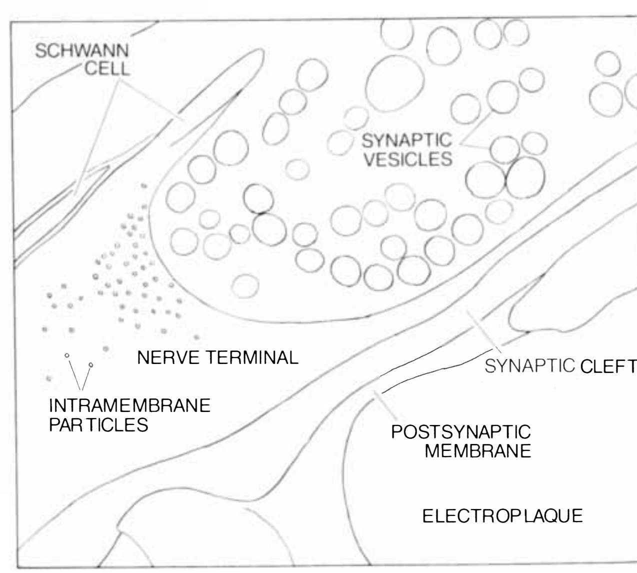
The local change in the potential difference of the membrane also induces the entry of calcium ions into the neuron. When the action potential reaches the axon terminal, the calcium entering the terminal triggers the release of acetylcholine from the terminal into the synaptic cleft. The acetylcholine diffuses across the synaptic cleft, and part of it becomes momentarily attached to receptors in the membrane of the downstream cell. There it causes another local change in electrical resistance. In muscle fibers and in some nerve cells the change is simply the first step in the generation of another action potential; in short, the muscle fiber contracts or the nerve signal continues on its way. In other postsynaptic cells the response can be inhibitory. All the acetylcholine is then broken down into acetate and choline by the enzyme acetylcholinesterase found on the outer surfaces of cell membranes as well as in the synaptic cleft.

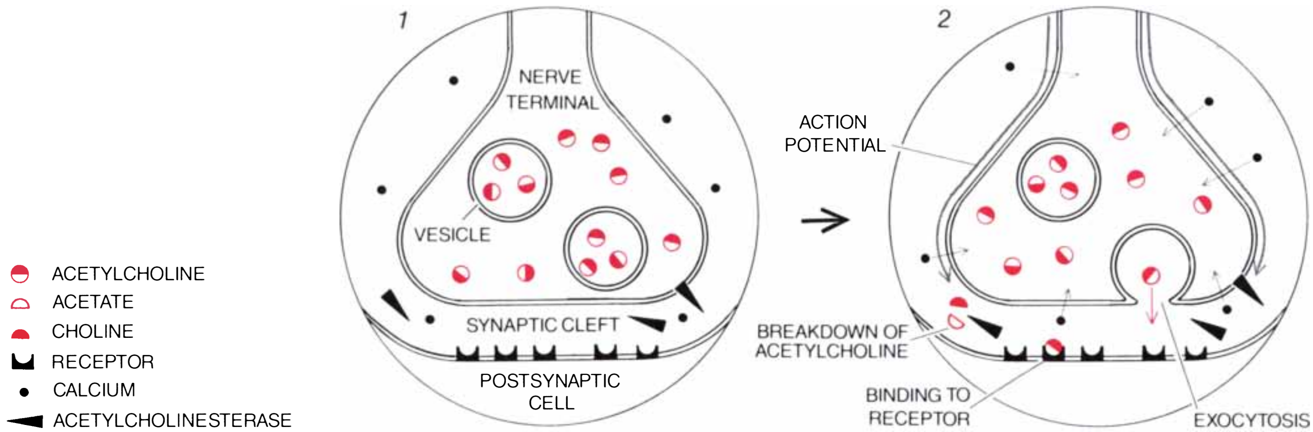
If the neuron is to transmit more than one signal, the nerve terminal must renew its supply of acetylcholine. Acetylcholine is synthesized by the transfer of an acetyl group from the compound acetylcoenzyme *A* to a choline molecule. The chemical precursors of acetylcholine circulate through many tissues and readily find their way inside the axon terminals. Indeed, in the endplate, or neuromuscular junction, and in certain other kinds of synapse both the acetate and the choline that result from the breakdown of acetylcholine outside the axon terminal are recycled back into the terminal. The enzyme that catalyzes the synthesis of acetylcholine is called choline acetyltransferase, and it is carried down the axon from the body of the neuron to the terminal, where it is stored in the cytoplasm. Thus all the ingredients needed for the synthesis are concentrated in the small space of the axon terminal. It is not surprising that the concentration of acetylcholine seems to be highest there as well.

The idea that the vesicles could be

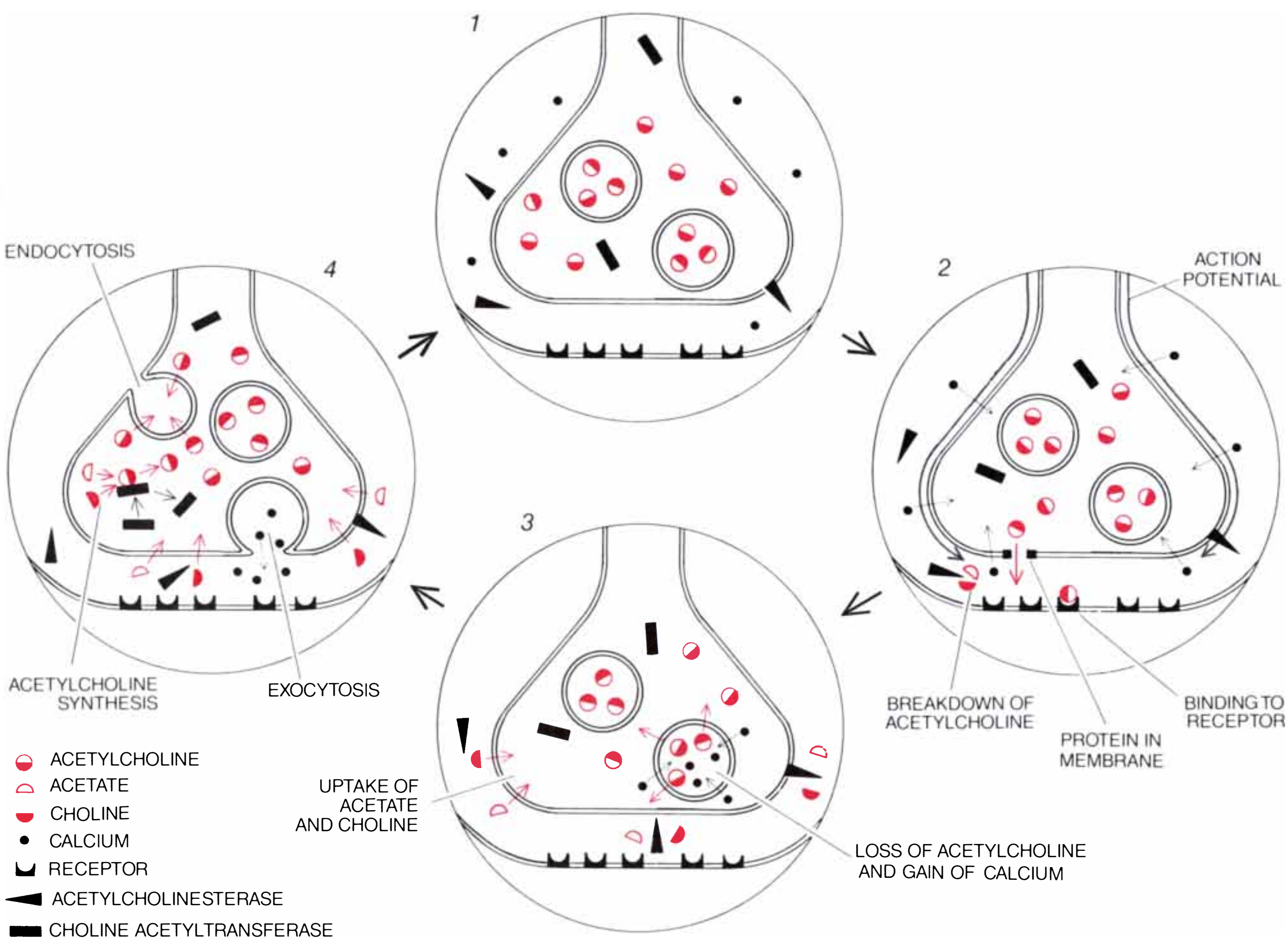


SYNAPSE in the electric organ of the fish *Torpedo marmorata* is magnified 106,000 times in the electron micrograph above; the diagram at the right identifies major components of the synapse. When the nerve cell is stimulated, acetylcholine is released by the nerve terminal and crosses the synaptic cleft to reach the cells of the electric organ, which are called electroplaques. The acetylcholine induces the rapid entry of sodium ions into the electroplaque, which gives rise to a sudden electric discharge; the discharge defends the fish and stuns its prey. The interior surface of the membrane of the nerve terminal is exposed by a freeze-fracture technique used to make the electron micrograph. It is rich in particles between five and 20 nanometers across. The particles may indicate the presence of proteins in the membrane; the ratio of the number of smaller particles to the number of larger ones changes when acetylcholine is released. An oblique fracture of the nerve terminal reveals a number of larger organelles in the nerve cell called synaptic vesicles. The vesicles store acetylcholine, and it has long been thought that acetylcholine is released when the vesicles fuse with the cell membrane and release their contents into the synaptic cleft. Work by the authors now suggests this view may have to be changed or abandoned. The electron micrograph was made by Luis-Miguel Garcia-Segura, Lelio Orci and one of the authors (Dunant) at the University of Geneva.





VESICULAR HYPOTHESIS for the release of acetylcholine is shown in this schematic diagram. In the resting state of the nerve cell (*1*) about half of the acetylcholine in the nerve terminal is stored in the vesicles. According to the hypothesis, when the nerve is stimulated (*2*), calcium ions enter the terminal and cause the vesicles to fuse with the cell membrane, releasing their contents into the synaptic cleft. Some acetylcholine becomes attached briefly to receptors in the postsynaptic cell, where it induces a characteristic response; all released acetylcholine is then broken down into choline and acetate by the enzyme acetylcholinesterase, found outside the terminal.



CYTOPLASMIC HYPOTHESIS has been proposed by the authors and their colleagues to account for the origin of the acetylcholine released by a stimulated nerve cell. The acetylcholine is initially divided between the vesicles and the cytoplasm, or ground material of the cell, as it is in the vesicular hypothesis (*1*). A nerve impulse causes calcium to enter the terminal (*2*). According to the new hypothesis, the calcium acts on specific proteins in the cell membrane that form channels for acetylcholine. A jet of acetylcholine is then emitted from the cytoplasm into the synaptic cleft and follows the same course as it does in the vesicular hypothesis. If the electrical stimulation of the nerve is prolonged, the acetylcholine within the vesicles begins to leak into the cytoplasm; the vesicles can then sequester the calcium that has entered the cytoplasm (*3*). The nerve terminal recovers its resting state when the calcium in the vesicles is expelled, perhaps by exocytosis (*4*). Acetate and choline reenter the terminal, where they are synthesized into new acetylcholine by the catalytic action of the enzyme choline acetyltransferase. New vesicles form, and the acetylcholine in the nerve terminal is distributed in such a way that the balance between the vesicular acetylcholine and the acetylcholine in the cytoplasm is fully restored.

the immediate source of the acetylcholine released at the synapses originated in the 1950's. The introduction of the electron microscope then allowed the vesicles to be observed for the first time. Moreover, in 1952 Paul Fatt and Bernhard Katz of University College London discovered small variations in the potential difference across the membrane of a muscle cell in its resting state. The variations are below the threshold needed to stimulate a contraction in the muscle fiber. They are almost certainly caused by the release of small amounts of acetylcholine from the motor-nerve terminals, which act on receptor proteins in the muscle membrane. The measured variations in the potential difference are called miniature endplate potentials.

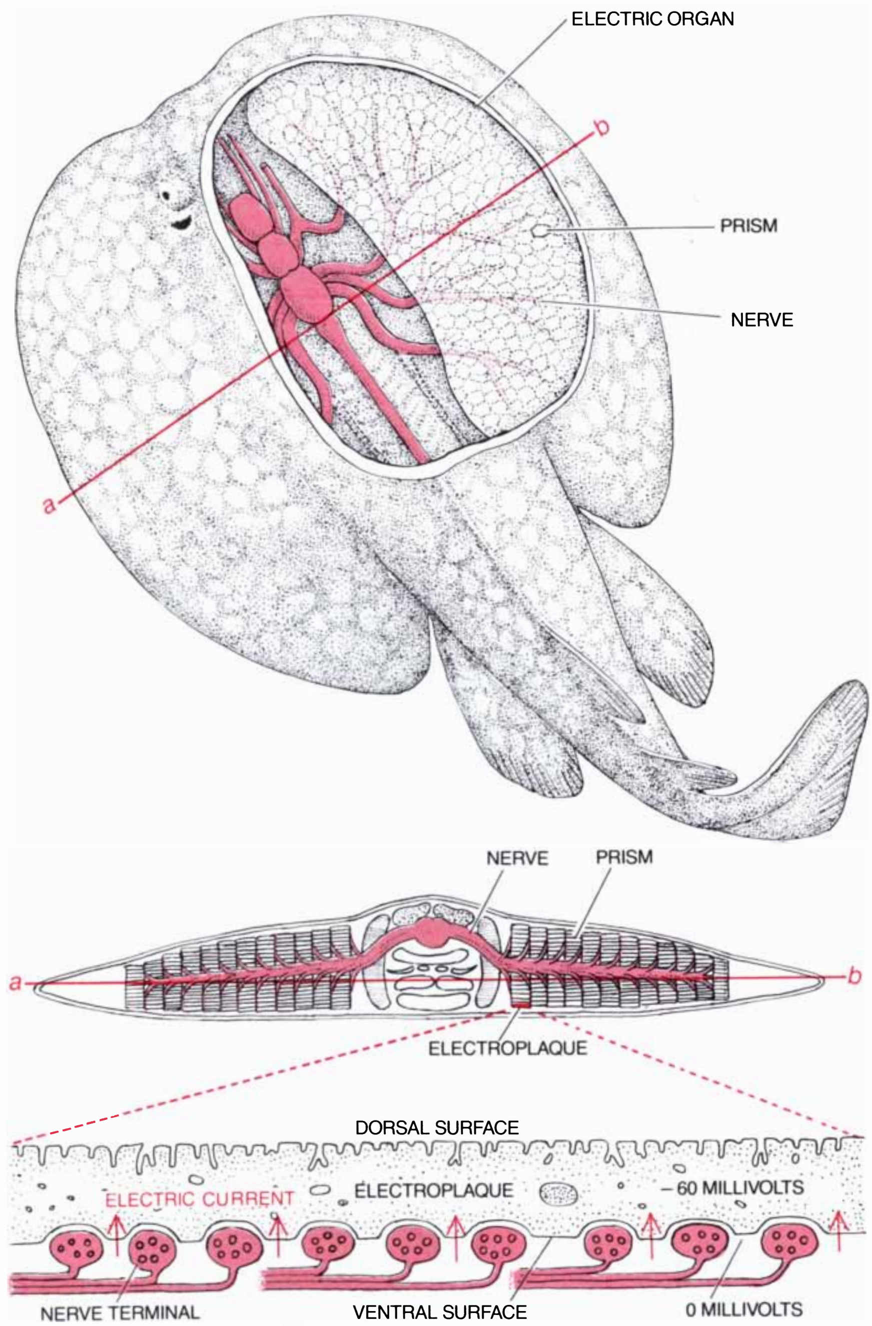
Katz and his colleagues proposed that each miniature endplate potential corresponds to the release of the acetylcholine packed into one synaptic vesicle. There were several good reasons for believing the vesicular hypothesis. First, the miniature endplate potentials all appeared to have roughly the same size and time curve, or change in potential difference recorded as a function of time.

Moreover, apart from their small size, the miniature endplate potentials have the same properties as the much larger endplate potential caused by the full-scale stimulus of a neuron. For example, the miniature potentials have the same time curve as the larger potentials, and they are detected in the same places on the muscle membrane. In fact, the full endplate potential is thought to be made up of an integral number of miniature potentials. In a large terminal that number may be as high as 400, and so according to the vesicular hypothesis as many as 400 vesicles would contribute to the transmission of a single impulse to the muscle fiber. Finally, each miniature endplate potential is caused by the nearly simultaneous action of a few thousand acetylcholine molecules, a quantity that could reasonably be sequestered in a single vesicle.

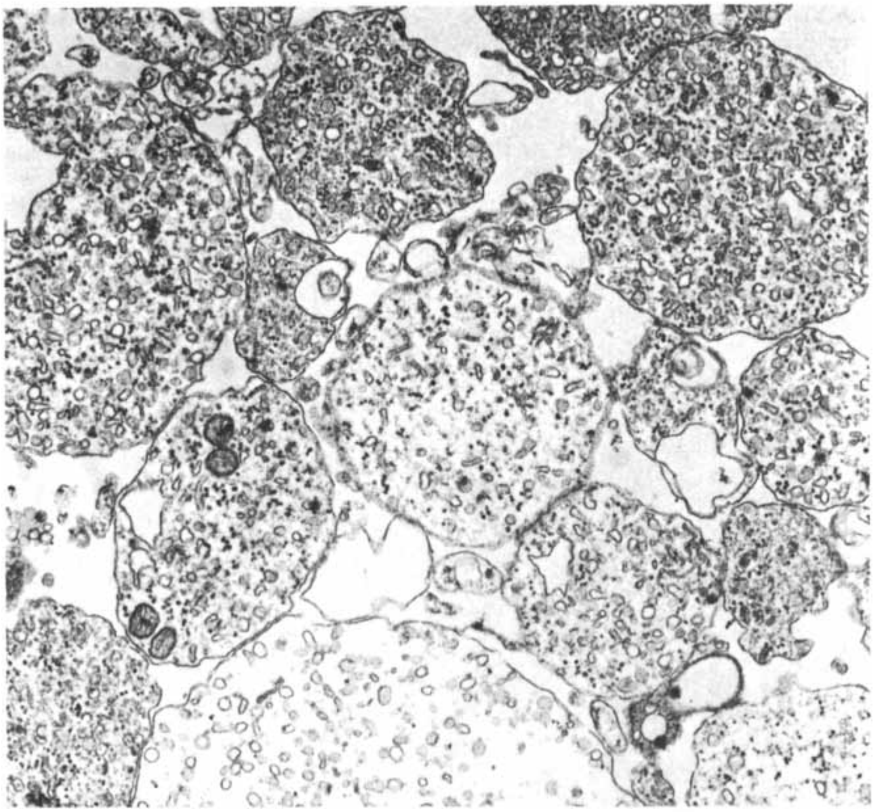
In the mid-1970's this picture began to lose part of its attractive simplicity. Mahlon E. Kriebel and his colleagues at the State University of New York at Syracuse Upstate Medical Center demonstrated that the previously recognized miniature potential is probably built up of elements of even smaller size. These subminiature endplate potentials are on the average about a tenth the size of the miniature potentials; they had not previously been measured because they could not be resolved by earlier electrical recording techniques. It took some time for the new finding to be accepted because it raised the difficult question of whether a vesicle is responsible for the release of one quantum of acetylcholine or of an even smaller subunit.

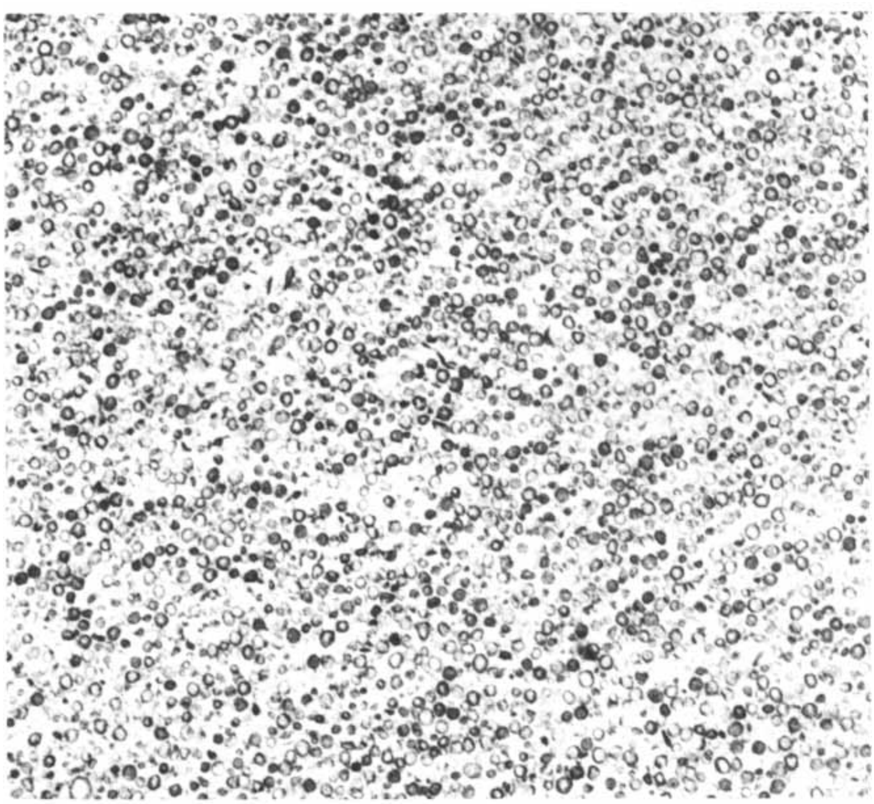
To address this question methods had to be devised whereby one could more directly analyze the level of acetylcholine and its turnover inside the nerve terminal. Electrophysiological methods record only what happens in the postsynaptic membrane. The activity of the transmitting axon terminal is thus viewed indirectly: through the effects of the acetylcholine it releases.

Between 1959 and 1964 Victor P.



NEUROANATOMY OF ELECTRIC FISH *Torpedo marmorata* **is shown in three views. At the top the skin covering the dorsal surface of the right electric organ of the fish has been removed to expose the tops of the prisms that make up the organ. The nerves leading into the electric organ are shown in color. The prisms consist of the electroplaques, which are stacked from the ventral to the dorsal side of the fish; they are shown in the middle illustration as a vertical section through the line in color. At the bottom is a highly magnified schematic drawing of the synapses of an electroplaque. The electric potential inside the electroplaque is −60 millivolts in its resting state, and there is a large resistance to the flow of electric current only on its ventral side. When acetylcholine is released by the nerve terminals, it suddenly lowers the electrical resistance on the ventral side of the electroplaque. A large current is generated by the rush of positively charged sodium ions into the cell.**





SYNAPTOSOMES, or isolated nerve terminals that release acetylcholine, are shown in a preparation made from the electric organ of the *Torpedo* (*upper electron micrograph*). The terminals are pinched off from the nerve axons and then resealed; the resulting synaptosomes retain the capacity to synthesize and release acetylcholine. Synaptosomes were first isolated in 1976 by one of the authors (Israël), Nicolas Morel, Robert Manaranche and Paule Mastour-Frachon at the Laboratoire de Neurobiologie Cellulaire in Gif-sur-Yvette, France. The lower electron micrograph shows the synaptic vesicles, which were first isolated from the electric organ in 1968 by Israël, Jean Gautron and Bernard Lesbats, then at the Hôpital de la Salpêtrière in Paris. The magnification of each image is 18,000 diameters.

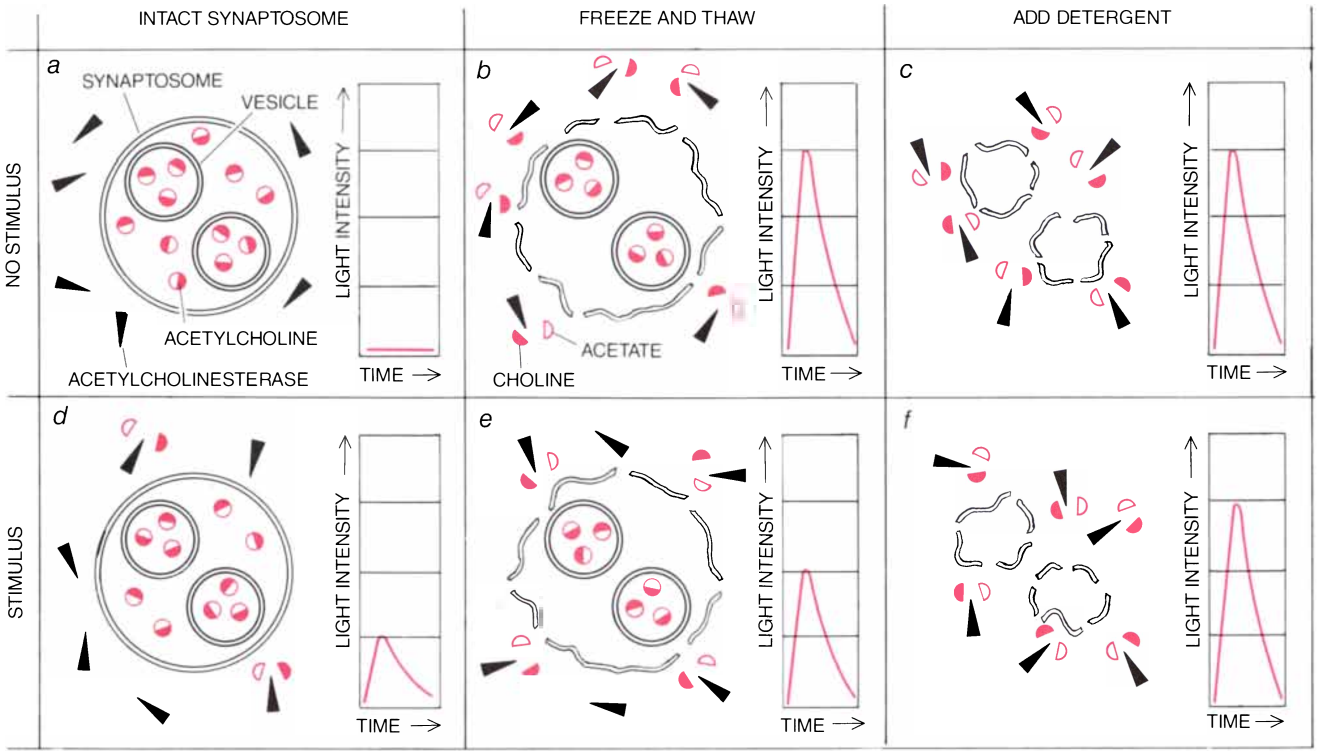
Whittaker of the University of Cambridge and Eduardo de Robertis of the University of Buenos Aires and their respective colleagues took the first steps toward the direct analysis of the terminals. They were able to show that when the brain tissue of rodents is homogenized, the axon terminals are pinched off and resealed. The pinched-off terminals can then be separated from the other fragments of the tissue by centrifugation in a ground material whose density increases with the distance from the center of the centrifuge. The nerve terminals extracted from the centrifuge are recovered in a solution having a characteristic density; they are called synaptosomes. An essential part of the neuron can thereby be manipulated and analyzed separately from the rest of the brain.

Whittaker and de Robertis independently took a further step when they found that the outer membrane of the synaptosome could be broken and the vesicles inside it could be isolated. They found that the vesicles do indeed contain acetylcholine, as they had expected. They also found, however, that not all the acetylcholine was sequestered in the vesicles. Part of it was apparently derived from the cytoplasm. Roger M. Marchbanks of the University of Cambridge then suggested that cytoplasmic acetylcholine also has an important functional role in neurotransmission.

To make further progress on the study of acetylcholine and its release, it was highly desirable to find a more convenient tissue than the rodent brain. Only a small proportion of the nerve terminals in the brain are cholinergic. In principle the junction between nerve and muscle cells in vertebrate animals would provide a more homogeneous material because the motor nerves are cholinergic. The nerve terminals at such junctions, however, are extremely small compared with the mass of the muscle fibers.

It turns out that for studies of the synaptic transmission of acetylcholine the saltwater electric fish *Torpedo* provides an ideal experimental material. *Torpedo* is commonly found in shallow coastal waters. Each fish has a pair of specialized organs called electric organs that can generate short bursts of current that serve to defend the fish and to stun its prey.

The electric organ is well suited to experiments with acetylcholine because it is made up of a vast number of synapses. These synapses were found to be cholinergic in 1940 by Alfred Fessard, Wilhelm Feldberg and David Nachmansohn at the Marine Biological Station in Arcachon, France. In



CHEMILUMINESCENT ASSAY makes it possible to monitor continuously the release of acetylcholine from synaptosomes in suspension (*a*). The synaptosomes are mixed with the enzymes acetylcholinesterase, choline oxidase and horseradish peroxidase and with the light-emitting substance luminol. Whenever the acetylcholine from the synaptosomes comes in contact with this mixture, it is converted into choline; the choline is then oxidized and a by-product of the reaction triggers the emission of light. If the synaptosome is frozen and then thawed, its outer membrane is destroyed, but the membranes of the vesicles are left intact. The extravesicular acetylcholine leaks out of the synaptosome and gives rise to the emission of light (*b*). A detergent then destroys the vesicular membranes and releases the vesicular acetylcholine, which leads to a second emission of light (*c*). If the experiment is repeated after the synaptosomes are stimulated, acetylcholine is released, and it gives rise to a light emission (*d*). The cytoplasmic pool of acetylcholine (*e*) is then reduced by the amount of acetylcholine released during the stimulation. The vesicular acetylcholine remains unchanged (*f*).

fact, the electric organ is actually a transformed neuromuscular system in which the mechanism for the transmission of a nerve stimulus is much like the mechanism in an ordinary system. For example, miniature and subminiature endplate potentials are also recorded in the electric organ. Nevertheless, the experimental preparation made from the electric organ possesses neither the complicated contractile structure of the muscle cell nor the complex neuronal circuitry and heterogeneous content of neurotransmitters found in the central nervous system.

A typical *Torpedo marmorata*, which is the European species of the fish, is 40 centimeters long and weighs about 1,500 grams. The two electric organs are found in flat "wings" to the left and right of the main body, and they comprise about a fifth of the body weight. Each organ is made up of some 400 prisms whose axes run from the dorsal to the ventral surface of the fish.

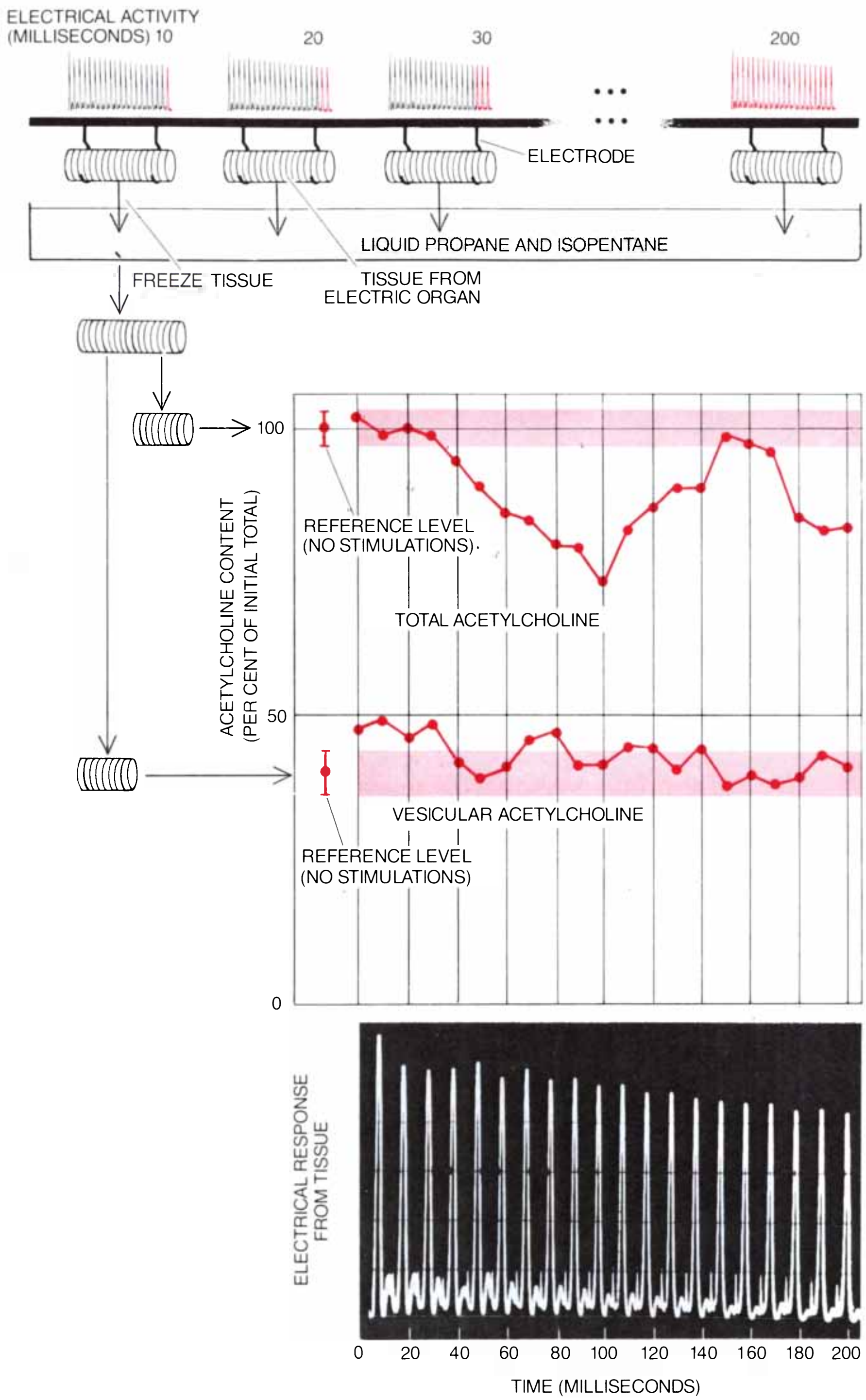
The prism is a stack of thin flat cells called electroplaques, which generate the electric current. The electroplaque has a high electrical resistance on its ventral membrane and a low resistance on its dorsal membrane. The synapses are found only along the ventral side of the cell. We estimate there are more than 500 billion nerve terminals in each electric organ.

Acetylcholine is emitted by all the terminals simultaneously, and it acts on receptor molecules of the electroplaques by opening membrane channels only on the ventral side, or in other words the high-resistance side, of the cell. The open channels allow positively charged sodium ions to enter rapidly through each ventral membrane, and the movement of the sodium ions constitutes an electric current. The effect is rather like the sudden discharge of a large set of small batteries wired in series. A substantial electric current is propagated into the surrounding seawater because the ions derived from salts dissolved in the water are induced to move by the electric field. For a *Torpedo* of average size the current from both electric organs can be higher than seven amperes.

Vesicles from the electric organ of *Torpedo* were first isolated in 1968 by Jean Gautron, Bernard Lesbats and one of us (Israël) at the Hôpital de la Salpêtrière in Paris. Between 1976 and 1977 Israël also collaborated with Nicolas Morel, Robert Manaranche and Paule Mastour-Frachon at the Laboratoire de Neurobiologie Cellulaire in Gif-sur-Yvette, France, in the work that led to the isolation of synaptosomes from the electric organ. The methods for preparing synaptic vesicles resembled the ones that had been successful for brain tissue. Because of the homogeneity of the initial tissue, however, the vesicles obtained from the electric organ were 1,000 times richer in acetylcholine than the vesicles obtained from the brain.

One of our first objectives was to compare the amount of acetylcholine in the vesicles with the amount of acetylcholine in the nerve terminal as a whole. If the nerve terminal is treated with an acid that inactivates enzymes, both the outer membrane of the nerve terminal and the inner membranes of the vesicles are broken up. The acetylcholinesterase outside the terminal is also destroyed by the treatment, but the acetylcholine is not. Hence all the acetylcholine inside the terminal is left behind by the treatment and can be measured.

On the other hand, if the nerve ter-



EXPERIMENTAL PROCEDURE whereby the variation of acetylcholine released by intact nerve terminals can be determined over short intervals of time is shown in this schematic diagram. Pieces of tissue from the electric organ of *Torpedo* are stimulated by short bursts of electrical impulses for varying intervals of time (*stimulus patterns in color*). The transmission of acetylcholine initiated by the stimuli is then abruptly stopped at various stages in the process when all the tissues are simultaneously plunged into a bath of very cold liquid (*top*). All the acetylcholine released by the nerve terminals is destroyed by the action of acetylcholinesterase outside the terminals. Each piece of frozen tissue is then split in two. In one part the total acetylcholine is measured; in the other part only the vesicular acetylcholine is measured. In both cases the measurements are similar to the ones that determine the amount of acetylcholine in the various compartments of a synaptosome (*see illustration on preceding page*). The graphs in the middle show that the total acetylcholine in the nerve terminal falls during the stimulations, but the vesicular acetylcholine remains roughly constant. Hence the released acetylcholine must be derived from some extravesicular source, which is presumed to be the cytoplasm. The measured electrical response of the tissue to the stimulation is quite similar to the electrical activity in the living electric organs (*bottom*).

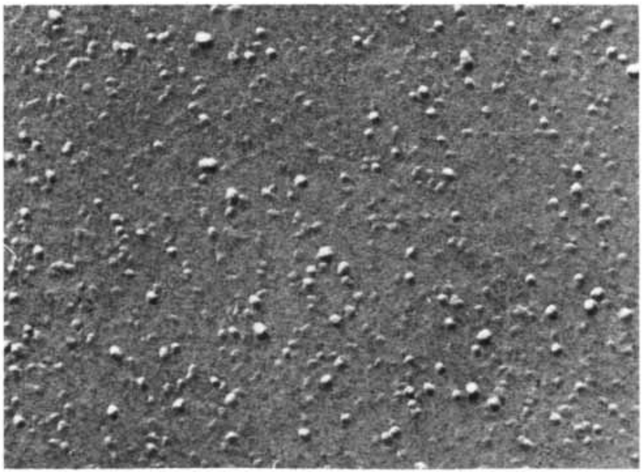
minal is quickly frozen and then rethawed, or if it is otherwise mechanically disrupted, only its outer membrane is opened. The action of acetylcholinesterase is not inhibited, and so the acetylcholine in the cytoplasm outside the vesicles is broken down by the acetylcholinesterase. The vesicular acetylcholine, however, is left intact because it is protected by the membrane of the vesicle from the action of acetylcholinesterase. It can be measured once it is liberated by an acid, which destroys the acetylcholinesterase and the vesicular membranes.

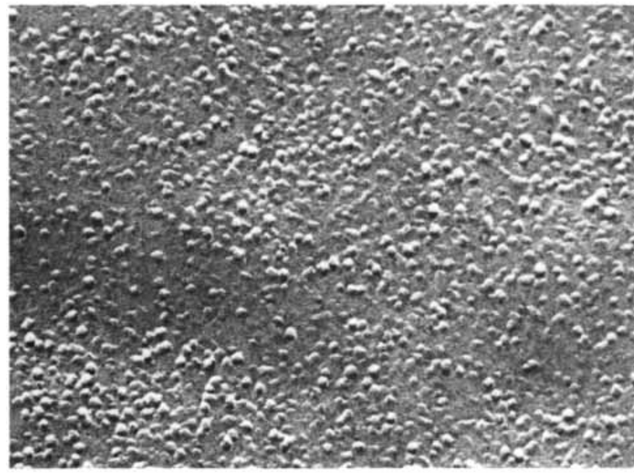
In 1968 we corroborated the finding that synaptic vesicles store a large amount of acetylcholine. We found that vesicles in the electric organ, like the vesicles in the brain, hold about half the acetylcholine content of the entire nerve terminal. Taken together, the two results appeared to give strong support to the vesicular hypothesis. It is also important to remember, however, that the balance of the acetylcholine is probably just floating freely in the cytoplasm of the terminal. The existence of cytoplasmic acetylcholine is not surprising since choline acetyltransferase, the enzyme responsible for synthesizing acetylcholine, is also found in the cytoplasm.

It seemed clear at this stage in our investigations that we could further strengthen the vesicular hypothesis. If the hypothesis is correct, the stimulation of a nerve terminal should release vesicular acetylcholine before releasing acetylcholine from any other part of the cell. We could test this prediction in the concentrated mass of cholinergic nerve terminals present in the electric organ of *Torpedo*.

In 1970 we began by stimulating the nerves that lead to the organ until the exhaustion of the electrical response. We were convinced at the time that this treatment would lead to a marked decrease in the number of vesicles as well as in their content of acetylcholine. To our great surprise it turned out that the extravesicular acetylcholine was the first to be used up and then renewed during the stimulation. In contrast, the number of vesicles and the amount of acetylcholine in them remained constant for several minutes; only later did the vesicular acetylcholine begin to be depleted.

To probe this finding in greater detail we and our colleagues have developed several specialized experimental techniques. For example, to measure changes in the level of acetylcholine in various compartments of the synaptosomes over short time intervals one of us (Israël) and Lesbats developed a new chemical assay. A mixture of the





PARTICLES within the membrane of a nerve terminal in the electric organ are shown before the nerve is stimulated (*electron micrograph at left*) and three milliseconds after a stimulus is applied (*electron micrograph at right*). The micrographs show that the density of the particles increases dramatically during the passage of a nerve impulse. A few milliseconds later the density returns to its resting state. The change may reflect the workings of the mechanism that releases acetylcholine. Micrographs were made by Dominique Müller, Garcia-Segura, Arpad Parducz and one of the authors (Dunant) at the University of Geneva. Magnification is 125,000 diameters.

enzymes acetylcholinesterase, choline oxidase and horseradish peroxidase and the light emitter luminol is introduced into the solution surrounding the synaptosomes. When acetylcholine is released in the medium, a chain of chemical reactions gives rise to an emission of light proportional to the amount of acetylcholine present. We have applied the new assay in the study of acetylcholine released from *Torpedo* synaptosomes.

Synaptosomes are difficult to stimulate electrically, and so another set of experimental techniques was needed. For example, if certain peptide molecules called ionophores are added to the solution, the ionophores form artificial channels in the synaptosome membrane that allow certain ions to pass through. One such ionophore opens the membrane to calcium ions, which then cause the release of acetylcholine. The released acetylcholine can be monitored by the chemiluminescent assay.

Our findings with synaptosomes were in accord with the earlier experimental results with intact tissue: the cytoplasmic pool of acetylcholine was the first to be released. Indeed, even when the synaptosomes are emptied of most of their vesicles and then refilled with commercially prepared acetylcholine, the calcium stimulus causes the acetylcholine to be released. The pool of vesicular acetylcholine is apparently not needed for the release.

In order to make fast biochemical measurements of intact tissues during the release of acetylcholine, we constructed a machine at the University of Geneva that can stimulate several isolated prisms of the electric organ in rapid sequence [*see illustration on opposite page*]. The prisms thereby reach different stages in the release of acetylcholine at the same time, and the release is abruptly halted when the prisms are plunged into a cold liquid.

In one series of experiments the stimulus was a short burst of impulses given every 10 milliseconds over a period of 200 milliseconds. We chose the stimulus pattern because it resembles the natural discharge of a *Torpedo* when it is alive in the sea. Again—this time with a little less surprise—we found that the acetylcholine in the synaptic vesicles does not appear to change with the stimulus.

Moreover, we found that the level of extravesicular acetylcholine fell during the first 100 milliseconds, rose during the next 100 milliseconds and then fell again. The re-formation of acetylcholine in intact nerve terminals is evidently an extremely rapid process. The level of acetylcholine finally drops presumably because the terminals can no longer synthesize enough acetylcholine to replace the large quantity that has been released.

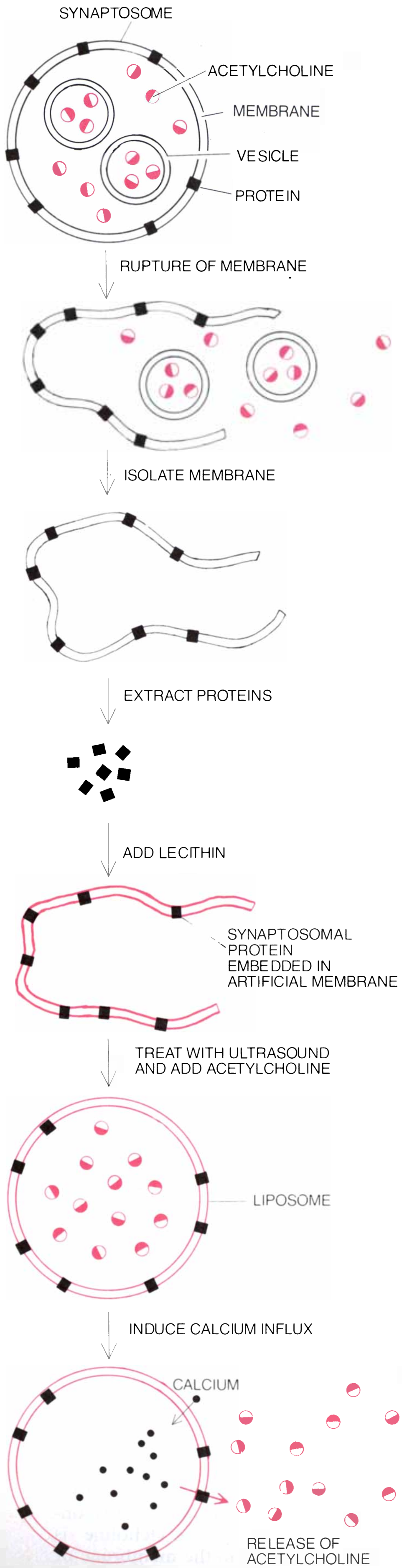
It seems difficult to believe acetylcholine can cross the membrane of the transmitting neuron at any high rate without causing some transient but visible structural change. If the synaptic vesicles are immediately involved in the process of acetylcholine release, one would expect to see openings in the membrane wherever fusion takes place. Such openings would also be expected during every release of acetylcholine from the neuron.

Several electron micrographs have now been obtained that show the fusion of a vesicle with the membrane of the nerve terminal at a neuromuscular synapse. The vesicles, however, were not found under all conditions in which acetylcholine was released. We also have obtained a few electron micrographs that show the exocytosis of vesicles in the nerve terminals of the electric organ. The fusion of the vesicles in our experiments, however, does not always accompany the release of acetylcholine, and when it does, it seems to take place a short time after the release.

Since the fusion of the vesicles is not reliably associated with the release of acetylcholine, we have become keenly interested in another structural change in the membrane of a nerve terminal. The change is always simultaneous with the liberation of acetylcholine. When the membrane is rapidly frozen, it can be split open to expose its inner or its outer leaflet. At high magnification one can see particles on the internal surfaces of the leaflets, ranging in size from five to 20 nanometers across.

Whenever acetylcholine is released from synaptosomes, we found that the number of particles between five and 11 nanometers across decreased, whereas the number of particles between eight and 20 nanometers across increased. In the intact prisms we could not detect any change in the number of the smaller intramembrane particles, but there was a significant increase in the number of the larger ones. The timing of their appearance was precisely correlated with the time course of the release as recorded by the electrical response.

Taken together, the observations suggest that the mechanism responsible for releasing acetylcholine is somehow present in the membrane of



the transmitting nerve terminal. The mechanism is probably a change in the configuration of proteins lodged in the membrane, and it may be that the effects of the change become manifest as a transient increase in the number of large intramembrane particles.

If this view is correct, it should be possible to extract the protein from the native terminal membranes and then insert it into artificial membranes. The artificial membranes should thereby gain the ability to release acetylcholine when they are exposed to the triggering action of calcium ions.

We have recently attempted such an experiment in one of our laboratories, and the results, although preliminary, appear to be successful. Synaptosomes from the *Torpedo* are opened, and the synaptic vesicles and all other cytoplasmic constituents are cleaned out of the membranes. The membranes are freeze-dried and their components are incorporated into artificial membranes made from commercial lecithin.

The artificial membranes form small closed sacs when they are treated with ultrasonic waves; the artificial membrane resembles the membranes of the original *Torpedo* synaptosomes, but the sacs formed from the artificial membranes do not hold any internal organelles such as synaptic vesicles. When the sacs are filled with acetylcholine and calcium is introduced, the acetylcholine is released into the external solution. The large intramembrane particles are observed in the artificial membrane during the artificially stimulated release of acetylcholine.

If the mechanism that releases acetylcholine is found in the membrane of the nerve terminal and emits cytoplasmic acetylcholine, what is the role of the vesicles? One function is to store acetylcholine in reserve, which could be called on after much nervous activity in order to compensate for losses in the cytoplasmic pool.

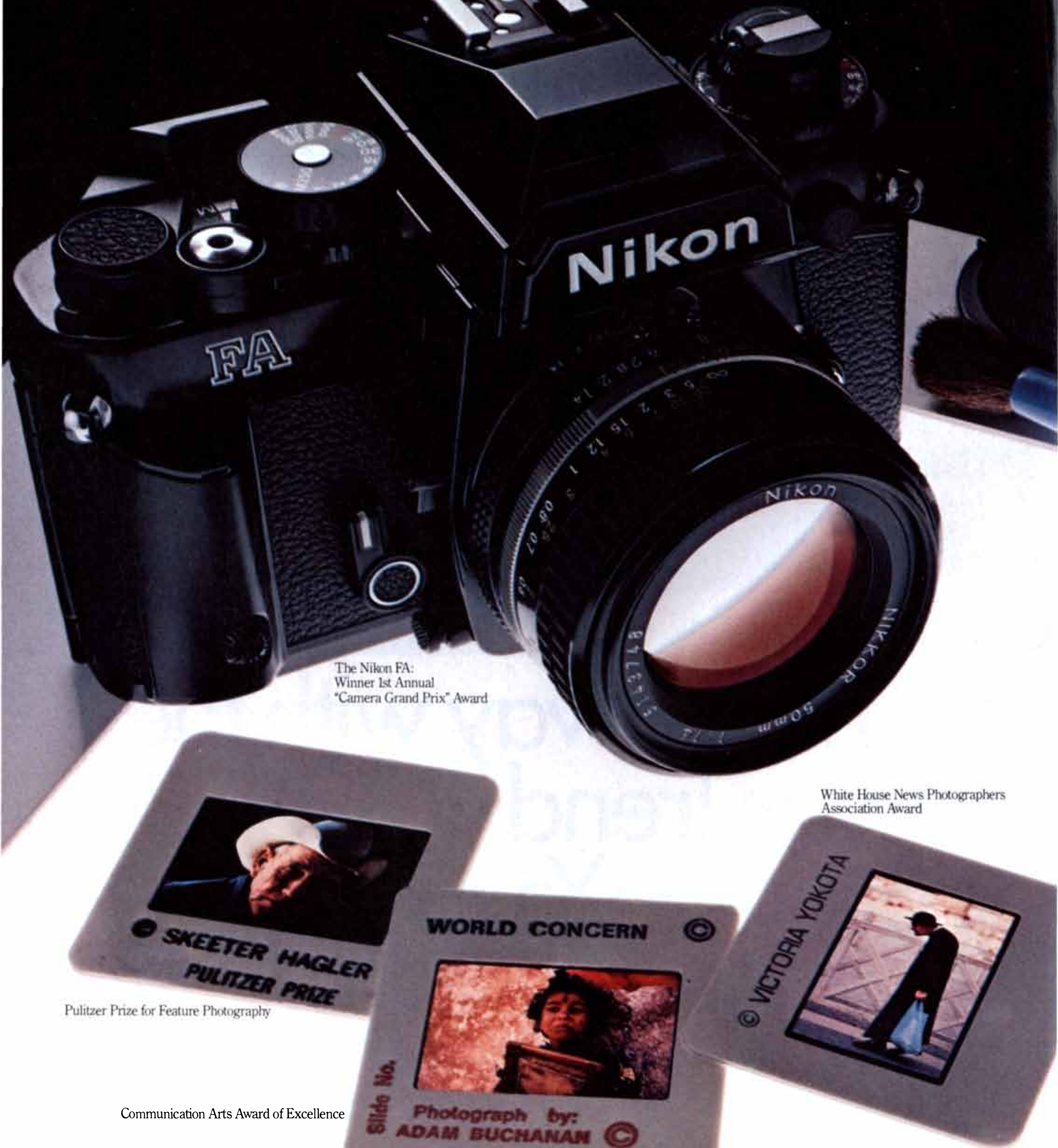
The synaptic vesicles have another important property that has been demonstrated in one of our laboratories. They can accumulate calcium and thereby help to terminate the release of acetylcholine. After the prolonged stimulation of the nerve, acetylcholine begins to leak from the vesicles, and the vesicles then fill up with calcium. One can imagine that in order to remove the calcium from the cell the vesicles fuse every now and then with the synaptic membrane and expel their contents by exocytosis. Such a chain of events has so far been only partially demonstrated, but it is probably safe to say the vesicles are essential for regulating the levels of acetylcholine and calcium at the center of the synapse.

There is a final question that such a releasing mechanism might resolve: How can the miniature endplate potentials and their subunits be explained? Imagine the protein in the membrane is a polymer, or in other words a molecule made up of elements that can be assembled or broken down. When the neuron is at rest, the elements are dissociated in the membrane, and in this state they are too small to be seen as particles in electron micrographs. When calcium enters the terminal, however, it leads to the formation of a polymer that can transiently appear as a large, intramembrane particle.

The process could be associated with the release of a package of acetylcholine molecules from the cytoplasmic pool. The releasing particle would act as a channel that is specific for acetylcholine. It would open for a short time, liberating a given amount of acetylcholine. For example, the packet might consist of the few hundred molecules that are thought to generate the subunit of a miniature endplate potential. The synchronized formation of a certain number of releasing polymers in a given region of the synapse would then be responsible for one miniature potential, and the same synchronized process in all the terminal branches of the incoming nerve fibers could give rise to a full nerve impulse.

Other models can be imagined, of course. The issue can probably be settled only if properties of the releasing mechanism are compared in two conditions. First, the mechanism must be separated and studied in vitro, where it can be manipulated and made to operate in artificial membranes. Second, it must be studied in the membrane of an intact transmitting neuron, where it can work in the most natural manner. What is exciting is that it is now within our grasp to learn more about the events taking place in the membranes of nerve terminals during the brief times they are activated.

PROTEINS of high molecular weight have recently been isolated from the membranes of nerve terminals in the electric organ and incorporated into the small sacs called liposomes, as shown in this schematic diagram. The liposomes are formed when artificial membranes made from commercial lecithin are treated with ultrasonic waves. The liposomes are filled with acetylcholine, and an influx of calcium into the liposomes causes acetylcholine to be released from them, just as it is from an ordinary nerve terminal. The experimental result indicates that the mechanism for releasing acetylcholine is bound up in the membrane of the nerve terminal.



For over 30 years Nikon cameras have helped photographers win awards for their pictures.

In fact, Nikon cameras have taken more Pulitzer Prize winning photographs than any other 35mm camera.

But our cameras win awards too.

Now, the Nikon FA has been selected winner of the 1st Annual "Camera Grand Prix" award. Given by a panel of judges consisting of outstanding professional photographers, equipment experts and editors, the FA received this honor primarily for its innovative Automatic Multi-Pattern metering system.

AMP metering virtually ensures optimum exposure even under difficult lighting

For more information write to Nikon Inc., Dept. 55, 623 Stewart Ave., Gar

We don't just help win awards for others, we also win them for ourselves.

conditions. Of course this isn't the only outstanding feature you'll find on the FA. It also has five modes of operation, through-the-lens flash metering with a Nikon Speedlight, a peak flash-sync of 1/250th of a second, and a top shutter speed of 1/4000th of a second.

And when you buy a Nikon you're eligible to become a member of the Nikon U.S.A. Club simply by applying for the Nikon Inc. U.S.A. warranty.

So visit your nearest Nikon Dealer and see the FA for yourself. After all, you're the judge whose opinion matters most.

Nikon®
We take the world's greatest pictures.®

SCIENCE AND THE CITIZEN

AIDS Update

The AIDS epidemic continues—at an increasing pace. Four years after the first cases were reported there is still no cure. There have been some interesting developments, however.

Isolation by Robert C. Gallo and his colleagues at the National Cancer Institute of HTLV-III, a virus that appeared to be the agent of AIDS, was announced last April. A year earlier, workers in the laboratory of Luc Montagnier at the Pasteur Institute in France identified a virus they called LAV and suggested it might cause AIDS. Soon after HTLV-III was reported Jay A. Levy and his colleagues at the University of California at San Francisco School of Medicine described an AIDS virus they called ARV. All three isolates are retroviruses: viruses whose genetic material is not the usual DNA but the related nucleic acid RNA, which is "reverse-transcribed" into DNA in the infected host by a viral enzyme, reverse transcriptase.

Now the full nucleotide sequences of the three viruses have been published. As expected, HTLV-III, LAV and ARV are variants of a single highly heterogeneous virus. The largest differences among the variants are in the sequences coding for proteins making up the outer envelope. Variability of the envelope proteins may complicate the development of a vaccine.

Now that the agent is known to be a retrovirus the effort in many laboratories is to test drugs that interfere with reverse transcription of the RNA. Pasteur Institute workers have reported some success in AIDS patients with a drug designated HPA-23, which inhibits its reverse-transcriptase activity. At the National Institute of Allergy and Infectious Diseases a different drug, suramine, has been shown to inhibit reverse transcriptase in cultured cells infected with HTLV-III and is now undergoing a preliminary clinical trial.

Meanwhile the epidemic waxes. By late February some 8,500 AIDS cases had been reported to the Centers for Disease Control; 4,100 victims, or 48 percent, had died. The rate of transmission is increasing: 1,485 cases were reported in the last half of 1983, 2,254 in the last half of 1984. Homosexual and bisexual men account for 73 percent of the cases among adults and adolescents. Other groups are also at special risk: users of intravenously administered drugs (17 percent), Haitian immigrants (3 percent) and heterosexual contacts of homosexual men or of female drug users, recipients of multiple blood transfusions and hemophiliacs treated with pooled-blood products (about 1 percent each). Among children less than 13 years old 97 cases have been diagnosed; most of them are babies born to women who have AIDS, but some are hemophiliacs or recipients of blood transfusions.

A screening technique should soon be available to keep the AIDS virus out of the blood supply: an ELISA (enzyme-linked immunosorbent assay) with which to test donated blood for the presence of antibodies to HTLV-III. Five groups now have ELISA kits ready to sell. The kits were expected to be licensed by the Food and Drug Administration for distribution to blood banks by February 15, but licensing was delayed.

One problem has been a controversy over the implications of a positive test result. A finding that antibodies to HTLV-III are present in a donor's blood indicates only that he or she has been exposed to the virus, which may or may not be present at the time of the test. The donor could have beaten off the disease, could be an asymptomatic carrier or could eventually develop the symptoms. Clearly the blood should not be distributed, but what should the donor be told? The FDA is expected to recommend that when the ELISA is positive, the donor should be told of the result only after the test has been repeated; he or she should then be referred to a physician for confirmation using a different test and for counseling about the implications of the test result.

A surprising discovery has emerged from research on AIDS. HPA-23, the antiretrovirus agent being tested at the Pasteur Institute, seems to have some effect on the activity of prions, the cause of Creutzfeld-Jakob disease, a rare, fatal neurological disorder in humans. Prions appear to consist only of protein; they seem to lack genetic material of any kind.

When Is a Planet?

It "is an object that is *not* a star, which is in orbit around an object that *is* a star. What should we call it? From the existing astronomical nomenclature, 'planet' is the obvious choice." What Donald W. McCarthy, Jr., of the University of Arizona means is that he and his colleagues may have made the first direct observation of another solar system.

Specifically, McCarthy, Frank J. Low, also of Arizona, and Ronald G. Probst of the National Optical Astronomy Observatories have observed a brown dwarf. A brown dwarf is thought to form as a star forms, out of collapsing interstellar gas and dust. Unlike a star, however, a brown dwarf does not have enough mass to sustain thermonuclear reactions. The radiation it emits, which is primarily in the infrared, comes chiefly from the energy of its collapse. The existence of brown dwarfs was postulated on theoretical grounds in 1975 by Jill Tarter of the National Aeronautics and Space Administration's Ames Research Center. This is the first generally recognized sighting of one.

The brown dwarf observed by McCarthy and his associates orbits around the faint star Van Biesbroeck 8 at a distance of about 6.5 astronomical units. (One astronomical unit, approximately 93 million miles, is the mean distance from the sun to the earth.) Van Biesbroeck 8 and its dwarf companion lie roughly 21 light-years from the earth. The investigators have named the brown dwarf Van Biesbroeck 8B, or VB 8B.

McCarthy's group found VB 8B while they were examining faint "astrometric binaries," stars whose motion against distant "fixed" stars shows periodic irregularities. Because of their perturbed motion such stars are presumed to be under the gravitational influence of companions. The investigators were able to distinguish the infrared light of VB 8B from the much brighter emission of Van Biesbroeck 8 by a method known as "speckle interferometry." In speckle interferometry the astronomer combines a collection of extremely short exposures in a way that averages out distortion caused by random turbulence in the earth's atmosphere.

Some astronomers are hesitant to call VB 8B a planet because it closely resembles a star in composition and history. McCarthy replies that VB 8B also has much in common with the planet Jupiter. The two objects have about the same radius, although VB 8B has between 10 and 70 times Jupiter's mass. Jupiter does fit some of the early definitions of "brown dwarf," he points out. Furthermore, if VB 8B occupied Jupiter's place in the solar system, it would appear to have the same brightness in visual wavelengths and would have been classified as a planet.

The detection of VB 8B is one of several recent discoveries that may in-

TANDY... Clearly Superior™

It's evident when you can cover the MS-DOS market with a line of computers unmatched in performance and value.

Tandy brings you not one, but three solutions to the personal computer dilemma.

Like the IBM PC-AT, the Tandy 2000 performs up to three times faster than the IBM PC. Disk storage is double the PC's. Our color graphics offer twice the resolution, twice the colors. And during March, the Tandy 2000 is only $1999. That's $500 off.

Or how about a system that does everything an IBM PC XT does—but costs at least $1900 less? Our new low-priced ten-megabyte Tandy 1200 HD is compatible with virtually all XT software and hardware . . . yet it's now only $2499.

Our new Tandy 1000 comes complete with software. We call it DeskMate™, and it's six programs on one disk.

The Tandy 1000 has many features that cost extra on the IBM PC. Like adapters to use a monitor, printer, joysticks and light pen, plus a Disk Operating System and BASIC—all for only $1199.

For the best in technology, support, service and value, stop by your local Radio Shack Computer Center. We invite comparison!

Available at over 1200 Radio Shack Computer Centers and at participating Radio Shack stores and dealers

Radio Shack® COMPUTER CENTERS
A DIVISION OF TANDY CORPORATION

TANDY Radio Shack ADVANCED TECHNOLOGY SERIES ...AND AFFORDABLE TOO™

Engineered for Excellence

We've introduced the latest in technology for over 60 years. The Tandy 1000 and 2000 represent the state of the art in performance, quality and price.

*Plus applicable use/sales tax. Tandy 2000, 1200 HD and 1000 prices do not include monitor and apply at Radio Shack Computer Centers and participating stores and dealers. IBM/registered TM International Business Machines Corp.

Send Me a New 1985 Computer Catalog!
Mail To: Radio Shack, Dept. 85-A-566, 300 One Tandy Center, Fort Worth, Texas 76102
NAME
COMPANY
ADDRESS
CITY
STATE
ZIP
PHONE

dicate the existence of other solar systems. In 1983 the *Infrared Astronomical Satellite (IRAS)* detected large gas clouds surrounding four other stars. Last fall Richard J. Terrile of the Jet Propulsion Laboratory and Bradford A. Smith of the University of Arizona photographed a flattened disk of matter surrounding one of those stars, Beta Pictoris, which lies approximately 50 light-years from the earth. From the configuration of the disk it seems that Beta Pictoris may already have entered the early stages of planet formation.

Red Shift

Why does the scarlet gilia change color? The plant is known for its bright red flowers, but specimens that have produced red blossoms early in the season sometimes grow pink or white ones later on. At high altitudes those that first bloom late in the season tend to produce pale flowers.

Ken N. Paige and Thomas G. Whitham of Northern Arizona University report in *Science* that the shift in flower color may be an adaptive response to a change in the population of pollinators. Paige and Whitham studied plants whose chief pollinators are the broad-tailed and rufous hummingbirds and the white-lined sphinx, a species of hawkmoth. Early in the flowering season, which lasts from July until early September, hummingbirds and hawkmoths are present in nearly equal numbers. In mid-August the hummingbirds begin to migrate, and by the end of the month there are few left.

The investigators hypothesized that hummingbirds, which forage in the daytime, would be more attracted to red flowers and that hawkmoths, which forage after sunset, would be drawn to whiter flowers. Paige and Whitham tested their hypothesis by placing nylon covers over 30 plants (of various shades) at night, so that hawkmoths could not visit them; these plants could be pollinated only by hummingbirds. As expected, the hummingbirds preferred to visit red plants: the red plants that had been covered at night set more fruit (a sign of pollination) than did whiter plants that had also been covered. The investigators placed covers over a different set of plants during the day to exclude hummingbirds and test the preference of hawkmoths. Of the plants covered in daylight, the lighter ones set more fruit than the red ones: hawkmoths prefer flowers that are lighter in color.

It appears that in late summer, when most of the hummingbirds have gone, a plant attracts more pollinators if it produces whiter flowers: individual plants that had changed from darker to lighter colors set more fruit than plants that had not changed. In late August, when all the hummingbirds had migrated, the investigators applied white paint to flowers from 105 red plants; as a control, the flowers of 115 red plants were painted red. Flowers that had been painted white set 38 percent more fruit.

Hummingbird visits red flowers in early summer (**left**)*; a hawkmoth visits white ones later*

Eocene Greenhouse

Between 50 and 55 million years ago during the early Eocene epoch temperate forests flourished at the latitude of Greenland; the earth as a whole was some five degrees Celsius warmer than it is today. The cause, say Robert M. Owen and David K. Rea of the University of Michigan, may have been a greenhouse effect, the result of carbon dioxide released into the atmosphere by a chain of tectonic and biological events in the ocean basins.

Writing in *Science,* the authors argue that much of the carbon dioxide released by the oceans is an indirect result of hydrothermal activity along the midocean ridges, where new oceanic crust is generated. There seawater penetrates the cracks and fissures of the newly formed crust, meets hot basalt and reemerges, forming hot springs on the ocean floor. In the process the ocean's content of metals, among them calcium, is enriched.

Dissolved calcium is ultimately removed from seawater by the activity of marine organisms, which incorporate it into their shells and skeletons in the form of calcium carbonate. The process consumes bicarbonate ions, also in solution in seawater; for every molecule of calcium carbonate formed in the reaction a molecule of carbon dioxide is freed. The rate of the reaction would increase, Owen and Rea say, in response to a rise in the concentration of marine calcium.

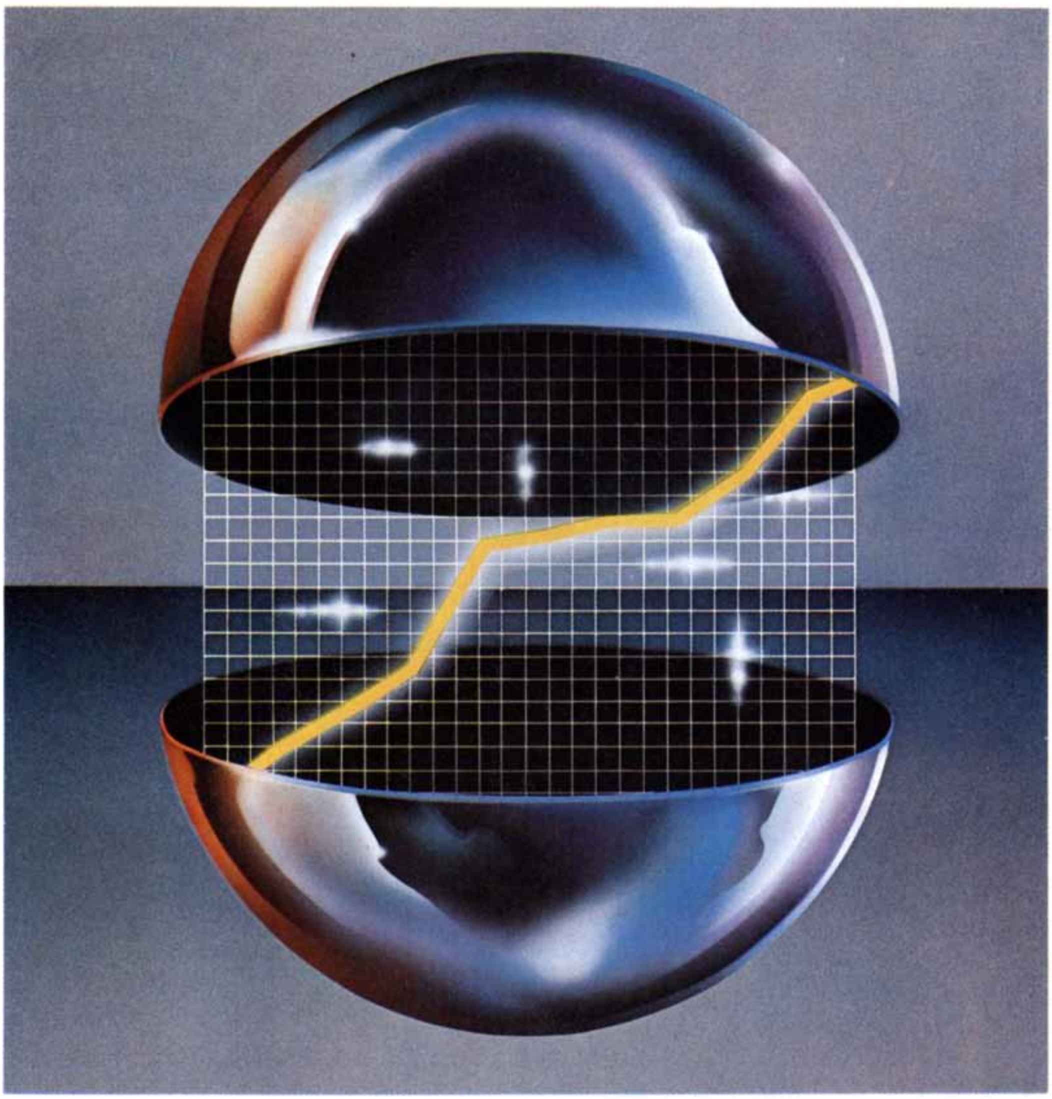
Converging evidence suggests the early Eocene witnessed a pulse of hydrothermal activity that sharply increased the oceans' content of calcium. Hydrothermal springs also carry quantities of iron and blanket the surrounding sea floor with iron-rich sediments. Consequently the vertical distribution of such sediments in core samples extracted from the ocean floor provides a record of past hydrothermal activity. The sediments are particularly abundant in samples dating from the early Eocene. More recent cores have shown that on a local scale hydrothermal activity climaxes during episodes of tectonic rearrangement, when a midocean ridge or transform fault shifts. A worldwide realignment of tectonic boundaries is known to have taken place in the early Eocene.

The sedimentary record confirms that the level of marine calcium increased at the time of the Eocene hydrothermal pulse. In deposits of marine fossils that are about 50 million years old the ratio of strontium to calcium, constant through much of geo-

pər·'fȯr·,məns

Performance

• *Something accomplished, as in nine straight years of earnings growth, with solid performances in aerospace, electronics, automotive and general industries.*
• *A coordinated action, the work of 17,000 engineers and scientists, 106,000 employees worldwide.*
• *The fulfillment of a claim or promise; e.g., delivering products of high technological quality.*



The ultimate test of performance is the bottom line. Ours shows 11 dividend increases in the last 13 years. Nine straight years of earnings growth. And a record average return on stockholders' equity of 20.3%. We're among the top industrial corporations in America on Return on Equity.

A tribute to our people. And to our diversity. To learn more about us, write: Rockwell International, Department 815S-64, 600 Grant Street, Pittsburgh, PA 15219.

Rockwell International

...where science gets down to business

Aerospace / Electronics / Automotive
General Industries / A-B Industrial Automation

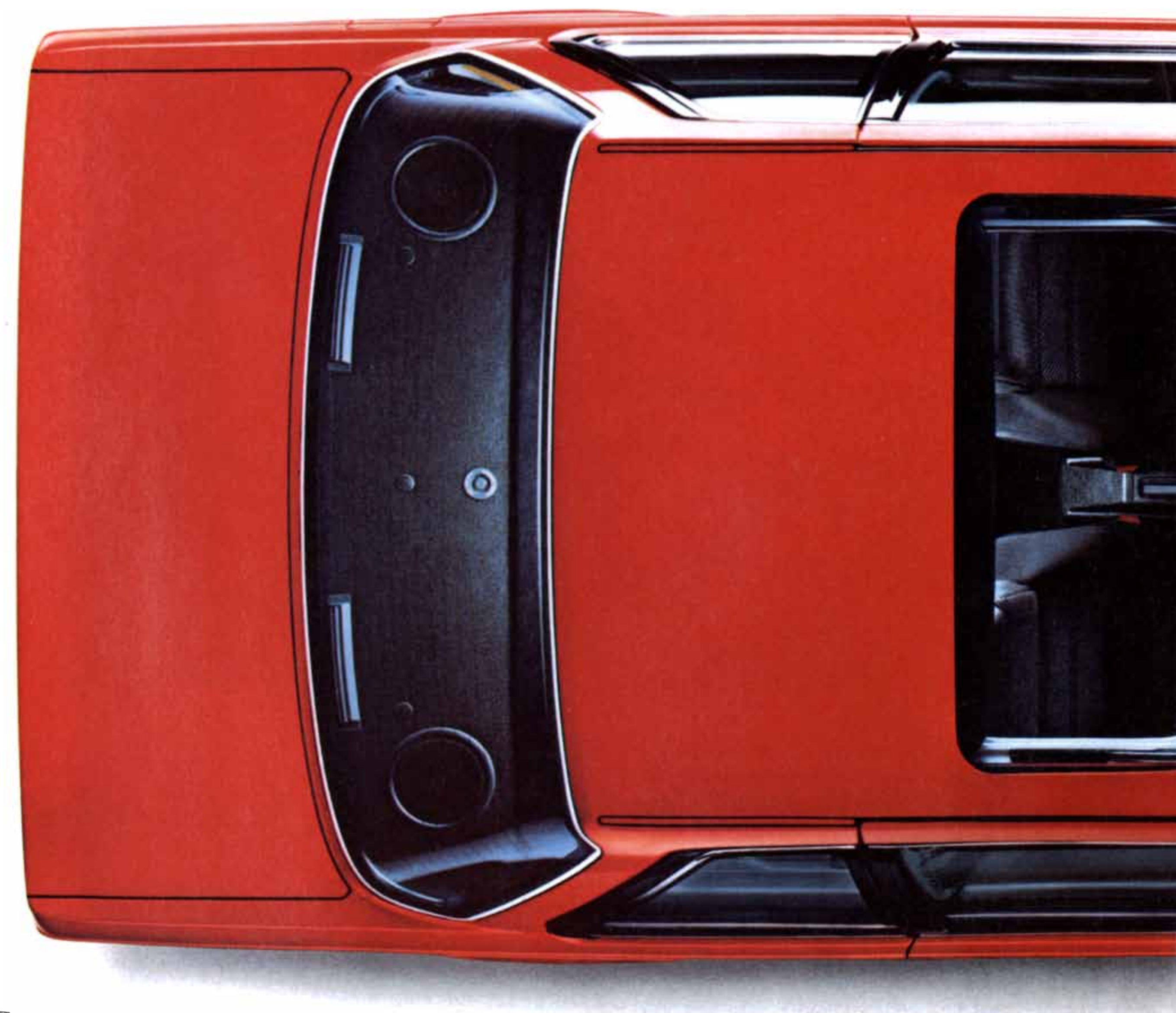
Touch a button on the instrument panel. A tinted-glass Moonroof opens wide to sky and clouds overhead. Still, there's more to opening up the Honda Prelude.

Try the accelerator. Put your foot down and feel the power of a dual-carbureted 1.8-liter engine. *Road & Track* (May 1984) clocked the Prelude from zero to sixty in 9.7 swift seconds.

The Prelude even looks fast standing still. It has a sleek wedge shape. And underneath, a double wishbone front suspension helps to handle curves with precision.

Inside, you will find a neatly designed

Drive it w

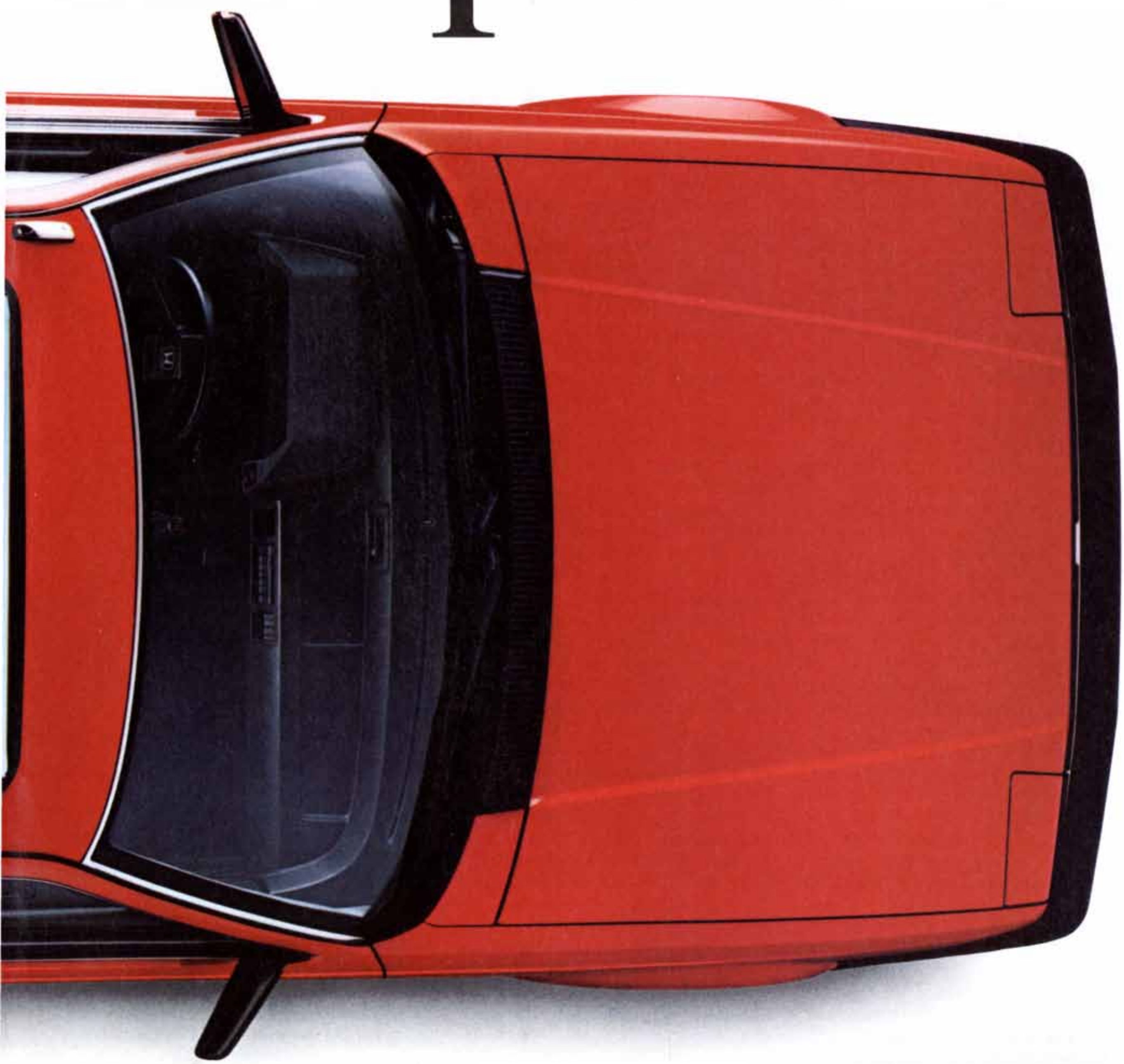


cockpit. Important controls are within easy reach of your fingertips. So is a sophisticated stereo system.

Now, settle into the Honda Prelude's contoured bucket seat. And you're ready for serious driving. Or star sighting through the Moonroof. Just open wide and say ahhh.

The Prelude HONDA

ide open.



logic time, drops sharply, probably reflecting an abnormal abundance of calcium in seawater. Bountiful marine calcium is also suggested by the very thickness of the deposits, which accumulated at about twice the current rate.

Owen and Rea believe a large increase in atmospheric carbon dioxide accompanied these marine conditions. As the carbon dioxide trapped infrared radiation the greenhouse effect warmed the earth's climate. The same sequence of events may well have occurred at other points in geologic history, the authors think. They also suggest the Eocene greenhouse may foreshadow the climate of a century from now, when the carbon dioxide released by human activity will perhaps have led to a comparable warming.

Getting Pickled

For some 30 years it has been known that the brains of chronic alcoholics tend to shrink; this change is accompanied by deficits in mental and physical performance. In recent years computerized axial tomography (the CAT scan) has also revealed shrinkage in people who drink moderately. A study of the effect over a period of several years by Lesley A. Cala of the Queen Elizabeth II Medical Centre in Australia and several associates shows the effects are at least partially reversible if the drinker abstains from alcohol for several months.

Cala's group first studied 240 alcoholics and 59 "heavy social drinkers," taking as the dividing point the ingestion of 120 grams (4.2 ounces) of pure alcohol per day. That is the figure employed by the World Health Organization to define an alcoholic. The CAT scans showed brain atrophy in 95 percent of the alcoholics and 67 percent of the heavy social drinkers.

The investigators then turned to light and moderate drinkers. They found that among 65 people with an average alcohol intake of 61 grams (2.1 ounces) per day, 85 percent were shown to have brain damage. Psychometric deficits appeared in 70 percent and liver damage in 10 percent.

To measure the effects of abstinence, Cala and his associates studied 26 of the light and moderate drinkers who had agreed to stop drinking for six months. The investigators report in the *Australian Alcohol/Drug Review* that brain atrophy diminished and that "improvements were found... in psychometric results and biochemistry."

Finally, the investigators scanned and tested 12 of the six-month abstainers who had been drinking again for six months (mean consumption 47 grams per day). They found "no significant deterioration... in the atrophy scores on CAT scan or in psychometric performance."

Cala and his associates conclude that consumption of about 40 grams of alcohol per day does not harm the brain and that abstinence reverses damage "in the social drinking range." It is not known, the investigators conclude, "how many times the brain could be re-exposed to alcohol at the original high level of intake followed by abstinence and still show repeated episodes of recovery."

Pocket Quasars

Two sources of radio emission have been observed in the Milky Way that may lead to a clearer understanding of how quasars, most of which lie at or near the edge of the visible universe, generate prodigious amounts of energy. The sources are among approximately 150 radio-emitting objects in the Milky Way that are thought to be remnants of supernova explosions. Analysis of the two sources demonstrates that they have a structure rather different from that of a supernova remnant.

The sources were mapped at the Very Large Array of radio telescopes of the National Radio Astronomy Observatory near Socorro, N.M., by Robert H. Becker of the University of California at Davis and David J. Helfand of Columbia University. One of the sources is designated G357.7–0.1. The "immediate impression," the workers report in *Nature,* "is that of a number of smoke rings of varying dimension, concentrically placed along a common axis." The axis is some 12 arc minutes long, or more than a third of the apparent diameter of the full moon. The intensity of the emission increases from east to west, and at the western end of the axis there is a compact, perhaps even pointlike, source of emission.

The second radio source, designated G5.3–1.0, also shows "axial symmetry, a gradient in surface brightness from west to east and a compact source at one extreme of the axis." In all three respects the sources differ from supernova remnants, which tend to be shells of radio emission (presumably the debris of the explosion), sometimes surrounding a compact central source (the collapsed core of the star that exploded). The structure of the sources suggests to Becker and Helfand that they consist of traveling binary systems, in which a normal star feeds matter to its companion, a neutron star or a black hole. In turn the companion emits high-energy subatomic particles at the western, leading end of each radio source.

The sources are in the Milky Way, some tens of thousands of light-years from the earth. Quasars, in contrast, are billions of light-years away. Yet both may embody the same phenomenon: the jetlike emission of energy by means of matter accreting onto a black hole. (In the case of the Milky Way sources the black hole would have roughly the mass of a star; in the case of the quasar the hole would have the mass of millions of stars.) The question is how the accretion leads to the production of energetic jets. At the end of March, Becker and Helfand will return to the Very Large Array and attempt to measure the structure of the sources in greater detail than before.

Gonorrhea Vaccine

A potential vaccine against the contagious venereal disease gonorrhea has been developed by a group of investigators at the Stanford University School of Medicine. The candidate substance, a synthetic protein fragment, appears to work by blocking the first step in the process by which gonococcal bacteria infect the cells lining the human urogenital tract: the adhesion of the bacteria to the cell walls by means of the hairlike filaments called pili. According to Gary K. Schoolnik, the leader of the Stanford group, the potential vaccine stimulates the immune system to seek out and inactivate the pili, thereby preventing the bacteria from binding to the cells. The workers describe their findings in a recent issue of *Proceedings of the National Academy of Sciences.*

Previous attempts to develop a vaccine to protect against gonorrhea have failed because the protein molecule that makes up the pili, called pilin, has a tendency to change its configuration continually, enabling the gonococcal bacteria to elude the body's highly specific immunological defenses. The key to the new vaccine's preliminary success, Schoolnik and his co-workers report, is that it stimulates immunity to a part of the pilin molecule that remains unchanged, even in different bacterial strains.

The invariant part of the pilin molecule was found by synthesizing pieces of the protein, using an automatic solid-phase technique, and looking for pieces that remained unchanged from one strain to another. The invariant pieces were then injected into rabbits, whose immune system reacted by making antibodies. Antibody-containing samples of blood serum taken from the rabbits were tested by adding them to human cells in tissue culture; the sample that most effectively prevented a variety of gonococcal strains from

We're often asked why a Leica produces such superior photographs.

The answers may surprise you.

It is generally conceded, by professionals and serious amateurs alike, that Leica® cameras produce superior photographs. Photographs more vibrant, more alive, more three-dimensional than other photographs.

So it is tempting to say that this is simply because Leica produces superior cameras.

Which is true, of course. But hardly a complete explanation.

Because it is only when you fully understand the depth of our commitment to perfection that you can truly appreciate our extraordinary cameras.

Ten tons of glass.

Leica lenses are famous. The world's finest. So it won't surprise you to learn that we ourselves make the optical glass for them.

What will surprise you is how much glass it takes. And how few lenses we end up with.

Every year, our scientists use over ten tons of raw glass. Mixing and melting special formulations in pure platinum crucibles. Refining glass to ultimate optical standards. Discarding what others would consider acceptable. Scrutinizing every finished lens individually. Making sure it combines high color saturation with delicate color differentiation and unexcelled sharpness. Making sure it is perfect.

Ten tons of glass to make a few thousand lenses. No wonder no other manufacturer dares try to imitate us.

Never plastic.

And the mechanical components of our lenses are made just as exactingly. Critical moving parts are brass and aluminum, never plastic. Lens helixes are lapped to ten-micron tolerances, then polished by hand with jeweler's rouge. Thus eliminating the need for heavy greases used by other manufacturers. Greases which can affect focusing performance.

Body talk.

Which leads us to another surprise. Leica camera bodies are not designed and built just to enhance Leica lens performance. They're designed and built to enhance yours.

Again, take focusing. Focusing depends on the photographer's ability to see the image. So we've built the brightest viewfinder possible. Using silver instead of aluminum coatings on our prism. And seventeen dichroic coatings on our primary SLR mirror. To reflect maximum undistorted light.

And for creative freedom, we've put an SLR exposure control at your fingertips. So you can switch instantly from manual or aperture priority with spot metering to program, shutter, or aperture priority with full scene integration.

The art beyond the science.

There are hundreds more examples we could give of Leitz's technical accomplishments. Of the specific superiority of Leica design and manufacture.

But at the very highest reaches of any discipline, there are areas beyond mere rational explanation.

And at Leitz®, a camera is more than just metal and glass and tests and formulations. It is pride and integrity and an almost fanatical dedication to perfection. It is skills and secrets passed on

from father to son. From generation to generation.

At Leitz, camera making is an art beyond a science.

But that shouldn't surprise you.

After all, so is photography.

For further information about Leica cameras call 1-800-223-0514 or write E. Leitz, Inc. 24 Link Drive Rockleigh, NJ 07647.

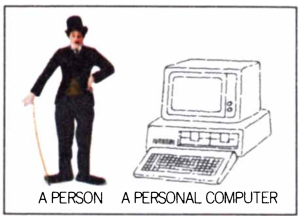
Leitz means precision. Worldwide.

Leitz and Leica are registered trademarks of E. Leitz, Inc.

Simple answers to your questions about IBM Personal Computers.

If you're personally interested in personal computers, but want to know more, these definitions, descriptions and details should help.

"Just what is a personal computer, and how can I use it?"



An IBM Personal Computer is a computer designed for a person. It's a tool to help accomplish just about anything a person needs to do with information. It can help you start a small enterprise at home just as surely as it can help a corporate planner solve complex problems.

"Suppose I've never had my hands on a computer. How 'easy' will it be?"

As with any new tool, you'll want to get comfortable with your IBM Personal Computer before getting down to work. The nice thing is that the computer is on your side, interacting with you as you learn. Then you're running programs and feeling good with the results. It becomes clear that you've made a good investment, and you'll probably be telling your friends why they should get one.

"Are IBM Personal Computers simple or sophisticated?"

Both. Our systems have many advanced design features; they are there to make your computer simple to operate and to help make you more productive. As with a well-designed car, the computer is designed around you, the user.

"What kind of software programs do you have to help me?"

Perhaps the world's largest and most up-to-date library of business programs has been written specifically for the IBM PC family. And among the best of this software is IBM's Personal Computer Software.

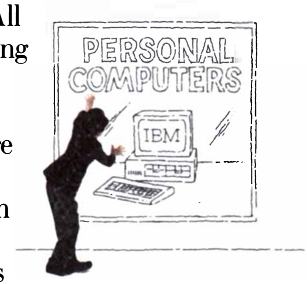
A great deal of it is compatible from one system to another or from office to home. You might be interested in help with your writing, filing, graphing, planning or reporting. And if you want to get all your ducks in a row, line them up with the IBM Assistant Series. You can work with each program individually, or together as a team. There is also software to help you with accounting, inventory and payroll—practically anything, including communications packages to connect you to a company mainframe or outside information services.

"How expensive are they? And what if my needs change?"

With all the quality, power and performance built into IBM Personal Computers (including their extraordinary expansion capabilities), you'll find they're surprisingly affordable. But the value doesn't end there, because if your needs change you can always expand or upgrade within the IBM PC family. It's a very extensive, very compatible family of products that can help you protect your initial investment.

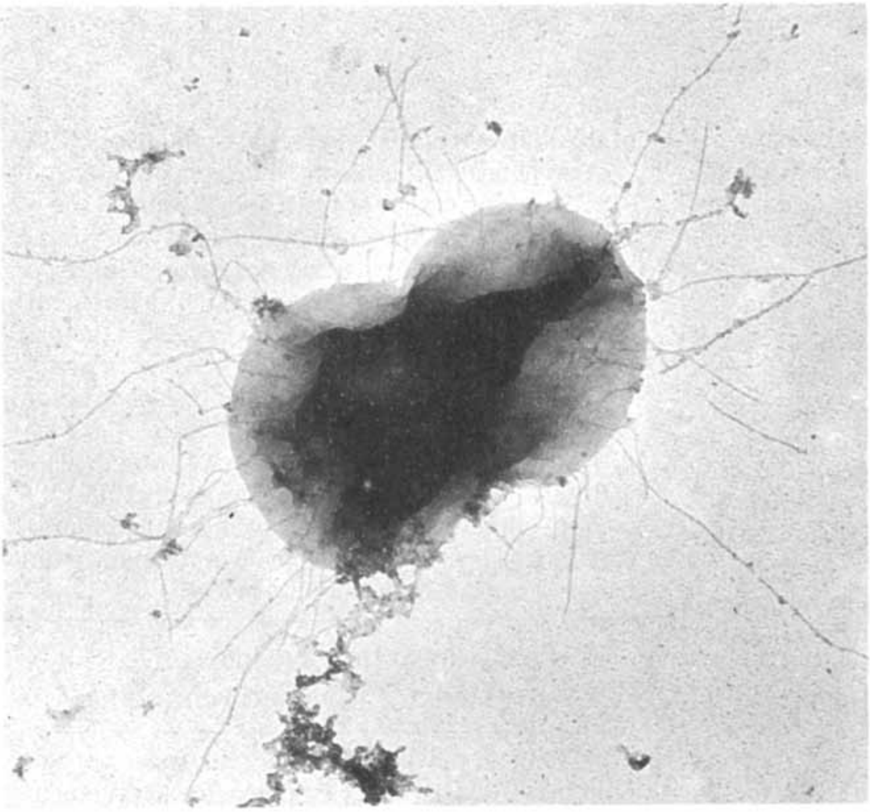
"If I want a demonstration, where do I go and who will show it to me?"

Go to any Authorized IBM Personal Computer Dealer or IBM Product Center, or contact your IBM marketing representative. All have received special training and you should find them quite helpful.



Ask to see the software programs that interest you most, and get your hands on the system yourself. Then you'll begin to see what this tool for modern times can do for you.

For a store near you, call 1-800-447-4700, Dept. HC. In Alaska or Hawaii, 1-800-447-0890. IBM®



A gonococcal bacterium with pili radiating from its surface

binding to the human cells served to identify the potential vaccine. The next step in the group's research program calls for tests of the vaccine candidate in human volunteers within a year.

The achievement of an effective vaccine against gonorrhea would be a significant medical advance. Although the disease can be treated with antibiotics, some gonococcal strains have become resistant to such drugs. In the U.S. alone gonorrhea is estimated to afflict more than a million people annually, making it the commonest communicable disease for which records are kept.

Tumor Toxin

In the late 19th century William B. Coley, a New York physician, observed that tumors in patients who contract a bacterial infection sometimes shrink or even disappear. Coley began to treat cancer patients with a mixture of killed bacteria that reproduced many of the features of an actual infection. The therapy sometimes succeeded, and unlike modern radio- and chemotherapy it did not damage healthy tissues.

For years the mechanism of the tumor regression noted by Coley remained a puzzle. An understanding of one basis of the effect may now be within reach. Over the past decade a substance that plays a crucial role in the regression has been identified, studied and, most recently, made in the laboratory by genetic-engineering techniques.

Lloyd J. Old and his colleagues at the Memorial Sloan-Kettering Cancer Center reported the existence of the substance in 1975. They discovered it by studying a well-known experimental counterpart to Coley's clinical observations. The investigators injected tumor-bearing mice with endotoxin, a substance extracted from the cell wall of certain strains of bacteria. The tumors blackened and died, a phenomenon known as hemorrhagic necrosis. Old's group found that the endotoxin acts indirectly: it causes the mice themselves to produce a substance that attacks the tumor. The workers named the substance tumor necrosis factor (TNF). TNF acts with remarkable selectivity. When mouse serum containing the substance was tested in culture, it killed or halted the development of malignant cells but did not affect normal cells.

Old and his colleagues also found out which cells make TNF. Before administering endotoxin to the mice, the workers had infected some of them with bacillus Calmette-Guérin (BCG). BCG is an attenuated tuberculosis bacillus known to have antitumor effects. Only the mice that had been primed with BCG before the injection of endotoxin produced detectable amounts of TNF (although even uninfected mice produced enough to cause hemorrhagic necrosis). Infection with BCG causes a rapid proliferation of macrophages, the immune system's scavenger cells; consequently the workers identified macrophages as the likely producers of TNF in mice.

Macrophages and other cells of the immune system are now known to produce TNF in human beings as well. Yet the basis of the factor's toxicity to cancer cells and of its capacity to distinguish malignant cells from normal ones has remained unclear. Almost certainly TNF acts in concert with other substances produced by the immune system; in 1983 Old and other workers at Sloan-Kettering noted that human TNF was most effective against tumor cells when it was tested together with interferon. A deeper understanding of the substance eluded investigators, however, largely because TNF was not available in a plentiful or pure form.

That problem has now been solved. Writing in *Nature,* a team at Genentech, Inc., headed by David V. Goeddel reports cloning the gene that codes for human TNF; two other biotechnology companies have duplicated the achievement. Detailed studies of TNF's molecular structure and the mechanisms by which it identifies and destroys malignant cells can now be made. Its role in the normal organism can be examined as well. The cloning of TNF has also opened the way to clinical trials, in which the promise of Coley's original observations will be put to the test.

Engineering Antibodies

A new approach to making antibodies in the laboratory may soon offer immunologists unprecedented flexibility in studying the structure and function of these proteins. The technique, which involves the splicing of mouse and human antibody genes, also promises to have important applications in the treatment of immune diseases. Its first medical application is planned for later this year. A group of researchers at Stanford University intends to conduct a clinical trial of a hybrid mouse-human antibody they believe may arrest the neurological damage of multiple sclerosis.

To investigate the properties of a particular antibody or to use it therapeutically one must isolate large amounts in pure form. First one injects an antigen into mice. Antibody-producing *B* lymphocytes are then extracted from the animals and fused with malignant myeloma cells to form


Ride off
into the Somerset.
It could be the ride of your life.
For the Buick Somerset is truly a very personal driver's car — a light, agile, front-wheel-drive coupe.
The Somerset's graceful appearance is enhanced by its graceful performance, owing in part to power-assisted rack-and-pinion steering, a highly capable 2.5-litre, four-cylinder engine and an array of driver controls and electronic digital readouts that are placed exactly where they belong — right in front of you, as you can see:
Being a Buick, the very untraditional Somerset also offers all the traditional amenities.
Visit your Buick dealer, buckle yourself in — and ride off into the Somerset.
To ask any questions, request a brochure or test drive, call the Buick Product Information Center, 8 a.m. to 8 p.m. Eastern time, weekdays:
1-800-85-BUICK (1-800-852-8425).
Wouldn't you
really rather have a Buick?
BUICK
GM
Buick Somerset

"hybridomas": cells that multiply freely in culture while secreting antibodies. Finally, the desired hybridoma is isolated and cloned. The clone provides a steady supply of a specific, "monoclonal" antibody.

The new genetic-engineering technique is an extension of this well-established procedure. Writing in *Proceedings of the National Academy of Sciences,* Sherie L. Morrison of Columbia University, Vernon T. Oi of the Becton-Dickinson Monoclonal Center and their colleagues report the successful splicing of antibody gene segments from mouse hybridomas with human antibody genes. The mouse fraction of the recombined genes codes for the variable region of the antibody—the end that binds to the antigen and gives the antibody its specificity; the human gene segment codes for the larger, constant region that determines the antibody's biological function. When the hybrid genes are "transfected" into mouse myeloma cells, the cells produce hybrid antibodies.

The hybrids are of predominantly human origin, which suggests their potential clinical benefits. Until now most of the monoclonal antibodies administered to human patients have been derived from mice. (Human hybridomas tend to be unstable in the laboratory.) Mouse antibodies, however, are often rejected by the human immune system. A rejection of near-human antibodies is thought to be less likely.

The Stanford group led by Lawrence Steinman intends to exploit these improved odds in treating multiple sclerosis. While the causes of multiple sclerosis remain obscure, a widely accepted hypothesis holds that it results from an abnormality among the patient's own "helper" *T* lymphocytes, which leads to an immune attack on the myelin sheath that protects nerve cells. In a recent issue of *Science* Steinman and his colleagues report that the course of an analogous disease in mice can be reversed by a monoclonal antibody that binds to a particular surface marker on helper *T* cells. When the antibody was injected after the appearance of early symptoms, the disease was halted in 90 percent of the mice.

For the forthcoming clinical trial the workers will link the variable-region segment of the mouse antibody gene with constant-region DNA from human cells. The resulting hybrid, they hope, will recognize and destroy helper *T* cells in multiple sclerosis patients without inducing an immune reaction.

The long-term benefits of recombining antibody genes are likely to be even more significant. According to Morrison and Oi, through selective genetic manipulations workers should eventually be able to determine how different antibodies carry out their diverse functions. Someday it may even prove possible to make new antibodies for specific medical purposes that improve on the natural models.

Backward Glance

Can two measurements, one made in the past and the other made in the future, give us more knowledge about the present state of affairs than a measurement made right now? Even more paradoxically, can a measurement made in the future specify aspects of the present that, without the future measurement, would have remained forever indeterminate? Three theoretical physicists have now presented a quantum-mechanical argument that suggests an affirmative answer to both questions.

The argument is put forward in a recent issue of *Physical Review Letters* by David Z. Albert and Yakir Aharonov of the University of South Carolina and Susan D'Amato of Furman University. It centers on a cornerstone of quantum mechanics called the uncertainty principle. According to the uncertainty principle, the precise values of two so-called noncommuting variables cannot be determined at the same time.

For example, suppose one wants to measure the spin components of an elementary particle. In quantum mechanics the x and y components of the spin are subject to the uncertainty principle. If one makes an exact measurement of the x component of the spin, one can have no knowledge at all about the y component; if one measures the y component instead, the x component must be unknown.

Albert, Aharonov and D'Amato consider a thought experiment first put forward some 50 years ago. Suppose the x component of a spin is precisely measured on Monday. The measurement completely obscures any knowledge about the value of the y component, but it can be done in such a way that the x component is unaffected. Furthermore, suppose the y component is measured on Friday in such a way that it too does not disturb the result a second measurement of the y component would give.

What can one say on Friday about the results of the two measurements? It seemed to the theorists who first formulated the thought experiment that one could say the following: If a measurement of the x component had been made on Wednesday, the result would have been the same as the value obtained for the x component on Monday. Moreover, if a measurement of the *y* component had been made on Wednesday, the result would have been the same as the value obtained for the *y* component on Friday.

In short, on Friday one would be in a position to say what the result of a measurement of either the *x* or the *y* component would have been on Wednesday. No single measurement of the two spin components on Wednesday could have put an observer into the same position: the measurement of either spin component would have disturbed the value of the other.

For technical reasons many physicists have explicitly rejected the conclusion of the thought experiment. Instead they have accepted an interpretation of the uncertainty principle in which the uncertainty is not merely a matter of measurement but extends to the description of physical reality. In this standard interpretation, not only is it impossible to measure the *x* and *y* components of a particle simultaneously but also it is impossible to state what the two components would be at a given time if one or the other were measured. In other words, the two conditional statements made on Friday in the thought experiment have been assumed to be ruled out of bounds.

The three theorists have now shown that the technical objections raised against the thought experiment were ill-founded. Hence the standard interpretation of the uncertainty principle must be rejected. It may be, however, that there is even more at stake. Some theorists have taken an even stronger view of the uncertainty principle, namely that only the observation of a quantum phenomenon can make a noncommuting variable take on a definite value. The rest of the time its value is not only unknown but also physically "smeared out": the quantity literally has no definite value.

If the strong view is independent of the interpretation of the uncertainty principle that has now been unseated, the conclusions are quite paradoxical. According to the strong view, it is not possible to say on Wednesday that both spin components have a definite value. On the other hand, it is possible to say on Friday that the two spin components would have had such and such values on Wednesday, if one or the other component had been measured. The measurement on Friday caused, in some sense of the word cause, the smeared-out values of the spin on Wednesday to collapse into some definite configuration. The logical puzzles about time and causality that this development engenders have not yet been fully explored.

Elementary Particles and Forces

A coherent view of the fundamental constituents of matter and the forces governing them has emerged. It embraces disparate theories, but they may soon be united in one comprehensive description of natural events

by Chris Quigg

The notion that a fundamental simplicity lies below the observed diversity of the universe has carried physics far. Historically the list of particles and forces considered to be elementary has changed continually as closer scrutiny of matter and its interactions revealed microcosms within microcosms: atoms within molecules, nuclei and electrons within atoms, and successively deeper levels of structure within the nucleus. Over the past decade, however, experimental results and the convergence of theoretical ideas have brought new coherence to the subject of particle physics, raising hopes that an enduring understanding of the laws of nature is within reach.

Higher accelerator energies have made it possible to collide particles with greater violence, revealing the subatomic realm in correspondingly finer detail; the limit of experimental resolution now stands at about 10^{-16} centimeter, about a thousandth the diameter of a proton. A decade ago physics recognized hundreds of apparently elementary particles; at today's resolution that diversity has been shown to represent combinations of a much smaller number of fundamental entities. Meanwhile the forces through which these constituents interact have begun to display underlying similarities. A deep connection between two of the forces, electromagnetism and the weak force that is familiar in nuclear decay, has been established, and prospects are good for a description of fundamental forces that also encompasses the strong force that binds atomic nuclei.

Of the particles that now appear to be structureless and indivisible, and therefore fundamental, those that are not affected by the strong force are known as leptons. Six distinct types, fancifully called flavors, of lepton have been identified. Three of the leptons, the electron, the muon and the tau, carry an identical electric charge of -1; they differ, however, in mass. The electron is the lightest and the tau the heaviest of the three. The other three, the neutrinos, are, as their name suggests, electrically neutral. Two of them, the electron neutrino and the muon neutrino, have been shown to be nearly massless. In spite of their varied masses all six leptons carry precisely the same amount of spin angular momentum. They are designated spin-1/2 because each particle can spin in one of two directions. A lepton is said to be right-handed if the curled fingers of a right hand indicate its rotation when the thumb points in its direction of travel and left-handed when the fingers and thumb of the left hand indicate its spin and direction.

For each lepton there is a corresponding antilepton, a variety of antiparticle. Antiparticles have the same mass and spin as their respective particles but carry opposite values for other properties, such as electric charge. The antileptons, for example, include the antielectron, or positron, the antimuon and the antitau, all of which are positively charged, and three electrically neutral antineutrinos.

In their interactions the leptons seem to observe boundaries that define three families, each composed of a charged lepton and its neutrino. The families are distinguished mathematically by lepton numbers; for example, the electron and the electron neutrino are assigned electron number 1, muon number 0 and tau number 0. Antileptons are assigned lepton numbers of the opposite sign. Although some of the leptons decay into other leptons, the total lepton number of the decay products is equal to that of the original particle; consequently the family lines are preserved.

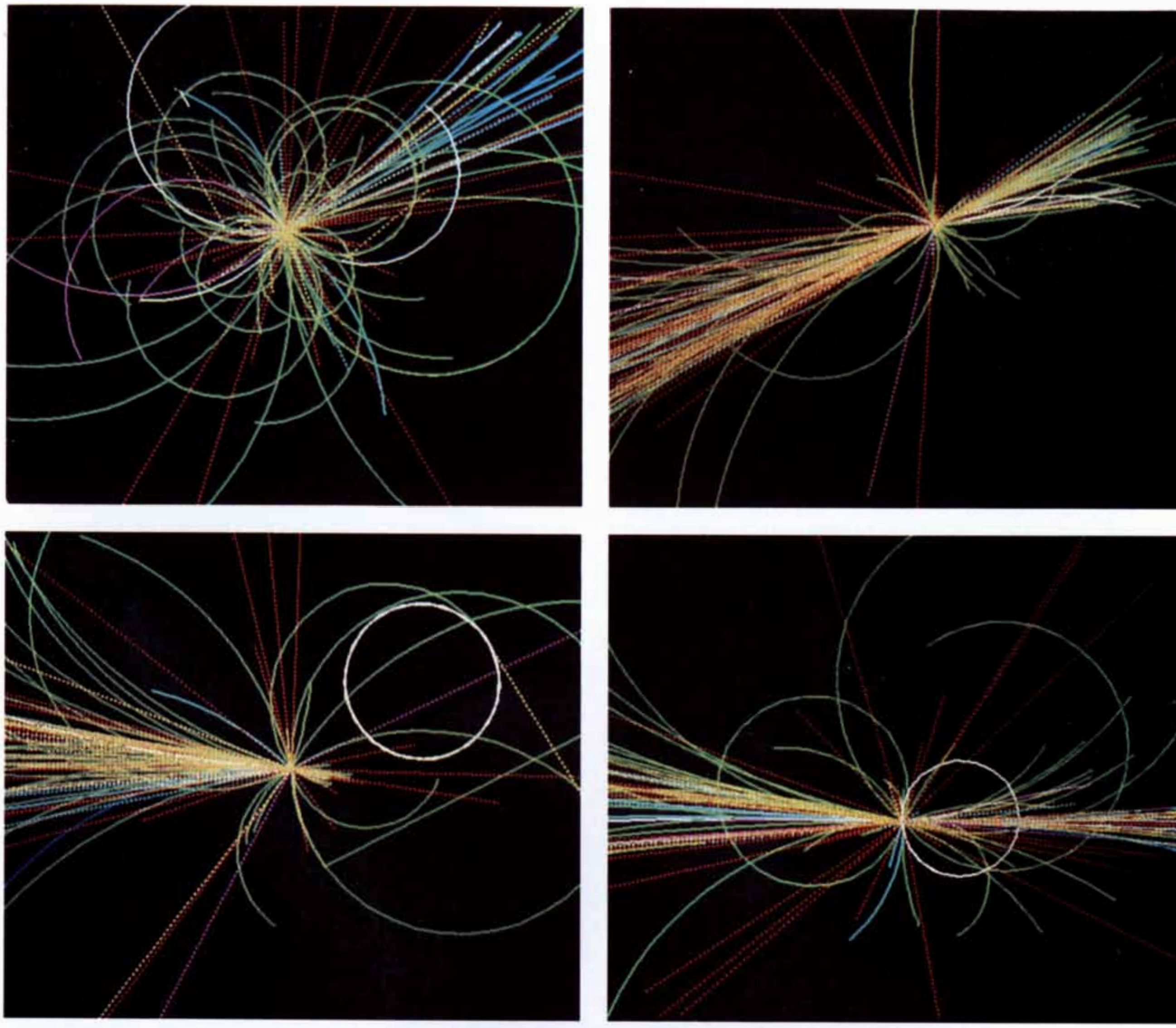
The muon, for example, is unstable. It decays after a mean lifetime of 2.2 microseconds into an electron, an electron antineutrino and a muon neutrino through a process mediated by the weak force. Total lepton number is unaltered in the transformation. The muon number of the muon neutrino is 1, the electron number of the electron is 1 and that of the electron antineutrino is -1. The electron numbers cancel, leaving the initial muon number of 1 unchanged. Lepton number is also conserved in the decay of the tau, which endures for a mean lifetime of 3×10^{-13} second.

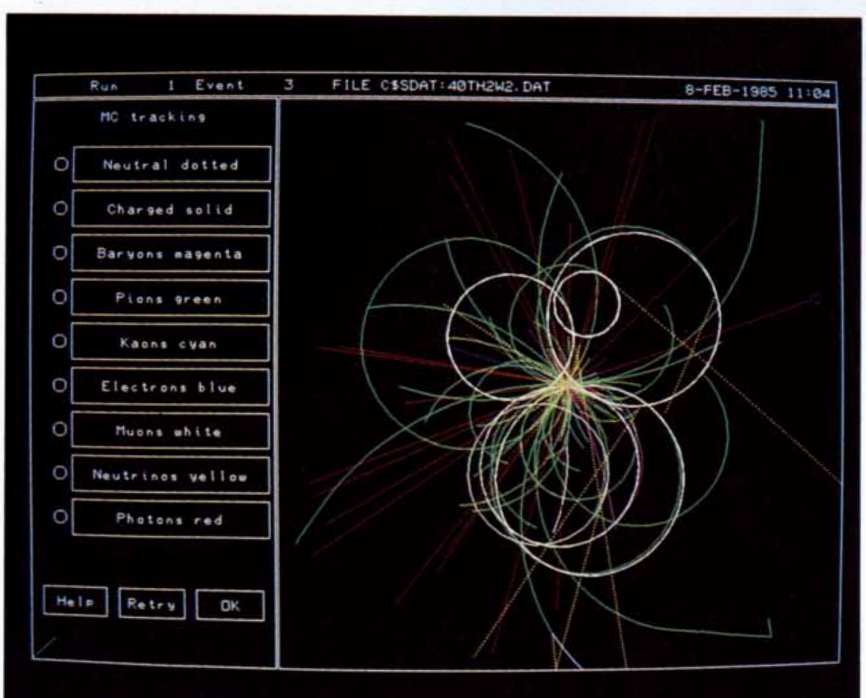
The electron, however, is absolutely stable. Electric charge must be conserved in all interactions, and there is no less massive charged particle into which an electron could decay. The decay of neutrinos has not been observed. Because neutrinos are the less massive members of their respective families, their decay would necessarily cross family lines.

Where are leptons observed? The electron is familiar as the carrier of electric charge in metals and semiconductors. Electron antineutrinos are emitted in the beta decay of neutrons into protons. Nuclear reactors, which produce large numbers of unstable free neutrons, are abundant sources of antineutrinos. The remaining species of lepton are produced mainly in high-energy collisions of subnuclear particles, which occur naturally as cosmic rays interact with the atmosphere and under controlled conditions in particle accelerators. Only the tau neutrino has not been observed directly, but the indirect evidence for its existence is convincing.

Quarks

Subnuclear particles that experience the strong force make up the second great class of particles studied in the laboratory. These are the hadrons; among them are the protons, the neutrons and the mesons. A host of other less familiar hadrons exist only ephemerally as the products of high-





DEBRIS of a hypothetical high-energy collision between two protons is depicted in computer simulations made in accordance with the known and conjectured behavior of elementary particles. James Freeman of the Collider Detector Group at the Fermi National Accelerator Laboratory (Fermilab) devised the simulation program using the ISAJET model devised by Frank E. Paige, Jr., of the Brookhaven National Laboratory. The collision, possible outcomes of which are shown, takes place at an energy of 40 TeV (trillion electron volts), far greater than can be reached in today's accelerators. The enormous energy is assumed to give rise to a Higgs boson, a massive particle that plays a crucial role in theory but has not been observed. The Higgs boson promptly decays into two *W* bosons, also short-lived and massive, which then decay by several routes. Some of the particles whose tracks are plotted are the products of the *W* bosons' decay; others emerged from the breakup of the incident protons. The electrons, muons and neutrinos are elementary particles; the baryons, pions and kaons are composites of fundamental constituents, and the photons are quanta of energy. A magnetic field is simulated, causing paths of charged particles to curve, whereas neutral ones are not affected.

energy collisions, from which extremely massive and very unstable particles can materialize. Hundreds of species of hadron have been catalogued, varying in mass, spin, charge and other properties.

Hadrons are not elementary particles, however, since they have internal structure. In 1964 Murray Gell-Mann of the California Institute of Technology and George Zweig, then working at CERN, the European laboratory for particle physics in Geneva, independently attempted to account for the bewildering variety of hadrons by suggesting they are composite particles, each a different combination of a small number of fundamental constituents. Gell-Mann called them quarks. Studies at the Stanford Linear Accelerator Center (SLAC) in the late 1960's in which high-energy electrons were fired at protons and neutrons bolstered the hypothesis. The distribution in energy and angle of the scattered electrons indicated that some were colliding with pointlike, electrically charged objects within the protons and neutrons.

Particle physics now attributes all known hadron species to combinations of these fundamental entities. Five kinds, also termed flavors, of quark have been identified—the up (u), down (d), charm (c), strange (s) and bottom (b) quarks—and a sixth flavor, the top (t) quark, is believed to exist. Like the leptons, quarks have half a unit of spin and can therefore exist in left- and right-handed states. They also carry electric charge equal to a precise fraction of an electron's charge: the d, s and b quarks have a charge of $-1/3$, and the u, c and the conjectured t quark have a charge of $+2/3$. The corresponding antiquarks have electric charges of the same magnitude but opposite sign.

Such fractional charges are never observed in hadrons, because quarks form combinations in which the sum of their charges is integral. Mesons, for example, consist of a quark and an antiquark, whose charges add up to -1, 0 or $+1$. Protons and neutrons consist respectively of two u quarks and a d quark, for a total charge of $+1$, and of a u quark and two d quarks, for a total charge of 0.

Like leptons, the quarks experience weak interactions that change one species, or flavor, into another. For example, in the beta decay of a neutron into a proton one of the neutron's d quarks metamorphoses into a u quark, emitting an electron and an antineutrino in the process. Similar transformations of c quarks into s quarks have been observed. The pattern of decays suggests two family groupings, one of them thought to contain the u and the d quarks and the second the c and the s quarks. In apparent contrast to the behavior of leptons, some quark decays do cross family lines, however; transformations of u quarks into s quarks and of c quarks into d quarks have been observed. It is the similarity of the two known quark families to the families of leptons that first suggested the existence of a t quark, to serve as the partner of the b quark in a third family.

In contrast to the leptons, free quarks have never been observed. Yet

LEPTONS			
PARTICLE NAME	SYMBOL	MASS AT REST (MeV/c²)	ELECTRIC CHARGE
ELECTRON NEUTRINO	ν_e	ABOUT 0	0
ELECTRON	e or e^-	0.511	−1
MUON NEUTRINO	ν_μ	ABOUT 0	0
MUON	μ or μ^-	106.6	−1
TAU NEUTRINO	ν_τ	LESS THAN 164	0
TAU	τ or τ^-	1,784	−1

QUARKS			
PARTICLE NAME	SYMBOL	MASS AT REST (MeV/c²)	ELECTRIC CHARGE
UP	u	310	⅔
DOWN	d	310	−⅓
CHARM	c	1,500	⅔
STRANGE	s	505	−⅓
TOP TRUTH	t	> 22,500: HYPOTHETICAL PARTICLE	⅔
BOTTOM BEAUTY	b	ABOUT 5,000	−⅓

FORCE	RANGE	STRENGTH AT 10^{-13} CENTIMETER IN COMPARISON WITH STRONG FORCE	CARRIER	MASS AT REST (GeV/c²)	SPIN	ELECTRIC CHARGE	REMARKS
GRAVITY	INFINITE	10^{-38}	GRAVITON	0	2	0	CONJECTURED
ELECTROMAGNETISM	INFINITE	10^{-2}	PHOTON	0	1	0	OBSERVED DIRECTLY
WEAK	LESS THAN 10^{-16} CENTIMETER	10^{-13}	INTERMEDIATE BOSONS: W^+	81	1	+1	OBSERVED DIRECTLY
			W^-	81	1	−1	OBSERVED DIRECTLY
			Z^0	93	1	0	OBSERVED DIRECTLY
STRONG	LESS THAN 10^{-13} CENTIMETER	1	GLUONS	0	1	0	PERMANENTLY CONFINED

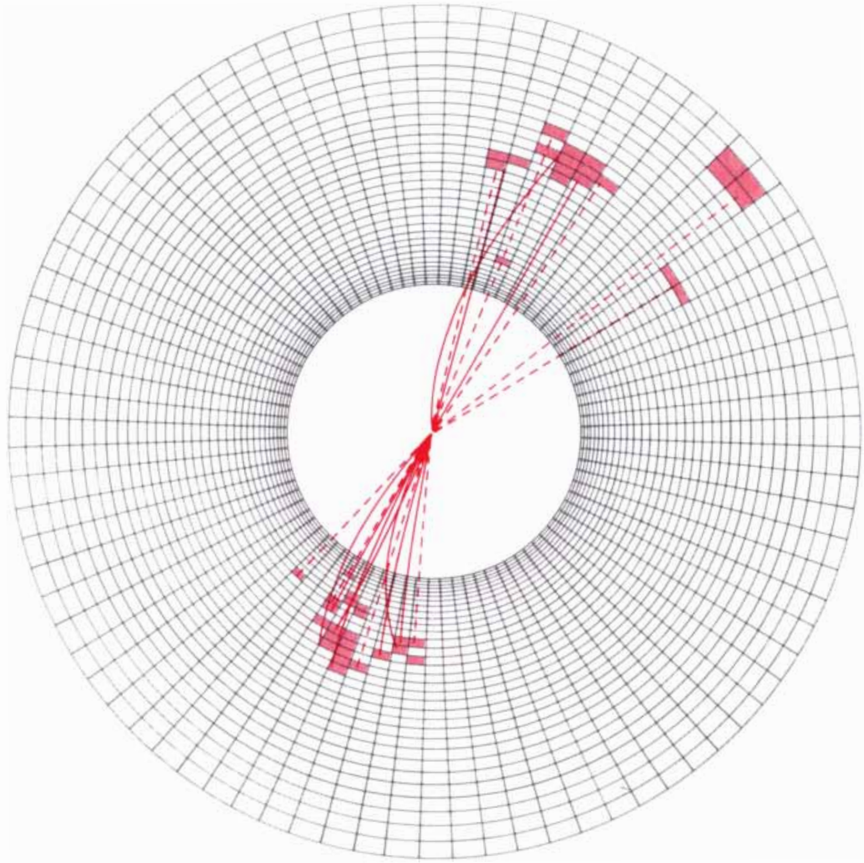
FUNDAMENTAL SCHEME OF NATURE, according to current theory, embraces 12 elementary particles (*top*) and four forces (*bottom*). All the particles listed are thought to be structureless and indivisible; among their properties are an identical amount of spin, given by convention as 1/2, and differing values of electric charge, color charge and mass, given as energy in millions of electron volts (MeV) divided by the square of the speed of light (c). Only the pairs of leptons and quarks at the top of each column are found in ordinary matter; the other particles are observed briefly in the aftermath of high-energy collisions. The four forces thought to govern matter vary in range and strength; although the strong force is the most powerful, it acts only over a distance of less than 10^{-13} centimeter, the diameter of a proton. All the forces are conveyed by force particles, whose masses are given in billions of electron volts (GeV) divided by the square of the speed of light. Because of its weakness, gravity has not been studied experimentally by particle physicists.

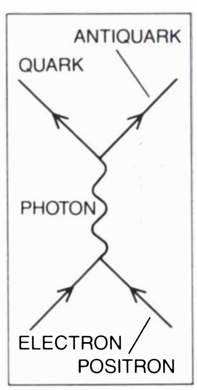
circumstantial evidence for their existence has mounted steadily. One indication of the soundness of the quark model is its success in predicting the outcome of high-energy collisions of an electron and a positron. Because they represent matter and antimatter, the two particles annihilate each other, releasing energy in the form of a photon. The quark model predicts that the energy of the photon can materialize into a quark and an antiquark. Because the colliding electron-positron pair had a net momentum of 0, the quark-antiquark pair must diverge in opposite directions at equal velocities so that their net momentum is also 0. The quarks themselves go unobserved because their energy is converted into additional quarks and antiquarks, which materialize and combine with the original pair, giving rise to two jets of hadrons (most of them pions, a species of meson). Such jets are indeed observed, and their focused nature confirms that the hadrons did not arise directly from the collision but from single, indivisible particles whose trajectories the jets preserve.

The case for the reality of quarks is also supported by the variety of energy levels, or masses, at which certain species of hadron, notably the psi and the upsilon particles, can be observed in accelerator experiments. Such energy spectra appear analogous to atomic spectra: they seem to represent the quantum states of a bound system of two smaller components. Each of its quantum states would represent a different degree of excitation and a different combination of the components' spins and orbital motion. To most physicists the conclusion that such particles are made up of quarks is irresistible. The psi particle is held to consist of a *c* quark and its antiquark, and the upsilon particle is believed to comprise a *b* quark and its antiquark.

What rules govern the combinations of quarks that form hadrons? Mesons are composed of a quark and an antiquark. Because each quark has a spin of 1/2, the net spin of a meson is 0 if its constituents spin in opposite directions and 1 if they spin in the same direction, although in their excited states mesons may have larger values of spin owing to the quarks' orbital motion. The other class of hadrons, the baryons, consist of three quarks each. Summing the constituent quarks' possible spins and directions yields two possible values for the spin of the least energetic baryons: 1/2 and 3/2. No other combinations of quarks have been observed; hadrons that consist of two or four quarks seem to be ruled out.

The reason is linked with the answer to another puzzle. According to the exclusion principle of Wolfgang Pauli, no two particles occupying a minute region of space and possessing half-integral spins can have the same quantum number—the same values of momentum, charge and spin. The Pauli exclusion principle accounts elegantly for the configurations of electrons that determine an element's place in the periodic table. We should expect it to be a reliable guide to the panoply of hadrons as well. The principle would seem to suggest, however, that exotic hadrons such as the delta plus plus and the omega minus particles, which materialize briefly following high-energy collisions, cannot exist. They consist respectively of three *u* and three *s* quarks and possess a spin of 3/2; all three quarks in each of the hadrons must be identical in spin as well as in other properties and hence must occupy the same quantum state.





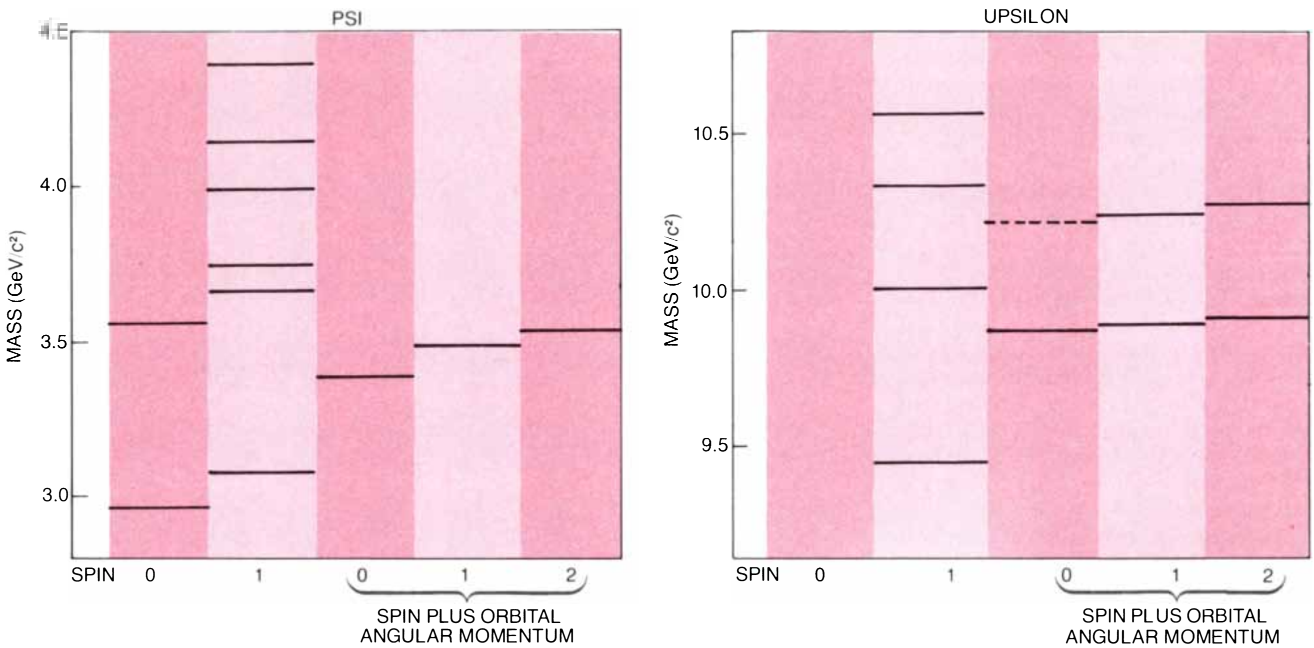
EVIDENCE OF QUARKS, two narrow jets of particles emerge from the collision and mutual annihilation of an electron and an antielectron, or positron. The annihilation releases energy, which gives rise to matter. The detected particles have a variety of masses and spins; some are neutral (*broken lines*) and some electrically charged (*solid lines*). If the particles arose directly from the annihilation, they would be expected to follow widely divergent paths. The focused character of the jets suggests instead that each jet developed from a single precursor: a quark or an antiquark. They are the immediate products of the photon of electromagnetic energy released in the collision, which is diagrammed at the left using arrows to represent the relative motion of the particles. The event shown was recorded in the JADE detector of the PETRA accelerator at the Deutsches Elektronen-Synchrotron (DESY) in Hamburg. The paths of the particles were reconstructed by computer from ionization tracks and from the pattern of energy (*color*) deposited as the particles struck the inner layer of the 2.4-meter-long cylindrical detector.

Colors

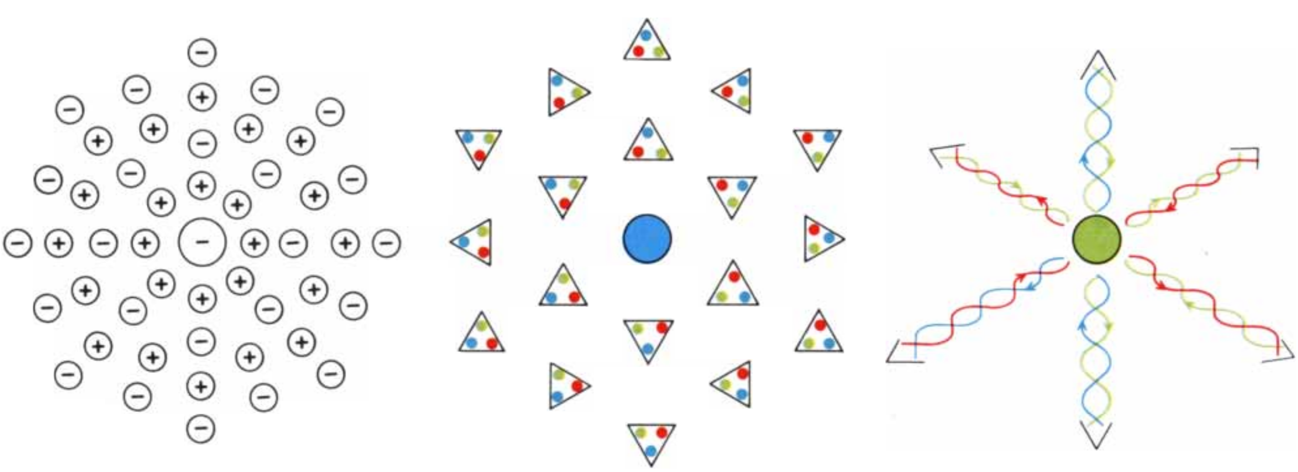
To explain such observed combinations it is necessary to suppose the three otherwise identical quarks are distinguished by another trait: a new kind of charge, whimsically termed color, on which the strong force acts. Each flavor of quark can carry one of three kinds of color charge: red, green or blue. To a red quark there corresponds an antiquark with a color charge of antired (which may be thought of as cyan); other antiquarks bear charges of antigreen (magenta) and antiblue (yellow).

The analogy between this new kind of charge and color makes it possible to specify the rules under which quarks combine. Hadrons do not exhibit a color charge; the sum of the component quarks' colors must be white, or color-neutral. Therefore the only allowable combinations are those of a quark and its antiquark, giving rise to mesons, and of a red, a green and a blue quark, yielding the baryons.

Colored states are never seen in isolation. This concealment is consistent with the fact that free quarks, bearing a single color charge, have never been observed. The activity of the strong



VARIETY OF MASSES at which two-quark systems known as the psi (*left*) and the upsilon (*right*) particles are observed reveal the energy states each can adopt. The psi particle consists of a *c* quark and its antiquark, bound by the color force; the upsilon particle is a similar combination of a *b* quark and its antiquark. Each column within a spectrum of masses corresponds to a different combination of the quarks' spins and orbital angular momentum. Different masses or, equivalently, energy levels within a column represent quantum levels of excitation. The resemblance of the spectra to the spectra of atoms indicates that particles such as the psi and the upsilon are also bound systems of smaller constituents. Such spectra offer insight into the behavior of the color force at short distances.



SCREENING AND CAMOUFLAGE EFFECTS modify the behavior of fundamental forces over distance. The left panel shows an electron in a vacuum; it is surrounded by short-lived pairs of virtual electrons and positrons, which in quantum theory populate the vacuum. The electron attracts the virtual positrons and repels the virtual electrons, thereby screening itself in positive charge. The farther from the electron a real charge is, the thicker the intervening screen of virtual positive charges is and the smaller the electron's effective charge will be. The color force is subject to the same screening effect (*center*). Virtual color charges (mostly quark-antiquark pairs) fill the vacuum; a colored quark attracts contrasting colors, thereby surrounding itself with a screen that acts to reduce its effective charge at increasing distances. An effect called camouflage counteracts screening, however. A quark continuously radiates and reabsorbs gluons that carry its color charge to considerable distances and change its color, in this case from blue to green (*right*). A charge's full magnitude can be felt only outside the space it occupies. Therefore camouflage acts to increase the force felt by an actual quark as it moves away from the first quark, toward the edge of the color-charged region. The net result of screening and camouflage is that at close range the strong interaction, which is based on the color charge, is weaker, whereas at longer ranges it is stronger.

force between colored quarks must be extraordinarily powerful, perhaps powerful enough to confine quarks permanently within colorless, or color-neutral, hadrons. The description of violent electron-positron collisions according to the quark model, however, assumes the quarks that give rise to the observed jets of hadrons diverge freely during the first instant following the collision. The apparent independence of quarks at very short distances is known as asymptotic freedom; it was described in 1973 by David J. Gross and Frank Wilczek of Princeton University and by H. David Politzer, then at Harvard University.

Analogy yields an operational understanding of this paradoxical state of affairs, in which quarks interact only weakly when they are close together and yet cannot be separated. We may think of a hadron as a bubble within which quarks are imprisoned. Within the bubble the quarks move freely, but they cannot escape from it. The bubbles, of course, are only a metaphor for the dynamical behavior of the force between quarks, and a fuller explanation for what is known as quark confinement can come only from an examination of the forces through which particles interact.

The Fundamental Interactions

Nature contrives enormous complexity of structure and dynamics from the six leptons and six quarks now thought to be the fundamental constituents of matter. Four forces govern their relations: electromagnetism, gravity and the strong and weak forces. In the larger world we experience directly, a force can be defined as an agent that alters the velocity of a body by changing its speed or direction. In the realm of elementary particles, where quantum mechanics and relativity replace the Newtonian mechanics of the larger world, a more comprehensive notion of force is in order, and with it a more general term, interaction. An interaction can cause changes of energy, momentum or kind to occur among several colliding particles; an interaction can also affect a particle in isolation, in a spontaneous decay process.

Only gravity has not been studied at the scale on which elementary particles exist; its effects on such minute masses are so small that they can safely be ignored. Physicists have attempted with considerable success to predict the behavior of the other three interactions through mathematical descriptions known as gauge theories.

The notion of symmetry is central to gauge theories. A symmetry, in the mathematical sense, arises when the solutions to a set of equations remain the same even though a characteristic of the system they describe is altered. If a mathematical theory remains valid when a characteristic of the system is changed by an identical amount at every point in space, it can be said that the equations display a global symmetry with respect to that characteristic. If the characteristic can be altered independently at every point in space and the theory is still valid, its equations display local symmetry with respect to the characteristic.

Each of the four fundamental forces is now thought to arise from the invariance of a law of nature, such as the conservation of charge or energy, under a local symmetry operation, in which a certain parameter is altered independently at every point in space. An analogy with an ideal rubber disk may help to visualize the effect of the mathematics. If the shape of the rubber disk is likened to a natural principle and the displacement of a point within the disk is regarded as a local symmetry operation, the disk must keep its shape even as each point within it is displaced independently. The displacements stretch the disk and introduce forces between points. Similarly, in gauge theories the fundamental forces are the inevitable consequences of local symmetry operations; they are required in order to preserve symmetry.

Of the three interactions studied in the realm of elementary particles, only electromagnetism is the stuff of everyday experience, familiar in the form of sunlight, the spark of a static discharge and the gentle swing of a compass needle. On the subatomic level it takes on an unfamiliar aspect. According to relativistic quantum theory, which links matter and energy, electromagnetic interactions are mediated by photons: massless "force particles" that embody precise quantities of energy. The quantum theory of electromagnetism, which describes the photon-mediated interactions of electrically charged particles, is known as quantum electrodynamics (QED).

In common with other theories of the fundamental interactions, QED is a gauge theory. In QED the electromagnetic force can be derived by requiring that the equations describing the motion of a charged particle remain unchanged in the course of local symmetry operations. Specifically, if the phase of the wave function by which a charged particle is described in quantum theory is altered independently at every point in space, QED requires that the electromagnetic interaction and its mediating particle, the photon, exist in order to maintain symmetry.

QED is the most successful of physical theories. Using calculation methods developed in the 1940's by Richard P. Feynman and others, it has achieved predictions of enormous accuracy, such as the infinitesimal effect of the photons radiated and absorbed by an electron on the magnetic moment generated by the electron's innate spin. Moreover, QED's descriptions of the electromagnetic interaction have been verified over an extraordinary range of distances, varying from less than 10^{-18} meter to more than 10^8 meters.

Screening

In particular QED has explained the effective weakening of the electromagnetic charge with distance. The electric charge carried by an object is a fixed and definite quantity. When a charge is surrounded by other freely moving charges, however, its effects may be modified. If an electron enters a medium composed of molecules that have positively and negatively charged ends, for example, it will polarize the molecules. The electron will repel their negative ends and attract their positive ends, in effect screening itself in positive charge. The result of the polarization is to reduce the electron's effective charge by an amount that increases with distance. Only when the electron is inspected at very close range—on a submolecular scale, within the screen of positive charges—is its full charge apparent.

Such a screening effect seemingly should not arise in a vacuum, in which there are no molecules to become polarized. The uncertainty principle of Werner Heisenberg suggests, however, that the vacuum is not empty. According to the principle, uncertainty about the energy of a system increases as it is examined on progressively shorter time scales. Particles may violate the law of the conservation of energy for unobservably brief instants; in effect, they may materialize from nothingness. In QED the vacuum is seen as a complicated and seething medium in which pairs of charged "virtual" particles, particularly electrons and positrons, have a fleeting existence. These ephemeral vacuum fluctuations are polarizable just as are the molecules of a gas or a liquid. Accordingly QED predicts that in a vacuum too electric charge will be screened and effectively reduced at large distances.

The strong interaction affecting quarks that is based on the color charge also varies with distance, although in a contrary manner: instead of weakening with distance the color charge appears to grow stronger. Only at distances of less than about 10^{-13} centimeter, the diameter of a proton, does it diminish enough to allow mutually bound quarks a degree of independence. Yet the explanation for this peculiar behavior is found in a theory that is closely modeled on QED. It is a theory called quantum chromodynamics (QCD), the gauge theory of the strong interactions.

Like QED, QCD postulates force particles, which mediate interactions. Colored quarks interact through the exchange of entities called gluons, just as charged particles trade photons. Whereas QED recognizes only one kind of photon, however, QCD admits eight kinds of gluon. In contrast to the photons of QED, which do not alter the charge of interacting particles, the emission or absorption of a gluon can change a quark's color; each of the eight gluons mediates a different transformation. The mediating gluon is itself colored, bearing both a color and an anticolor.

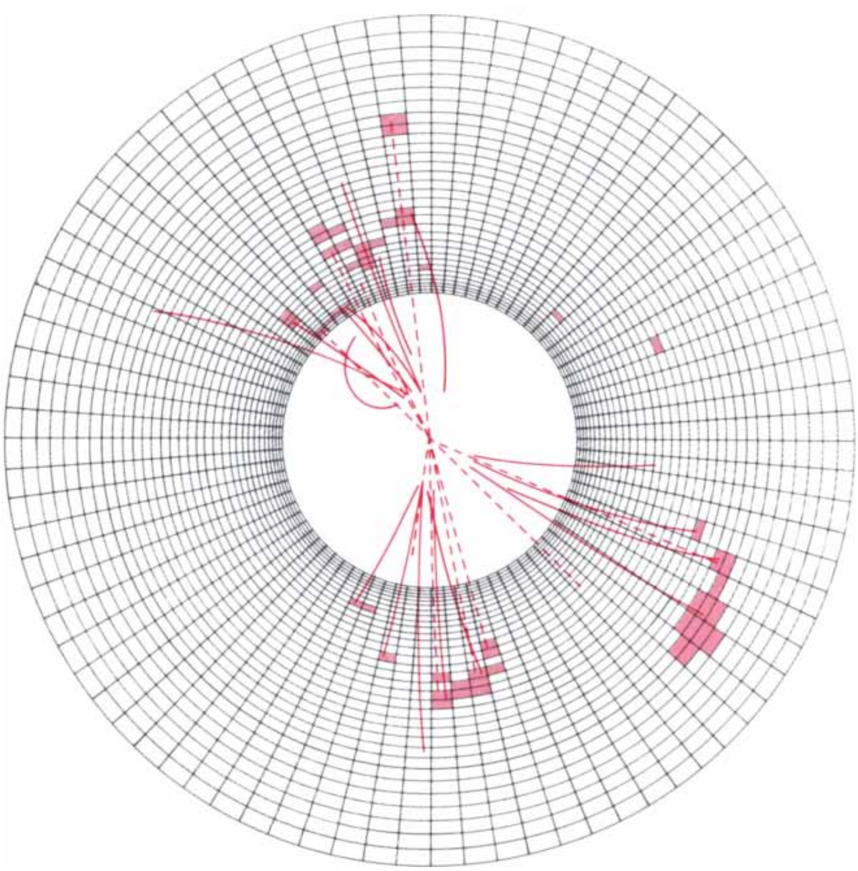
The fact that the gluons are color-charged, in contrast to the electrically neutral photons of QED, accounts for the differing behaviors over distance of the electromagnetic and strong interactions. In QCD two competing effects govern the effective charge: screening, analogous to the screening of QED, and a new effect known as camouflage. The screening, or vacuum polarization, resembles that in electromagnetic interactions. The vacuum of QCD is populated by pairs of virtual quarks and antiquarks, winking into and out of existence. If a quark is introduced into the vacuum, virtual particles bearing contrasting color charges will be attracted to the quark; those bearing a like charge will be repelled. Hence the quark's color charge will be hidden within a cloud of unlike colors, which serves to reduce the effective charge of the quark at greater distances.

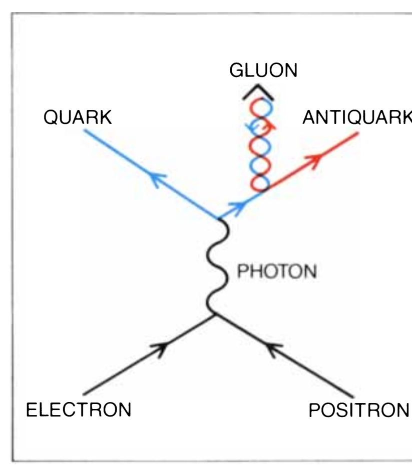
Camouflage

Within this polarized vacuum, however, the quark itself continuously emits and reabsorbs gluons, thereby changing its color. The color-charged gluons propagate to appreciable distances. In effect they spread the color charge throughout space, thus camouflaging the quark that is the source of the charge. The smaller an arbitrary region of space centered on the quark is, the smaller will be the proportion of the quark's color charge contained in it. Thus the color charge felt by a quark of another color will diminish as it approaches the first quark. Only at a large distance will the full magnitude of the color charge be apparent.

In QCD the behavior of the strong force represents the net effect of screening and camouflage. The equations of QCD yield a behavior that is consistent with the observed paradox of quarks: they are both permanently confined and asymptotically free. The strong interaction is calculated to become extraordinarily strong at appreciable distances, resulting in quark confinement, but to weaken and free quarks at very close range.

In the regime of short distances that is probed in high-energy collisions, strong interactions are so enfeebled that they can be described using the methods developed in the context of QED for the much weaker electromagnetic interaction. Hence some of the same precision that characterizes QED can be imparted to QCD. The evolution of jets of hadrons from a quark and an antiquark generated in electron-positron annihilation, for example, is a strong interaction. QCD predicts that if the energy of the collision is high enough, the quark and the antiquark moving off in opposite directions may generate not two but three jets of hadrons. One of the particles will radiate a gluon, moving in a third direction. It will also evolve into



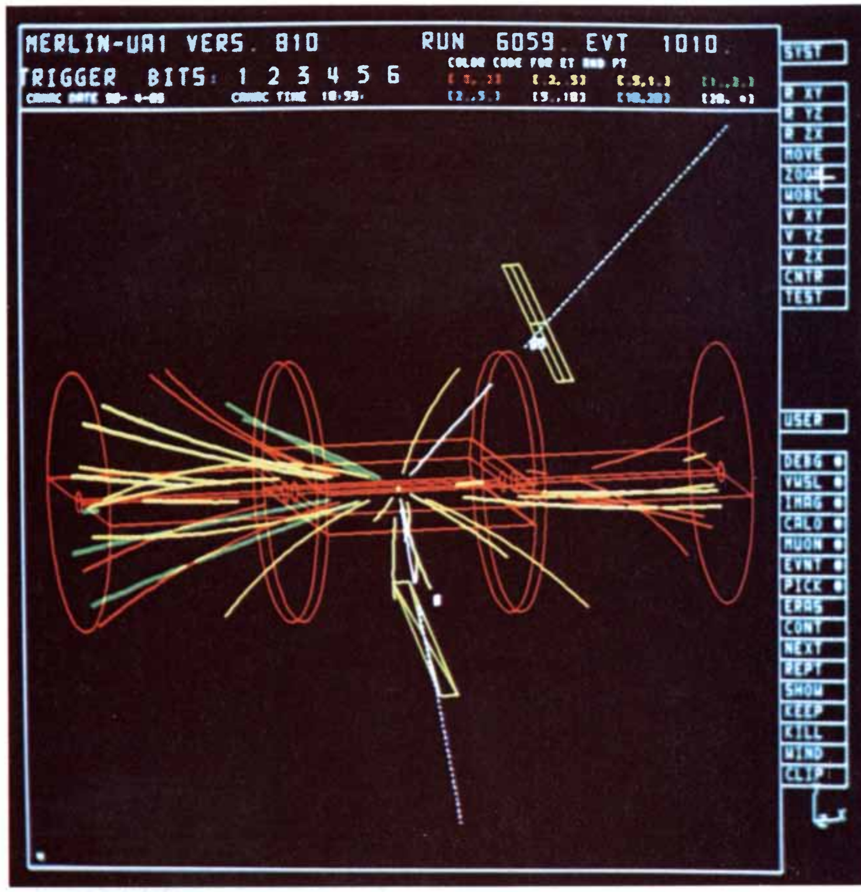


THREE-JET EVENT, recorded in the JADE detector, confirms the existence of the gluon, the mediating particle of the color force. An electron and a positron collided at high energy, creating a quark and an antiquark, as in the event shown on page 87. In this case one of the quarks radiated a gluon (*above*). The quarks and the gluon diverged; each promptly gave rise to a shower of particles, which preserved the trajectory of the original entity (*left*). The event reveals the asymptotic freedom of quarks and gluons: their ability to move independently within a very small region in spite of the enormous strength of the color force across larger distances.

hadrons, giving rise to a third distinct jet—a feature that indeed is commonly seen in high-energy collisions.

The three jets continue along paths set by quarks and gluons moving within an extremely confined space, less than 10^{-13} centimeter. The quark-antiquark pair cannot proceed as isolated particles beyond that distance, the limit of asymptotic freedom. Yet the confinement of quarks and of their interactions is not absolute. Although a hadron as a whole is color-neutral, its quarks do respond to the individual color charges of quarks in neighboring hadrons. The interaction, feeble compared with the color forces within hadrons, generates the binding force that holds the protons and neutrons together in nuclei.

Moreover, it seems likely that when hadronic matter is compressed and heated to extreme temperatures, the hadrons lose their individual identities. The hadronic bubbles of the image used above overlap and merge, possibly freeing their constituent quarks and gluons to migrate over great distances. The resulting state of matter, called quark-gluon plasma, may exist in the cores of collapsing supernovas and in neutron stars. Workers are now studying the possibility of creating quark-gluon plasma in the laboratory through collisions of heavy nuclei at very high energy [see "Hot Nuclear Matter," by Walter Greiner and Horst Stöcker; SCIENTIFIC AMERICAN, January].



SIGNATURE OF THE Z^0 PARTICLE is visible on a computer-synthesized display from the UA1 detector of the Super Proton Synchrotron Collider at CERN, the European laboratory for particle physics in Geneva, where the existence of the W^+ and W^- particles was also established. The Z^0, W^+ and W^- are the particles that transmit the weak force; their existence was predicted by the unified theory of the weak and the electromagnetic interactions, and their discovery vindicated the theory. The tracks depicted within the detector correspond to particles detected following the high-energy collision of a proton and an antiproton. The tracks displayed in white are those of an electron and a positron, the characteristic decay products of the Z^0, which disintegrated soon after it materialized in the collision.

Electroweak Symmetry

Understanding of the third interaction that elementary-particle physics must reckon with, the weak interaction, also has advanced by analogy with QED. In 1933 Enrico Fermi constructed the first mathematical description of the weak interaction, as manifested in beta radioactivity, by direct analogy with QED. Subsequent work revealed several important differences between the weak and the electromagnetic interactions. The weak force acts only over distances of less than 10^{-16} centimeter (in contrast to the long range of electromagnetism), and it is intimately associated with the spin of the interacting particles. Only particles with a left-handed spin are affected by weak interactions in which electric charge is changed, as in the beta decay of a neutron, whereas right-handed ones are unaffected.

In spite of these distinctions theorists extended the analogy and proposed that the weak interaction, like electromagnetism, is carried by a force particle, which came to be known as the intermediate boson, also called the *W* (for weak) particle. In order to mediate decays in which charge is changed, the *W* boson would need to carry electric charge. The range of a force is inversely proportional to the mass of the particle that transmits it; because the photon is massless, the electromagnetic interaction can act over infinite distances. The very short range of the weak force suggests an extremely massive boson.

A number of apparent connections between electromagnetism and the weak interaction, including the fact that the mediating particle of weak interactions is electrically charged, encouraged some workers to propose a synthesis. One immediate result of the proposal that the two interactions are only different manifestations of a single underlying phenomenon was an estimate for the mass of the *W* boson. The proposed unification implied that at very short distances and therefore at very high energies the weak force is equal to the electromagnetic force. Its apparent weakness in experiments done at lower energies merely reflects its short range. Therefore the whole of the difference in the apparent strengths of the two interactions must be due to the mass of the *W* boson. Under that assumption the *W* boson's mass can be estimated at about 100 times the mass of the proton.

To advance from the notion of a synthesis to a viable theory unifying the weak and the electromagnetic interactions has required half a century of experiments and theoretical insight, culminating in the work for which Sheldon Lee Glashow and Steven Weinberg, then at Harvard University, and Abdus Salam of the Imperial College of Science and Technology in London and the International Center for Theoretical Physics in Trieste won the 1979 Nobel prize in physics. Like QED itself, the unified, or electroweak, theory is a gauge theory derived from a symmetry principle, one that is manifested in the family groupings of quarks and leptons.

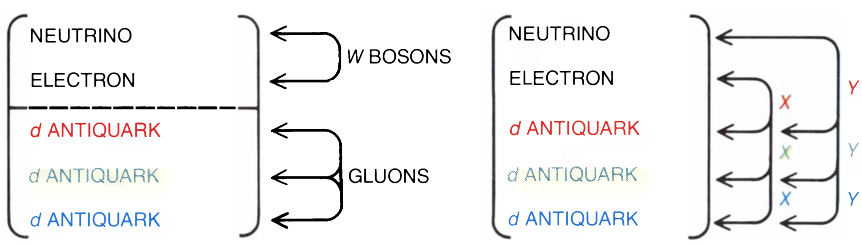
Not one but three intermediate bosons, along with the photon, serve as force particles in electroweak theory. They are the positively charged W^+ and negatively charged W^- bosons, which respectively mediate the exchange of positive and negative charge in weak interactions, and the Z^0 particle, which mediates a class of weak interactions known as neutral current processes. Neutral current processes such as the elastic scattering of a neutrino from a proton, a weak interaction in which no charge is exchanged, were predicted by the electroweak theory and first observed at CERN in 1973. They represent a further point of convergence between electromagnetism and the weak interaction in that electromagnetic interactions do not change the charge of participating particles either.

To account for the fact that the electromagnetic and weak interactions, although they are intimately related, take different guises, the electroweak theory holds that the symmetry uniting them is apparent only at high energies. At lower energies it is concealed. An analogy can be drawn to the magnetic behavior of iron. When iron is warm, its molecules, which can be regarded as a set of infinitesimal magnets, are in hectic thermal motion and therefore randomly oriented. Viewed in the large the magnetic behavior of the iron is the same from all directions, reflecting the rotational symmetry of the laws of electromagnetism. When the iron cools below a critical temperature, however, its molecules line up in an arbitrary direction, leaving the metal magnetized along one axis. The symmetry of the underlying laws is now concealed.

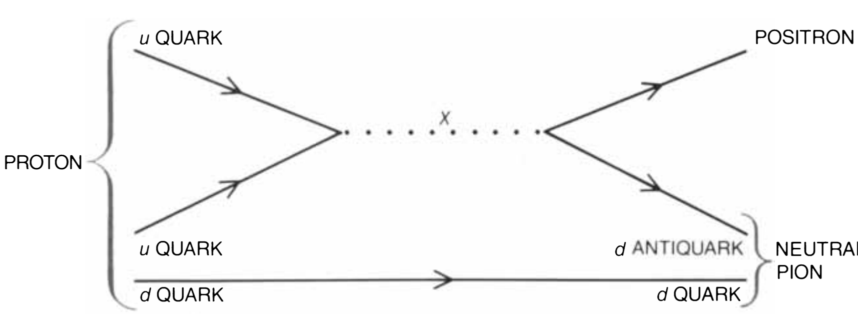
The principal actor in the breaking of the symmetry that unites electromagnetism and the weak interaction at high energies is a postulated particle called the Higgs boson. It is through interactions with the Higgs boson that the symmetry-hiding masses of the intermediate bosons are generated. The Higgs boson is also held to be responsible for the fact that quarks and leptons within the same family have different masses. At very high energies all quarks and leptons are thought to be massless; at lower energies interactions with the Higgs particle confer on the quarks and leptons their varying masses. Because the Higgs boson is elusive and may be far more massive than the intermediate bosons themselves, experimental energies much higher than those of current accelerators probably will be needed to produce it.

The three intermediate bosons required by the electroweak theory, however, have been observed. Energies high enough to produce such massive particles are best obtained in head-on collisions of protons and antiprotons. In one out of about five million collisions a quark from the proton and an antiquark from the antiproton fuse, yielding an intermediate boson. The boson disintegrates less than 10^{-24} second after its formation. Its brief existence, however, can be detected from its decay products.

In a triumph of accelerator art, experimental technique and theoretical reasoning, international teams at CERN led by Carlo Rubbia of Harvard and Pierre Darriulat devised experiments that in 1983 detected the W bosons and the Z^0 particle. An elaborate detector identified and recorded in the debris of violent proton-antiproton collisions single electrons whose trajectory matched the one expected in a W^- particle's decay; the detector also recorded electrons and positrons traveling in precisely opposite directions, unmistakable evidence of the Z^0 particle. For their part in the experiments and in the design and construction of the proton-antiproton collider and the detector Rubbia and Simon van der Meer of CERN were awarded the 1984 Nobel prize in physics.



KINSHIP OF ALL MATTER is implied by unified theories of the fundamental forces; one branch of a unified family of elementary particles is shown here. Particles that are equivalent within a theory can metamorphose into one another. Because leptons, such as the electron and the neutrino, respond to the electroweak force alone whereas quarks also respond to the strong force, the two kinds of particle are not equivalent in current theory, and transformations of one into the other have not been observed (*left*). If the simplest unified theories are correct and the fundamental forces are ultimately identical, then at some very high energy quarks and leptons are interconvertible (*right*). Known transformations are mediated by force particles such as the *W* bosons and the gluons; transitions between the quark and lepton groups would be mediated by new force particles, here given as *X* and *Y*.



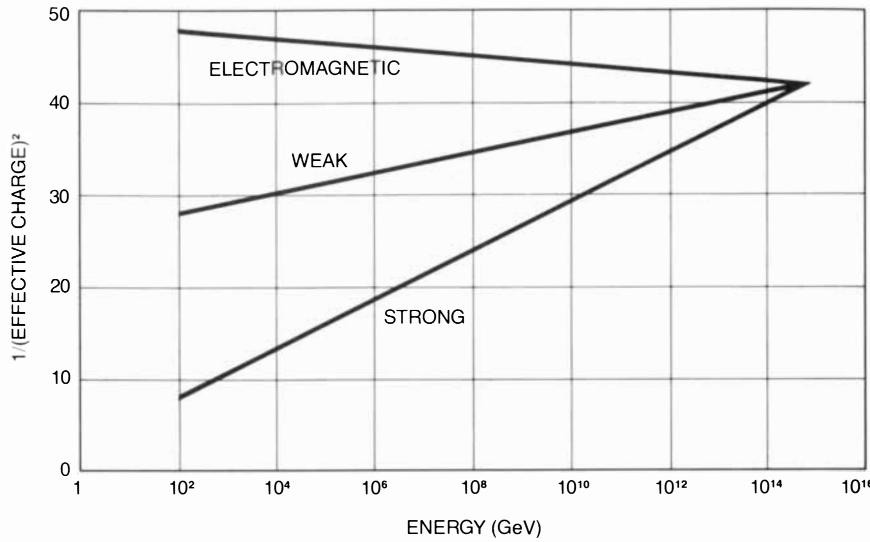
DECAY OF THE PROTON is a possible consequence of transformations of quarks into leptons, a phenomenon unified theories would allow. The diagram shows one of several proposed decay routes. The proton's constituent *u* quarks combine to form an *X* particle, which disintegrates into a *d* antiquark and a positron (a lepton). The *d* antiquark combines with the remaining quark of the proton, a *d* quark, to form a neutral pion. Because pions are composed of matter and antimatter, they are short-lived; the mutual annihilation of their constituents will release energy in the form of two photons. The positron is also ephemeral: an encounter with a stray electron, its antiparticle, will convert it into energy as well.

Unification

With QCD and the electroweak theory in hand, what remains to be understood? If both theories are correct, can they also be complete? Many observations are explained only in part, if at all, by the separate theories of the strong and the electroweak interactions. Some of them seem to invite a further unification of the strong, weak and electromagnetic interactions.

Among the hints of deeper patterns is the striking resemblance of quarks and leptons. Particles in both groups are structureless at current experimental resolution. Quarks possess color charges whereas leptons do not, but both carry a half unit of spin and take part in electromagnetic and weak interactions. Moreover, the electroweak theory itself suggests a relation between quarks and leptons. Unless each of the three lepton families (the electron and its neutrino, for example) can be linked with the corresponding family of quarks (the *u* and *d* quarks, in their three colors) the electroweak theory will be beset with mathematical inconsistencies.

What is known about the fundamental forces also points to a unification. All three can be described by gauge theories, which are similar in their mathematical structure. Moreover, the strengths of the three forces appear likely to converge at very short distances, a phenomenon that would be apparent only at extremely large energies. We have seen that the electromagnetic charge grows strong at short distances, whereas the strong, or color, charge becomes increasingly feeble. Might all the interactions become comparable at some gigantic energy?



CONVERGENCE OF FORCES at extremely high energies, which are equivalent to very small scales of distance, is expected in unified theories. The graph gives an inverse measure of the forces' intrinsic strength; that of the strong and weak forces diminishes with energy, whereas that of electromagnetism increases. The simplest unified theory predicts that the fundamental identity of the three forces is revealed in interactions taking place at an energy of more than 10^{15} GeV, which corresponds to a distance of less than 10^{-29} centimeter.

If the interactions are fundamentally the same, the distinction between quarks, which respond to the strong force, and leptons, which do not, begins to dissolve. In the simplest example of a unified theory, put forward by Glashow and Howard Georgi of Harvard in 1974, each matched set of quarks and leptons gives rise to an extended family containing all the various states of charge and spin of each of the particles.

The mathematical consistency of the proposed organization of matter is impressive. Moreover, regularities in the scheme require that electric charge be apportioned among elementary particles in multiples of exactly 1/3, thereby accounting for the electrical neutrality of stable matter. The atom is neutral only because when quarks are grouped in threes, as they are in the nucleus, their individual charges combine to give a charge that is a precise integer, equal and opposite to the charge of an integral number of electrons. If quarks were unrelated to leptons, the precise relation of their electric charges could only be a remarkable coincidence.

In such a unification only one gauge theory is required to describe all the interactions of matter. In a gauge theory each particle in a set can be transformed into any other particle. Transformations of quarks into other quarks and of leptons into other leptons, mediated by gluons and intermediate bosons, are familiar. A unified theory suggests that quarks can change into leptons and vice versa. As in any gauge theory, such an interaction would be mediated by a force particle: a postulated *X* or *Y* boson. Like other gauge theories, the unified theory describes the variation over distance of interaction strengths. According to the simplest of the unified theories, the separate strong and electroweak interactions converge and become a single interaction at a distance of 10^{-29} centimeter, corresponding to an energy of 10^{24} electron volts.

Such an energy is far higher than may ever be attained in an accelerator, but certain consequences of unification might be apparent even in the low-energy world we inhabit. The supposition that transformations can cross the boundary between quarks and leptons implies that matter, much of whose mass consists of quarks, can decay. If, for example, the two *u* quarks in a proton were to approach each other closer than 10^{-29} centimeter, they might combine to form an *X* boson, which would disintegrate into a positron and a *d* antiquark. The antiquark would then combine with the one remaining quark of the proton, a *d* quark, to form a neutral pion, which itself would quickly decay into two photons. In the course of the process much of the proton's mass would be converted into energy.

The observation of proton decay would lend considerable support to a unified theory. It would also have interesting cosmological consequences. The universe contains far more matter than it does antimatter. Since matter and antimatter are equivalent in almost every respect, it is appealing to speculate that the universe was formed with equal amounts of both. If the number of baryons—three-quark particles such as the proton and the neutron, which constitute the bulk of ordinary matter—can change, as the decay of the proton would imply, then the current excess of matter need not represent the initial state of the universe. Originally matter and antimatter may indeed have been present in equal quantities, but during the first instants after the big bang, while the universe remained in a state of extremely high energy, processes that alter baryon number may have upset the balance.

A number of experiments have been mounted to search for proton decay. The large unification energy implies that the mean lifetime of the proton must be extraordinarily long—10^{30} years or more. To have a reasonable chance of observing a single decay it is necessary to monitor an extremely large number of protons; a key feature of proton-decay experiments has therefore been large scale. The most ambitious experiment mounted to date is an instrumented tank of purified water 21 meters on a side in the Morton salt mine near Cleveland. During almost three years of monitoring none of the water's more than 10^{33} protons has been observed to decay, suggesting that the proton's lifetime is even longer than the simplest unified theory predicts. In some rival theories, however, the lifetime of the proton is considerably longer, and there are other theories in which protons decay in ways that would be difficult to detect in existing experiments. Furthermore, results from other experiments hint that protons can indeed decay.

Open Questions

Besides pointing the way to a possible unification the standard model, consisting of QCD and the electroweak theory, has suggested numerous sharp questions for present and future accelerators. Among the many goals for current facilities is an effort to test the predictions of QCD in greater detail. Over the next decade accelerators with the higher energies needed to produce the massive W and Z^0 bosons in adequate numbers will also add detail to electroweak theory. It would be presumptuous to say these investigations will turn up no surprises. The consistency and experimental successes of the standard model at familiar energies strongly suggest, however, that to resolve fundamental issues we need to take a large step up in interaction energy from the several hundred GeV (billion electron volts) attainable in the most powerful accelerators now being built.

Although the standard model is remarkably free of inconsistencies, it is incomplete; one is left hungry for further explanation. The model does not account for the pattern of quark and lepton masses or for the fact that although weak transitions usually observe family lines, they occasionally cross them. The family pattern itself remains to be explained. Why should there be three matched sets of quarks and leptons? Might there be more?

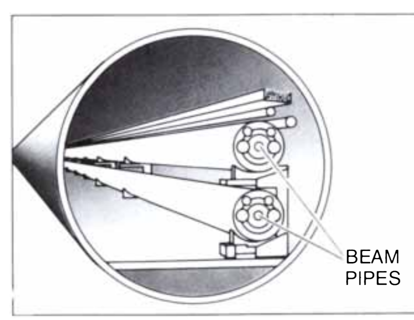
Twenty or more parameters, constants not accounted for by theory, are required to specify the standard model completely. These include the coupling strengths of the strong, weak and electromagnetic interactions, the masses of the quarks and leptons, and parameters specifying the interactions of the Higgs boson. Furthermore, the apparently fundamental constituents and force particles number at least 34: 15 quarks (five flavors, each in three colors), six leptons, the photon, eight gluons, three intermediate bosons and the postulated Higgs boson. By the criterion of simplicity the standard model does not seem to represent progress over the ancient view of matter as made up of earth, air, fire and water, interacting through love and strife. Encouraged by historical precedent, many physicists account for the diversity by proposing that these seemingly fundamental particles are made up of still smaller particles in varying combinations.

There are two other crucial points at which the standard model seems to falter. Neither the separate theories of the strong and the electroweak interactions nor the conjectured unification of the two takes any account of gravity. Whether gravity can be described in a quantum theory and unified with the other fundamental forces remains an open question. Another basic deficiency of the standard model concerns the Higgs boson. The electroweak theory requires that the Higgs boson exist but does not specify precisely how the particle must interact with other particles or even what its mass must be, except in the broadest terms.

The Superconducting Supercollider

What energy must we reach, and what new instruments do we need, to shed light on such fundamental problems? The questions surrounding the Higgs boson, although they are by no means the only challenges we face, are particularly well defined, and their answers will bear on the entire strategy of unification. They set a useful target for the next generation of machines.

It has been proposed that the Higgs boson is not an elementary particle at all but rather a composite object made up of elementary constituents analogous to quarks and leptons but subject to a new kind of strong interaction, often called technicolor, which would confine them within about 10^{-17} centimeter. The phenomena that would reveal such an interaction would become apparent at energies of about 1 TeV (trillion electron volts). A second approach to the question of the Higgs boson's mass and behavior employs a postulated principle known as supersymmetry, which relates particles that differ in spin. Supersymmetry entails the existence of an entirely new set of elusive, extremely massive particles. The new particles would correspond to known quarks, leptons and bosons but would differ in their spins. Because of their mass, such particles would reveal themselves fully only in interactions taking place at very



PROPOSED ACCELERATOR, the Superconducting Supercollider, will make it possible to study interactions at energies of more than 1 TeV. In the design depicted (one of many) the accelerator ring has a diameter of 30 kilometers and is buried 100 meters underground; smaller rings feed protons into the large ring. A cross section of the main tunnel (*above*) shows the two pipes, each about five centimeters in diameter, which will contain the counterrotating beams of protons. Superconducting magnets supercooled with liquid helium to increase their power and efficiency surround each of the pipes, focusing and confining the beams.

high energy, probably about 1 TeV.

Our best hope for producing interactions of fundamental particles at energies of 1 TeV is an accelerator known as the Superconducting Supercollider (SSC). Formally recommended to the Department of Energy in 1983 by the High Energy Physics Advisory Panel, it would incorporate proved technology on an unprecedented scale. A number of designs have been put forward, but all envision a proton-proton or proton-antiproton collider. High-energy beams of protons are produced more readily with current technology than beams of electrons and positrons, although electron-positron collisions are generally simpler to analyze; because protons are composite particles, their collisions yield a larger variety of interactions than collisions of electrons and positrons. Another common feature of the designs is the use of superconducting magnets, first employed on a large scale in the Tevatron Collider at the Fermi National Accelerator Laboratory (Fermilab) in Batavia, Ill. The technology increases the field strength and lowers the power consumption of the magnets that bend and confine the beam.

One of the more compact designs incorporates niobium-titanium alloy magnets cooled to 4.4 degrees Celsius above absolute zero. If the magnets generated fields of five tesla (100,000 times the strength of the earth's magnetic field), two counterrotating beams of protons accelerated to energies of 20 TeV (needed to produce 1-TeV interactions of the quarks and gluons within the protons) could be confined within a loop about 30 kilometers in diameter. In other designs magnetic fields are lower and the proposed facility is correspondingly larger.

It is believed such a device could be operational in 1994, at a cost of $3 billion. The Department of Energy has encouraged the establishment of a Central Design Group to formulate a specific construction proposal within three years and is currently funding the development of magnets for the SSC at several laboratories.

The SSC represents basic research at unprecedented cost on an unmatched scale. Yet the rewards will be proportionate. The advances of the past decade have brought us tantalizingly close to a profound new understanding of the fundamental constituents of nature and their interactions. Current theory suggests that the frontier of our ignorance falls at energies of about 1 TeV. Whatever clues about the unification of the forces of nature and the constituents of matter wait beyond that frontier, the SSC is likely to reveal them.

Compartmentalization of Decay in Trees

Animals heal, but trees compartmentalize. They endure a lifetime of injury and infection by setting boundaries that resist the spread of the invading microorganisms

by Alex L. Shigo

Trees have a spectacular survival record. Over a period of more than 400 million years they have evolved as the tallest, most massive and longest-lived organisms ever to inhabit the earth. Yet trees lack a means of defense that almost every animal has: trees cannot move away from destructive forces. Because they cannot move, all types of living and nonliving enemies—fire, storms, microorganisms, insects, other animals and later man—have wounded them throughout their history. Trees have survived because their evolution has made them into highly compartmented organisms; that is, they wall off injured and infected wood.

In that respect trees are radically different from animals. Fundamentally, animals heal: they preserve their life by making billions of repairs, installing new cells or rejuvenated cells in the positions of old ones. Trees cannot heal; they make no repairs. Instead they defend themselves from the consequences of injury and infection by walling off the damage. In a word, they compartmentalize. At the same time they put new cells in new positions; in effect, they grow a new tree over the old one every year. The most obvious results of the process are growth rings, which are visible on the cross section of a trunk, a root or a branch.

The defenses wielded by trees suggest a new view of tree biology, one in which the role of tree pathology is given full recognition. Trees have been guided through evolution by their need to defend against attack while standing their ground.

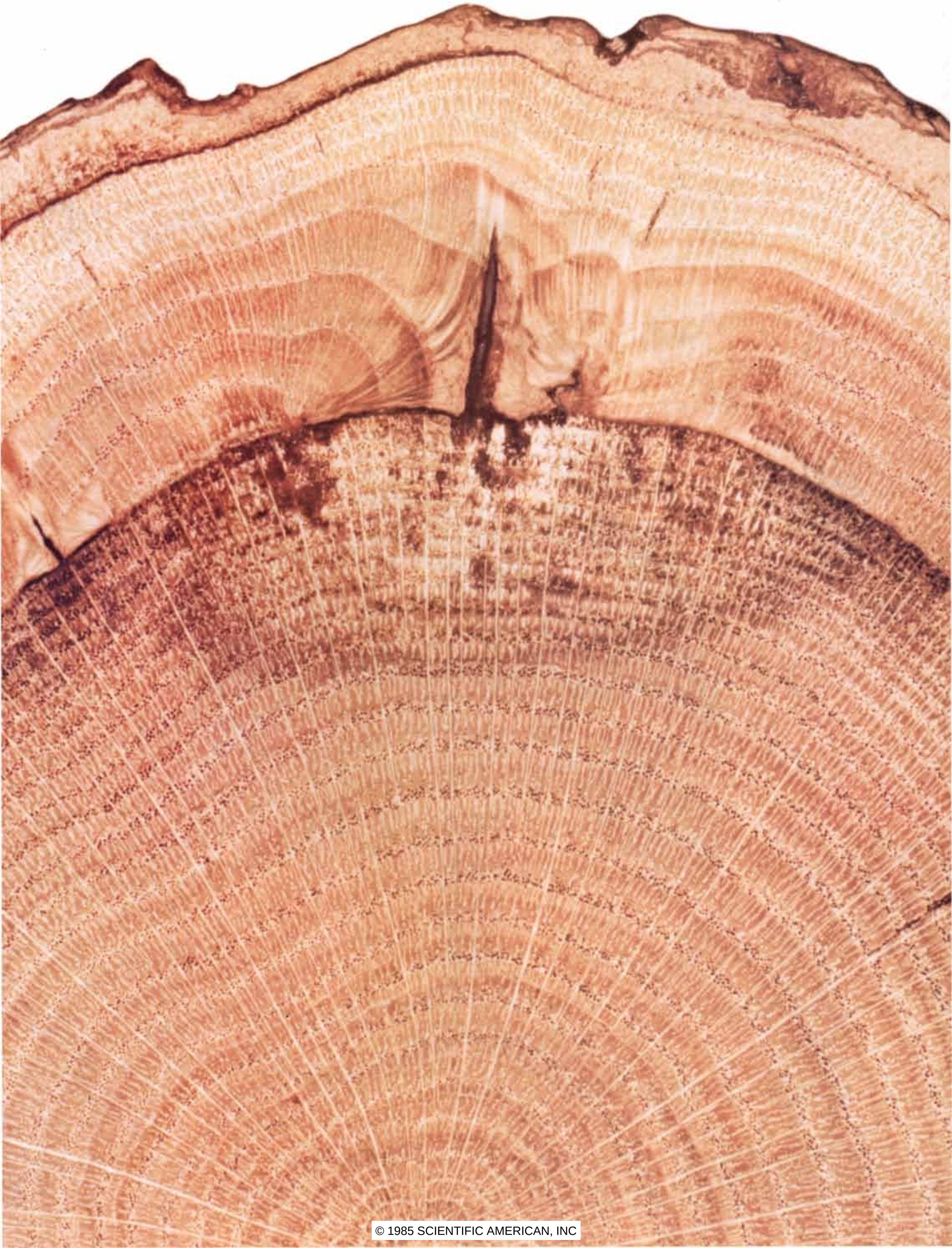
To understand disease and decay in trees it is essential to understand normal tree function and growth. Trees, together with grasses and algae, are the earth's major energy trappers. In particular, forests, which cover about a tenth of the planet's surface, trap about half of all the energy entering the biosphere. The energy, which enters as solar radiation, serves to transform carbon dioxide and water into carbohydrates, the chemical form in which energy is stored. In turn the carbohydrates power growth, maintenance, reproduction and defense.

Trees themselves—that is, the conifers and the hardwoods (angiosperms and gymnosperms)—are perennial, woody, compartmented, shedding plants. In addition most trees are long-lived, massive and tall. Their internal structure follows a characteristic pattern. The generation of new cells in the tree is the function of the vascular cambium, a thin cylindrical layer found in the trunk, the roots and the branches. In the outer direction the cambium lays down phloem, or inner bark, the layer that transports liquids downward in the tree. (Specifically, it carries substances produced by photosynthesis in the leaves.) In the inward direction the cambium lays down concentric layers of xylem, or wood, which transports water and water-soluble substances upward.

The inward cell production is worth a close examination. On a broad scale the cambium lays down one inward layer per year; these are the annual growth rings, prominent on a cross section of the tree. (In tropical trees, however, growth rings are indistinct.) Viewed on a finer scale the cambium forms two basic types of cells: those with their long axis perpendicular to the axis of the trunk, a root or a branch and those with their long axis parallel to the axis of trunk, root or branch. The perpendicular cells become ray parenchyma, which form radial partitions in the wood. Meanwhile the longitudinal cells fill the compartments between the rays.

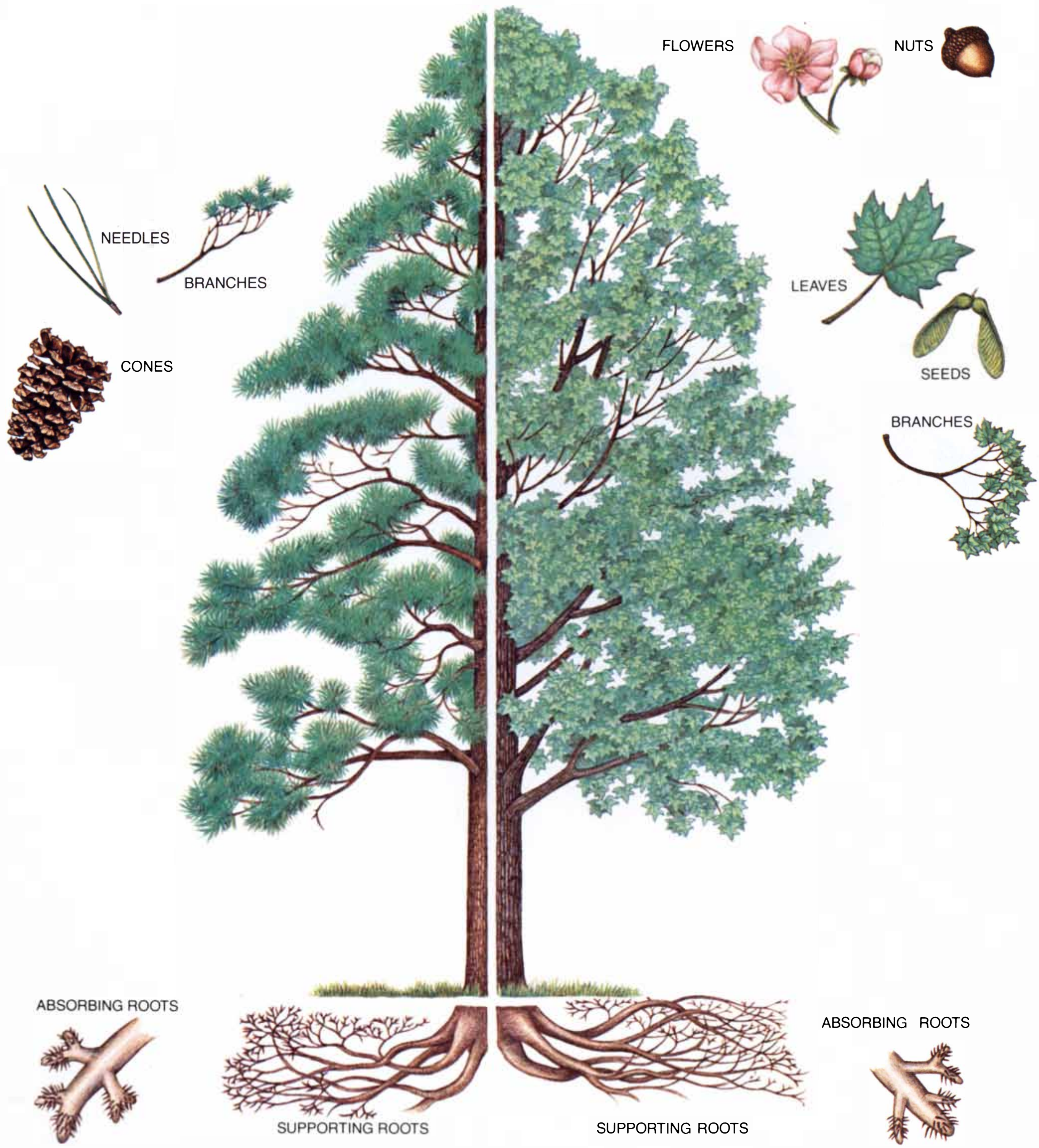
The longitudinal cells are of three varieties. In some the living contents die in a few days or weeks, leaving only a tubular cell wall. In hardwoods such cells are called vessels; in conifers they are called tracheids. They serve to transport liquids. In others the cell wall is thick. Such cells, called fiber cells or fiber tracheids, provide mechanical support to the wood. Finally, the cells called parenchyma retain their living contents behind a thin cell wall. (In contrast, vessels and tracheids have a hollow interior under a thick cell wall.) Parenchymal cells store nutrients and other materials. In them the cytoplasm can remain living for years, indeed sometimes for more than a century. The complex interlocking of the various cells in wood combines with the tough cell walls themselves to give

A TREE'S DEFENSES against injury and infection produced the pattern on this cross section of the trunk of a red oak. Nine years before the section was made the tree was wounded by buckshot. Microorganisms that established themselves in the wound caused the wood there to decay, making it whiter than elsewhere. In response the tree mounted a chemical defense: the dark boundary of discolored wood surrounding the decay indicates the production of antimicrobial substances. The dark line curving over the wounded wood is a further defense: the cambium, or growth layer, alive at the sides of the wound produced a cellular wall protecting the wood that formed after the time of the injury. The pattern of growth rings outside the wound reveals the subsequent history of the tree. After five years the wound closed, and after that four more growth rings formed. The tree was sectioned by the author and photographed by Kenneth R. Dudzik of the U.S. Forest Service in Durham, N.H.



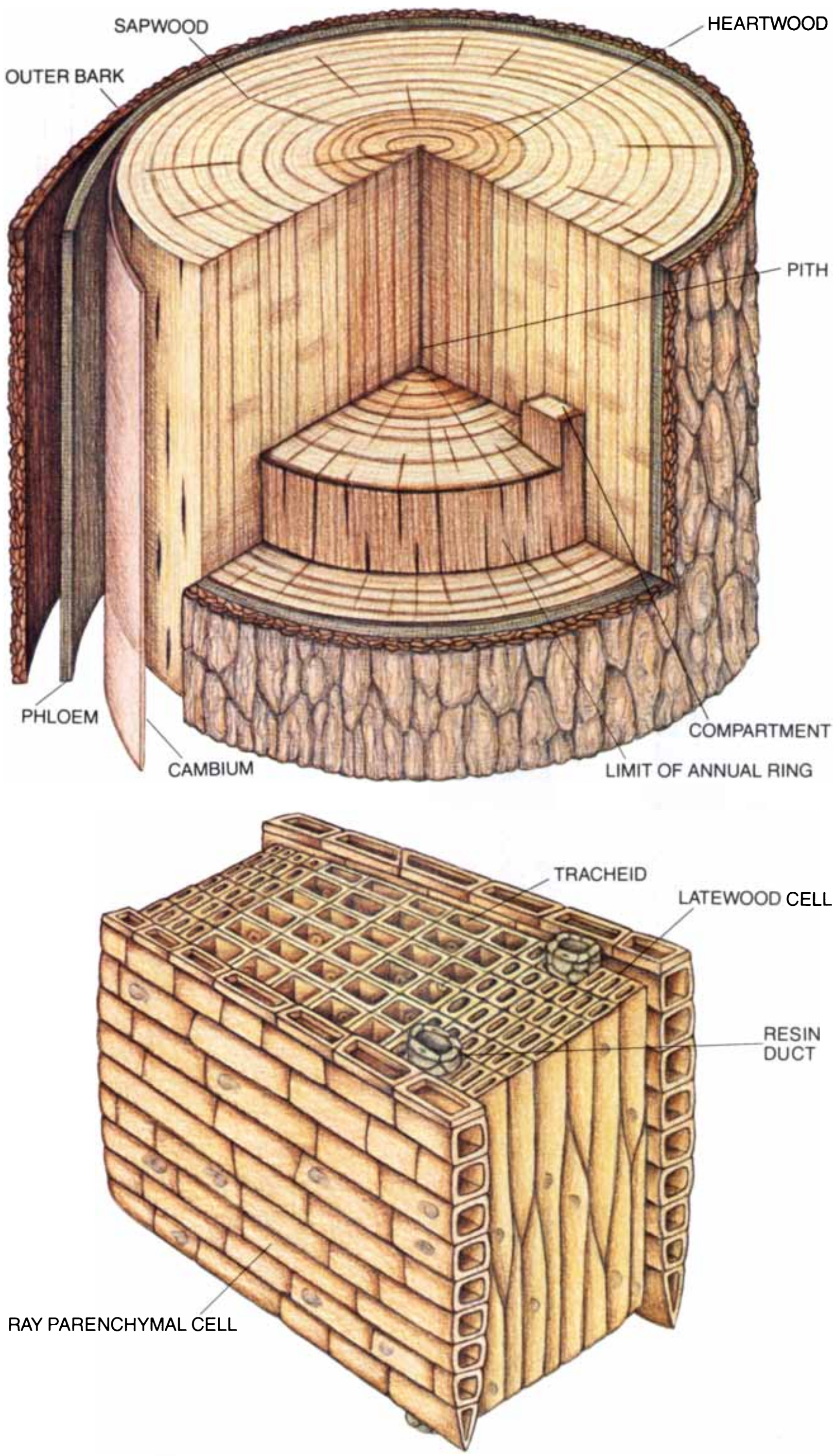
wood tissue its strength. The interlocking design continues into the molecular construction of the cell walls, and even into the individual cell-wall molecules: cellulose and lignin.

At any given time in the life of the tree the more recent layers of wood, the ones in which the parenchyma still retain their living contents, form the tree's sapwood. In many trees the older layers, nearer the axis of the tree, form a district called heartwood, which often is darker in color. Some of the darkness is due to the deposition of substances known collectively as extractives. Heartwood has a high degree of mechanical self-support; thus it continues to serve the tree. On the other hand, heartwood has no capacity to store nutrients and transport substances.



COMPARTMENTS OF A TREE enable it to resist the spread of infection after an injury; they also enable it to wall off parts of itself when the parts reach a genetically programmed stage of senescence. At the left a composite tree is shown. The left side of the composite is a loblolly pine, typical of conifers; the right side is a sugar maple, typical of hardwoods. Certain compartments, shown in the drawing, are shed when their function is completed. At the upper right the trunk of the pine appears in cross section, revealing more

The study of tree anatomy suggests a number of ways in which trees can be viewed conceptually. First, trees are tissue generators. In essence the germination of a seed is the activation of a cell generator. It has the capacity to proliferate cells, but always in new locations; the tree has no capacity to restore or regenerate tissues already in place. The generator is the cambium. Second, wood is a highly ordered arrangement of different types of cells in different stages of aging. Third, a tree is a hierarchy of compartments. Within the trunk, the roots and the branches the largest compartments are the annual rings. Then come groups of cells in each ring, compartmented by rays, and next the individual cells.



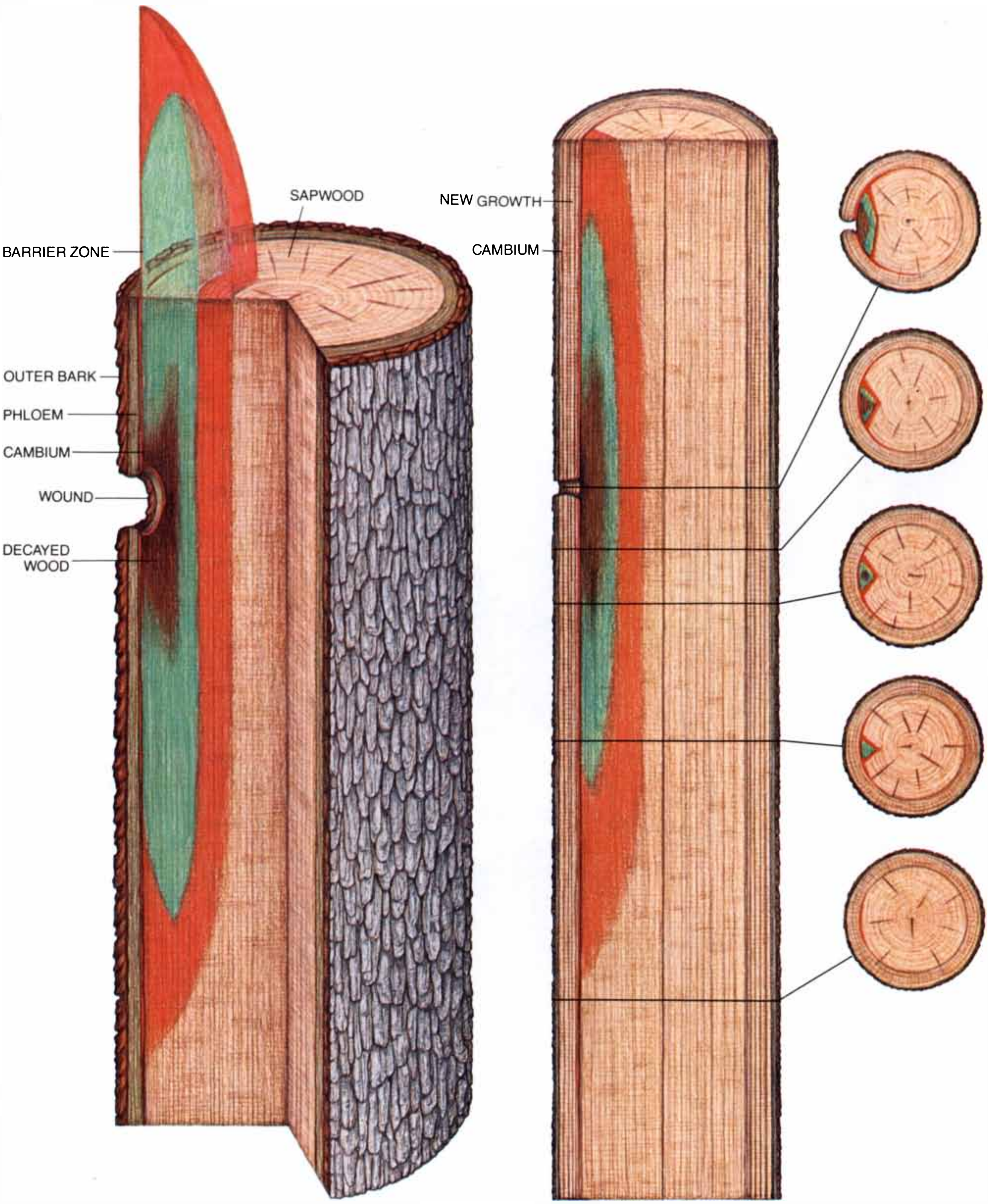
compartments. The cambium generates phloem, or inner bark, and xylem, or wood. The latter is compartmented by annual rings; in turn the rings are compartmented by partitions called parenchymal rays. At the lower right is a wood compartment. It includes parenchyma, or energy-storing cells; tracheids, or cellular tubes, and thick-walled latewood cells.

The study of how trees respond to injury or infection suggests a further concept. Trees respond by compartmentalizing: they attempt to wall off the injured or infected region. They neither kill nor arrest the activity of microorganisms in the compartments that get walled off. Nor do they respond in specific ways to specific microorganisms; the compartmentalization comes in response to the fact of the injury.

Broadly speaking, the tree makes three responses to injury and infection. In the first of them the boundaries of compartments already in place are strengthened to resist the spread of destruction. For the most part the strengthening is achieved by chemical means. In sapwood the metabolism of the living parenchymal cells changes in ways that alter the contents of the cells. In heartwood enzymatic reactions take place in the otherwise nonliving tissue.

The details are poorly understood. Still, under normal circumstances the cells in wood devote their biochemical pathways to the storage of chemical energy in the form of carbohydrates. Usually the molecules are in a reduced form: their content of electrons is relatively great. After injury to the wood the biochemical activity in the cells surrounding the injury is shunted into new metabolic paths, so that molecules tend to be oxidized; that is, protons, or hydrogen ions, are attached. On a broad scale the cellular contents undergo a chemical process much like the tanning of leather. Thus molecules such as gallic acid and tannic acid appear. They share the property of being rich in phenol (six-carbon rings bearing hydroxyl, or OH, groups). The phenols occupy the interior of the cells; they also impregnate the cell walls. The phenols discolor the wood in shades of red, green or blue, depending on the details of the chemical pathways, which are determined by the genetics of a given species of tree. More important to the tree's defenses, phenol compounds tend to be antimicrobial.

In the second response the tree makes to injury and infection the tree creates a new wall by anatomical and chemical means. First the cambium changes the pattern by which it generates new cells. In the wake of the injury few pipeline cells are produced. Fiber cells too are made in lessened quantity. On the other hand, parenchyma, or



RESPONSES TO A WOUND take two forms: the strengthening of existing compartments and the creation of a wall to protect the cambium. The sectioned maple trunk at the left includes an infected wound. A succession of microorganisms (*green*) has become established in the sapwood (the layers of living wood under the cambium), and at the center of the infection the wood has decayed (*brown*). At the margins of the infection parenchymal cells are producing antimicrobial substances (*red*), so that the walls already present in the tree are being strengthened against the infection. In addition cellular tubes above and below the infection are being plugged. (In some trees the ballooning of parenchymal cells constricts the tubes.) The trunk at the right shows the tree in a subsequent growing season. The wall created by the cambium has protected new growth. Meanwhile the internal defenses have compartmentalized the infection.

cells that retain their living contents, are made in increased quantity. Now, however, they are smaller, and their metabolic activity is altered so that their chemical contents resist microorganisms. The new wall, or barrier zone, is the cause of many of the defects found in wood products. For example, it can cause the wood in a living tree to separate along a circle. The defect is known as ring shake.

The third response the tree makes is to continue growing. Trees survive after injury and infection if they have enough time, energy and genetic capacity to recognize and compartmentalize the injured and infected tissue while generating the new tissue that will maintain the life of the tree. To an extent, therefore, an injured or infected tree resembles a heavily compartmented ship or an old type of submarine under attack. When the ship is hit by a torpedo, the crew rushes to seal off the damaged area. The faster the action of the crew and the stronger the walls circumscribing the damaged compartment, the smaller the spread of damage. But after the damage is contained the damaged compartment or compartments are no longer accessible. At this point the analogy ends. The tree survives by growing what amounts to a new tree (with a new set of compartments) over itself during the next growing season.

It must be said that the tree's reaction zones (its chemically strengthened boundaries) are not absolute: they may retreat, rapidly or slowly, from the infection, as certain microorganisms overcome the chemical defense. It should also be said that phenols are poisonous not only to microorganisms but also to the tree. In effect, the tree poisons part of itself in an attempt to stall an invasion. The tree survives because while it is strengthening barriers it is also creating a new tree.

Remarkably, the capacity of a tree to shed parts of itself is much like the response the tree makes to injury and infection. That is, the shedding is an aspect of compartmentalization. In particular, needles, leaves, reproductive structures and absorbing roots that have fulfilled their genetic program are walled off from the tree. (Absorbing roots are the fine, nonwoody roots that absorb substances from the soil.) Twigs, branches and large roots may also be walled off, say in the wake of an injury or an infection, or after the twig, branch or root has reached a particular stage of senescence. (It is hard to know just what stage; the genetic programming of aging in trees is poorly understood.) The foregoing requires a qualification: trees do not actively cast off parts of themselves. Wind, snow, ice, animals and other agents cause the walled-off parts to fall away from the tree. Meanwhile walled-off absorbing roots are digested by soil microorganisms.

I turn now to the other side of the strife between the tree and its enemies. I shall begin with the wounding of a tree, say by an animal. The wound provides new space and nutrients for an array of invading organisms including insects, nematodes, bacteria and fungi. While they compete among themselves for the new space and nutrients, the living cells in the sapwood underlying the wound are reacting to the invasion by undergoing a variety of biochemical changes leading to the production of phenol-based chemical defenses.

Microorganisms attack tree wounds in several ways. Certain bacteria and fungi infect inner bark and stay there, creating the diseases known as annual cankers. Other microorganisms invade wounds and remain in the wounded sapwood tissue, creating so-called wound rots. Still other microorganisms infect inner bark, become established and then infect wood. By the annual repetition of the process they create perennial cankers. Finally, some microorganisms attack a wound, first infecting sapwood and then infecting inner bark. Again the process repeats itself each growing season. These are the so-called canker-rot fungi. They are notably insidious. When the canker rot has progressed from wood to bark, it forms a hard pad of material, similar to a wedge, that kills the subjacent cambium. The tree responds by reactivating its compartmentalizing defenses. The fungus in turn invades the new wound, from which it grows another wedge. The seesaw activity can continue until the tree is girdled.

The first microorganisms to succeed in invading the tree are termed pioneers. They may simply be able to tolerate the chemical alterations brought on in the wood by the wounding of the tree. In some cases, however, their genetic makeup renders them capable of digesting the chemical defenses. Typically, but not always, the pioneers are bacteria, along with certain species of fungi. Notable among the latter are the Hymenomycetes, which cause decay in trees, and the Fungi Imperfecti and Ascomycetes, which for the most part do not themselves cause decay.

A crucial point about the infection is that microorganisms establish themselves in a particular sequence. When the pioneer microorganisms surmount the inhibitory chemicals, they pave the way for other invaders, ones that would have succumbed to the tree's defenses. Thus the invasion takes the form of a succession of organisms, in a pattern essential for the survival of the invaders. To be sure, no microorganism "eats poison" to help its successors. Each organism acts in a way that furthers its retention of space and energy. Hence some of the pioneers may actually create or preserve conditions that inhibit infection by aggressive wood-decaying fungi. Such pioneers may ultimately be the basis for the biological control of decay in trees. Studies of certain fungi, such as species of *Trichoderma*, already show that the fungi thrive in wound-altered sapwood but do not detoxify the chemicals that keep out decay-causing fungi.

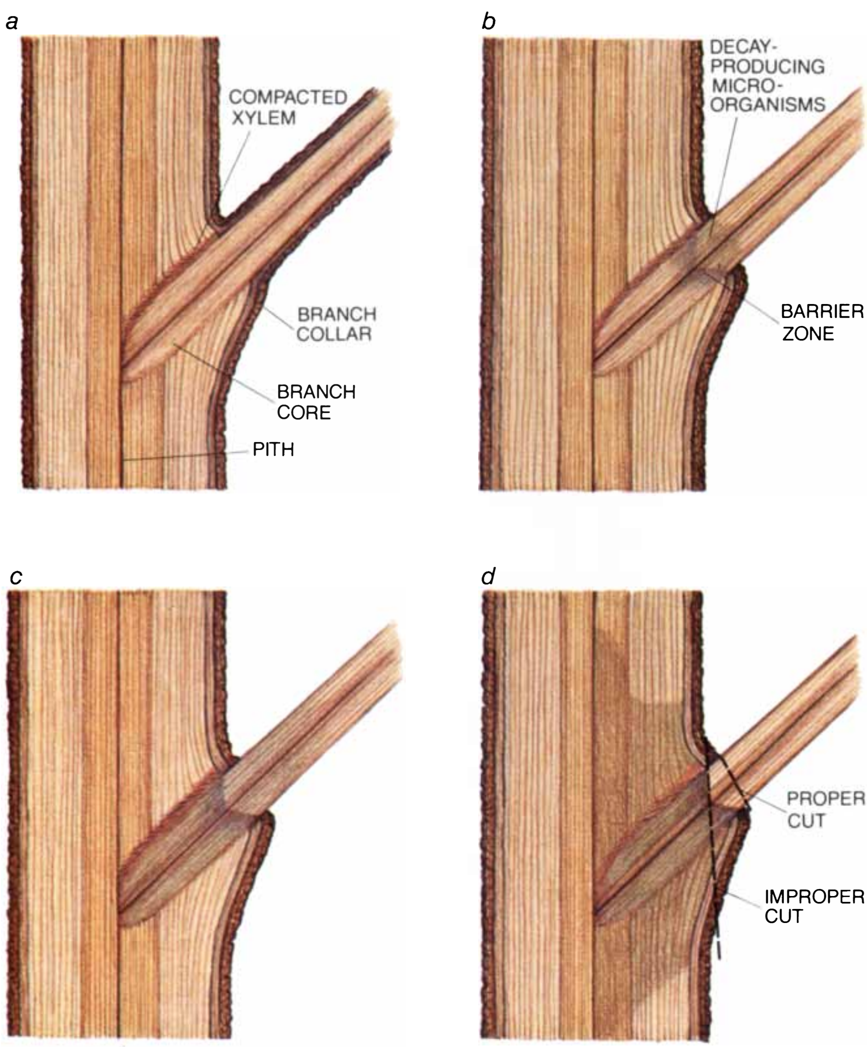
Suppose a succession of invading microorganisms is successful at digesting the wound-altered sapwood. Their success does not necessarily doom the tree. For one thing, the tree is generating new cells around the old ones. If the tree can generate new cells faster than old ones are being digested, the tree has a good chance of survival. The death of a tree, if it happens, can be mechanical or biological. The tree will die if the trunk breaks. Alternatively, it will die if the cambium—the cell generator—is killed. Further still, it will die if so much of the tree's tissue has been walled off during a lifetime of defense against injury and infection that the remaining compartments are insufficient to store the tree's energy reserves.

The new understanding of trees as compartmentalizing organisms did not arise at once. Indeed, it came as a contradiction of earlier notions, some of which were developed soon after the underpinnings of modern biology were established a century ago. It seems a trite thing to say, but trees are fundamentally different from animals, and much of the failure to understand trees derives from unconsciously confusing the two.

Before 1845 the preferred explanation of life was that it originates spontaneously (that is, by spontaneous generation) from inorganic material. Much was known about microorganisms; in particular the association of decayed wood and fungi had been recognized. The idea, however, was that decay gives rise to the fungi. In the wake of Heinrich Anton De Bary, Louis Pasteur and the development of the germ theory, which attributes disease to minute, harmful organisms, the German tree pathologist Robert Hartig reversed the idea by proposing that fungi give rise to decay. Hartig showed that the sporophores, or fruiting bodies, found in tree wounds and the mycelia, or strandlike organisms, found in decayed wood represent different life stages of the same fungal organism. The observation, and the reversal of the hypothesis, set the stage for the science of tree pathology.

Subsequent studies of decay were made by investigators interested primarily in the deterioration of lumber. Accordingly the studies relied on the removal of wood from trees, followed by laboratory investigation of changes in the tissue. The studies centered on heartwood, or in any case wood darker than sapwood. The strategy underlying the studies was justified on the ground that the heartwood in a tree is dead, unresponsive tissue. To a degree the studies succeeded: chemists and pathologists came to understand how fungi (in particular their enzymes) digest wood. Still, the processes in the living tree were not considered. Although heartwood is dead by animal standards, it is reactive to injury and infection. Some texts still state that decay in trees cannot be considered a disease because only dead heartwood gets infected. Decay in trees is in fact the major disease of all trees worldwide.

Early investigators were not able, therefore, to recognize the defenses activated by injury and infection, which alter wood when the tree is wounded. In a living tree most wounds do not produce decay because the microorganisms infecting the wound are faced by wood that has changed. In contrast, the decay-producing fungi attacking lumber face no countering force from the wood. The microorganisms simply compete among themselves.



DEATH OF A BRANCH follows a course much like the reaction of the tree to injury or infection. The living branch (*a*) is separated from the trunk by an external feature called the branch-bark ridge and an inner partition of compacted xylem along the angle of the ridge. The swollen base of the branch is called the collar. As the branch dies (*b*) it is invaded by decay-producing microorganisms. The tree marshals chemical defenses in a protection zone in the collar. The sequence is essential for the shedding of the branch. In most cases the decay stops at the collar. Sometimes, however, the decay invades the rest of the core of the branch (*c*). If the dead branch is pruned improperly (*d*), so that the cut is through the collar, the protective zone is removed and the sapwood in the trunk is opened to an invasion.

The crucial technology that enabled investigators to develop a new understanding of decay in trees was the chainsaw. By the 1940's the device was powerful enough so that an individual operating it in a forest could easily cut longitudinal sections through trees, exposing columns of discolored and decaying wood under the growth rings the tree had developed after the injury. Before 1940 trees had been dissected, but most of the cuts had been crosscuts. Only a few hardy investigators, such as the pathologist George H. Hepting, working in the Mississippi Delta in the 1930's, proceeded otherwise. Hepting, using crosscut saws and axes, was able to see what others saw later: that throughout long stretches of trees the wood generated after a wound was not invaded by the fungi infecting the wound itself.

In the 1960's and 1970's I myself had the opportunity to disect thousands of trees, first in the northeastern U.S. and then in Europe, India, Puerto Rico and Australia. I was impressed on the one hand by the orderly patterns of discoloration and on the other hand by the orderly succession of the invading microorganisms. My colleagues and I then made experiments in living trees. Still later, biochemical studies were undertaken by Walter C. Shortle, my colleague at the Forest Service of the U.S. Department of Agriculture in Durham, N.H.

In an effort to define a commonality among the responses a tree makes to injury and infection and account for the patterns of decay and discoloration in wounded trees, my colleagues and I have devised a model we call CODIT (an acronym for Compartmentalization of Decay in Trees). The first part of the model represents the responses the tree makes at the time of the injury. In essence the tree strengthens walls that are already in place in the wood, at least in large part. Wall 1 resists the vertical spread of infected wood, wall 2 the inward spread and wall 3 the lateral spread. The second part of the model represents a response the tree makes later. In essence the tree raises a wall not in place at the time of the injury: the cambium generates wall 4, which separates the infected wood from newly forming, healthy wood.

Wall 1 does not really exist as an anatomical entity before the infection. It is primarily a plugging wall, which resists the vertical spread of infection by anatomical and chemical means. The tree has vertical plumbing, consisting, as I have noted, of vessels in hardwoods and tracheids in conifers. In the wake of an infection the tubes must be plugged; the tubes are an easy way for invaders to spread. Trees plug the tubes in several ways. The cells surrounding the tubes (called contact parenchyma) can balloon into the bore of the tubes. Encrustations can develop at the ends of the tubes. Granular or crystalline material can fill the tubes. Pits between tubes can close. Sometimes air bubbles develop, impeding the transport of liquids. The wall-1 defense is relatively weak, but then the tree cannot survive if it plugs all its vertical plumbing. Moreover, the vertical spread of infection is relatively unimportant: the entire core of a tree can be infected, yet the new trees formed by the cambium in subsequent years will keep the tree alive.

Wall 2 and wall 3 exist before the infection: they are formed by the annual rings, which resist the inward spread of infection, and by the parenchymal rays, which resist the lateral spread. After the infection the walls are chemically strengthened. In part, however, wall 2 is anatomical. In some trees, such as the maple (but not, for example, the elm), the end of each season of growth is marked by the production, on the part of the cambium, of a final layer of what are called marginal or terminal parenchyma in hardwoods and thick-walled latewood cells in conifers. The cells form a barrier of cellular masonry at the perimeter of each growth ring. Wall 2 is moderately strong; wall 3 is the strongest of the three. If the latter fails, decay can spread like an opening fan. This is the cause of hollow trees, used for protection and dwelling by many animals. The tree itself can survive, with full crowns of healthy leaves, owing to the work of the cambium in growth seasons subsequent to the injury.

The crucial resistance to the infection is the resistance to its outward spread—in particular the defense of the cambium against destruction from within the tree. Here wall 4, the cambium's wall, comes into play. It is quite weak in the sense of strengthening the structure of the tree but quite strong in the sense of being a barrier against microorganisms, a barrier isolating tissue that is outside the injury and therefore formed after the injury. Indeed, it is impervious to most of the fungi and bacteria that inhabit wood or bark. Recent work by R. B. Pearce, P. J. Holloway and Jill Rutherford at the University of Oxford establishes that the cells of the wall-4 barrier are lined with suberin, the fatty acid that gives outer bark its resistance to microbial invaders. (The invaders almost never have enzymes capable of acting on suberin.)

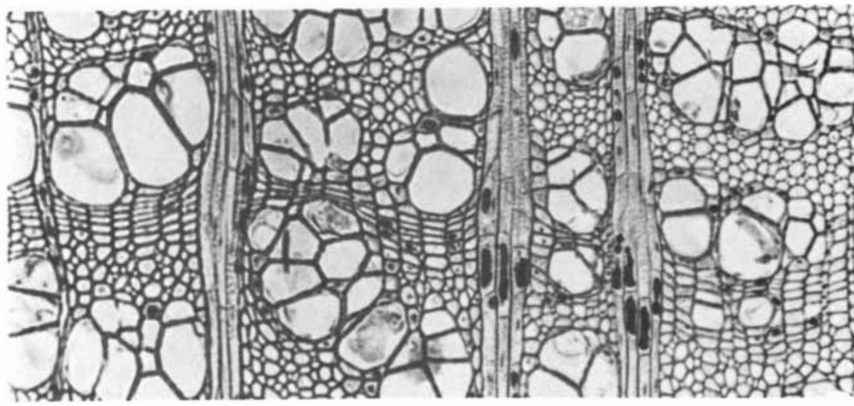
My colleagues and I have also invented a device that detects decayed wood in living trees without damaging the trees. The device capitalizes on a circumstance of decay: as wood decays, its content of electric-charge carriers (chiefly potassium ions) increases, so that the electrical resistance of the tissue, measured in ohms, decreases. Our device therefore consists of a battery-operated pulsed-current generator, a probe and an ohmmeter. To test for decay, a hole 3/32 of an inch in diameter is drilled eight to 12 inches into the tree. The probe is slowly inserted. A sudden decrease in the electrical resistance encountered at the tip of the probe (the tip lacks insulation) signals decay. The device is also being employed to detect decay at the groundline in utility poles.

My colleagues and I are now collaborating with geneticists in an effort to learn which individuals within a given species of tree have the best capacity for compartmentalizing. (The capacity is under strong genetic control.) Armed with our new concepts of tree biology we are also reexaming tree diseases. We have found, for example, that elms compartmentalize the wood infected by the fungus that causes Dutch elm disease. The defense can lead to starvation as tissues that normally store energy get walled off. Moreover, we are devising corrections in many of the standard tree-care procedures, such as pruning.

Much of the misunderstanding that underlies improper tree care comes about because concepts developed to explain animal biology are applied, almost unconsciously, to trees. In many ways trees are treated like animals or, worse, like people. Dressings are put on wounded trees in an effort to stop decay and promote healing, much as a parent dresses a cut on a child. Cavities of decay are cleaned out beyond the decay and into healthy wood, much as a dentist cleans out a decayed tooth. Branches are pruned flush with the trunk, and in some countries the bark of the trunk at the base of the branch is scribed in the shape of a diamond; the subsequent callus, or scar, is considered a sign that the tree is healing well.

None of these treatments is beneficial to trees; indeed, all of them can be harmful. No scientific data show that any substance applied to a tree wound will stop decay. Dressings are primarily cosmetic. (Perhaps too they are psychological medicine for the owner of the tree.) The cleaning of a cavity past the decay and into healthy wood promotes further decay. It is in fact one of the worst things one can do to a tree. The cavity exists because the tree has walled off the decay. Finally, improper pruning enables decay to take hold in injured cells in the trunk. Around the base of every branch is a swelling known as a collar. Within it is the protective zone of the branch, that is, the place where the branch's chemical defenses are established. The collar should not be injured, much less removed, by pruning.

Decay in trees is a natural process. In some cases it can be regulated: it can be quickened, slowed or stalled. When it cannot be influenced, it can at least be detected, nondestructively, in living trees. Moreover, it can be predicted, in both rate and configuration. Furthermore, trees can be selected for their resistance to the spread of decay. Scientific forestry is approaching a new understanding of trees and how they survive under many pressures.



CROSS SECTION OF SAPWOOD from an American elm includes a wall, or barrier zone, created by the cambium in response to an injury. Within the zone the parenchymal cells are smaller. Moreover, the parenchymal rays are swollen (three rays cross the field of view), and many of the ray parenchymal cells have dark inclusions consisting of antimicrobial substances. The larger cells throughout the cross section are vessels, the hardwood's equivalent of tracheids. The injury that provoked the tree's defenses was an infection by the fungus that causes Dutch elm disease. Elms react to the infection by walling off the fungus. In doing so, however, they wall off some of their capacity to store reserve energy.

The Mercedes-Benz 300D Sedan, 300TD Station Wagon and 300CD Coupe: with their Turbodiesel performance, they ***are***

The Mercedes-Benz Turbodiesels for 1985: still the most powerful line of diesels sold in America.

THE MERCEDES-BENZ 300D Sedan, 300TD Station Wagon and 300CD Coupe represent three variations on a radical theme: the idea that dramatic over-the-road performance can be blended with diesel efficiency and stamina.

The idea works. These Mercedes-Benz Turbodiesels *move*. With accelerative energy and cruising ease worthy of gasoline-powered cars. With power enough to flatten hills and make quick work of sudden passing maneuvers.

TURBODIESEL POWER, DIESEL DURABILITY

Yet consider the bottom line. The Turbodiesel you will be living with and maintaining and paying the bills for, year in and year out, is a true-blue diesel. No complex electrical system. No conventional tune-ups. A durability factor that has become part of automotive folklore.

The key to the Mercedes-Benz Turbodiesels' performance is less the *turbo* than the *diesel*–its three-liter, five-cylinder engine.

It is unique, a high torque powerhouse so advanced that it even oil-cools its own pistons as they move.

Turbocharging any engine boosts its power. Turbocharging this engine boosts its power–by 42 percent in models sold on the West Coast, by 45 percent in models sold elsewhere.

Many makers have aped the Turbodiesel idea since Mercedes-Benz pioneered it in production automobiles in 1978. Scant surprise that no maker has yet aped the Mercedes-Benz Turbodiesels' vivid level of performance.

The Turbodiesels rank not only as the most powerful but also the most *varied* line of diesels sold in North America today.

SEDAN, STATION WAGON AND COUPE

The four-door 300D Sedan accommodates five persons and a gaping 12-cu.-ft. trunk within a wheelbase of just 110 inches, helping lend near sports-sedan agility to this familiy-sized automobile.

"The 300D's success in striking a balance between ride comfort and handling response," reports one automotive journal, "is equalled by less than a handful of other cars in the world."

The 300TD Station Wagon interlaces the driving pleasures of a Mercedes-Benz with the workhorse utility of a five-door carryall. Total cargo capacity well exceeds 100 cu. ft. A hydropneumatic *leveling* system is integrated with the rear suspension, to help keep the vehicle riding on an even keel–whether the load is heavy or light.

EXOTIC, YET PRACTICAL

The 300CD Coupe is the world's only limited-production two-plus-two diesel touring machine. It sits on a taut 106.7-inch wheelbase–one secret of its quick-witted

1 9 8 5 T U R B O D I E S E L S

diesels apart. With their handling agility and riding comfort and obsessively fine workmanship, they are automobiles apart.

agility. Its graceful coupe body-work, sans central door pillars, is formed in a process involving intensive handworkmanship. The 300CD is that rarity of rarities, an automobile both highly exotic *and* relentlessly practical.

Sedan or Station Wagon or Coupe, Mercedes-Benz Turbodiesel power is harmonized with high standards of performance in every sense of the word.

From suspension to steering to brakes, every Turbodiesel is engineered to be a precision driving instrument. "There's a cornucopia of driving delights at your disposal," concludes *Car and Driver*–suggesting that in driving precision there is driving pleasure.

From biomechanically correct seats, to a superb automatic climate control system, to the dulling of the outside wind noise to an almost inaudible murmur, remarkable comfort prevails. Virtually every *useful* driving amenity is standard, including an uncannily precise electronic cruise-control unit.

Safety precautions are remarkably comprehensive–both in helping avoid trouble, and in protecting the occupants should trouble occur.

MORE THAN POWER

Ultimately, the Turbodiesels' appeal extends beyond their performance and driving pleasure. There is no more powerful line of diesels sold in North America–and there may be no more versatile, more competent, more timely line of automobiles. In North America, or the world.

Engineered like no other car in the world

SEE YOUR AUTHORIZED MERCEDES-BENZ DEALER

MAGELLANIC CLOUDS, prominent features of the southern sky, are galaxies extraordinarily close to each other and to our own galaxy, the Milky Way. The Large Magellanic Cloud, or LMC (*upper photograph*), is the type specimen of a class of galaxies called Magellanic irregulars, which are ill-defined spirals. The Small Magellanic Cloud, or SMC (*lower photograph*), is classified as peculiar as well as irregular; it has no readily apparent spiral structure. Astronomers have identified a bridge of hydrogen gas extending between the two clouds, which may represent evidence that they nearly collided some 200 million years ago. The SMC was apparently split in two by the encounter; the SMC consists of two misshapen galaxies superposed along the line of sight. A long stream of gas trailing behind the clouds probably results from their interaction with the Milky Way. Two-billion-year-old stars emit most of the light in the brightest regions of the clouds; a red glow visible elsewhere comes from emission nebulas, where hotter stars have ionized the hydrogen gas that permeates the system. The photographs are composites of images made with various wavelengths of light. The original photographic plates were made with the U.K. Science and Engineering Research Council (S.E.R.C.) Schmidt telescope at the Siding Spring Observatory in Australia, and the composites were created by David Malin of the Anglo-Australian Observatory.

The Clouds of Magellan

The nearest neighbors of our galaxy have had a turbulent history. Close encounters with each other and with the Milky Way have split the smaller cloud and drawn out an enormous stream of hydrogen gas

by Don Mathewson

In the southern sky, unobservable from the Northern Hemisphere, float the nearest galactic neighbors of the Milky Way. They are the Large Magellanic Cloud (LMC), which is some 160,000 light-years away, and the Small Magellanic Cloud (SMC), which lies at a distance of 200,000 light-years. The next-closest galaxy to the Milky Way is the great spiral in Andromeda, more than two million light-years distant. The serene appearance of the Magellanic clouds belies a turbulent history and perhaps an even more violent future.

To early observers the Magellanic clouds were simply a prominent feature of the southern sky; they were first described in 1521 by a member of Ferdinand Magellan's circumglobal expedition. Later, variable stars in the clouds served as a yardstick for measuring intergalactic distances, which led to a new conception of the overall structure of the universe.

In more recent years astronomers have found a number of surprises in the structure of the Magellanic clouds themselves. Tides raised by their interaction with each other and with the Milky Way have distorted their form, and there is now a long tendril of gas and stars extending between the LMC and the SMC. One reconstruction of the history of the clouds suggests they collided or at least had a near-miss some 200 million years ago. My colleagues and I have proposed that the SMC was torn apart by such an encounter and now consists of two misshapen galaxies superposed along our line of sight. Perhaps the most remarkable of the recent discoveries is the Magellanic Stream, a plume of hydrogen gas that trails behind the clouds, extending over more than a quarter of the sky. The stream may indicate a violent future for the Magellanic clouds: they could be destined to fall into the core of the Milky Way.

The clues to these events, both past and future, lie in the relations between the two clouds and in their interaction with the Milky Way. The Magellanic clouds are often described as "dwarf companions" of the Milky Way, as if they were analogous to the moons of a planet. Actually it is not certain that the clouds are gravitationally bound satellites of our galaxy. Furthermore, the clouds are not dwarfs; they are galaxies of medium size and are small only in comparison with the Milky Way, which is a quite large spiral galaxy. The LMC has about half the diameter of the Milky Way (roughly 50,000 light-years), about a tenth the mass and about a seventh the brightness. The diameter of the SMC is about a third that of the Milky Way, the mass is about a fortieth and the brightness is about a twenty-fifth.

Both Magellanic clouds are classified as irregular galaxies. They lack the symmetry of the ellipticals and the bold catherine-wheel patterns of the classic spirals. Indeed, the LMC is the prototype of a class of galaxies called Magellanic irregulars, which have a subtle, ill-defined spiral structure. The SMC is considered "peculiar" as well as irregular, meaning that it does not fit into any of the established categories. Both galaxies have a bar near the center, but neither has a readily identifiable nucleus.

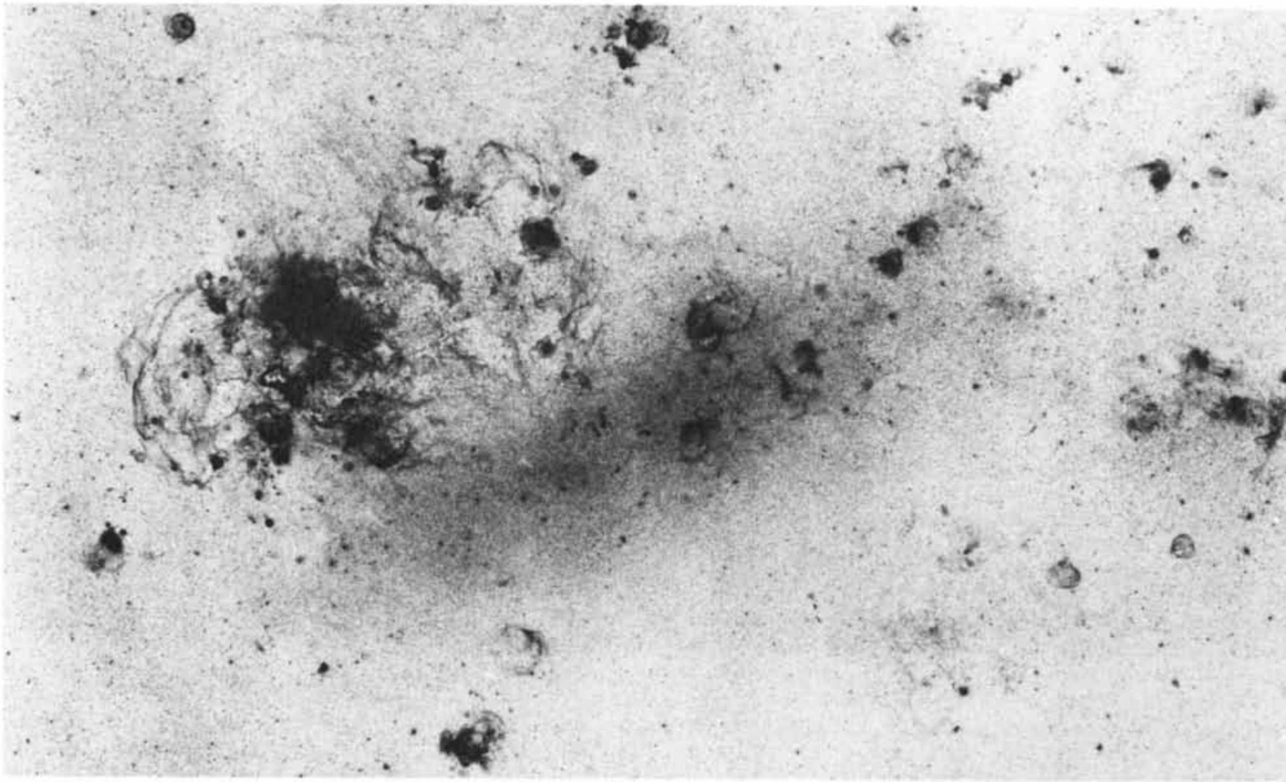
The most conspicuous feature of the clouds is a glittering array of very young and luminous blue-white stars; they are about 100 times as massive as the sun and about a million times as luminous. Many of these stars are surrounded by an emission nebula: a glowing arc of hydrogen gas ionized by the intense ultraviolet radiation from the central star. The LMC includes the largest of all known emission nebulas, called 30 Doradus, or the Tarantula.

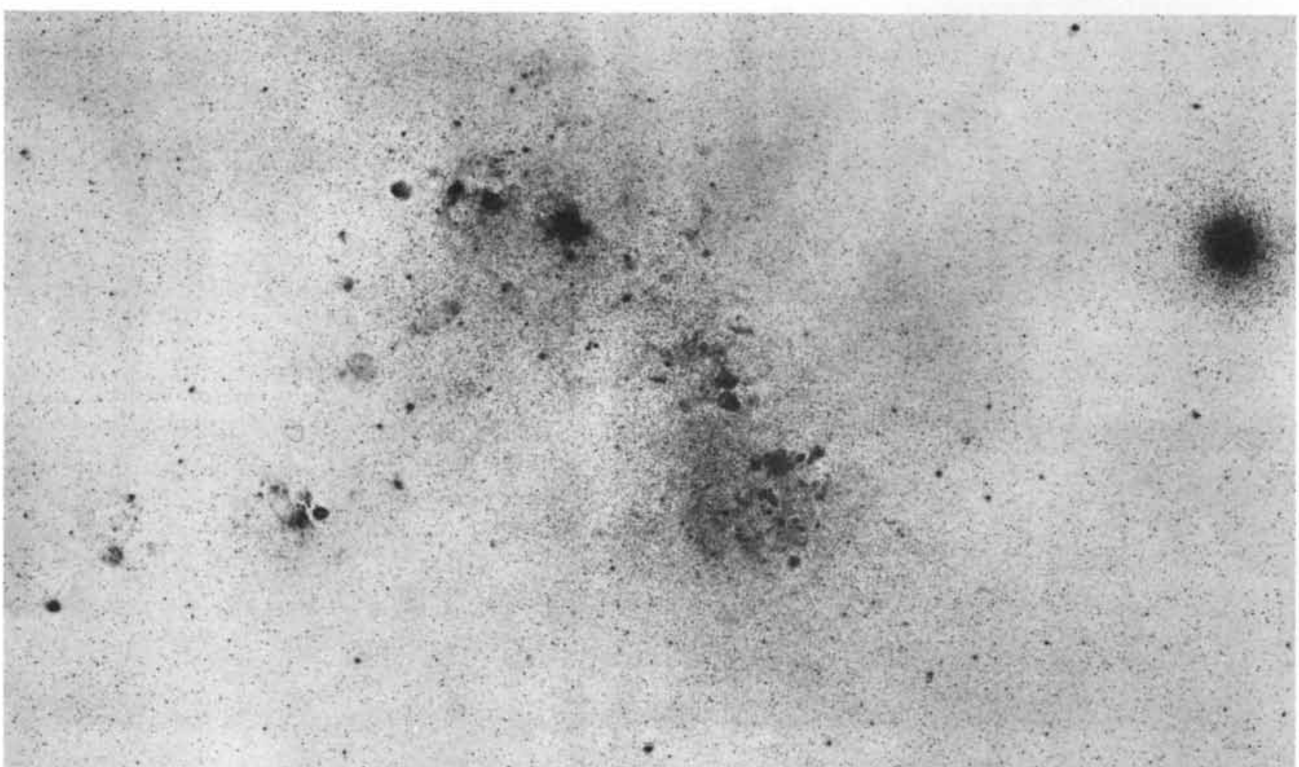
Underlying this decorative icing are more numerous stars of lower mass and luminosity. Some of them are concentrated in small, roughly spherical aggregates called globular clusters. In the Milky Way most of the stars in globular clusters are quite old: roughly 10 billion years, or the same order of magnitude as the age of the universe. They were among the first stars to form in the galaxy. The globular clusters of the Magellanic clouds, in contrast, have a wide range of ages, and some of them are less than 100 million years old. The difference in age suggests a difference in history. It is generally accepted that the Milky Way coalesced rapidly—within about 200 million years—from a cloud of primordial hydrogen and helium. The first generation of stars, including those of the globular clusters, condensed before the galactic disk had fully taken shape, and so they have orbits inclined to the plane of the disk, forming a spherical halo. Later generations of stars are confined mainly to the disk.

The Magellanic clouds probably began their initial collapse at about the same time as our galaxy, but relatively few globular clusters were formed, and the first episode of star making was notably inefficient. Until about two billion years ago the clouds remained largely gaseous. Even today they are gas-rich systems: 10 percent of the mass of the LMC and 30 percent of the mass of the SMC consists of un-ionized, or neutral, atomic hydrogen, compared with only 2 percent of the mass of the Milky Way.

Two billion years ago some large-scale event triggered a burst of star-forming activity in both clouds. Most of the stars visible today were born in this period. The instigating event may have been the passage of the clouds close to our galaxy.

Although observations of the clouds have been made at many wavelengths, much of what is known about the structure and history of the clouds comes from radio astronomy. The abundant neutral hydrogen of the





IONIZED HYDROGEN IN THE MAGELLANIC CLOUDS is detected by means of the red light emitted in a narrow band of wavelengths centered on the spectral feature called the hydrogen-alpha line. In the LMC (*upper photograph*) numerous emission nebulas appear as dark blobs, swirls and filaments. The most prominent of the nebulas, at the left side of the image, is called 30 Doradus, or the Tarantula; it is the largest emission nebula known anywhere in the universe. In the SMC (*lower photograph*) there are fewer emission nebulas and they are generally smaller, but a string of nebulas at the left is notable in that it marks the start of the bridge of hydrogen gas that extends between the clouds. Only hydrogen that has been ionized by the intense ultraviolet radiation from a nearby hot star shows up in these photographs; the clouds include vastly larger quantities of hydrogen in un-ionized, or neutral, form. The photographs were made by Roderick Davies, E. Kenneth Elliott and John Meaburn with the S.E.R.C. Schmidt telescope at the Siding Spring Observatory. The emulsions were exposed for five hours through a filter that excluded all light outside the narrow band of wavelengths centered on the hydrogen-alpha line. The images are printed as negatives to enhance faint features.

clouds can readily be detected by radio telescopes as a result of its emission at a wavelength of 21 centimeters. The strength of the 21-centimeter radiation yields an estimate of the number of hydrogen atoms present. Equally important, the Doppler shift in the wavelength of the emission indicates the radial velocity of the hydrogen: its speed toward or away from us.

Early surveys of the neutral hydrogen in the Magellanic clouds were made by James V. Hindman, Frank J. Kerr and Richard X. McGee of the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO). They showed that the clouds are embedded in an extensive envelope of hydrogen and that a bridge of gas extends between them. Higher-resolution surveys were later made with the 210-foot radio telescope at Parkes in New South Wales; the LMC was examined by McGee and Janice Milton and the SMC by Hindman. They found that the bright emission nebulas lie in dense concentrations of gas.

For the LMC the Parkes survey provided a complete picture of the velocity structure of the galaxy, confirming earlier patchy measurements of the radial velocity of stars. These measurements show that the LMC is a rotating disk inclined by about 30 degrees to our line of sight, although the velocity pattern is not the one expected for an undistorted rotating disk of gas. Two of my colleagues at the Mount Stromlo Observatory near Canberra, Philip Schwarz and Vincent Ford, and I concluded that the pattern fits a badly warped disk. The distortion of the disk is particularly severe in the region nearest the SMC, where the bridge of gas joins the LMC.

If the velocity pattern of the LMC is distorted, that of the SMC is chaotic. Indeed, the velocity measurements at first seemed impossible to interpret. Over almost the entire area of the galaxy the hydrogen shows two distinct peaks in velocity that differ by as much as 30 to 50 kilometers per second. A rotating disk would have a range of radial velocities, but at any one point all the hydrogen would be moving at the same speed. A common explanation of multiple values in a velocity measurement is the presence of an expanding shell of gas: one component of the velocity derives from the approaching part of the shell and the other from the receding part. A hypothesis of this kind was put forward 20 years ago for the SMC: it suggested the existence of three expanding shells of hydrogen within the galaxy. The hypothesis was accepted until recently, but it is now untenable. The two velocity peaks have been found in a variety of objects that lie well outside the supposed boundaries of the gas shells.

Radial velocities in the SMC have been measured for young stars, for emission nebulas and for planetary nebulas as well as for neutral hydrogen. In every case, and throughout the area of the galaxy, the same bimodal distribution is observed. It follows that the two peaks in hydrogen velocity must derive from separate entities, which have their own stellar and nebular populations. In other words, the material associated with each velocity component includes all the ingredients of a complete galaxy. Furthermore, the pattern of radial velocities within each component is fairly similar to that of a rotating disk of gas, although there are irregularities. The irregularities mirror those in the LMC: they are most pronounced on the side facing the companion galaxy, at the junction with the intercloud bridge of gas.

To interpret the Doppler shifts of the Magellanic clouds it is necessary to subtract the velocity of the solar system in its orbit around the nucleus of the Milky Way. When the adjustment is made, it turns out the lower-velocity component of the SMC is moving toward us at 25 kilometers per second and the higher-velocity component is moving away at 15 kilometers per second. Thus the components have a relative velocity of 40 kilometers per second. There remains an ambiguity: the components could be either converging or separating.

The ambiguity was resolved by examining the spectra of bright stars in the SMC. The spectral lines emitted by some of the stars were found to be shifted by an amount equivalent to a velocity of 15 kilometers per second away from us; the stars are therefore part of the higher-velocity component. The same stellar spectra showed absorption lines due to interstellar calcium with a Doppler shift equivalent to a velocity of 25 kilometers per second toward us. The absorbing material must be in front of the emitting star, and so the two components must be separating.

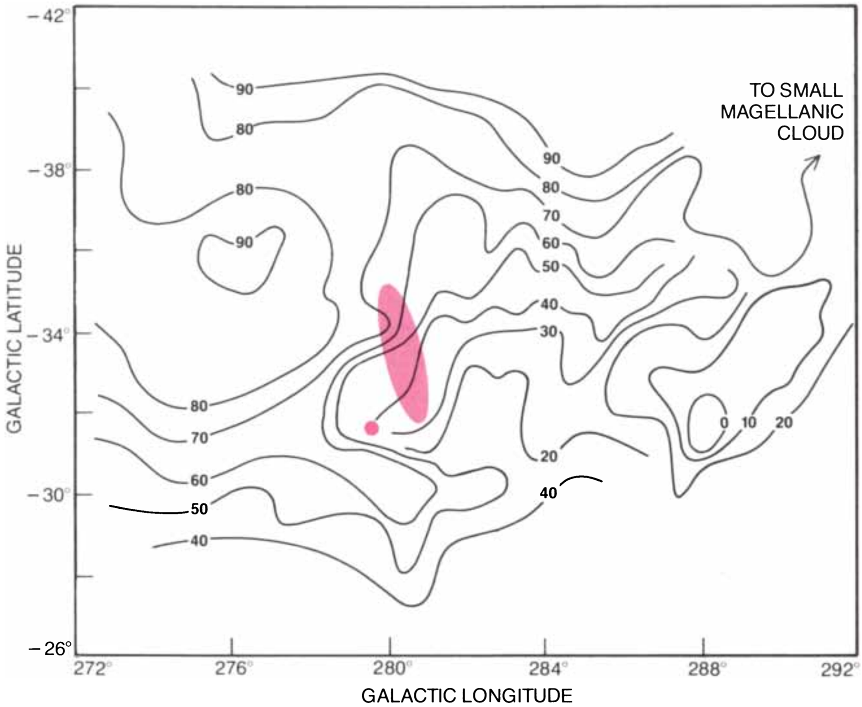
This finding is particularly interesting in view of numerous recent reports in the astronomical literature suggesting the SMC is an object of great depth. Measurements of the distances of young, bright stars indicate they are spread over a range of more than 30,000 light-years, much more than the thickness of a typical galaxy.

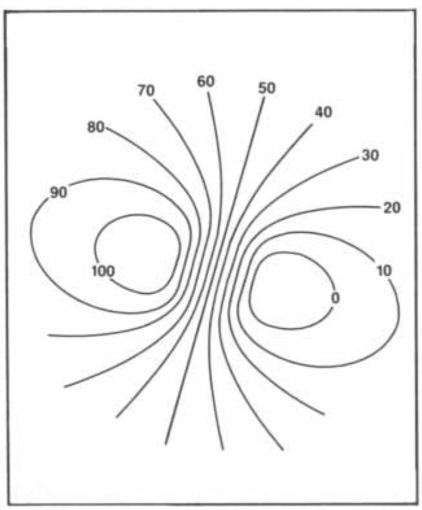
It was on the basis of the evidence outlined above that Ford and I concluded the SMC consists of two galaxies superposed along our line of sight. They must be fragments of what was once a single galaxy; the SMC has been torn in half! The two fragmentary galaxies cover about the same region of the sky, but the center of the lower-velocity (nearer) component lies to the southwest of the higher-velocity (farther) component. The population of older globular clusters is also concentrated in the southwest, and so it appears that the main remnant of the SMC should be identified with the lower-velocity component. I have named the new galaxy associated with the higher-velocity component the Mini-Magellanic Cloud.

The bridge of gas joining the LMC and the SMC apparently has connections to both the SMC Remnant and the Mini-Magellanic Cloud. The characteristic double peaks in the velocity of hydrogen have been observed along the bridge. Moreover, Mary Brück of the Royal Observatory in Edinburgh has found that the bridge has stars embedded along its entire length. Ford and I have measured the radial velocity of some of the stars, again finding two similar peaks. Thus the bridge is composed of two streams of gas that, like the two components of the SMC itself, appear to be one because they are superposed along the line of sight. Outside the bridge in the region between the clouds the distribution of hydrogen has a complex structure, with long spurs extending from the bridge. Nevertheless, the velocities measured in this region merge smoothly with those of the clouds, strongly suggesting that the LMC and the two pieces of the SMC form a bound system.

Unquestionably the biggest surprise in the exploration of the Magellanic clouds is the discovery of the Magellanic Stream. This is the filament of hydrogen gas that spans more than a quarter of the sky. The first part of the stream to be discovered extends through about 35 degrees, from the neighborhood of the Andromeda galaxy toward the Milky Way's south galactic pole. This stretch of the filament was mapped between 1965 and 1972 by the Northern Hemisphere radio astronomers Nannielou H. Dieter, Aad Hulsbosch, Ernst Raimond and J. van Kuilenburg. In 1972 Peter G. Wannier and George T. Wrixon, using a sensitive radio telescope at Bell Laboratories in Holmdel, N.J., extended the length to 60 degrees and discovered that the radial velocity of the gas increases toward the south galactic pole. The variation in velocity can be described by a sinusoidal curve, that is, the velocity varies in proportion to the sine of the angular position.

Reading Wannier and Wrixon's account of these findings, I noticed that if





STRUCTURE OF THE LMC is suggested by a map of radial velocities (speeds toward or away from the solar system). The measured values, which have been adjusted to compensate for the motion of the solar system, are all positive, indicating movement away from the Milky Way. At the left is the velocity profile of an ideal rotating disk of gas inclined at an angle of 30 degrees to our line of sight. The velocity contours for the LMC have the same pattern, but their distortion suggests the disk is badly warped. The shaded region represents the central bar and the dot indicates the position of 30 Doradus. Velocities were determined from the Doppler shift of radiation emitted by neutral hydrogen at a wavelength of 21 centimeters. The measurements were made by Vincent Ford, Philip Schwarz, John Murray and the author, using the 60-foot radio telescope at Parkes in New South Wales.

the filament were extrapolated down past the south galactic pole, it would pass close to the center of mass of the Magellanic clouds. Furthermore, if the sinusoidal increase in radial velocity were also extrapolated, the value obtained at the position of the Magellanic clouds would not be very different from the velocity of the clouds themselves. Both observation and extrapolation strongly associated the stream with the clouds. It was an exciting moment.

It took only three hours of observing with the 210-foot radio telescope at Parkes to establish that along the extrapolated path of the filament there is indeed hydrogen gas with the predicted velocities. It took seven long months, however, to complete a survey of the entire southern sky and make certain the gas we were detecting was not a more broadly distributed background unconnected with the stream. Martha N. Cleary, John D. Murray and I made the survey with the 60-foot radio telescope at Parkes, which we had taken out of mothballs for the task.

Since then the stream has been studied by many other astronomers. It is now known to have an angular extent of at least 110 degrees. In addition six major concentrations of hydrogen gas have been identified in it. The stream rather resembles a string of sausages. The density of the concentrations decreases with distance from the Magellanic clouds. The presence of the stream is clear evidence of an interaction of some kind between the clouds and our galaxy. The challenge is to define the nature of that interaction.

Any theory of the origin of the Magellanic Stream must explain four key observations. First, the path of the stream is nearly a great circle in the coordinate system of the Milky Way and passes close to the south galactic pole; in other words, the stream lies in a plane that passes through the center of the Milky Way and is almost perpendicular to the galactic disk. Second, the stream appears to be associated more with the SMC than with the LMC, although its source is probably among the spurs of gas found in the region between the clouds. Third, the radial velocity of the material in the stream varies almost sinusoidally from 200 kilometers per second toward the solar system at the tip of the stream to about zero where it joins the clouds. Finally, the stream, unlike the bridge between the LMC and the SMC, has no stars associated with it.

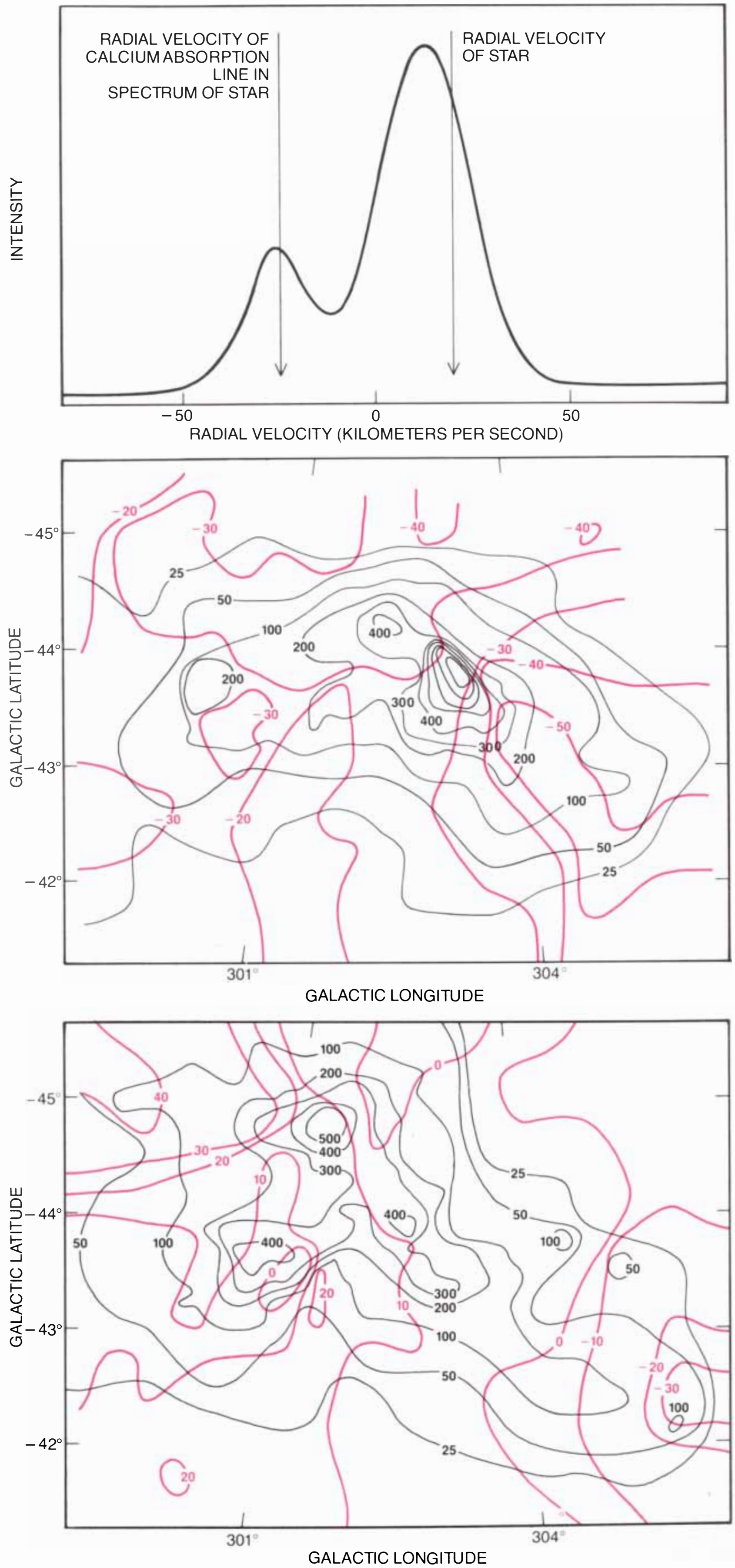
The most popular explanation of the Magellanic Stream has been the "tidal" theory, which proposes that tides raised in the Magellanic clouds during a close encounter with the Milky Way drew out a long tail of gas. In the early 1970's Alar and Juri Toomre of the Massachusetts Institute of Technology had much success in modeling the bridges and tails of many distant galaxies as tidal effects. The tides are created in exactly the same way as the familiar ocean tides induced by the moon. A point on the earth directly under the moon experiences a greater gravitational pull toward the moon than a point on the opposite side of the earth; in galaxies there is a similar difference between the forces at nearer and farther points. The more extended a galaxy, the greater the distortion caused by tides. For this reason the gaseous component of a galaxy, which is spread over a larger volume than the stars are, is more readily pulled out into a stream.

The initial papers proposing a tidal explanation of the stream were incorrect; they made the wrong assumption about the direction of the clouds' motion around the Milky Way. Later models, particularly those of Douglas Lin and Donald Lynden-Bell of the Cambridge University Institute of Astronomy and Tadayuki Murai and Mitsuaki Fujimoto of Nagoya University, reversed the direction of the orbit. Computer simulations were able to account for the basic features of the stream as tidal effects.

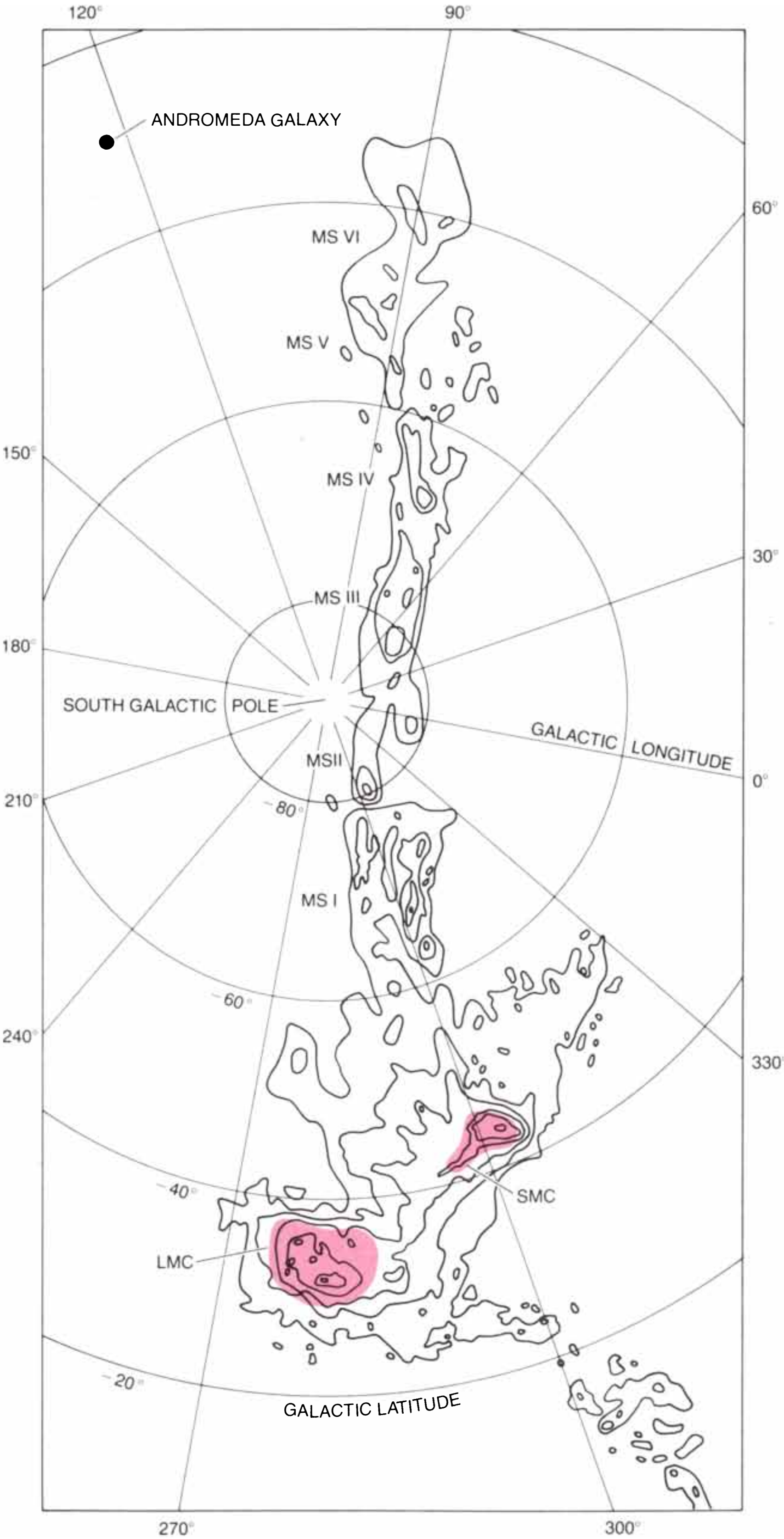
Murai and Fujimoto considered not only the tidal interaction between the clouds and our galaxy but also tides between the LMC and the SMC. They assumed the clouds are in an elliptical orbit (that is, they are gravitationally bound satellites of the Milky Way), are currently just past their point of closest approach and are moving counterclockwise as viewed from the solar system. The computer simulation begins two billion years ago, when the clouds were in roughly the same position as they are now, at the beginning of their last orbit. Murai and Fujimoto find that the tidal interaction between the LMC and the SMC forms a bridge between them and a short tail on the SMC, which then evolves into a much longer tail under the influence of the gravitational field of the Milky Way. In order to reproduce the correct distribution of velocities along the Magellanic Stream they must postulate that the Milky Way has a dark halo of 10^{12} solar masses, about 10 times the visible mass of the galaxy. Such an assumption is reasonable; indeed, there are other reasons for supposing the Milky Way has a massive halo.

An important feature of Murai and Fujimoto's model is that the LMC and the SMC nearly collided some 200 million years ago. In the aftermath of the near-miss the SMC was split into two concentrations that differed in radial velocity by about 60 kilometers per second and were separated by 40,000 light-years. Thus the model reproduces in remarkable detail the current configuration of the three galaxies, and in particular that of the SMC.

Murai and Fujimoto's model yields a satisfying result, but a serious problem remains. The model works only if the simulation is started at the beginning of the last orbit. If the Magellan-



VELOCITY MEASUREMENTS in the SMC give evidence of a complex structure with two components. The distribution of velocities (*top*) has two peaks, with some material approaching the solar system at 25 kilometers per second and other material receding at 15 kilometers per second. The spectra of stars in the receding component have absorption lines associated with the approaching component, and so the absorbing gas must be in front of the receding material. The evidence suggests the SMC consists of two galaxies moving apart at 40 kilometers per second. The nearer galaxy is called the SMC Remnant and the farther one is called the Mini-Magellanic Cloud. By measuring the Doppler shift of 21-centimeter radiation the two galaxies can be distinguished. The maps of the SMC Remnant (*middle*) and the Mini-Magellanic Cloud (*bottom*) show the density of neutral hydrogen (*black*) and radial velocity (*color*).



MAGELLANIC STREAM is a lumpy trail of neutral hydrogen that follows a great-circle path from the neighborhood of the Andromeda galaxy past the south galactic pole to the clouds. Its total angular extent, as seen from the solar system, is 110 degrees, or more than a quarter of the sky. The contours represent the large-scale distribution of neutral hydrogen determined by radio-frequency observations. Shaded areas indicate the optical extent of the LMC and the SMC. Six concentrations of gas are designated MS I through MS VI.

ic clouds are actually satellites of the Milky Way, the association presumably dates from 10 billion years ago, when the galaxies formed. If the simulation is started that early, however, it fails to reproduce the stream. To overcome this difficulty Murai and Fujimoto suggest that the LMC and the SMC were not bound to each other 10 billion years ago. Instead the LMC captured the SMC two billion years ago, after which the sequence of events followed that of the model. The capture might explain the sudden surge of star formation in the clouds two billion years ago.

There is little doubt that the fissioning of the SMC was caused by the tidal influence of the LMC. It is also clear that the origin of the bridge between the clouds lies in their tidal interaction. I have nagging doubts, however, about the tidal origin of the Magellanic Stream. The proposed mechanism requires a number of events of rather low probability, such as the fortuitous capture of the SMC by the LMC. Moreover, why are there no stars in the stream? In similar filaments of hydrogen in other tidally interacting systems stars are mixed with the gas. Since the tidal explanation does not seem entirely satisfying, my colleagues and I at Mount Stromlo have put forward two models of the origin of the stream. We call them the primordial and the hydrodynamic models. It is fair to note that they too require special circumstances.

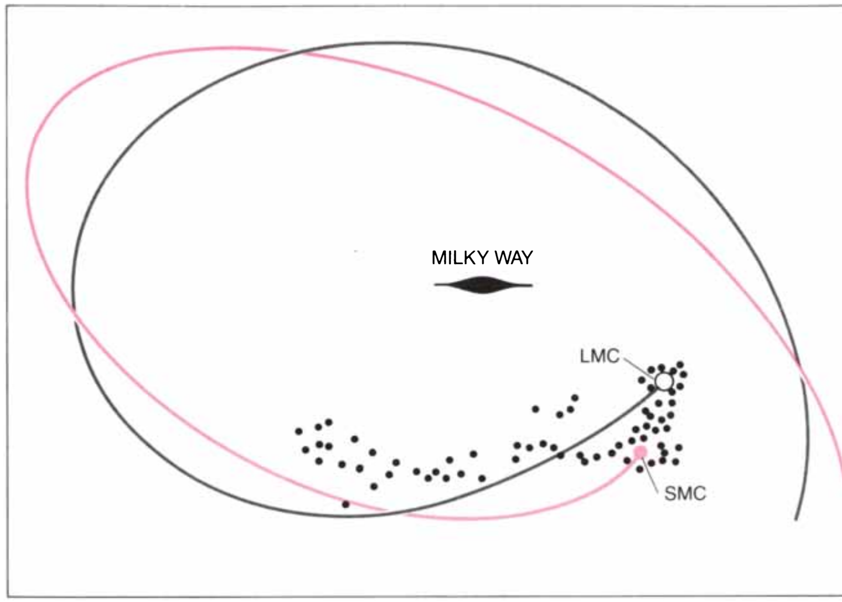
The primordial model is the simplest of all the models. It supposes the stream is a trail of material left behind by the clouds along their orbit. Having the material laid down along the orbital path leads to some appealing simplifications. In the celestial mechanics of Kepler and Newton the path followed by an orbiting body has certain special geometric properties when it is viewed from a focus of the orbit; in this case the focus is the nucleus of the Milky Way, but on an intergalactic scale of distances the solar system is reasonably close to the focus. From a focus the radial velocity of the orbiting body varies sinusoidally as a function of angular position along the path. Furthermore, the orbit seen from a focus invariably projects on the sky as a great circle. Of course, these are two salient properties of the Magellanic Stream.

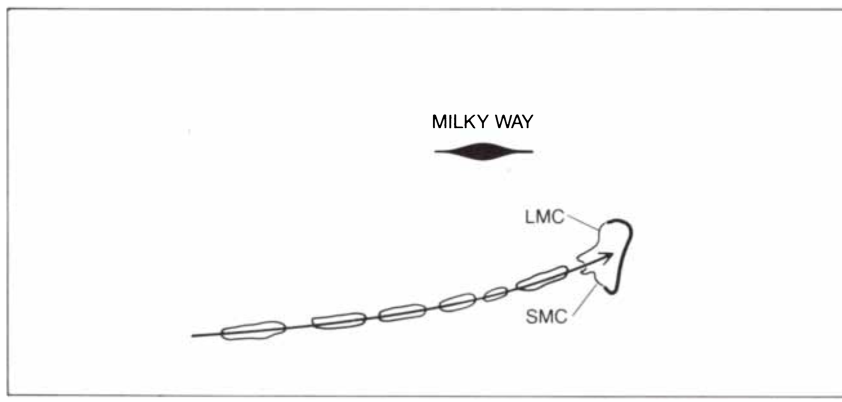
From measurements of the position and the velocity of the clouds and the stream, the geometry of the clouds' orbit can be calculated in the primordial model. It turns out the orbit is hyperbolic. In other words, it is not an orbit at all in the sense of a closed path; it is an open trajectory on which the clouds make one close approach to the Milky Way and, if their motion is not disturbed, proceed on to other regions of space. According to the primordial model, the clouds came from the direction of the Andromeda galaxy and are now just beyond their point of closest approach, which took them to within 160,000 light-years of the Milky Way. Their velocity at closest approach was 300 kilometers per second.

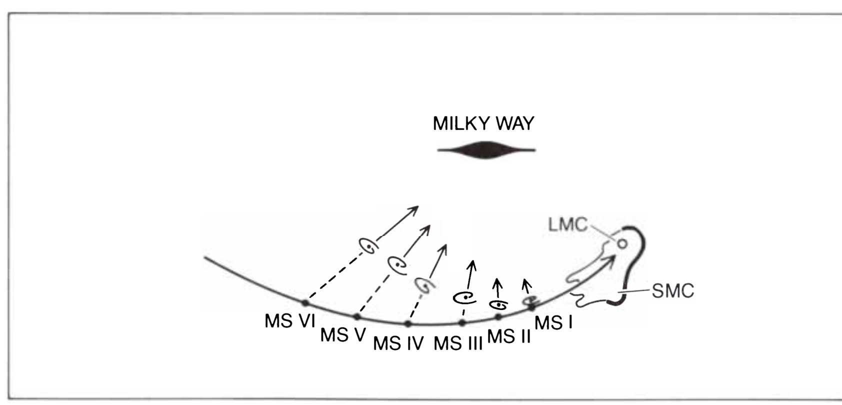
The weakness of this theory is that it offers no explanation of how the Magellanic Stream, with its six condensations of hydrogen, has been drawn out of the clouds. The answer may lie in the effect of ram pressure, exerted by gas in the halo of the Milky Way on gas in the region between the LMC and the SMC. The tail of hydrogen would be pulled out in much the same way as a cyclist's cap is swept away by his own slipstream. The ram-pressure mechanism appears to operate in dense clusters of galaxies, where many galaxies are embedded in a gaseous medium; because of their motion through the medium, the galaxies become depleted in gas. I have estimated that if the halo of the Milky Way has a gas density of 200 atoms per cubic meter and if the speed of the Magellanic clouds is more than 200 kilometers per second, some gas could be drawn out of the region between the LMC and the SMC. Under certain conditions it could form a long tail of gas without stars.

There is evidence that ram pressure is at work elsewhere in the Magellanic clouds: compression along the leading edge of the clouds has created a steep gradient there in the density of neutral hydrogen. A gas density of 200 atoms per cubic meter in the Milky Way's halo would account for the observed compression. With three of my colleagues at the Mount Stromlo and Siding Spring Observatories—Geoffrey V. Bicknell, Robert Gingold and Michael Dopita—I am currently attempting to develop a more detailed and quantitative model of the Magellanic Stream based on the ram-pressure mechanism.

If ram pressure is responsible for the stream, there is a further reason to think that the clouds are following a hyperbolic orbit and that this is their first encounter with our galaxy. There are now some 500 million solar masses of weakly bound gas in the intercloud region; this gas constitutes about half of the total gas content of the Magellanic system. If the clouds were gravitationally bound to the Milky Way and had been orbiting within its halo for 10 billion years, ram pressure would have swept the gas away long ago. It is also worth noting that the 28 members of the Local Group of galaxies (of which the Milky Way and the Andromeda







MODELS OF THE MAGELLANIC STREAM suggest three possible explanations of its origin. The tidal model (*top*) assumes the clouds have been in an elliptical orbit around the Milky Way for at least 1.8 billion years, the time needed to complete one revolution. A computer simulation done by Tadayuki Murai and Mitsuaki Fujimoto of Nagoya University shows that tides raised in the galaxies could pull out a tendril of gas along the observed arc of the stream. Most of the material comes from the SMC. The primordial model (*middle*) proposes that the material of the stream is strung out along the orbital path of the clouds. It could have been stripped away by ram pressure: the drag caused by the supersonic passage of the clouds through the gaseous halo of the Milky Way. If the stream marks the trajectory of the clouds, their orbit must be a hyperbola rather than a closed path, and this must be their first encounter with the Milky Way. The hydrodynamic, or turbulent-wake, model (*bottom*) also assumes a hyperbolic orbit, but the gas in the stream is not extracted from the clouds; instead it consists of whirls or vortexes created in the gaseous medium of the Milky Way's halo. Vortexes formed in the wake of the moving clouds become denser and cooler than the surrounding gas, and so they sink toward the nucleus of the Milky Way; the sinking could explain the observed distribution of radial velocity along the stream.

galaxy are the most prominent members) lie in a plane; the Magellanic clouds and the stream are in the same plane, suggesting they are members of the Local Group in their own right and not mere satellites of the Milky Way.

The hydrodynamic, or turbulent-wake, theory of the Magellanic Stream was put forward by Schwarz, Murray and me. One of its notable achievements is that it can account for the six separate concentrations of gas that appear at roughly equal intervals along the stream.

In the hydrodynamic model, as in the primordial one, the gaseous halo of the Milky Way is a major player, but it has a much different role. The stream does not consist of gas from the Magellanic clouds torn loose by ram pressure; rather, the material of the stream is gas from the halo itself, perturbed by the passage of the clouds. The six large aggregations of hydrogen are vortexes formed as the clouds hurtled through the halo at supersonic speed.

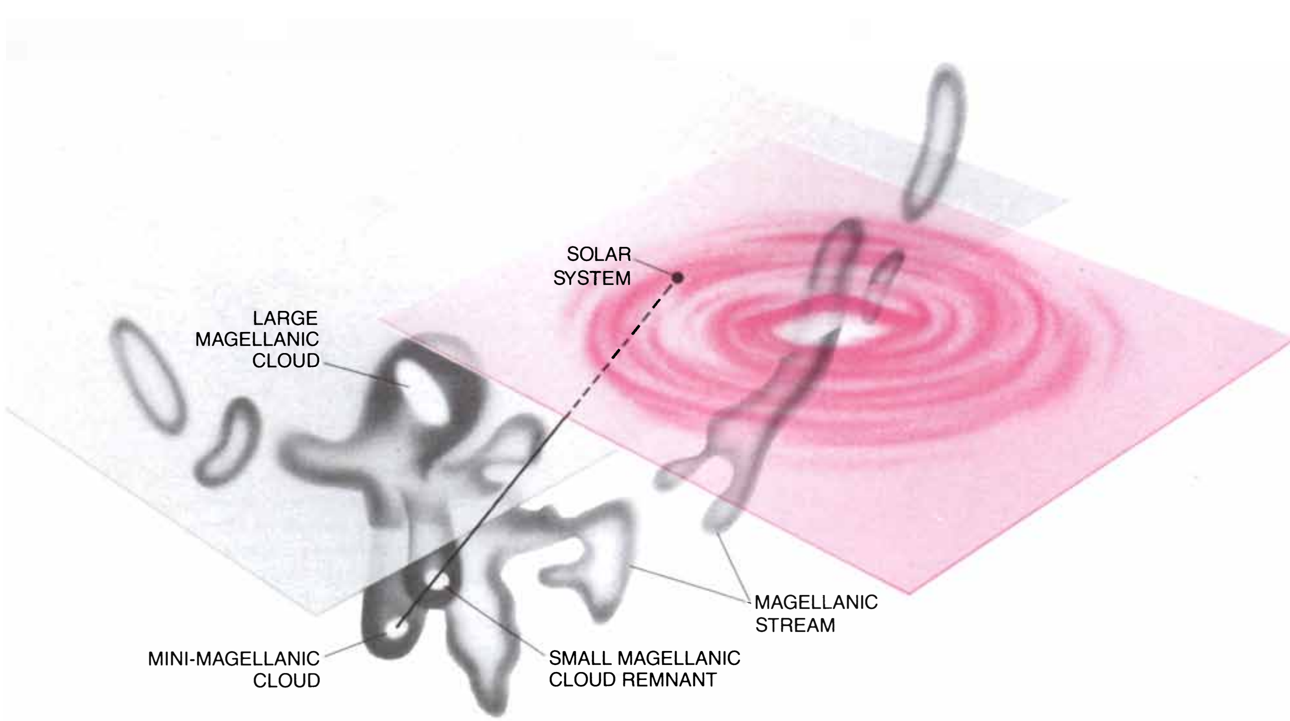
The center of each vortex becomes a region of higher than average density, which leads to rapid cooling. The cool, dense gas begins to sink toward the center of the galaxy. It is the sinking that gives rise to the observed sinusoidal distribution of radial velocities along the stream. The condensation of gas nearest the Magellanic clouds has almost zero velocity because it is the youngest and has just begun to fall. The velocity increases steadily with distance from the clouds because the older parts of the stream have had longer to accelerate in their free fall toward the galactic core.

The hydrodynamic model is in good agreement with the radial-velocity measurements. It also explains why all parts of the stream are approaching us and have little transverse motion. Still another observation that fits readily into the model is the fading of the stream as it ages: as the turbulent energy is dissipated, the density difference that defines the stream diminishes. On the other hand, some hydrodynamicists have criticized the model because they are uncertain whether the vortexes would cause a density increase.

What will be the fate of the Magellanic clouds? The answer depends on how much orbital energy they are losing to friction as they plunge through the galactic halo. Even if their original orbit was a hyperbolic one, it will not necessarily remain hyperbolic. Friction is generated as gas in the halo is deflected by the passage of the clouds, effectively braking their motion. Scott D. Tremaine of M.I.T. thinks the clouds may already have been captured by our galaxy. Tremaine maintains that the rate of energy loss is great enough to cause the clouds to spiral inward and collide face on with the Milky Way in another two billion years. The collision would not be a solid impact. The clouds would be stripped of their gas, but the stars and the globular clusters would pass through largely unscathed.

In the aftermath of such a collision the remnants of the clouds would continue to circulate around the Milky Way in polar orbits (perpendicular to the plane of the disk). The globular clusters would be their most prominent component. In this connection an interesting observation has been made by William E. Kunkel of the Cerro Tololo Inter-American Observatory in Chile. He has drawn attention to the fact that the outlying globular clusters and the slightly larger stellar aggregates called dwarf spheroidals are distributed in a nearly polar ring. It is likely they are the debris left by the collision of another galaxy with the Milky Way some six billion years ago. François Schweizer, Bradley C. Whitmore and Vera C. Rubin of the Carnegie Institution of Washington have recently shown that such polar rings are not uncommon in galaxies.

If the Magellanic clouds turn out to be victims of our galaxy's cannibalism, they may not be the first such victims.



SYSTEM OF GALAXIES made up of the Milky Way and the Magellanic clouds is shown in an artist's rendering of what the system might look like if it could be seen from the outside and if the gaseous hydrogen that permeates and surrounds the clouds were visible. The Milky Way is seen to be a majestic spiral galaxy, with a lenticular nucleus and pinwheel arms. The clouds are smaller and less symmetrical; moreover, a higher fraction of their mass is hydrogen gas. The two components of the SMC are clearly distinguished, as is the hydrogen bridge connecting the clouds. The Magellanic Stream is shown in the configuration assumed in the tidal model.

The above spillway (at LG III powerhouse site) is part of Hydro-Québec's La Grande Complex, Phase I generating facility at James Bay, Québec. With the complex's guaranteed production of 62.2 billion kilowatt hours per year, it is the largest hydroelectric project in the North America. It is an apt symbol of Canada's vast energy resources.

ENERGY FUEL POWER

CANADIAN ENERGY

With the dominance in world markets of conspicuous consumer products from Asia it may be overlooked that the largest two-way trade in the world is between the United States and Canada – $150 billion (Canadian) in 1984. For Canada much of its export is energy.

In an address to the Seventh World Energy Engineers Conference in Atlanta this last November, Allan E. Gotlieb, Ambassador of Canada to the United States, outlined the magnitude of Canadian trade: "We sell over $90 billion (Canadian) in goods and services to the world each year, much of it in the form of resource-based exports."

"Indeed," Ambassador Gotlieb continued, "the United States imports more energy from Canada than from any other country in the world."

While the proportion of Canada's contribution to the United States' overall energy needs is statistically low – 4% for oil and gas and 2% for electricity – this supply translates into virtually 100% of all U.S. imports of gas and electricity and about 10% of oil imports, making Canada the second largest supplier of petroleum to the United States, after Mexico, and before any OPEC country. Strategically, these imports are significant because their regional distribution permits the displacement or reallocation of U.S. domestic reserves which may be more costly to transport or produce. California, for example, relies on Canadian gas for as much as 40% of its supply.

Ten billion kilowatt hours of electricity currently flow into New York from Québec and Ontario. These exports are scheduled to increase dramatically with the completion of the Marcy South transmission line in 1988 and with the New York Power Authority's contract to purchase 111 billion kilowatt hours of electricity from Hydro-Québec through 1997.

The New Brunswick Electric Power Commission is selling more than one-third of the output from its 630 megawatt Point Lepreau nuclear reactor to Maine and Massachusetts. Boston Edison Company has the largest contract: 100 megawatts of electricity from Lepreau 1. The decision to develop a second nuclear plant at Lepreau will depend almost entirely on the New England export market.

ENERGY FUEL POWER

On November 1, 1984, the first direct exports of Canadian gas to states in the northeast began flowing across the border at Niagara Falls. The contract between TransCanada PipeLines and Boundary Gas (a consortium including Brooklyn Union Gas) calls for importing 40 million cubic feet a day of western Canadian gas increasing to 92.5 million in a second phase by 1986 or 87.

Canada is well endowed with energy: proven oil/gas resources in the Western Sedimentary Basin, Beaufort Sea, and Grand Banks; vast hydraulic power potential of the Laurentian Shield and generous reserves of coal and uranium.

This enormous energy supply potential is, however, scattered over a vast country. While some resources are easily accessible, others are remote and costly to develop.

The new federal government, which took office in September 1984, intends to maintain traditional energy policy objectives. An immediate priority, however, is stimulation of the economy and the creation of jobs. Ottawa believes that energy can play a major role in achieving these objectives. It is reviewing thoroughly energy policy and has indicated its intention of adopting a more market-oriented approach to energy pricing. Greater emphasis is being placed on encouraging increased participation by the Canadian private sector and on greater co-operation with provincial governments. The federal government's agreement with Newfoundland over the bitterly disputed right of development of the giant Hibernia oil field, 175 miles off Newfoundland, is the best example of this new spirit. A satisfactory resolution of negotiations with Alberta and Saskatchewan over the decontrol of oil and gas prices is also expected. At the same time, the federal government has signalled that foreign investment will play an important role. Federal Energy Minister Patricia Carney has stated that "energy is the engine of the economy."

Thus, Canada today looks forward to a decade or more of development of remote or difficult resources. High technology resource development is no longer a matter of manpower – it is capital and knowledge intensive. It will need risk capital and technology imports to realize its considerable energy supply potential.

TIDAL

The province of Nova Scotia is almost entirely surrounded by saltwater, and completion of each of its two major energy developments depends upon greater understanding of the ocean.

One is the Minas Basin tidal power project, to harness the 40 foot rise and fall of the tides at the end of the Bay of Fundy. A $500,000 pilot project at Annapolis, Nova Scotia is already in operation to test the efficiency of a new Canadian-developed turbine. Over 100 of these turbines would be used in a projected five-mile dam across the outlet of Cobequid Bay, at the end of the Minas Basin on Nova Scotia's northwest coast. Engineers estimate that up to 4,800 megawatts of electricity could be generated from the head of water trapped at high tide, twice as much as is now used in the entire province.

The low-head turbine is seen as an important feature of the project. It differs from the "bulb"-type turbines used at the only other tidal power plant in the world, at La Rance, near Cherbourg, France, because its compact shape and high efficiency at low heads of water will minimize the size and number of power houses on the dam.

The Nova Scotia Tidal Power Corp., a provincial development corporation, estimates that when fully developed, such a power plant could return up to $2.3 billion annually, paying for itself within ten years.

NATURAL GAS

Off the opposite shore of Nova Scotia lies a 40 mile long sandbar called Sable Island.

Mobil Oil Canada was first to strike a possibly commercial gasfield called "Venture," which it estimates has between 2-2.5 trillion cubic feet of reserves. The exigencies of building offshore platforms for gas production and treatment, however, and then laying a pipeline 250 miles along the ocean bed to shore near the Strait of Canso, require assurance of very large quantities of saleable gas – the threshold of economic viability is now estimated between 2.5 and 3.0 trillion cubic feet.

Fortunately, Shell Canada has discovered a second field, about 50 miles south of Venture, called the Glenelg field, which the company believes to have recoverable reserves of over one trillion cubic feet. While each field would require its own production platforms, they could share a single pipeline to shore, thus reducing the cost of each.

Most of the natural gas delivered by the two-phase line will go on via a trunk line to the New England states. Conditional contracts, involving 72% of Venture's output, have already been negotiated by the major interest holders – Petro-Canada and Mobil – and the New England States Pipeline group and Tenneco.

WATER

The sharp fall of water from the Laurentian highlands of Québec to sea-level has provided that province with vast reserves of hydraulic power, the world's most important renewable energy resource. Over 40,000 megawatts of this resource, it is estimated, is still undeveloped. The opportunities offered by export of hydroelectric power to areas of the United States previously dependent on expensive and wasteful combustion of fossil fuels, or nuclear energy, has been an important part of the economic strategy of Québec and its provincial publicly-owned utlity, Hydro-Québec. It provides a means to facilitate early exploitation of these resources, and lowers local costs of electric power, thus encouraging Québec's own industrial development.

Jacques Guevremont, Hydro-Québec's vice president, external markets, describes how the utility has changed its export policy: *"In view of the changes which have occurred in the Québec and American markets, Hydro-Québec has revised its export sales strategy. One fundamental change is that the exportation of electricity has become one of the utility's prime objectives. It is no longer a question of selling temporary surpluses as in the past. During the next few years (say, until 1988) efforts will be made to export surplus production; over a longer period, the objective will be guaranteed electricity contracts."*

Electricity exports to the U.S. took a leap in 1979, following completion of the 735 kilowatts transmission line from Chateauguay, near Montreal, to Massena, N.Y. In March 1982, the New York State Power Authority contracted for 111 billion kilowatt hours of interruptable power between 1984 and 1997, and seasonal imports of 1,200 megawatts of firm power to 1992. At less than 4 cents per kilowatt hour, it will be 30-40% cheaper than nuclear and oil-fired generation. The 33-member New England Power Pool agreed in August to buy 33 billion kilowatt hours over the decade to 1995. This year, the State of Vermont will receive 150 megawatts of firm power on a year-round basis from 1985 to 1995. The contract, worth $50 million a year plus escalation, represents an important first, as Québec's first export of firm power on a 12-month basis. Some U.S. observers, noting the surge of electricity exports, have questioned whether this doesn't leave consumers just as exposed to energy price increases as before. But the export contracts signed by Hydro-Québec are based on the actual ratio of fuels used by the utilities in the areas affected, recalculated either annually or on a 12-month moving average. This means that, as costly oil (at about $50 per megawatt hour generated) is replaced by cheaper coal (at about $20 per megawatt hour) the price of imported electricity decreases relatively, so that there are ample savings to be shared by both importing utilities and the ultimate consumers.

This massive and growing trade in elec-

QUÉBEC PULLS THE HYDRO HAT TRICK

♣

Hydro-Québec
Generating, transmitting and distributing 99% hydroelectric power (22,000 MW of installed capacity with an additional hydroelectric potential of 29,000 MW) over some 4000 miles of high-voltage lines throughout Québec and to other parts of Canada and the U.S.

♦

IREQ
The largest research and test center of its kind in North America with unique facilities (including a high-voltage laboratory for testing at up to 1500 kV AC and 1200 kV DC, a high-power laboratory with a testing capacity of up to 40 GVA and system simulators) and a decade of successful electrical research projects for U.S. and other major organizations behind it.

♥

Hydro-Québec International
Linking Hydro-Québec's expertise with the experience of Québec construction and engineering firms and manufacturers to export technical assistance to developing countries throughout the world.

Hydro-Québec
Public Relations
75 Dorchester W.,
Montréal, Québec H2Z 1A4
Tel.: (514) 289-2341
Telex: 055-61047

Electrifficiency

ENERGY FUEL POWER

tricity has created a need for new technology. IREQ, the research and development arm of Hydro-Québec, pioneered extra-high voltage transmission at 735,000 volts many years ago, and is today studying systems rated up to 1.5 million volts. It is also pioneering the technology interconnecting with other systems via a direct current link, thus minimizing requirements for phase matching and smoothing load adjustments. The New York State inter-connection just completed consists of 100 kilometres of DC line at 435,000 volts.

Another problem being tackled at IREQ is the inevitability of surplus generating capacity in a system with very large units of generation. When Hydro-Québec's James Bay complex is completed this year, at a cost of $8.2 billion for generating facilities and $2.5 billion for transmission, the utility will have a 31% surplus of capacity over demand. Most of this energy can be saved, however, by water storage, only about 10% being "spilled" without reducing power.

One logical solution would be to produce electrolytic hydrogen using this surplus of power. A present project, sponsored by both federal and Québec governments, involves an *"action concertée"* between IREQ, Noranda, and Electrolyzer Corp. for a joint $50 million demonstration plant using an improved electrolysis process to generate the 17 tonnes per day of hydrogen required for a 100 tonne per day ammonia plant at Shawinigan, Qué. It is intended to demonstrate that electrolytic hydrogen, in large scale production, can compete with hydrogen derived from pyrolysis of natural gas.

WIND

Hydro-Québec also considers local as well as export needs in power technology development. In partnership with the National Research Council, it is supervising construction of "Eole," a $30 million, 4-megawatt wind-powered generator of the Darrieus turbine type at Cape Chat, on the Gaspé Peninsula, remote from provincial load centres. An earlier prototype is already contributing 200,000 kilowatts of electricity in the remote Magdalen Islands, in the Gulf of St. Lawrence. With many of Canada's northern and coastal communities remote from electricity grids, the development of large wind-powered generators is seen as a fuel conserving replacement for diesel generators, and also a possible high-technology export item.

NUCLEAR

Ontario has built up a substantial capacity with its present CANDU (heavy water) reactors which, in contrast to nuclear plants in many other countries, are working reliably and providing economical power. The final four-reactor complex in the present series, at Darlington, Ont. will come on line in the early 1990s, providing 3,400 megawatts of power.

The earliest commercial CANDU reactor in the province, Douglas Point, was commissioned over 20 years ago, however, and is now being retired. Others are middle aged, and require major maintenance work, such as replacement of all or part of their pressure tubes, one of the hallmarks of the CANDU design. Maintenance work on a reactor which has been in service for several years necessitates carrying out precise, complex operations in a high-radiation environment.

For this reason, the Ontario utility has been working with Spar Aerospace Ltd., the developer of the "space arm" used in the U.S. shuttle program, to develop a remote controlled hydraulically actuated "arm" to retube CANDU reactors and perform other complex repair and maintenance operations in high radiation environments, to avoid exposing human employees. The development of an articulated mechanical arm six metres long with a lifting capacity of six tonnes was announced to a U.S.-Canadian conference on "Robotics and Remote Handling in the Nuclear Industry" in Toronto last September.

HEAVY OILS AND BITUMENS

The key to the energy future of Canada's prairie provinces, Alberta and Saskatchewan, is development of "unconventional" oil resources, defined as liquid hydrocarbon deposits with a reservoir energy of zero. Within this group are "heavy oils" with a low enough viscosity to be pumped to the surface, and "bitumens" existing in "oil sands" which are virtually solid at ambient temperatures and must be dug up, if near the surface, or liquified by heat or other means before pumping, if deeply buried.

Provincial authorities in Alberta estimate total oil sands content of crude bitumen in place to be 250 billion cubic metres, of which only 10% is mineable; in Saskatchewan recoverable reserves of heavy oil are estimated at 30 billion cubic metres. These volumes are 1,000 times greater than the remaining (and rapidly depleting) reserves of light and medium crude.

G.J. Willmon, Vice President for Oil Sands and Coal of Esso Resources Canada, says: *The status of oil sands development, specifically in-situ development, is that it's a bright spot which will continue. In fact, my company sees its future as a heavy oil company. A standard in-situ process has been developed for in-situ development. First, deviated wells are drilled into the reservoirs from central sites. For weeks or even months, steam is injected down the wells to heat the bitumen to get it flowing. Then the steam is turned off and the heated oil is pumped out. This is the standard in-situ steam injection process. Since average recovery is only about 20% of the bitumen in place, there is obviously scope for improvement. Research to increase recovery is continuing throughout the industry and by government agencies.*

Esso is presently developing its Cold Lake bitumen resources by a phased in-situ process. Phase Two will go into operation this year, doubling current production to 19,000 barrels per day. Two more phases added in 1986 will again double production, for a total cost for facilities of $450 million. This project, and others like it, are initially export-oriented, owing to the limited capability of domestic refiners to transform bitumen and heavy oil into products which are in high demand in Canada, such as gasoline.

The massive U.S. highway reconstruction program sparked by the U.S. government in 1983 has created an unprecedented demand for asphalt. Esso is diluting its Cold Lake crude with condensate from gas wells and pipelining it to northern tier states as asphalt feedstock.

Alberta's remaining gas reservoirs are becoming drier, however, as exploitation continues, and supplies of condensate, which must be extracted from the gas at processing plants before pipelining it, will therefore diminish as time goes on. This, anticipates Hans Maciej, technical director of the Canadian Petroleum Association in Calgary, will create a major problem in the shipment of heavy oil in the near future.

One company, Bitech Ltd. of Calgary, is offering a solvent de-asphalting process, which it claims can be carried out at the wellhead, which reduces viscosity of the heavy crude by 90%. The process is called the Angelov/Shibley process, after the two mining engineers who invented it.

Another group is approaching the problem by attempting that proverbial impossibility – mixing oil and water. The work, still in the laboratory stage, seeks to find a detergent which would carry globules of heavy oil in a water suspension, yet could be deactivated at the receiving point to allow easy recovery of the oil by settling or centrifugation.

From a long-term viewpoint, however, it is equally important to use technology to upgrade heavy crude and bitumen so that it can be refined to produce the normal run of petroleum products, rather than asphalt and a little heavy fuel oil. Both the Suncor and Syncrude oil-sands mining operations have such upgrading facilities. These are based on fluid coking technologies, which cook the surplus organic carbon out of the

ENERGY FUEL POWER

bitumen at high temperatures and pressures, producing "synthetic" crude, plus a good deal of high-sulphur coke, which, so far, is a waste product. A new company, New Grade Energy, has been formed with federal and provincial support to treat heavy Saskatchewan crude by a hydrogenation process developed by Union Oil. The treated crude will then be processed by the Consumers' Co-operative Refinery in Regina. Husky Oil Ltd., which pioneered the early heavy oil fields in Saskatchewan, has proposed a $1.5 billion upgrader specifically to treat the peculiarly heavy crude of the Lloydminster area.

Shell's Scotford refinery near Edmonton began operating in the summer of 1984. This $1 billion plant incorporates state-of-the-art technology, and is the first refinery in the world to rely exclusively on high quality synthetic crude oil, mined and upgraded from the Alberta oil sands. Scotford therefore produces higher yields of marketable light petroleum products than conventional refineries. With an initial capacity of 8,000 cubic metres per day, it will produce a wide variety of products, mainly transportation fuels, as well as feedstock for Shell's nearby styrene plant.

Like the in-situ extraction process, only 20% efficient, the technology of upgrading heavy oils and bitumens is still in its infancy. It still isn't demonstrated, for instance, whether it is preferable to improve the hydrogen-to-carbon ratio of the crude by rejecting carbon or incorporating hydrogen. But plenty of research is going on in both eastern and western Canada.

GAS LIQUIFACTION

Imports of Canadian natural gas have been making headlines in the United States recently. After years of regulatory rigidity on both sides of the border, when export sales of Canadian gas slumped because of federal government insistence that it be unrealistically priced, at the BTU equivalent of oil, the new atmosphere of deregulation has allowed growing imports into the USA at competitive prices negotiated by individual buyers and sellers. But there are still vast quantities of Canadian gas shut into already discovered fields because of lack of market.

The Canadian gas industry has been seeking ways to export natural gas in other modes than by pipeline. This, essentially, means the conversion of natural gas to "liquid energy." One of the most obvious ways is simply by liquifaction. The energy required to cool and compress the gas can be supplied by the medium itself. One very large project started by Dome Petroleum in its high-flying days, and now being carried on by a consortium including Union Oil of Canada, Pan-Alberta gas, and other gas products of Edmonton, is to build a large liquifaction plant on the west coast, near Prince Rupert, B.C., for export of liquified natural gas (LNG) to Japan. The plant would provide the equivalent of 3 trillion cu. ft. of natural gas over a 20-year period to five Japanese utilities, which have no domestic access to natural gas, via half a dozen dedicated LNG tankers. Cost of the whole plan, including tankers, would be about $4 billion. Negotiations are proceeding in Tokyo to determine the exact price and considerations of delivery; a $650 million terminal and regasification plant for the LNG is already under construction in Japan, and the utilities are hoping for the first tanker delivery in January, 1987.

Another way of converting natural gas – which is largely composed of methane – to liquid energy is by transforming it into methanol. Ocelot Industries has built a 1,200 tonne a day methanol plant near the aluminum port of Kitimat, on the B.C. coast, strategically placed to export to the Pacific rim.

OIL

In Canada's far north, over a decade of active exploration activity has paid off in numerous discoveries of large reservoirs of oil and natural gas which would, in easier climates, have been developed years ago. Most of these reserves lie offshore, either on the Arctic Islands or at the bottom of the Beaufort Sea. To bring these hydrocarbons to market will require adaption of existing technology to extreme climatic conditions, and a choice of two possible routes to market: either via pipeline, first under the year-round ice cover, and then across hundreds of miles of tundra, or via ice-strengthened tankers, during the couple of months of ice breakup which is as close as the region gets to summer.

The solution varies, deposit by deposit, depending upon its proximity to the mainland. But the decision also implies an orientation to the end market: tankers will have to go around Alaska or through the Davis strait, and their convenient ports will not be in the energy demand centres of Canada. A pipeline, on the other hand, would tie into the existing network somewhere in northern Alberta, and would therefore be easily assimilated into Canada's domestic energy supply system.

Recently, the pipeline seems to be favoured. Only 400 miles now separate the oil and gas resources of the Mackenzie Delta and its near off-shore from the head of pipe at Norman Wells, NWT, where Esso Resources' northernmost oilfield will be tied into the Alberta pipeline network by this summer.

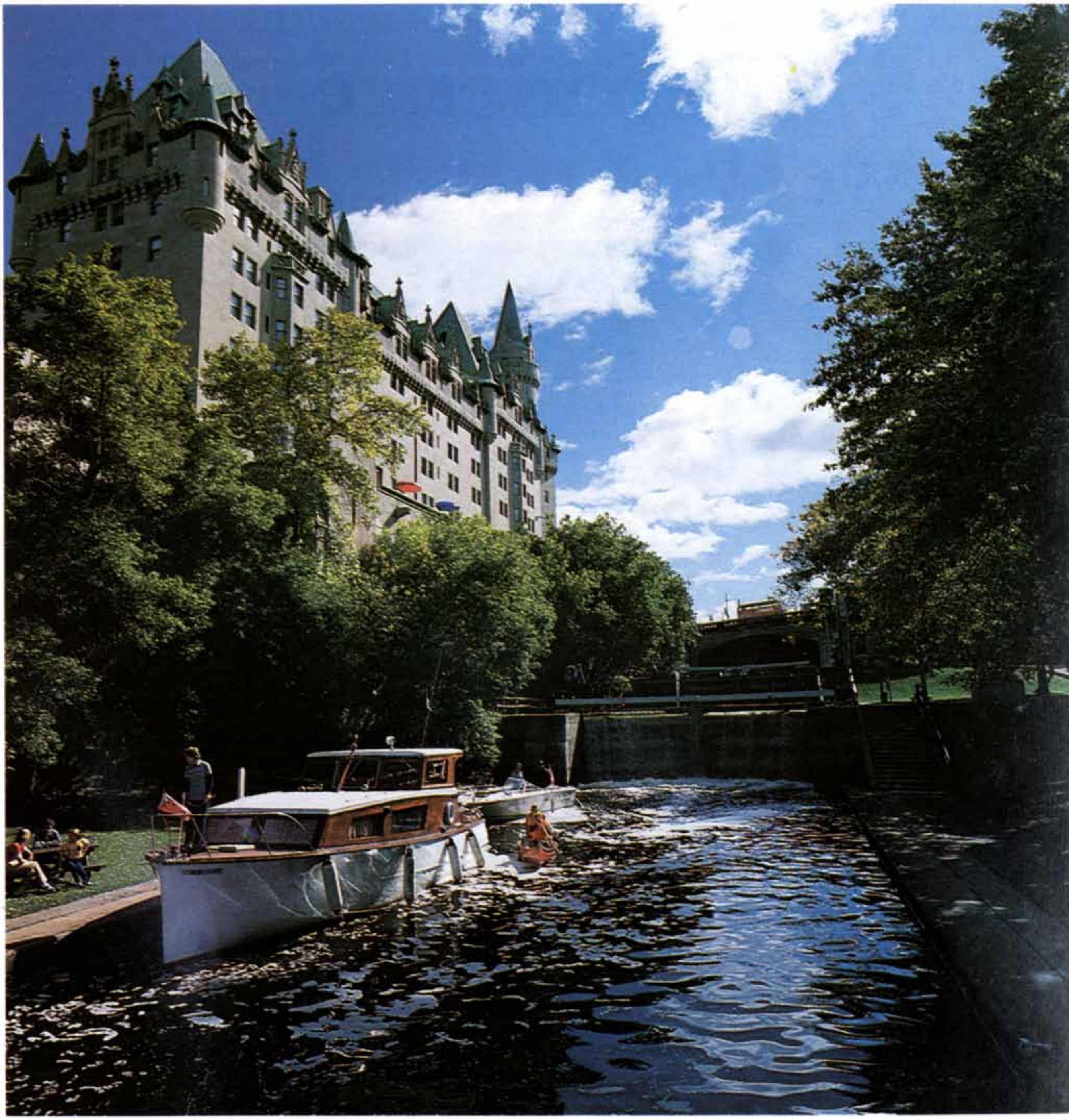
SYNERGY

One lesson which energy administrators and economists are learning from the recent oscillations of supply and demand is not to neglect the synergy between various energy forms in seeking solutions to development problems. One recent example is the construction of a pipeline to carry process steam from Ontario Hydro's Bruce nuclear generating station to a nearby industrial park. Hydro was stuck with the excess steam because of slower than normal electricity demand increases, and public opposition to the building of a transmission line from its nuclear facility to load centres to the south. The line had been stalled for lack of firm customers for the steam; potential customers were waiting for assurance that the line would be built. Finally a local group of citizens broke the deadlock. Their plan for a fermentation plant to make fuel ethanol was accepted by Ontario Hydro as sufficient justification to start the line, which should be completed by the end of the year. Thus, a citizen-based biomass alternative energy project became the means for utilizing surplus nuclear steam, and the catalyst for an important industrial development opportunity.

Similar synergies on a larger scale are likely to emerge as the patterns of Canadian-U.S. energy trade become more complex. To extract and exploit its more inaccessible energy resources, Canada needs both to import and to develop new technology. To earn the investment capital it needs to realize the energy projects of the future, it has to export existing energy resources. It is a cycle to which Canada appears willingly committed.

Hugh C. McIntyre, the author of this report, has covered Canadian science and technology in general and energy in particular for the last twenty-two years. He wrote the lead article in the October, 1975 issue of SCIENTIFIC AMERICAN on "Candu," Canada's heavy-water reactor. He is presently a technical writer for Ontario Waste Management Corporation.

Art direction, supervision of text and illustration and coordination of production were provided by Camp Associates Advertising Limited, Toronto.



Ottawa, Ontario/Canada

Experience wears a white hat.

To the people of Ontario, the white Ontario Hydro hard hat is a familiar, everyday reminder of a plentiful supply of dependable electricity delivered right to the doorstep . . . at a cost that's traditionally been amongst the lowest in the world.

It is worn in the control rooms of Canada's largest generating stations, by line crews stringing high voltage lines through miles of wilderness terrain, as well as the engineers working at our numerous construction sites. It is worn by our scientists looking into tomorrow and the environmental engineer safeguarding today.

In so many ways, it symbolizes the experience that has made Ontario Hydro an acknowledged international leader in utility management.

Now, something new has been added . . . **Ontario Hydro's New Business Ventures Division.** Their mission . . . to offer training, technology, skills and expertise to sister utilities around the world . . . to seek opportunities to help others solve energy problems . . . to share the innovative experience and know-how that travels with the familiar white hat.

For more information on all facets of utility operations write to: Mr. John H. Walker, International Marketing, New Business Ventures Division,
Ontario Hydro, 700 University Avenue,
Toronto, Ontario, Canada, M5G 1X6.
Telex: 06-217786 Telephone: (416) 592-2939.

We do a lot more than make electricity.

ontario hydro

Have you Heard?

Dineh, Allan Houser

Where ancient tradition
and modern expression
share one roof.

Where the vital Native American spirit
breathes and inspires,
evolves and expands.

Where over 75,000 Indian artifacts
have come to live,
and contemporary artists and cultural programs
live on.

Where life becomes culture
becomes art becomes life.

Have you Heard?

The Heard Museum
Phoenix

Anesthesiology

Rendering the patient insensitive to pain is still an essential part of the anesthesiologist's work; another major function is to maintain the stability of the patient's vital-organ systems

by Peter M. Winter and John N. Miller

Less than 140 years ago, before the development of anesthesiology, a surgeon was largely limited to performing amputations, setting fractures and repairing superficial wounds. The primary characteristic of a good surgeon was speed: a leg could be amputated in as little as 25 seconds. No successful way of reducing the agony of the patient was available, and surgical treatment of diseases within the abdomen, the chest or the skull was considered virtually impossible.

The groundwork for the introduction of effective anesthetics began to be laid in the last quarter of the 18th century. Joseph Priestley and Antoine Lavoisier discovered oxygen in the 1770's. Humphrey Davy in 1808 described his studies of nitrous oxide ("laughing gas") and mentioned its possibilities as an aid to surgery. Ether and chloroform were known to medical practitioners by 1831, but strangely the first efforts to use them as anesthetics were made by dentists. Indeed, it was a dentist, W. T. G. Morton, who demonstrated the surgical effectiveness of a general anesthetic. In 1846 he administered ether to Gilbert Abbott at the Massachusetts General Hospital so that the surgeon, John Collins Warren, could remove a tumor from the unconscious patient's neck.

Thereafter the transformation of surgery and the improvement of anesthesiology were rapid and profound. Today, although making the patient insensible to pain during surgery is still of obvious importance, it no longer constitutes the major part of the anesthesiologist's task. The major activity consists in keeping the functions of the vital-organ systems in equilibrium, counteracting the disequilibrating effects of the patient's disease, the surgery and the anesthetic agents. The anesthesiologist does this by what is termed titrated management: the continuous or frequently repeated measurement and adjustment of heart rate, blood pressure, respiratory rate and volume, blood acidity, levels of oxygen and carbon dioxide in the blood, temperature, kidney function and other signs that reflect the function of the patient's vital organs.

This process can probably be described best by an example such as that of an otherwise healthy middle-aged woman who is undergoing gallbladder surgery in an up-to-date U.S. hospital. The anesthetic procedure will be carried out by or under the direct supervision of an anesthesiologist—a physician with a minimum of three years of specialized training in anesthesiology after his or her graduation from medical school.

Because it has been demonstrated that patients who are adequately informed and instructed before surgery tend to have fewer postoperative complications and may have a shorter stay in the hospital, the anesthetic care begins the night before surgery. The anesthesiologist visits the patient to get information about her, the illness that requires surgery, other medical problems, the medications she may be taking and her understanding of the coming procedure and her anxieties about it. This is also the time for the patient to have questions answered about both the procedure and the postoperative course, to have potential complications explained to the extent that explanation seems desirable and to be instructed in the requirements for informed participation in postoperative care. During this time the anesthesiologist will also decide on the anesthetic technique that seems most appropriate for the type of surgery, the patient's condition and other medical problems affecting her status and the anesthesiologist's own preference based on experience.

On the day of the operation the anesthesiologist begins care by setting up in the operating room the monitoring devices needed to provide information about the function of vital-organ systems. The information must be delivered at a rate and a degree of accuracy that will allow moment-by-moment control of the patient's status.

At a minimum the monitoring devices would include a stethoscope, either placed on the patient's chest or inserted in her esophagus (after she has been anesthetized) to continuously monitor the functioning of her heart and lungs; a sphygmomanometer for measuring arterial blood pressure; a temperature probe in the esophagus; an electrical stimulator (to monitor muscle function) at either a wrist or the temple, and an electrocardiograph continuously displayed on an oscilloscope. In addition the anesthesiologist puts a small catheter in a peripheral vein (or in more than one vein) to gain direct access to the patient's circulation for the injection of medications, the maintenance of blood volume and electrolytes and (if it is necessary) blood transfusion.

If the case were more complex, the surgery more invasive or the patient in poorer health, the anesthesiologist might also elect to place a catheter in an artery to directly monitor arterial pressure and the function of the lungs, another catheter in the right atrium of the heart or in the pulmonary artery (which carries blood from the right ventricle of the heart to the lungs) and

GENERAL ANESTHESIA in surgery was first demonstrated at the Massachusetts General Hospital in 1846. This painting was made many years later by Robert Hinckley; it is in the Countway Library of Medicine in Boston. The patient, Gilbert Abbott, was having a neck tumor removed. Nearby with a glass container holding an ether-soaked sponge is W. T. G. Morton, a dentist, who administered the anesthetic. The surgeon is John Collins Warren.

electroencephalograph leads on the scalp (to monitor the activity of the brain). Signals from these devices are processed and displayed in digital or analogue form, either for immediate use or for keeping track of trends in the patient's physiology.

A standard preliminary step preceding the induction of general anesthesia is to give the patient 100 percent oxygen (O_2) through a face mask for a short time. The oxygen displaces the nitrogen (N_2) that is normally in the lungs and increases (from about 30 seconds to two minutes) the length of time during which the patient can be left not breathing during the introduction of anesthesia.

In administering anesthesia the anesthesiologist will be drawing on two groups of drugs. The first group consists of agents that act rapidly but briefly. They serve to induce anesthesia and to facilitate the establishment of control over the patient's airway. They are put directly into the venous circulation, a procedure that partially accounts for their rapid onset. They could maintain anesthesia, but for several reasons—the main one being that the drugs tend to have toxic effects if the doses are repeated frequently—they are less useful for the purpose than the drugs in the second group.

The drugs of the second group serve to maintain the anesthetic state. They include all the inhalation anesthetics (either gaseous or vaporized) as well as longer-acting muscle relaxants, narcotics and a few other agents. In general these drugs take effect less rapidly, and they must be delivered in such a way that they come into action as the induction agents are wearing off.

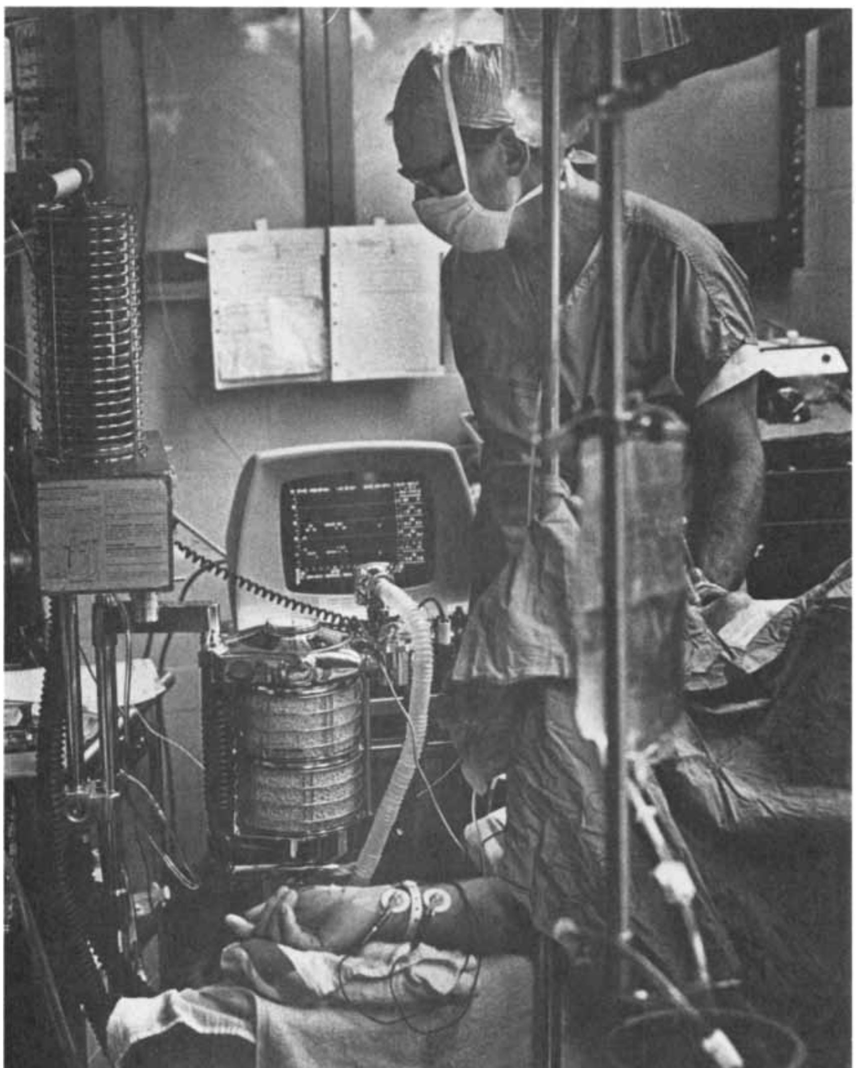
In drawing on the first group of drugs for the gallbladder patient the anesthesiologist typically proceeds through four stages. First, he administers a small dose of a muscle-paralyzing agent such as curare. Its purpose is to prevent the painful muscle contractions that would otherwise be induced by the paralyzing dose of succinylcholine that will soon follow.

Second, the anesthesiologist delivers intravenously a small dose of a barbiturate (such as sodium pentothal) that acts with great speed. The purpose is to ascertain whether a subsequent anesthetizing dose will unduly depress blood pressure and consciousness. The anesthetizing dose of sodium pentothal (about three milligrams per kilogram of body weight) is the third step. Because the drug is rapidly redistributed throughout the body, it causes anesthesia for only a few minutes.

Following the barbiturate is the paralyzing dose of succinylcholine. It acts for only a few minutes, in which time it relaxes all the skeletal muscles. This

profound relaxation is required so that the anesthesiologist can employ instruments to see the trachea (windpipe), making it possible to position a tube through which breathing can be controlled during surgery. The step is necessary because the muscle relaxants and inhalation anesthetics the patient will subsequently receive make spontaneous breathing impossible. To complete the installation the anesthesiologist inflates an elastic cuff at the lower end of the tube to ensure an adequate seal between the trachea and the tube. The seal keeps the contents of the stomach from reaching the lungs (a potentially disastrous event) and enables the anesthesiologist to control the patient's breathing by intermittently inflating the lungs with positive pressure.

With these four steps completed the anesthesiologist is in a position to begin the pharmacologic interventions that lead to surgical anesthesia. He connects the tube in the patient's trachea to the anesthesia machine, which usually delivers a mixture of nitrous oxide (N_2O) and oxygen (O_2). Since both anesthesia and surgery impede the function of the lungs, the concentration of oxygen in the mixture is seldom less than 30 percent—an increase of 50 percent over the amount of oxygen in normal air. In this case, because the surgery is in the upper abdomen, we in our practice would set the oxygen level at 50 percent to provide an extra margin of safety against difficulties of respiration that might arise.



EQUIPMENT FOR ANESTHESIOLOGY includes the devices shown in this photograph made at the medical center of the University of California at San Francisco. The cylindrical machine at the left, carrying a set of printed instructions, is the ventilator that maintains the anesthetized patient's breathing. The display screen at the right of the ventilator is the output from a mass spectrometer; it gives information on the gases, including the anesthetic, in the patient's lungs. The cylindrical apparatus below the display screen is a soda-lime canister, which absorbs carbon dioxide the patient exhales through the ventilator. The two circular leads on the patient's wrist stimulate peripheral nerves to determine the effect of muscle-paralyzing drugs such as curare. The anesthesiologist is John W. Severinghaus.

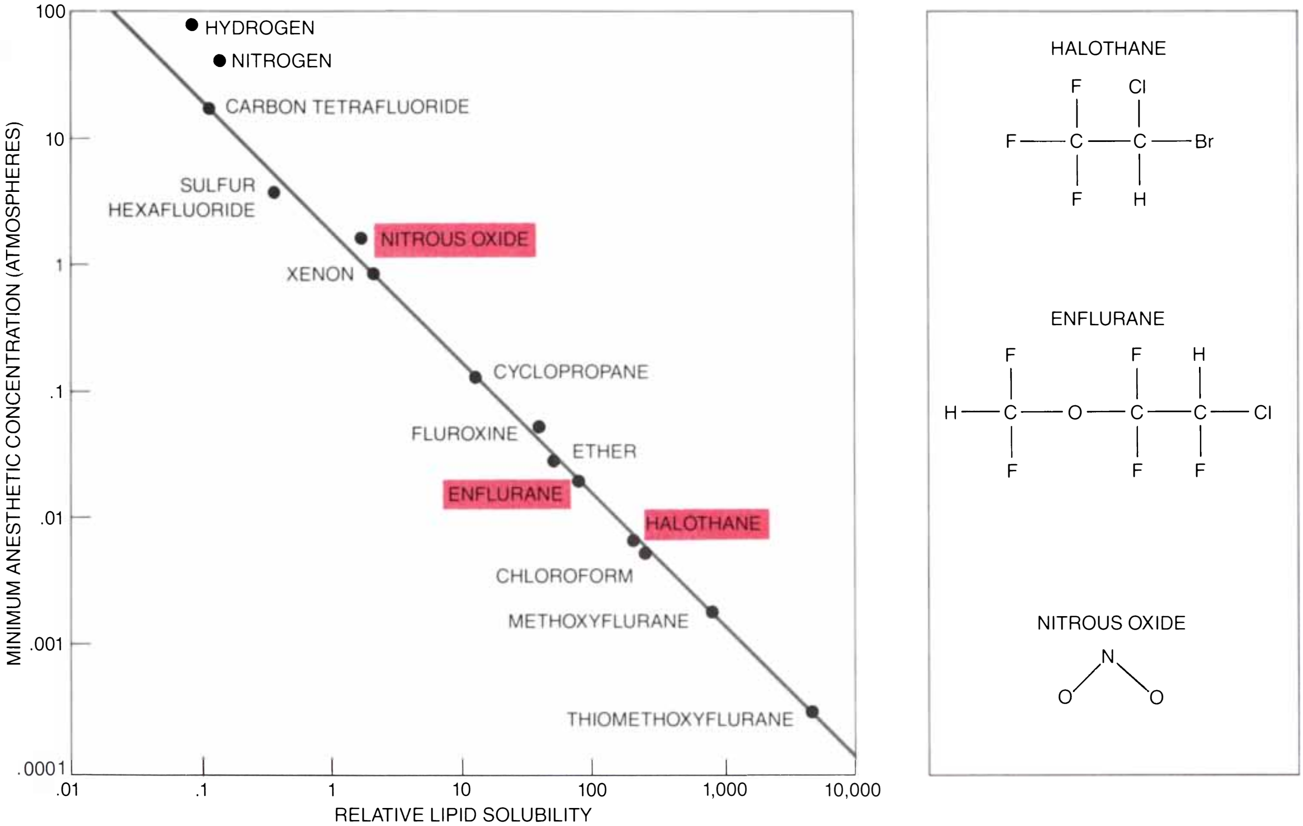
As the surgeons of the 19th century learned to their dismay, nitrous oxide alone is not sufficiently potent to produce adequate surgical anesthesia. Indeed, at a pressure of one atmosphere a mixture that is 50 percent nitrous oxide approximates only 50 percent of the total amount of anesthetic required. Therefore it is necessary to supplement this gas with a more potent agent. The common additives include halothane, enflurane and isoflurane, all organic molecules containing the halogens bromine, chlorine and fluorine. They have significant effects on the cardiovascular and pulmonary systems, decreasing cardiac output, blood pressure, the resistance of peripheral blood vessels and the volume of ventilation. Hence such anesthetics must be delivered with great precision. Variations as small as a fraction of 1 percent can cause potentially lethal disturbances of the function of vital organs. In patients with significant heart disease or other indications large doses of narcotics may be substituted.

Next the patient receives a paralyzing dose of curare or a similar drug to create enough relaxation of the abdominal muscles so that the surgeon will have good access to the surgical field. Since the respiratory muscles are also paralyzed, the patient's breathing is taken over completely, usually by connecting the respiratory circuit to a mechanical ventilator.

The site of action of these paralyzing agents is the synapse: the junction between nerve and muscle. In normal circumstances an electrochemical impulse sweeps down the motor nerve to the synapse. There it causes the release of the neurotransmitter acetylcholine, which crosses the synaptic cleft and stimulates muscular contraction. Succinylcholine and curare, although they differ in their mechanisms of operation, paralyze by interfering with the action of acetylcholine.

Once the patient is anesthetized the surgeon can start operating. Throughout the course of the surgery the anesthesiologist "fine tunes" the patient and her responses to the disequilibrating effects of the anesthesia and the surgery on vital systems.

Toward the end of the surgical procedure the anesthesiologist must prepare to reverse or eliminate the effects of virtually all the drugs he has used. The process is as complex and hazardous as the induction of anesthesia was. Paralyzed muscles must be returned to a state in which the patient can breathe



POTENCY OF ANESTHETICS correlates with their solubility in lipids. A standard measure of potency is the minimum anesthetic concentration required to abolish movement in response to a painful stimulus. The linear correlation between lipid solubility and anesthetic potency implies that the site of action of an anesthetic is at the lipid membrane of the cell. Three of these substances (*color*) are regularly used as anesthetics. Hydrogen and nitrogen deviate from the linearity shown by the other substances because they are agents so weak that in order to cause anesthesia under experimental conditions they must be given at high pressure, and the pressure in itself partially counteracts the anesthetic effect. The chemical structure of the three commonly employed anesthetics is represented at the right; their constituent atoms are fluorine (*F*), chlorine (*Cl*), carbon (*C*), bromine (*Br*), hydrogen (*H*), nitrogen (*N*) and oxygen (*O*).

for herself. The various agents employed to produce and maintain the absence of both pain and consciousness must be eliminated at the correct rate and in the correct order. Blood volume and composition must be checked to ensure that they are normal. Finally, the patient must be awakened enough to respond to the ministrations of the specially trained nurses in the recovery room.

Part of this procedure entails administering drugs that counteract the anesthetic agents. For example, the action of muscle relaxants is counteracted by drugs such as neostigmine that reestablish the normal function of acetylcholine at the synapse between nerve and muscle. In other instances the anesthesiologist merely has to be certain that natural processes are taking over—for example, that the inhalation anesthetics are removed from the circulation by respiration.

Gallbladder surgery is a fairly simple procedure. The same principles apply in much more serious situations: an elderly person having major blood vessels removed and replaced, a newborn child with congenital heart disease (for which the treatment requires stopping the heart), a patient receiving a number of medications for unrelated diseases, a person with multiple injuries or someone in whom many organ systems are severely impaired. Perhaps the most extreme example is a patient receiving a liver transplant. Such a patient is near death from liver failure and has many physiological derangements affecting almost all the organ systems. The surgery can last for 24 hours or more, and the transfusion of 50 gallons of blood (35 times the normal blood volume) is not unheard of.

The fact that the anesthesiologist can maintain the patient's physiological equilibrium in the face of several simultaneous onslaughts indicates that he provides a qualitatively different kind of medical care. Most medicine is practiced by prescription: the physician makes a diagnosis, prescribes a treatment and reexamines the patient after an interval of time (usually days) to ascertain how well the therapy is working. The anesthesiologist practices by titration: he continuously measures and adjusts the functions of interdependent systems of vital organs. He substitutes expertise for the homeostatic mechanisms that ordinarily maintain this delicate balance.

Such an approach depends entirely on accurate measurement of how each organ system is functioning. This fact largely accounts for the dependence of modern anesthesiology and surgery on high-technology instrumentation. It also helps to explain why increasingly radical surgical procedures are increasingly expensive. Such care is both personnel- and technology-intensive. This capacity to titrate pharmacology against the pathology of illness has been described as toxicological brinkmanship. Remarkably, it is done in hospitals throughout the U.S. with a mortality rate of approximately one patient in 10,000.

Although the anesthetic procedure we have described for the woman in gallbladder surgery is the commonest method, it is not the only one. Excellent surgical analgesia can also be achieved by a number of regional techniques. They entail blocking, by the administration of local anesthet-

Software that

Summary:

GTE computer scientists are producing software to help develop a variety of things, from telephone networks to integrated circuits—software that writes other software; software that designs microcircuits; even software that has its own intelligence.

Modern telephone systems are essentially special-purpose computers. Their software represents more than half of their cost, and this figure is rising.

GTE research is aiming to improve both productivity in software design and its quality. Two different compiling systems are helping us (compilers aren't new, of course, but these are special).

New-generation software compiler.

Our General-Purpose Compiling System (GPCS) does not work with only one language and computer as ordinary compilers do. GPCS was designed to be independent of host and target machine, and support Pascal, Ada and CHILL. It was also designed to permit automatic language translation between these supported languages.

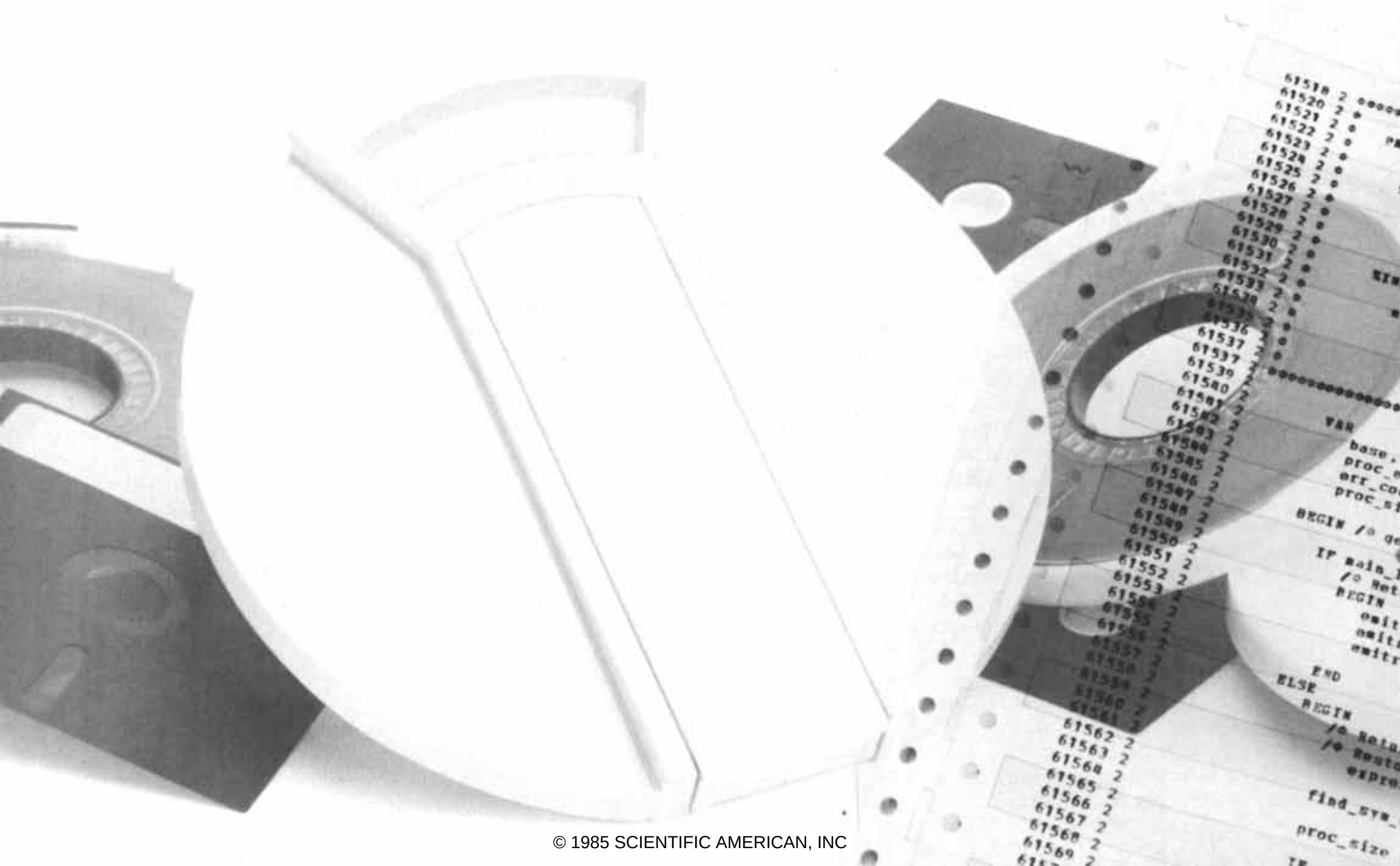
Our researchers have completed work on GPCS, and it is now in our software development facilities. There, it will write software for new switching systems at increased productivity levels.

Designing hardware with software.

The demand for more and more function has inevitably led to higher costs for designing very large-scale integrated circuits. Designing a 32-bit microprocessor with today's tools, for instance, can take over 100 man-years. Our silicon compiler project promises to reduce this time by 90 to 95 percent.

Using this design-automation tool, designers will describe what the circuit should do functionally, rather than graphically. The tool will translate (compile) requests into appropriate geometry without the laborious circuit layout and routing formerly required.

It is interesting to note that this project may result in the use of custom-logic circuits where microprocessors are used today. After all, when VLSI circuits become cheap and easy enough to produce, it may be preferable to integrate them into systems rather than write software for a microprocessor.



creates software.

Simplifying database access.

With the advent of communications networks like GTE Telenet, many databases have become accessible at relatively low cost. Each, however, may have its own access control mechanism, command language, and output format; and users may want to switch from one database to another with a transparent interface.

We have been working on a way for data-network subscribers to use natural language, which is translated into the various formats required by the different database services. It is called FRED—Front End for Databases. This technology incorporates natural-language processing and the expert systems areas of Artificial Intelligence.

FRED acts like a librarian. It recognizes the meaning of natural-language input and sends a request to the appropriate database.

Not all our computer science projects have the drama of artificial intelligence research, of course. However, our goal in all these investigations is to create software to help improve quality and productivity for advanced communications products and services.

The box at the right lists some of the pertinent papers GTE people have published on software and related subjects. For any of these you are invited to write GTE Marketing Services Center, Department TPIIIA, 70 Empire Drive, West Seneca, NY 14224 or call 1-800-828-7280 (in N.Y. State 1-800-462-1075).

Pertinent Papers.

The System Compiler, 1983 IEEE International Symposium on Circuits.

The MacPitts Silicon Compiler: A View from the Telecommunications Industry, VLSI Design, May-June, 1983.

An Intelligent Communication Assistant for Databases, Proceedings of the IEEE COMPSAC 83.

A Natural Language Interface for Medical Information Retrieval, AAMSI Congress, Computer Applications in Medicine.

Separate Compilation for Block Structured Languages: A Comparative Study of CHILL and Ada Compiling Systems, Proceedings of the IEEE Symposium on Application and Assessment of Automated Tools for Software Development.

An Efficient Compilation Strategy for Very Large Programs, ACM SIGPLAN 82 Symposium on Compiler Construction, June, 1982.

GTE

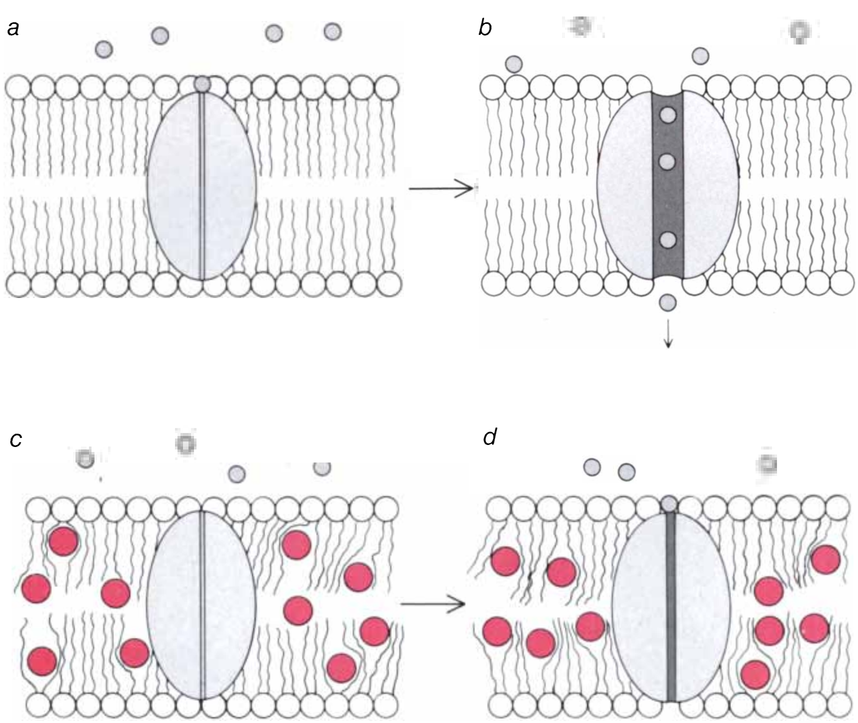
ics such as lidocaine and tetracaine, the propagation of neural impulses in structures distant from the brain. For example, the physician can render a fractured wrist insensitive to pain by injecting a local anesthetic near the nerves as they pass through the neck or the axilla (armpit).

Similarly, local anesthetics injected near the spinal cord by any of several approaches will make structures beyond this point insensitive to pain. The technique has revolutionized the anesthetic management of labor and childbirth. Until some 15 years ago labor was largely characterized by uncontrollable or poorly controlled pain. The reason was that such drugs as inhalation anesthetics and narcotics, which would be useful in reducing maternal pain, pass through the placenta and depress vital functions in the fetus. Regional analgesia provides excellent relief of abdominal and pelvic pain. The mother remains awake, able to take part in and to enjoy the process of labor and birth. The local anesthetics used for such purposes are relatively nontoxic; only small amounts reach the fetus, whose vital functions at birth are therefore not disturbed.

Experience in the management of pain has led anesthesiologists into several other areas of practice and research having little to do with surgical patients. They include the management of chronic pain, critical-care medicine, investigation of how anesthetics produce anesthesia and the technology of deep-sea diving.

Chronic pain severe enough to interfere with productivity and enjoyment of life afflicts an estimated 35 million people in the U.S. The cases vary widely, ranging from intractable headaches to the phantom pain that seems to come from an amputated limb. Chronic pain appears as a large group of syndromes, best managed by a team of specialists including anesthesiologists, neurologists, neurosurgeons, internists, psychiatrists and others.

Another subspecialty of anesthesiology is critical-care medicine. Continual titrated management is the key to survival for patients who are critically ill because of the simultaneous failure of several vital-organ systems. For example, a patient suffering from acute, reversible neuromuscular paralysis or multiple injuries can be cared for in much the same way as an anesthetized patient, except that the intensive titrated management will continue for days or weeks. Moreover, some forms of surgery such as cardiac and transplantation procedures have a major impact on the homeostasis of vital-organ systems and create the need for long-term intensive care. Anesthesiologists usually work with other specialists as part of a team caring for such patients.



ANESTHESIA may arise from the action of the anesthetic in the lipid membrane of the cell, blocking the transmission of nerve impulses. In a resting cell (*a*) the sodium ions (*light gray*) involved in the transmission of impulses are largely outside the cell. On the arrival of an electrochemical signal a channel opens in a membrane protein (*b*), allowing sodium ions to pass through to the interior of the cell. When molecules of an inhaled anesthetic (*color*) dissolve in the membrane (*c*), the channel does not open on the arrival of an impulse (*d*) and the sodium ions are blocked. The membrane may change in fluidity or volume or both.

Notwithstanding the fact that anesthetics are administered to millions of patients each year in the U.S., the pharmacologic mechanism producing unconsciousness and analgesia is poorly understood. A better understanding would contribute to the design of improved anesthetics and would yield insights on the action of such heavily abused drugs as alcohol and barbiturates.

How do anesthetics work? Morphine, codeine, heroin and other drugs with anesthetic properties have specific molecular structures that fit receptor sites on the surface of cells, much as a key fits into a lock. This cannot be true of inhalation anesthetics. Their molecular structure varies widely, from noble gases such as nitrogen and argon to complex halogenated hydrocarbons such as halothane and enflurane.

There are some clues to how these anesthetics act. It has been known since the 19th century that the potency of an anesthetic can be predicted with great accuracy on the basis of the agent's solubility in lipids: the greater the solubility, the more potent the anesthetic and the lower the concentration required to produce anesthesia. Although this relation does not explain how anesthetics act, it strongly suggests they exert their effect on the lipid membrane of neuronal cells. The currently accepted hypothesis is that the anesthetic dissolves in the lipids of the membrane and, by acting either on the lipids or on protein complexes embedded in them, changes the fluidity or the volume of the membrane.

This hypothesis gains a degree of support from the remarkable phenomenon known as pressure antagonism. High pressure, as our laboratory and others have demonstrated, nullifies the effect of an anesthetic. The action is linear and quantitatively predictable, regardless of the potency of the anesthetic involved. For example, if a mouse is anesthetized with halothane in a high-pressure chamber, it returns to the unanesthetized state if the pressure is raised by 50 atmospheres (about 735 pounds per square inch). The administration of an additional amount of anesthetic can be antagonized by a highly predictable increment of pressure. No other known pharmacologic or physical manipulation can accomplish such antagonism.

Pressure antagonism has been demonstrated for all inhalation anesthetics and for other gases that are not generally regarded as anesthetics. One of them is nitrogen, which constitutes 79 percent of normal air. The solubility of nitrogen in lipids predicts that it would produce anesthesia at a pressure of 28 atmospheres. Indeed, the nitrogen narcosis ("rapture of the deep") experienced by divers is merely a small degree of anesthesia from the increased amount of nitrogen the diver breathes. A diver breathing air at a depth of 150 feet in seawater is receiving an amount of nitrogen equal to approximately 15 percent of the anesthetizing dose of nitrous oxide or halothane.

This effect materially interferes with deep commercial diving, which is why helium is substituted for nitrogen in a diver's air supply. Based on its solubility in lipids, helium should become anesthetic at a pressure of approximately 80 atmospheres. Actually, however, helium is an important exception to the correlation between anesthetic potency and solubility in lipids: it does not produce anesthesia at any pressure. Instead it brings on the antithesis of anesthesia, namely excitation of the central nervous system, manifested by tremors that progress to epileptic seizures. It is important to note that both the anesthetic antagonism and the nervous excitation are produced by the pressure rather than the helium.

In sum, anesthesia is created by anesthetics of varying potency. The potency of a given agent is determined by its lipid solubility. For a relatively potent agent such as halothane only 1 percent of one atmosphere is required to produce the anesthetic state. Nitrous oxide, which is less soluble in lipids, requires about one atmosphere to produce anesthesia. Nitrogen is still lower on the scale of lipid solubility: its solubility predicts a dose of 28 atmospheres of nitrogen to create anesthesia.

On the basis of these findings a hypothesis can be advanced relating a continuum of behavior to the critical volume of a lipid membrane. (The concept of volume is an oversimplification of much more complex changes taking place in the membrane.) The normal dimensions of the membrane equate with normal consciousness. Compression of the membrane by high environmental pressure reduces the volume of the membrane—a condition that correlates with excitation of the central nervous system leading to seizure. If molecules of anesthetic are dissolved in the membrane, the result is anesthesia from the expanded volume of the membrane. The more soluble the anesthetic, the smaller the amount required. If pressure is then applied, the volume of the membrane is restored to normal, whereupon normal behavior returns, notwithstanding the continued presence of the anesthetic.

MORTON'S EQUIPMENT for administering the first surgical anesthetic was this glass container. It held a sponge soaked in ether. Abbott inhaled through the mouthpiece at the right, making air pass over the sponge and pick up ether vapor. Fortunately Morton's anesthetic has a wide safety margin, so that Abbott survived in spite of the primitive equipment.

This understanding has important implications for activities far afield from patient care: deep-sea diving and offshore-oil development. We stated above that commercial divers avoid the anesthesia of nitrogen narcosis by breathing mixtures of helium and oxygen. As depth increases, divers encounter excitation of the central nervous system. This imposes a limit on the depth at which productive work can be done, particularly in offshore-oil development.

At the Duke University Medical Center, Peter B. Bennett and one of us (Miller), together with our colleagues, have addressed this problem in a recent experiment. Volunteers spent four days in a high-pressure chamber complex at pressures of up to the equivalent of 2,500 feet of seawater (69 atmospheres). Calculations had indicated that the addition of 25 percent of an anesthetic dose would return membrane volume to normal. The safest "anesthetic" to use in this case is nitrogen. Accordingly the breathing mixture consisted of helium, nitrogen and oxygen in proportions of 89.3, 10 and .7 percent respectively. The volunteers were essentially unaffected by the excitation that would have appeared with standard helium mixtures.

The problem is not solved, however; the density of the highly compressed gas increases the mechanical work of breathing and may limit the depth to which divers could descend. About half of the density of the breathing mixture was due to the nitrogen. The depth limitation imposed by the dense gas could be minimized by substituting a more potent anesthetic such as nitrous oxide (in much smaller quantities) to achieve the antipressure effect.

A further aspect of research on anesthesia is of paramount importance. In experiments done in association with Edmond I. Eger II and Raymond Smith we have demonstrated that inhalation anesthetics cause pharmacologic effects very similar to those of alcohol and such sedatives as barbiturates. In experimental animals they give rise to the development of tolerance and withdrawal syndromes that are similar in both quantity and kind to those seen with such drugs. Moreover, we have shown that a tolerance to alcohol confers a cross-tolerance to anesthetics. These findings provide strong circumstantial evidence suggesting a similar mechanism or site of action (perhaps both) for the two groups of drugs.

Alcohol and other sedative hypnotics are the most abused drugs known, with incalculable costs to both the addicted and society. It is difficult in research on alcohol to correlate aspects of behavior with the concentration of the drug. This problem is largely overcome by the use of inhalation anesthetics as substitutes. We foresee that anesthetics will become convenient probes for the study of these societally important problems.

The Parsons Steam Turbine

Invented in 1884 by Charles A. Parsons, son of an aristocrat and scientist, it revolutionized shipping and the generation of electric power. Many of Parsons' innovations are still in use

by W. Garrett Scaife

By the late 1800's the Industrial Revolution had reached an important turning point. During the preceding century and a half steam engines had made it possible to employ the chemical energy of various fuels in a great variety of mechanical devices. In one of the steam engine's most significant uses it had been linked to another technological development, the electric dynamo, making a large flow of electrical power available. As the demand for power increased so did the size of steam engines, until they began to push their mechanical limits. If industrial development was to continue, a new source of power was necessary.

In 1884 Charles Algernon Parsons provided that source by developing the world's first practical turbogenerator. Ten years later he turned his attention to the task of applying his invention to transportation; several years of intense development were crowned by success when his turbine-powered ship *Turbinia* reached a speed of 35 knots. This was appreciably faster than the top speed of any ship in the Royal Navy at that time. Turbines are more compact than piston-driven, or reciprocating, steam engines, and they are simpler. Therefore as turbines grew in size and efficiency they replaced the older kind of engine. Today steam turbines are the world's largest source of electric power, although their near-monopoly of passenger-ship propulsion during the first half of this century has been ended by the advance of the diesel engine. The modern steam turbine still embodies features that were first developed by Parsons in 1884.

The two principles underlying the steam turbine, reaction and deflection, are themselves quite old. In about 130 B.C. Hero of Alexandria described a machine called an aeolipile. This hollow sphere, filled with steam, had two hollow arms on opposite sides bent so that they pointed along opposite tangents of the sphere. Expanding steam escaped in jets from the two arms and set the sphere spinning by reaction as the steam accelerated.

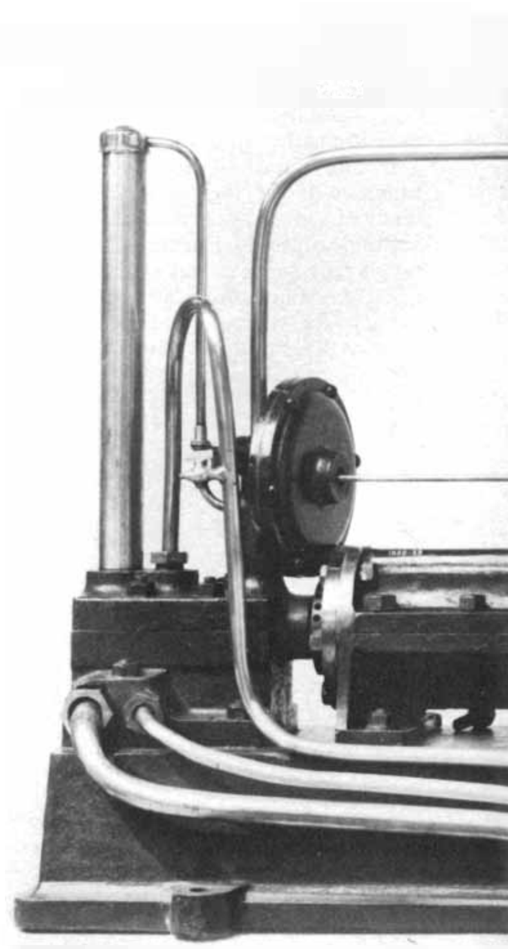
The other principle employed in turbines, deflection, is demonstrated by a machine proposed by Giovanni Branca in 1629. In Branca's machine a jet of steam spun a wheel fitted with blades shaped like those of a water wheel.

Steam turbines embody both of these effects. In a turbine a jet of high-pressure steam is directed against a set of fanlike blades. When it strikes the blades, the flow of steam is deflected; deflection of the flow spins the blades. The steam expands and accelerates as it flows through passages between the blades, converting the energy locked up in high pressure into kinetic energy.

Early turbines, such as Branca's, would not have been able to produce any useful power because the steam boilers of the time could not sustain significant amounts of pressure. The first successful steam engines, developed by Thomas Savery, Thomas Newcomen and others, did not require high-pressure steam. Instead they used low-pressure steam to displace the air under a piston. The steam was then condensed, creating a vacuum under the piston, and the weight of the atmosphere acted on the piston to perform useful work. Experience in building and using boilers for these so-called atmospheric engines eventually enabled engineers to design boilers capable of sustaining and withstanding pressures well above the pressure of the atmosphere.

The availability of high-pressure steam led inventors to think again of the turbine. Designers tried various approaches. In 1815 the locomotive engineer Richard Trevithick tried attaching two jets to the rim of the wheel of a railroad engine and supplying them with steam from the boiler. Trevithick's device failed. A sawmill driven by a similar reaction device was built in 1837 by William Avery at Syracuse, N.Y. In England alone more than 200 patents relating to turbines were issued in the century preceding 1884, half of them in the 20 years after 1864.

None of these efforts produced a practical machine. In part this was owing to a failure to understand the phys-



PARSONS TURBOGENERATOR built in 1884 was the first practical steam turbine. High-pressure steam entered through the rectangular opening visible near the turbine shaft's midpoint. There it was split into two

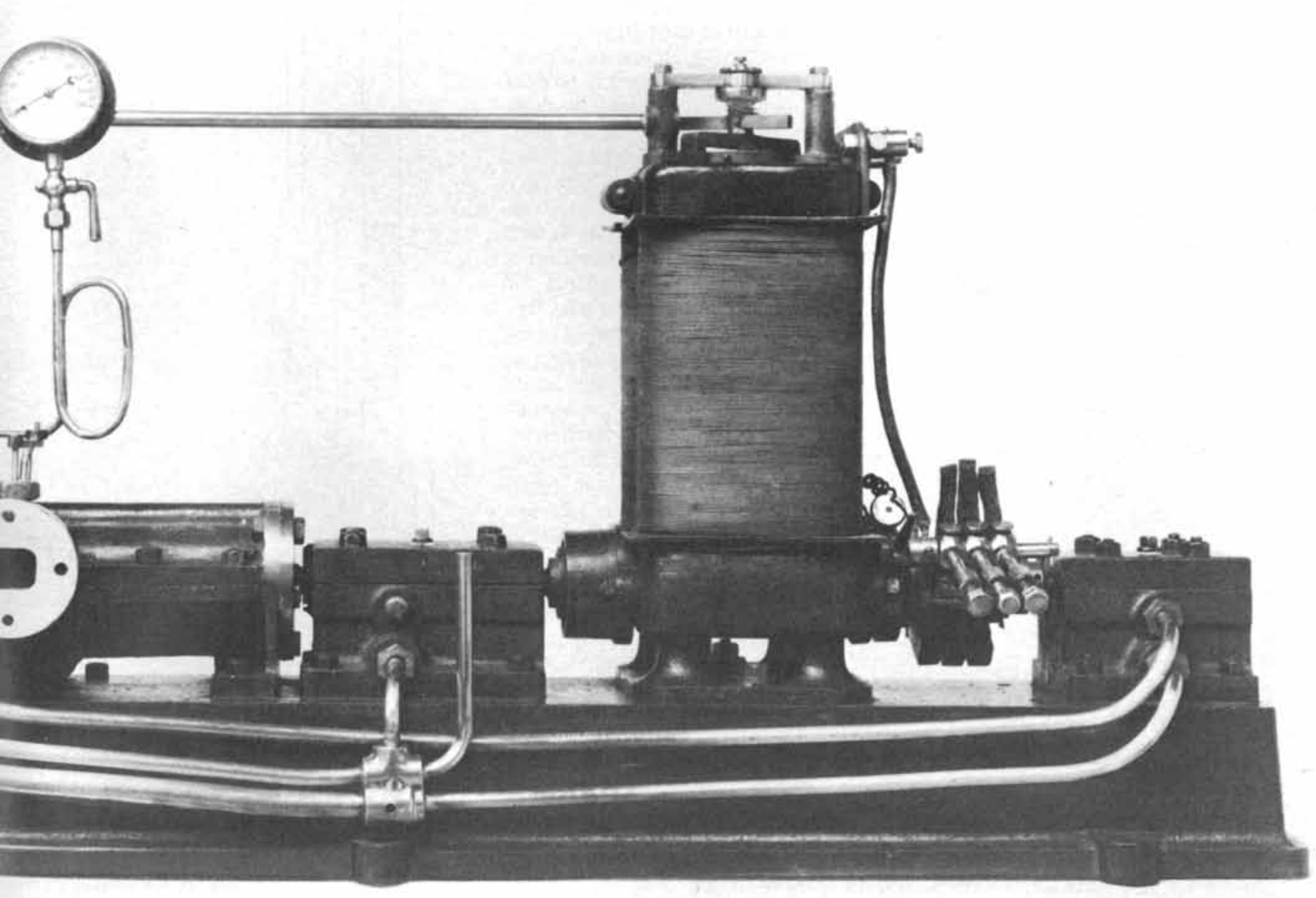
ical laws governing the expansion of steam. Steam is much less dense than water and much more elastic; consequently the fluid velocities in steam turbines are far higher than they were in the water turbines with which these designers were familiar. It had been determined that the blade speed needed to be at least half that of the fluid if a turbine was to operate at reasonable efficiency; early turbines therefore had extremely high rotational speeds.

High rates of rotation caused several difficulties, not the least of which was that the rapidly spinning machines tended to fly apart. Rotational velocities could have been lower if turbine blades had been mounted on wheels with large diameters. This was not possible, however. The steam throughput of the early machines was necessarily small, and so the cross-sectional area through which the steam passed also had to be small. The first experimental turbines therefore had small diameters and short blades.

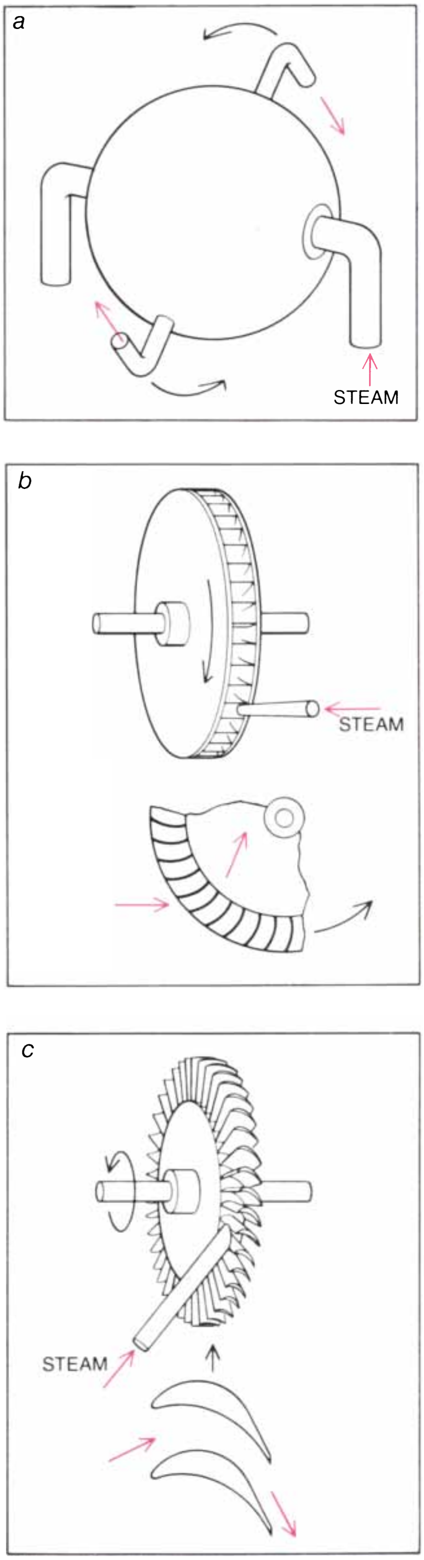
Another troublesome property of steam is subtler. When a compressible fluid such as steam is forced through a nozzle, the speed of the flow increases in proportion to the ratio of inlet pressure to outlet pressure. When the ratio reaches about two to one, the flow reaches a peak; further increases in the pressure ratio effect no further increase in fluid velocity. Early designers were thus unable to take full advantage of high-pressure steam, because there was a limit to the amount of pressure energy that could be converted into kinetic energy and then passed on to the blades. In 1889 the Swedish engineer Carl Gustaf de Laval employed a nozzle fitted with a trumpet-shaped mouth. This could develop much greater velocities; as a consequence the rotational speed of de Laval's machine was also very high.

Parsons developed a radically different approach to turbine design, one that enabled him to reduce the high rotational velocities while taking full advantage of the energy in high-pressure steam. He was able to sidestep the pressure problem by allowing the steam to expand gradually as it passed through 15 pairs of bladed turbine wheels. Each pair consisted of one wheel of "fixed" blades, attached to the inside of the turbine casing, and one of "moving" blades, fitted onto the rotating shaft [*see middle illustration on page 136*]. The fixed and the moving blades faced in opposite directions; that is, if each set of blades had been free to rotate, steam passing through the ensemble would have caused the two types of wheel to rotate in opposite directions.

As the steam passed through the fixed blades it expanded, speeding up the flow, and was redirected to strike the faces of the moving blades, spinning them by the force of deflection. Within the moving blades the steam expanded once again, emerging from them as an accelerated jet and



streams, which flowed toward each end of the shaft through bladed wheels. Expanding steam spun "moving" wheels, which were connected to the central shaft; these alternated with "fixed" wheels attached to the inside of the turbine casing. Fixed wheels directed the steam so that it would strike the faces of the moving blades. Steam expanded as it passed through each wheel. By allowing the steam to expand in a cascade of many stages, Parsons was able to take full advantage of the energy contained in high-pressure steam and still avoid the high rotational velocities necessary in other steam turbines. The spinning shaft powered a dynamo, or generator (*right*).



REACTION AND DEFLECTION are the two principles underlying operation of the steam turbine. Reaction is demonstrated by Hero's aeolipile (*a*), in which a steam-filled sphere is spun by jets of expanding steam that escape from two hollow arms. In (*b*) a wheel spins when blades deflect a jet of steam. A turbine's blades (*c*) also deflect the flow of steam; in addition the steam expands and accelerates as it passes between blades, propelling them by reaction force.

driving them by the force of reaction.

The cascade of bladed wheels made extremely high rotational velocity unnecessary. In each of the 30 wheels that made up Parsons' 15 stages the steam expanded by a small amount, giving up some of its energy. The pressure fell by only about 10 percent across each pair of fixed and moving blades, and so the maximum steam velocity was about a fifth of that of a single-stage turbine. Parsons reasoned that for such small pressure drops serious errors would not result if steam were considered to be an incompressible fluid, like water. He was thus able to estimate steam velocities, ideal efficiencies and appropriate blade shapes with quite good accuracy. Parsons' system of letting the steam expand in many gradual steps, which is a fundamental aspect of modern turbines, was only one of many innovations he introduced in his machine of 1884.

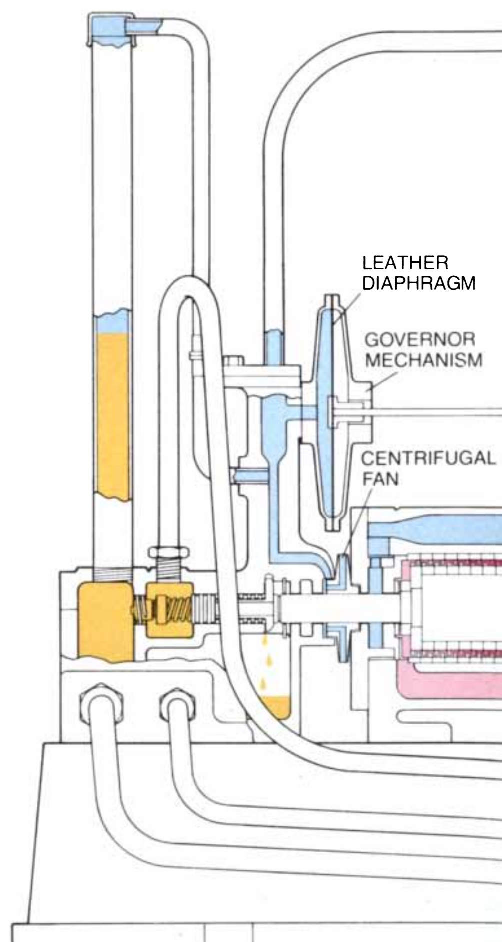
Another innovation was a new type of bearing designed specifically for shafts rotating at high rates. In spite of Parsons' success in reducing the speed of his turbine, he still required the shafts to turn at least 10 times faster than any existing engine or motor. This made it necessary for him to deal with the phenomenon known as shaft "whirling." The engineers of the period knew that each shaft had a characteristic critical speed, at which any slight imbalance would cause it to bow outward violently. It had been found that the critical rate of rotation corresponded to the natural frequency of lateral vibration of the shaft (the frequency at which the shaft would resonate when struck). Parsons and de Laval found independently that a shaft would run smoothly if it was speeded past the critical rate. Nevertheless, the shaft still deflected slightly owing to minor imbalances. It was therefore necessary, in order to prevent damage to the shaft, for it to run in bearings that would allow some lateral movement.

Parsons first tried resting an ordinary bearing on spring supports, but he found that such an arrangement only made vibration worse. In the end he devised a bearing supported by an assembly of washers. The washers were of two sizes; washers that were tight on the bearing shell (in which the shaft ran) but did not touch the bearing casing alternated with others that were tight in the casing and loose on the bearing shell. The entire assembly was compressed by a spring. Hence some lateral movement of the shaft was possible, but vibrations tended to be damped out by the large amount of friction between the two types of washers.

The arrangement was very successful, and those who saw the prototype turbine running at the 1885 Inventors Exhibition in London remarked on how smoothly it ran compared with contemporaneous steam engines. The latter were inevitably the source of a considerable amount of vibration, which was transmitted through their foundations for great distances.

In Parsons' prototype steam was fed through a governing valve to the center of the shaft. There the steam split, half flowing toward one end of the shaft and half toward the other. Each separate stream passed through a set of turbine wheels.

One advantage of splitting the flow was that any axial force produced by the steam when it pushed against the



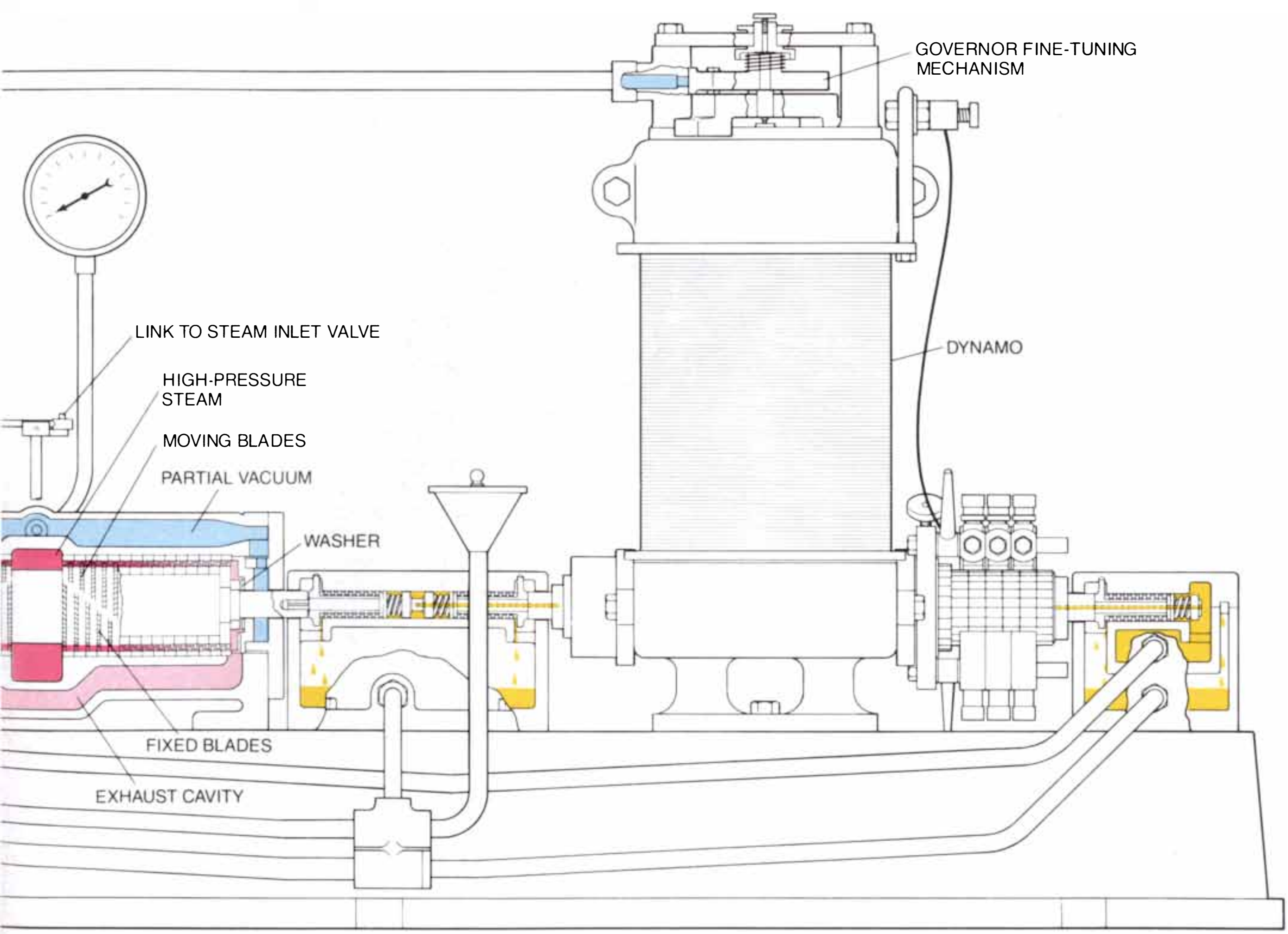
CROSS SECTION of Parsons turbine illustrates the flow of steam, air and oil in the 1884 prototype. High-pressure steam (*dark red*) enters at the center of the shaft and flows through turbine blades toward each end. Exhaust steam (*light red*) flows into two cavities joined by an outlet passage in the bottom of the casing. Farther out along the

blades was neatly balanced out; it was thus unnecessary to equip the shaft with a special thrust bearing. This configuration is still a feature of many modern steam turbines.

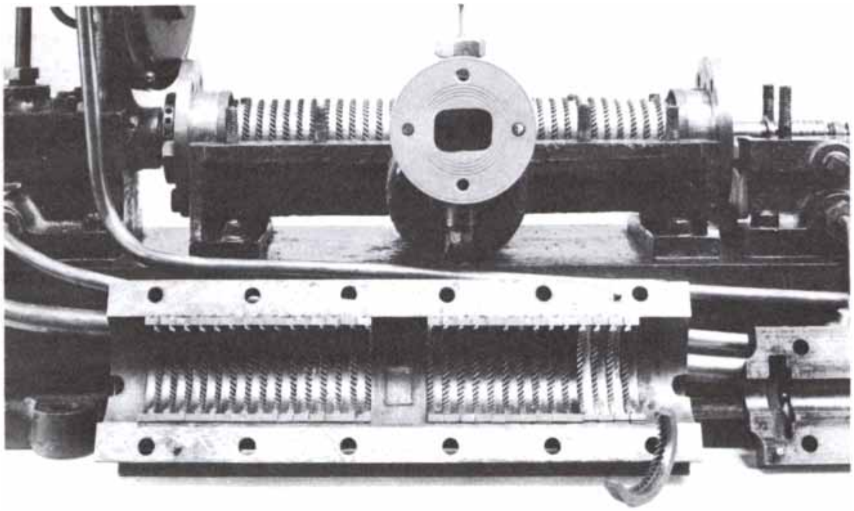
Even with Parsons' innovation of letting the steam expand in many stages, his first prototype ran at the high rate of 18,000 revolutions per minute. Consequently the centrifugal force on the rotor blades was nearly 13,000 times the force of gravity. To keep the wheels from flying apart, he employed a very simple design: each wheel was made from a single brass ring, and the blade passages consisted of 45-degree slots cut into the ring's circumference. The moving rings were keyed onto the steel shaft and locked against a central shoulder by screwed collars. The fixed rings were split in half, and each half was keyed onto either the top or the bottom of the turbine casing. To allow for the increase in the volume of the steam as it expanded, the blades increased in length, in three increments, from .2 inch to .28 inch. Each blade's leading edge was beveled to improve the flow of steam.

The turbine's high rate of rotation had other implications as well. None of the means for reducing speed (such as gear trains) available at that time were suitable for such high speeds. Consequently the simple flyball governors used on steam engines could not be used; they would have been torn apart by centrifugal forces. Instead Parsons developed a completely new type of governing device. A centrifugal fan, mounted on the turbine shaft, was connected to a sealed set of pipes that were filled with air. As the fan spun it evacuated air from the system of pipes, creating a suction effect. The suction effect acted on a leather diaphragm at one end of the system. The diaphragm was in turn linked by a shaft to a governor valve, which controlled the flow of steam into the turbine. When the turbine spun faster, the centrifugal fan created a greater suction on the diaphragm; the valve linked to the diaphragm then reduced the amount of steam going into the turbine and thereby slowed the turbine's rate of rotation.

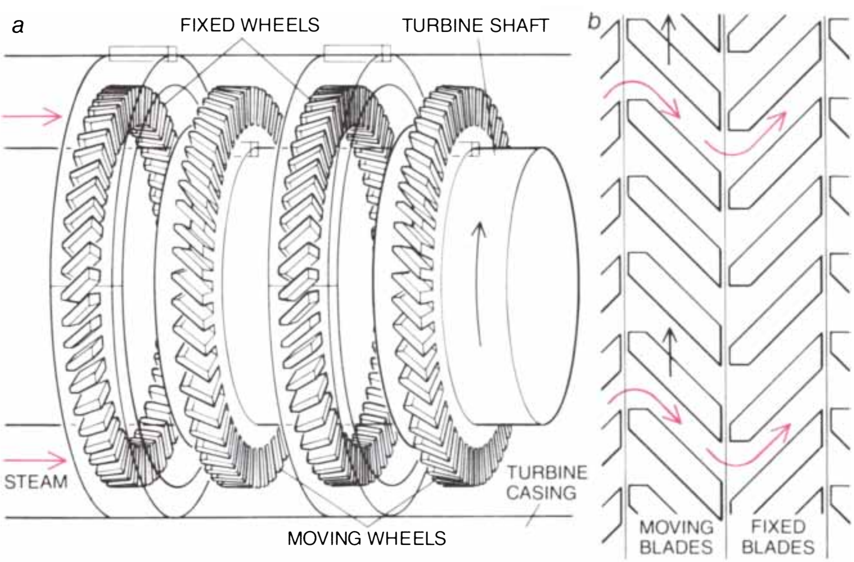
This governor worked well, but it was not very precise. Parsons' prototype turbine powered a dynamo (an electric generator). At the time that Parsons designed his first turbine incandescent lamps cost roughly the



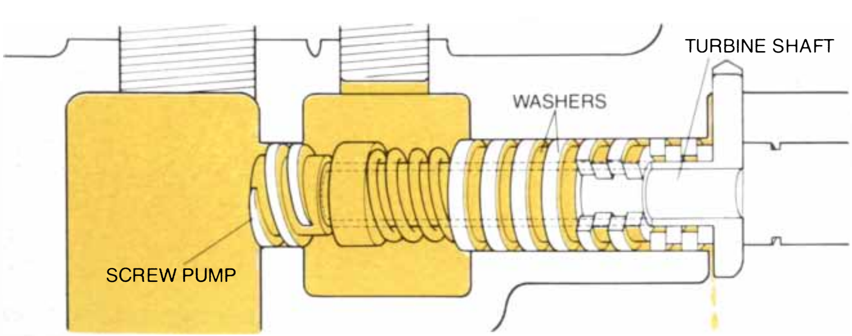
shaft are two other cavities, joined in the upper half of the casing, in which there is a partial vacuum (*blue*). Washers, held tight against the inside of the casing by suction between the exhaust chambers and the partial vacuum, prevent exhaust from escaping along the spinning shaft. Lubrication is accomplished by a screw (*far left*) that forces oil (*yellow*) under pressure into the shaft bearing. The oil flows through pipes to other bearings. It reaches the central bearings by way of the dynamo shaft, which is hollow (*center and far right*). The governor is controlled by a centrifugal fan (*left*), which creates a partial vacuum (*blue*) in a system of pipes. The vacuum causes suction to act on a leather diaphragm, which is in turn linked to the valve controlling the amount of steam let into the turbine. The governor is fine-tuned by a mechanism on top of the dynamo. This mechanism controls air flow into the sealed system of pipes in accordance with dynamo voltage. Suction in these pipes also draws oil that has returned from bearings into a standpipe reservoir (*far left*).



BLADED WHEELS were made out of solid brass rings cut with 45-degree slots. The moving wheels were keyed onto the turbine shaft, whereas the fixed wheels were cut in half and attached to the two halves of the turbine casing (*here the upper half is removed for display*).



ALTERNATING SEQUENCE of fixed and moving blades (*a*) channeled the flow of steam. Steam expanded and accelerated as it passed between fixed blades; they directed the accelerated flow onto moving blades. Steam expanded in the moving blades as well, propelling them by reaction and deflection. Detail (*b*) shows one of 15 pairs of fixed and moving blades.



SHAFT BEARING allowed some lateral movement of the shaft, but it damped vibrations. A spring compressed a sequence of washers in which those that fit tightly on the bearing shell (in which the shaft ran) but did not touch the casing alternated with others that were tight in the casing but loose on the shaft. The shaft required some freedom of movement so that minor imbalances would not cause it to be damaged; friction between the washers damped out vibration. A screw (*left*) fed oil (*yellow*) into the bearing from a reservoir.

price of a quarter-ton of coal; if they were not to be burned out by power surges (as frequently happened where steam engines were used), the voltage generated by the dynamo could not vary by more than about 1 or 2 percent. Parsons therefore equipped his governor with a fine-tuning mechanism that depended directly on the dynamo voltage.

The voltage at the terminals of a dynamo is proportional to the magnetic field that surrounds its pole pieces. Parsons placed an arm of soft iron, which was connected to an adjustable spring, at the top of the dynamo poles. The iron arm tended, against the resistance of the spring, to align itself with the magnetic field; the degree of alignment depended directly on the strength of the field, which in turn depended on the dynamo voltage. As the arm moved it carried a brass finger with it. The finger progressively blocked off more or less of the end of a pipe that was part of the air system connected to the centrifugal fan.

As the magnetic field strength increased, the finger progressively closed off the end of the pipe. This prevented air from leaking into the pipe system and enhanced the suction effect created by the centrifugal fan. The leather diaphragm, which was connected to the governing valve, was then sucked in, closing the steam inlet valve and allowing less steam into the turbine. The speed of the turbine was thus directly governed by the output voltage of the dynamo. Parsons' fine-tuning mechanism was one of the first servomotors, a class of feedback devices that require only small amounts of power to regulate the flow of far greater amounts of power.

The centrifugal fan, which was an important part of Parsons' governor, also had a role in his lubrication system. Because the turbine shaft spun at such a high rate, lubrication had to be absolutely reliable. Parsons attached a screw to the end of the shaft and placed it in the opening of an oil reservoir. The screw forced oil under pressure into the shaft bearing. The oil was also passed through a system of pipes to the far end of the dynamo shaft. The dynamo shaft was hollowed out so that the oil could flow through it to the central bearings, cooling the dynamo's armature in the process. The oil returned, under the influence of gravity, to a central junction. The main oil reservoir was connected by a standpipe to the system of air-filled pipes that was attached to the centrifugal fan. The suction generated by the fan served to draw oil up from the central junction back into the main reservoir so that it would reach a level sufficient to feed the screw pump.

Another innovation of Parsons', one that is still used in modern turbines, was his method of ensuring that steam did not leak from the ends of the turbine shaft into the engine room. Any attempt to make the casing a close fit on the shaft would have resulted in severe rubbing as the shaft passed the "whirling" speed. He therefore designed a close-fitting collar that would allow the shaft to move relatively freely until the steady running speed was reached; when the turbine was up to speed, the collar acted as an effective seal, keeping the exhaust steam within the turbine.

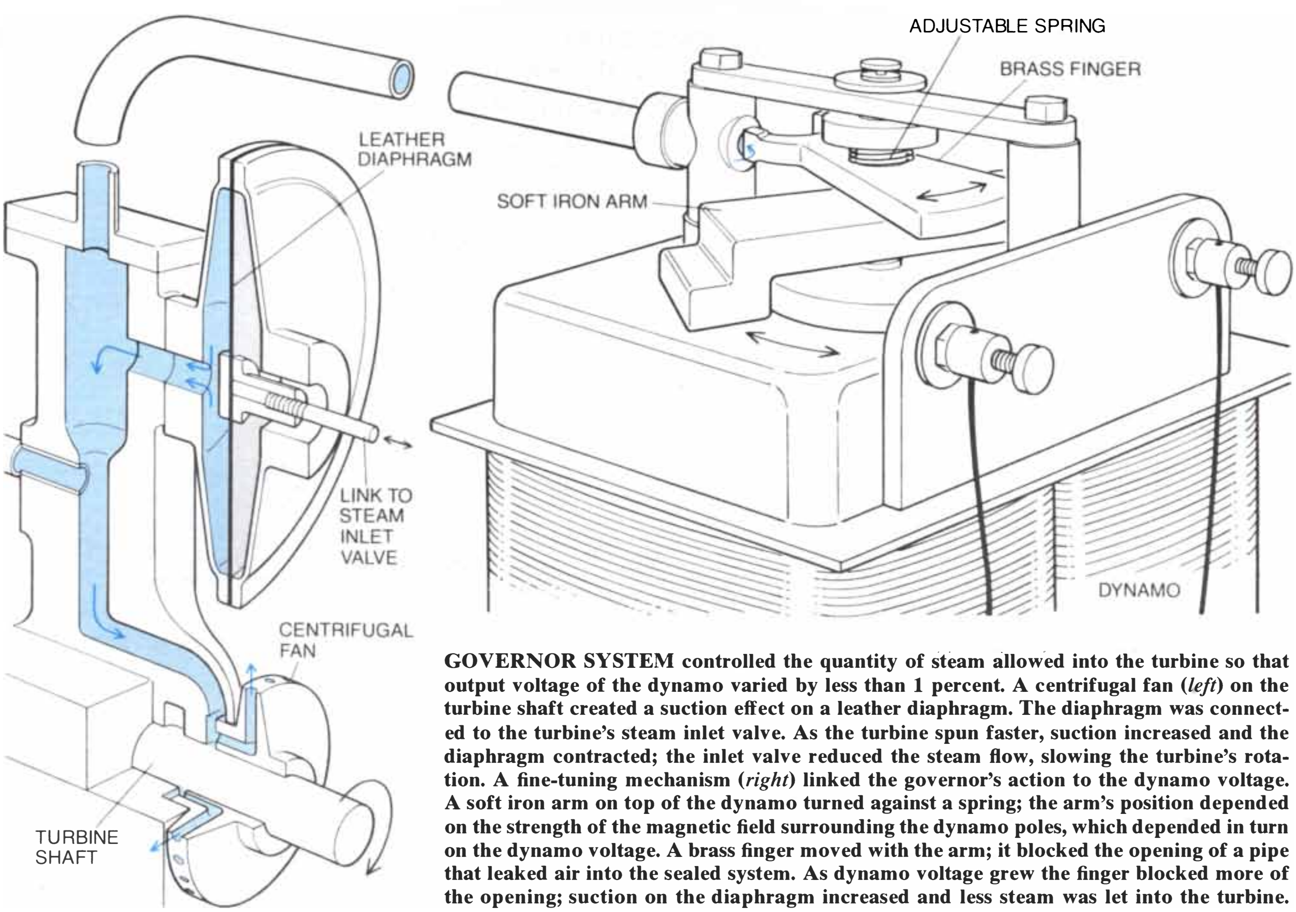
The collar was held in place, once the running speed was reached, by a suction effect established between the exhaust channel and a chamber containing a partial vacuum. The exhaust steam ran from two cavities (one at each end of the shaft) through an outlet passage cast in the lower section of the turbine casing. Farther out along the shaft than each of the exhaust cavities was another cavity. A passage cast into the upper section of the casing connected these outer cavities to each other. Parsons placed a collar, which was a tight fit on the shaft, inside each of the inner cavities. He also used a steam ejector pump to create a partial vacuum in the outer cavities. When the turbine was running at low speed, the collar was free to move with the shaft. As the running speed was approached, however, a pressure differential built up between the inner cavities (which were receiving low-pressure exhaust steam from the turbine) and the outer cavities (in which there was a partial vacuum). The pressure difference between the cavities held the collar tightly to the turbine casing and sealed the cavities from each other.

What kind of upbringing and education shaped Parsons' talents, enabling him to solve the problems that had thwarted development of the turbine? Some influences are clear. Parsons was the youngest son of a family that had been given land at Birr, in County Offaly, Ireland, early in the 17th century. Charles's father, the third earl of Rosse, was himself an accomplished scientist. He made significant contributions to the techniques of casting and grinding large telescope mirrors, and in 1845 he constructed a reflecting telescope in his estate workshops that was to remain the largest in the world for several decades. He discovered spiral nebulas, and he served as president of the Royal Society of London from 1849 until 1854. He sat in Parliament, and in order to attend to his parliamentary duties he kept a house in London. The family used to live there for a part of each year and often entertained members of the scientific community.

The Parsons children were not sent away to school. Instead the astronomers employed to work with the telescopes at night were also expected to tutor the children by day. In addition the boys were encouraged to use the estate's workshops and foundries for themselves. The manual, practical skills Charles thereby acquired were to be of crucial importance when he constructed his prototype turbine.

Charles enrolled at Trinity College, Dublin, and later transferred to St. John's College at the University of Cambridge, from which he was graduated in 1877. He studied mathematics there under such teachers as Edward J. Routh, who at that time was writing on the stability of steady motion and was particularly concerned with the behavior of engine governors.

Up to this point Parsons had enjoyed an exceptionally privileged upbringing. It was completed in an extraordinary way, considering his social station: he apprenticed himself to George Armstrong, the famous manufacturer



GOVERNOR SYSTEM controlled the quantity of steam allowed into the turbine so that output voltage of the dynamo varied by less than 1 percent. A centrifugal fan (*left*) on the turbine shaft created a suction effect on a leather diaphragm. The diaphragm was connected to the turbine's steam inlet valve. As the turbine spun faster, suction increased and the diaphragm contracted; the inlet valve reduced the steam flow, slowing the turbine's rotation. A fine-tuning mechanism (*right*) linked the governor's action to the dynamo voltage. A soft iron arm on top of the dynamo turned against a spring; the arm's position depended on the strength of the magnetic field surrounding the dynamo poles, which depended in turn on the dynamo voltage. A brass finger moved with the arm; it blocked the opening of a pipe that leaked air into the sealed system. As dynamo voltage grew the finger blocked more of the opening; suction on the diaphragm increased and less steam was let into the turbine.

of naval guns, at his Elswick factory in Newcastle upon Tyne. It would be interesting to know what led to Parsons' decision, because in those days the sons of the wealthy rarely turned to engineering as a career.

Parsons earned the reputation of being Armstrong's most industrious apprentice. During his training he got permission to work on a revolutionary steam engine in which the cylinders rotated, and he took out several patents between 1877 and 1882. Examination of these patents reveals that he used the technique of forced lubrication more than a decade before its use by A. C. Pain, who is generally credited with its invention. Until this time bearings had been drip-fed and required constant attention; the development of forced lubrication was of great importance for high-speed machinery and was to prove essential for turbines.

It seems likely that Parsons' thoughts first turned toward turbines while he was still a student. Lord Rayleigh quotes a contemporary at Cambridge who had been shown a toy paper engine in Parsons' rooms, the wheels of which "simply flew around" when Parsons blew on it. Parsons said of it that it would "run twenty times faster than any machine."

His serious experimentation with turbines began while he was at Armstrong's. From 1881 to 1883, the years just after he finished his apprenticeship, Parsons worked in partnership with James Kilson of Leeds on gas-propelled torpedoes. Much of Armstrong's business was concerned with naval weapons, and it is likely that Parsons' employer would have encouraged experiments on a novel form of torpedo propulsion. In these torpedoes a propellant was burned to produce jets of high-pressure gas. The jets impinged on the blades of a spinner, setting it in motion. The spinner in turn powered the torpedo's screw propeller.

In Parsons' notebooks there is no clear indication of the configuration of the spinners, but we can gain some idea of how they were constructed from a small boat he built at about this time out of copper sheeting. The boat was driven by a three-bladed screw propeller mounted under the hull. The propeller was mounted inside a large ring and the circumference of the ring was cut to form passages for the impinging jets of gas. The outer ring contained 44 passages; each passage was cut as a spiral. When gas was forced through the passages, the resulting deflection force made the ring turn. As the ring turned, the propellor turned with it, driving the boat forward.

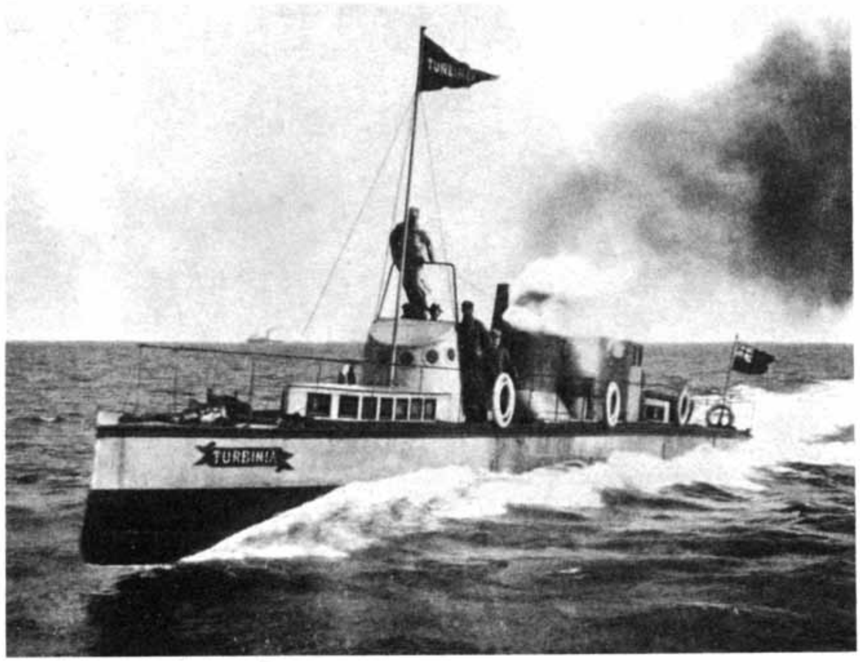
Parsons' early turbine experiments thus involved gas, not steam, turbines. He abandoned work on the gas turbine in 1883, although his patent specification of 1884 does describe the modern gas-turbine cycle. Years later he explained that "experiments made many years ago, some of which were directed to ascertain the possibility of a gas turbine, convinced me that with the metals now at our disposal... to cause flame, whether at full heat or cooled by an admixture of steam or water, to impinge on the blades, was the wrong method." His judgment was a sound one: suitable materials for components of gas turbines were not available until a decade or so after his death.

Early in 1884 Parsons joined Clarke Chapman and Company as a junior partner at Gateshead, where he set to work on the design of a steam turbine. As late as August, 1883, his notebooks on the torpedo experiments showed that he had not then grasped the importance of matching the speed of the turbine blades to that of the gas jets. Nor did he appreciate the problems of nozzle design with large pressure ratios. Yet by April, 1884, he had filed two provisional patents and in October and November of that year he had filed the detailed final specifications.

It was a period of incredible creativeness. Not only did Parsons have to experiment with high-speed shafts and other features of the turbine but also he had to devise a suitable means of utilizing the output of his motor. Running at 18,000 r.p.m., it was unsuited to any conventional tasks. He chose to develop a dynamo capable of speeds that few modern electrical machines can match, which could be coupled directly to the turbine. Parsons often said in later years that this achievement was as great as the construction of the turbine itself. Even today the chief application of steam turbines is to drive directly coupled electric generators.

The early machines were not efficient. Until these "steam eaters" had grown in output to the point where they matched the economy of the conventional steam engine, they needed special features to attract customers. These were provided in their small physical size, the steadiness of their electrical output, their ability to run unattended and their low maintenance costs; all these characteristics were present in the first machine.

In November, 1884, when his prototype was completed, the Honorable Charles A. Parsons was just 30 years old. Engineering genius and an instinct for the needs of the market were not in themselves sufficient gifts to ensure that his invention would be brought quickly to maturity. On a number of occasions he had to draw on his own personal wealth to fund his work. In 1898, during judicial proceedings to extend the life of his patents, it was computed that up to that point the development of the turbine had cost him the net sum of 1,107 pounds, 13 shillings and 10 pence.



***TURBINIA* was the first turbine-powered ship. Completed in 1894, the *Turbinia* achieved a record speed of 34 knots. Turbine power became standard for propulsion of larger craft.**

THE AMATEUR SCIENTIST

Experiments with the external-combustion fluidyne engine, which has liquid pistons

by Jearl Walker

The fluidyne engine is a remarkable device that takes its name from the fact that it has liquid pistons. Like its close relative, the solid-piston Stirling engine invented in 1816 by the Reverend Robert Stirling of Scotland, it is useful as a laboratory or workshop aid in studying aspects of thermodynamics. Some of the engines might also serve as inexpensive pumps for irrigation. Colin D. West of the Oak Ridge National Laboratory has investigated several fluidyne engines; his work provides the basis for much of my discussion.

You can picture the operation of a fluidyne engine by envisioning a pair of U tubes side by side [*see upper illustration on opposite page*]. The tube on the left is called the displacer, the tube on the right the output. Each one contains water or some other liquid. A smaller tube across the top connects the two openings at the top of the displacer tube with the left-hand opening at the top of the output tube, so that air can move back and forth between the two U-shaped tubes. The engine delivers the energy it develops from the right-hand opening at the top of the output tube.

The engine develops energy because the upper left-hand end of the displacer is kept hot by an external source of heat whereas the companion end on the right side is kept cold, either by an external source of refrigeration or simply by the fact that the liquid there is not heated. Hence the left side of the displacer is in effect a hot cylinder and the right side is a cold cylinder.

You start the engine by turning on the heat source. When the temperature gets high enough, the water becomes unstable and begins to oscillate slightly from side to side in the displacer. When water flows from the hot cylinder to the cold one, air is pushed through the connecting tube to the hot cylinder. As that air warms, the pressure in the entire connecting tube increases, driving down the column of water in the left side of the output unit, so that the water in the right side rises. Here is the energy that can be harnessed to do work.

The cycle is completed as the weight of the water in the cold cylinder forces water to flow from that cylinder into the hot one. The motion of the water drives air to the cold end, where it cools, decreasing the air pressure. By then the weight of the water is forcing the level in the right side of the output unit down from its highest point. The low pressure in the displacer unit also helps to pull the water in the output unit toward the displacer. If the water in the displacer continues to oscillate, the variation in air pressure produces corresponding oscillations in the water in the output unit; as a result the water level on the right side of that unit sweeps through a large volume. The water in the unit can then function as a piston to do work.

In order for the system to function efficiently and yield a net output of energy, the oscillation in the output must be tuned to match the frequency of the oscillation in the displacer. The task is accomplished by adjusting the size of the output tube and the amount of water it contains.

If the output tube is too narrow or contains too much water, the increase in air pressure in the displacer is insufficient to make the water on the right side of the output sweep through a large volume. If the amount of water is too small, the water is too easy to move. When air is transferred from the cold cylinder, some of it can flow into the space made available in the output tube. This air is not heated, and the air pressure does not increase much. When the air pressure in the displacer unit barely changes, the inward and outward flows of the water in the output unit require almost equal amounts of energy; consequently there is little energy for the output piston.

Some means must be provided to keep the water in the displacer oscillating. One method is to mount the engine on a pivot and connect a spring between the output tube and the floor. As the water oscillates in the output unit, its shifting weight sets the spring in motion; the spring action helps to maintain the oscillation of the water in the displacer unit.

Another technique relies on a feedback of energy from the water oscillating in the output [*see upper illustration on page 142*]. That unit is connected to the displacer below the side of the tube that constitutes the hot cylinder. As a result the hot and cold cylinders contain unequal amounts of water. (In each case the volume is measured from the water level in the cylinder to the connection with the output unit. In the hot cylinder that is a short distance; therefore the volume of water it contains is easier to move than the larger volume in the cold cylinder.)

The frequency of oscillation is the same in the displacer and output units but the oscillations are about 90 degrees out of phase. In the first stage the water level is at the maximum in the hot cylinder and between the midpoint and the minimum in the cold cylinder. The level in the output unit is falling through the midpoint.

As the level in the hot cylinder falls, the level in the cold one rises rapidly because it is receiving water from both the hot cylinder and the output unit. Air is displaced into the hot cylinder, where it warms, increasing the air pressure in the displacer unit. Since the water in the hot cylinder is easy to move, it is driven downward rapidly by the increasing air pressure. Some of the water pushes into the output, driving the water on the right side upward. When the water in the hot cylinder reaches its minimum level, the water in the output continues to rise because that unit is now drawing water from the cold cylinder. The continued rise in the output unit's level promotes a drop in the cold cylinder's level.

By the time the fluid level in the output unit reaches its maximum the level in the cold cylinder is falling, driving the level in the hot one upward. At the same time air is being moved to the cold cylinder of the displacer, where it cools. As a result the pressure in the displacer falls. When the level in the output unit begins to drop, it pushes water into the displacer. Again the water in the hot cylinder is easy to move, so that the water arriving from the output pushes into the hot cylinder, helping the level to rise there.

The interaction between the output unit and the displacer unit helps to maintain the continued oscillation of water in the displacer even though there is a loss of energy owing to viscosity. When the level in the output unit approaches its maximum, the output tube is pulling water from the cold cylinder. When the level in the output unit is falling toward the midpoint, the water there is pushing water into the hot cylinder, causing the level to rise in that cylinder.

Has there been a gain of power? Can one call this system an engine? Although power was supplied to the output unit during the outward push of water, power was consumed in pushing water back into the displacer. If the amounts of power involved in the two motions are equal, there is no net gain.

The answer is that the amounts of power are not equal, because the air pressure in the displacer unit changes. When the water in the hot cylinder is propelled downward and into the output tube, the air pressure is high. Hence a good deal of power is pushing the water in the output tube upward. When water is pushed back into the displacer unit, it moves against a lower air pressure. Less power is needed. The system achieves a net gain of power, and the output could be coupled to something else to do work.

In some fluidyne engines the output of the heat source is sufficient to vaporize some of the water. Even though this consumes quite a bit of energy, the increased amount of water vapor in the air of the displacer leads to greater variations in air pressure. Therefore the engine generates more power.

Much of the interest in fluidyne engines focuses on their potential as pumps. One technique is called a series arrangement. A vertical pumping tube is attached to one end of the output unit [*see illustration at lower left on next page*]. It contains two ball valves positioned so that water can flow only upward through them. At the bottom of the pumping tube is the supply of water that is to be pumped.

When the air pressure in the displacer unit is low, water moves from the output into the displacer. Consequently water is drawn up from the supply and passes through the bottom valve. When the air pressure in the displacer is high, it pushes water from the displacer into the pumping tube, forcing a discharge through the top valve. This cycle, repeated indefinitely, constitutes the pumping action.

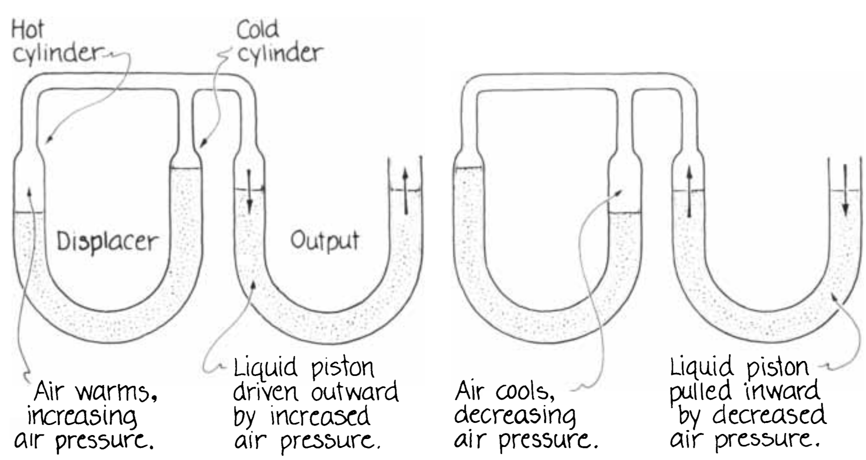
A better design is the parallel arrangement [*see illustration at lower right on next page*]. The vertical pumping tube is mounted between the displacer unit and another vertical tube that is a continuation of the output unit. The volume of water moving through the output is greater than the volume going through the valves. Because the output is tuned to match the frequency of oscillation in the displacer, the oscillations of the output provide a feedback that helps to maintain the oscillations in the displacer. The arrangement produces steadier pumping.

A fluidyne engine can be made with glass tubing. David Herbert of the Atomic Energy Research Establishment at Harwell in England has created such a device [*see top illustration on page 143*]. High temperatures and large variations in temperature require that it be made with borosilicate glass. The valves in the pumping tube are glass beads or balls taken from a bearing. A 16-watt quartz halogen projector bulb provides the heat for the displacer unit.

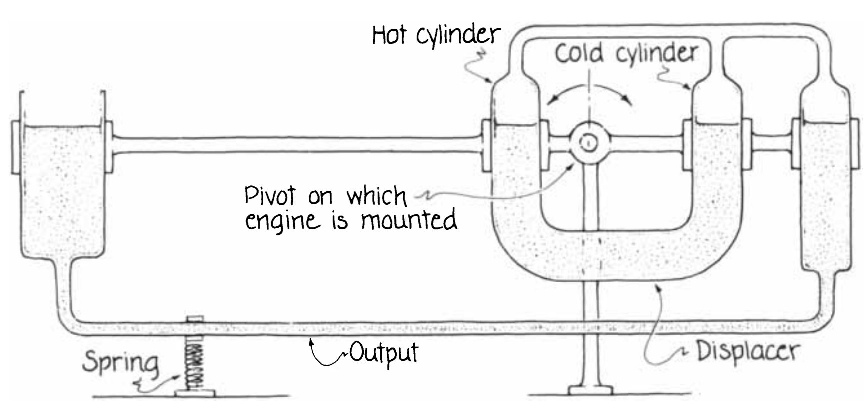
Similar engines (even larger ones) can be driven by sunlight if the beam is focused on the displacer arm by a lens. (An inexpensive Fresnel lens made of plastic will do.) The focal point should be at approximately the mean water level in the hot cylinder. A reflector mounted on the back of the cylinder intensifies the effect by sending some of the light back through the water. Blackening one side of the cylinder might also amplify heat absorption. Alternatively, you could darken the water by adding ink to it.

Even a simple flame can drive this type of fluidyne engine. John Geisow of the U.K. Atomic Energy Research Establishment has designed an engine heated by an alcohol lamp or a candle. A short length of thin aluminum extends upward from just above the flame to the hot cylinder. The strip is tilted against the cylinder so that friction holds it in place. This strip funnels heat to the cylinder by convection and also delivers heat by conduction through the metal.

A group in the same laboratory has designed a fluidyne engine based on a glass fruit jar (two-pound capacity). A coiled tube inside the jar serves as the output unit. It is 2.14 meters long and has an outside diameter of nine millimeters. (The tube could be as much as 14 centimeters shorter or 36 centimeters longer.) The tubing is coiled eight times in circles that have an outside diameter of 8.5 centimeters. One end of the coil fits loosely into the hot cylinder of the displacer, which is an inverted U tube mounted on the lid of the jar. The other end is linked to the



The action of a fluidyne engine



A fluidyne engine mounted on a mechanical rocker pumping tube by means of a plastic tube that will not kink.

You can make the displacer tube by blowing the glass, or you can connect two glass cylinders with a short length of plastic tubing. A hair dryer that runs at a high temperature can serve as the heat source; otherwise an industrial hot-air blower is best.

Without a pumping tube the engine needs little adjustment. With a pumping tube you have to set the water levels in the cylinders carefully. The length of output tube inserted into the hot cylinder is also critical. Properly adjusted, the engine can pump about 19 liters (five gallons) per hour.

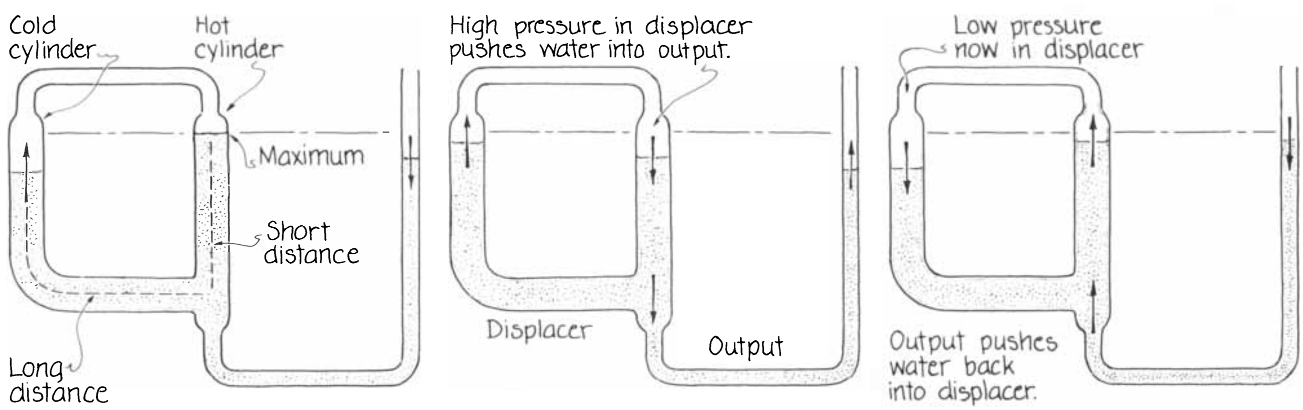
West has designed a plastic fluidyne engine that works on the liquid-feedback principle [*see upper illustration on page 144*]. Its front and back are made of clear acrylic 1/8 inch thick. Between them are 3/8-inch acrylic strips. The heat source is a three-watt, 20-ohm, wire-wound resistor. Leads from the resistor pass through the rear cover plate (surrounded by glue to prevent leakage) and are connected to a low-voltage transformer of the type employed in doorbells. It supplies alternating current to the resistor at a voltage level of nine volts and a power level of four watts. That exceeds the power rating of the resistor, but the resistor is in water and so stays sufficiently cool. The leads from the resistor are not insulated because there is little leakage of current between them in the liquid.

West sometimes adds sugar (30 percent by weight) to the water to increase the efficiency of the machine. Although the sugar increases the viscosity of the liquid, making the oscillations in the displacer harder to maintain, it also dramatically reduces the loss of heat from the liquid in the hot cylinder. When the system contains sugar, West applies 12.6 volts to the resistor.

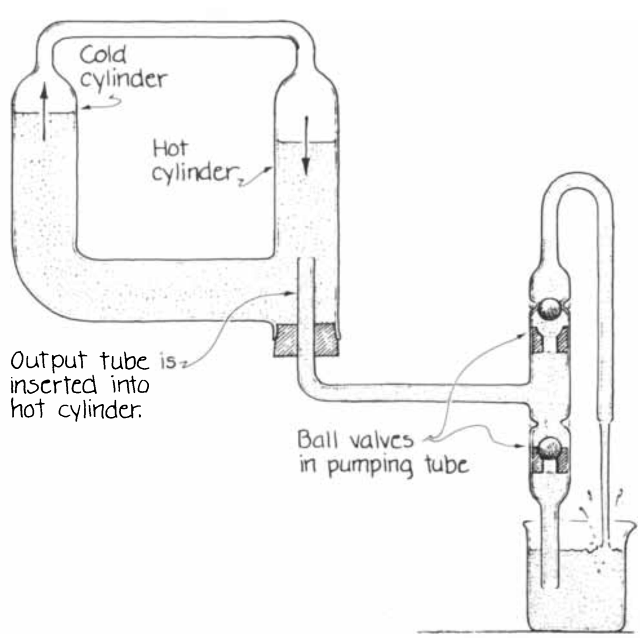
Why does sugar enhance the machine's efficiency? Hot water moving downward during the oscillations establishes a thin boundary layer next to the cylinder walls. Viscosity keeps the boundary layer from moving. Without sugar the boundary layer is thin enough so that the water can transfer heat to the wall, whence the heat is lost to the ambient air. It is therefore difficult to maintain a suitable temperature in the hot cylinder. Sugar increases the viscosity, thickens the boundary layer and decreases the loss of heat.

West urges caution in the use of sugar. It can make a mess, particularly if the engine is not operated for a while (the sugar comes out of solution and forms a cake) or if the resistor is overheated (the sugar caramelizes). You might substitute some other substance, such as glycerin.

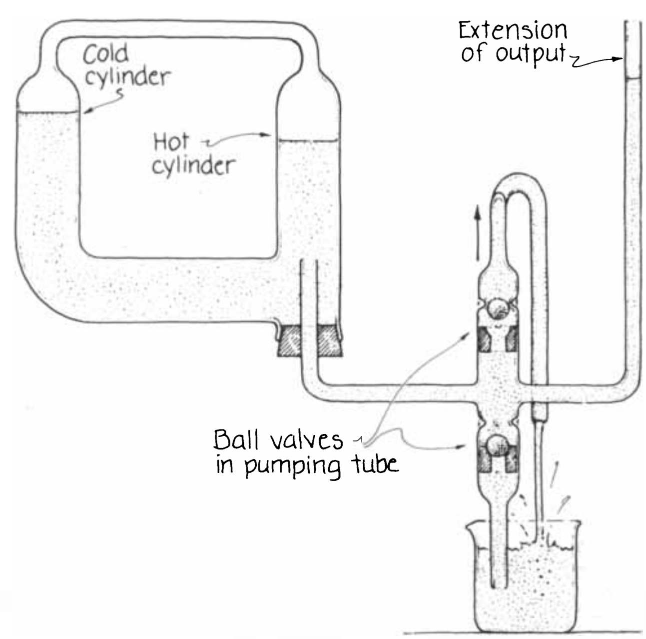
West has built another fluidyne en-



Interactions between the displacer and the output tubes



A series arrangement for pumping water



A parallel arrangement for pumping water gine of wood [*see lower illustration on next page*]. Each front and back cover plate is 13 by six inches in length and width and 3/4 inch thick. The spacers between the cover plates are 1.5 inch wide and 3/4 inch thick. Brass screws and waterproof glue hold the engine together.

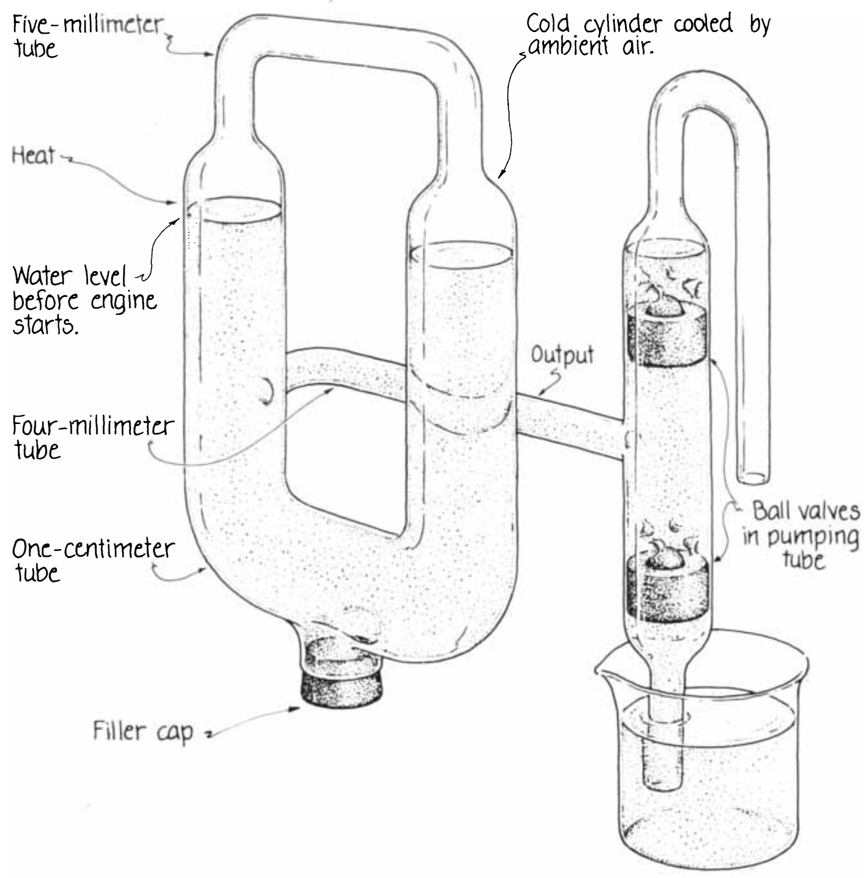
This engine's hot cylinder is heated by a 10-ohm resistor (bought from Radio Shack) rated at 10 watts. Leads run from it to small bolts driven through the wood. They are connected to a transformer operated at 12.6 volts, so that the power delivered to the resistor is about 15 watts.

The top of the cold cylinder contains a folded sheet of thin aluminum foil that serves to maintain a uniform temperature there; the foil receives heat from the hot air at one stage of the cycle and gives it up to the water at another stage. A can of water attached to the cold cylinder provides extra cooling. When the air pressure in the engine increases, some of the water in the cold cylinder is forced into the can. It mixes with the water already there and loses the heat it has picked up from the foil. When the air pressure in the engine decreases, the cooled water is pulled back from the can.

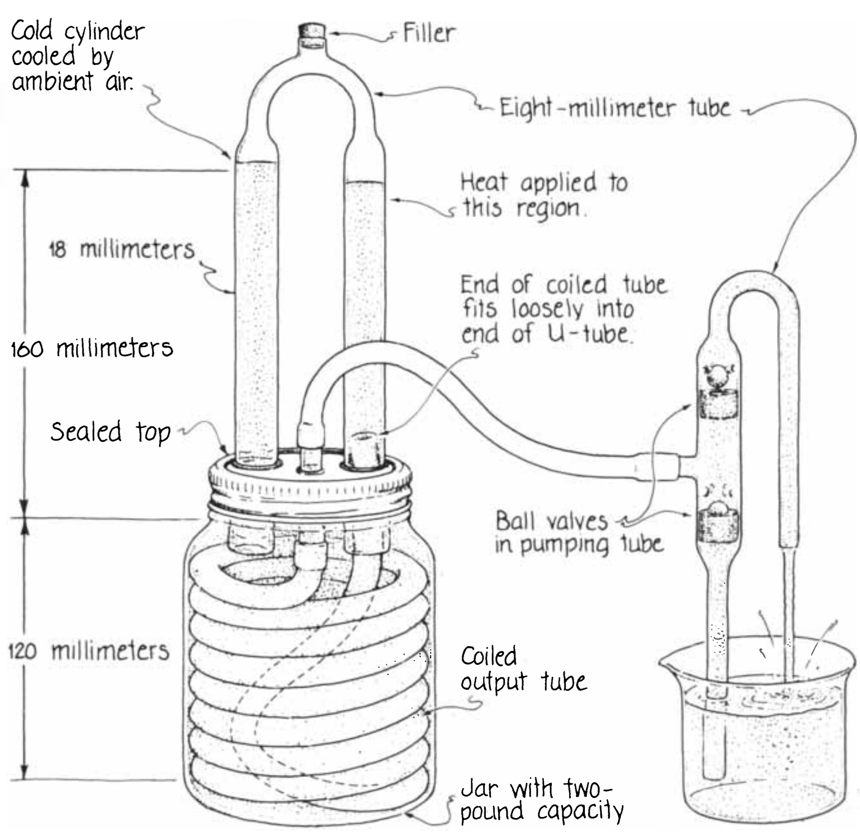
The engine's output unit is made of clear plastic tubing. The engine is filled through a funnel on top of the unit. West mounts a hypodermic needle in the top of the engine, holding it in place with silicone rubber. The needle keeps the average air pressure in the engine above the output unit near atmospheric levels. Without the needle you would have to open the engine occasionally to let off pressure.

A folded aluminum or ceramic honeycomb inserted in the top of the cold column enhances the operation of many fluidyne engines, particularly the larger ones. Even short lengths of drinking straws can serve the purpose. The material is intended to take up the heat from the hot air arriving from the hot cylinder. The material is cooled later as the water level rises in the cold cylinder. (Avoid material that has holes small enough to trap water by surface tension; it will keep the water level from oscillating properly.) In larger engines a similar material may be needed in the tube that passes air between the hot and cold cylinders. In engines where evaporation of the water has a role, however, the material in the connecting tube is unimportant.

A fluidyne engine is generally self-starting. The oscillations begin when the temperature in the hot cylinder gets high enough. If the engine depends on a mechanical rocking mechanism, you might have to start the oscillations by rocking the unit yourself. A pump-

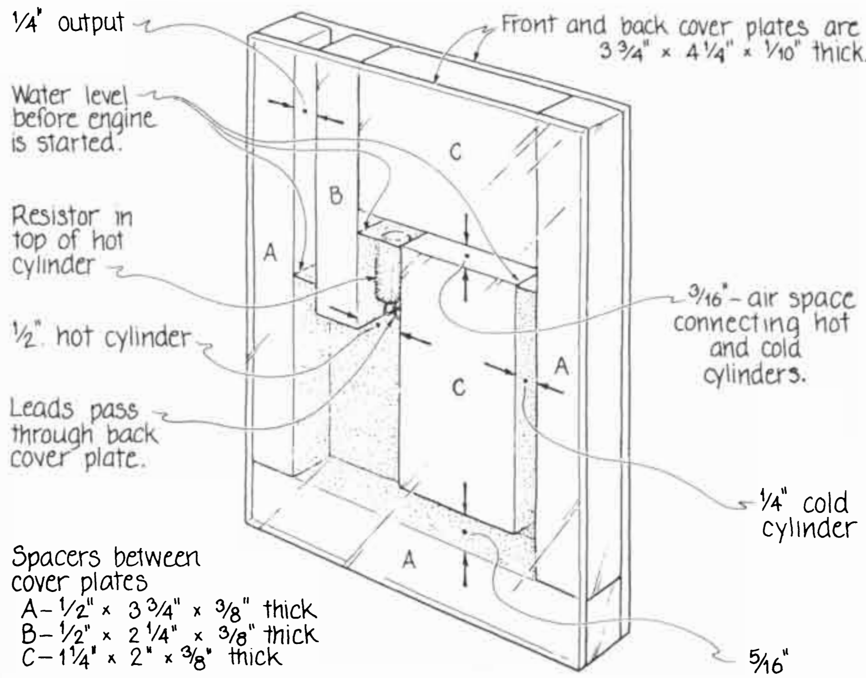


David Herbert's fluidyne pump

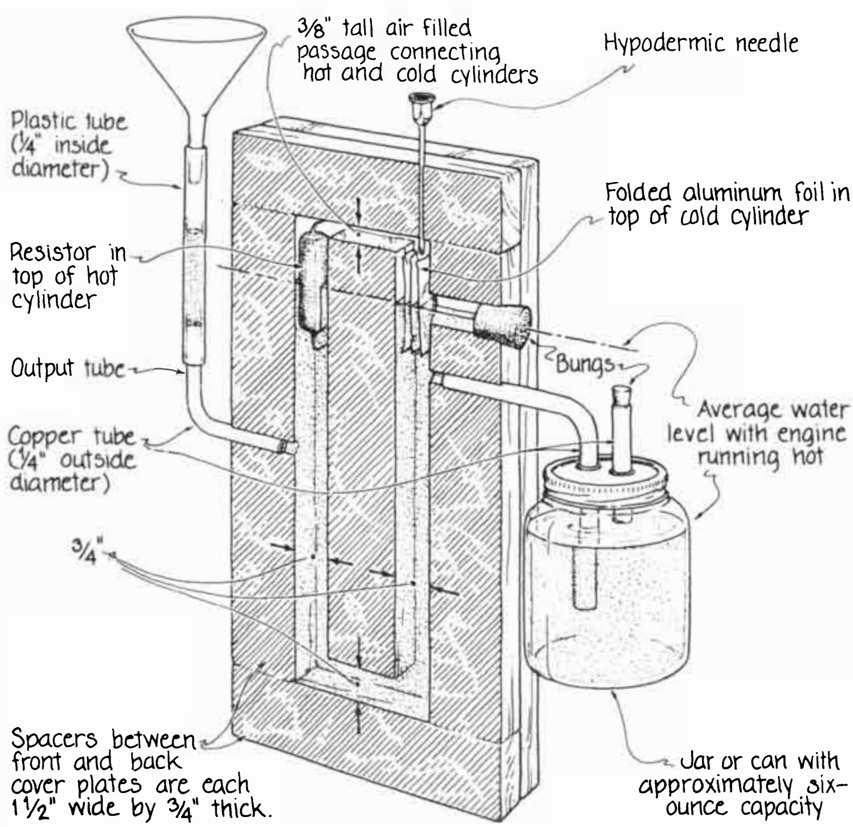


A fluidyne engine based on a fruit jar ing engine may need similar assistance.

West says that fluidyne engines offer many experimental opportunities. What parameters of size and heating are best? If the engine is not designed for pumping, would a fluid other than water be an improvement? A book by West on fluidyne engines is listed in "Bibliography" [*page 146*]. The book presents designs for more engines and different pumping configurations, together with discussions of the physical principles that explain how the engines work.



A plastic version built by Colin D. West



West's wood engine

THE SCIENCE OF WILDLIFE MANAGEMENT:

The Future of our Wildlife Depends on It

#1

Research

The precarious future that many species of North American wildlife faced around the turn of the century provided the impetus for the establishment of our first wildlife parks and refuges. Though initially effective, these early efforts aimed at helping wildlife soon developed serious shortcomings. The concept of providing complete protection, including the elimination of natural predators, to certain species was successful in building up threatened herds of animals, including elk and deer; however, as early as the 1920s, populations in many areas were outstripping their available food supplies.

Such problems helped spur the rapid growth of the modern science of wildlife management. Early wildlife management professionals were the first to recognize the vital importance of vegetation and other aspects of the natural environment that supported wild animal populations. This new understanding of the relationship between wildlife and habitat helped lead to the practical steps necessary to ensure the long-term abundance and health of certain kinds of wildlife.

Extensive biological research is the foundation on which all management programs are built. Studies on animal numbers, their distribution, food preferences and the like provide a detailed picture of a species' needs and habits.

Bird banding projects, such as these, help determine a species' seasonal and local movements and can provide information on age, longevity and other vital characteristics important in developing successful conservation programs.

Over the years, it has been the American hunter who, through license fees and excise taxes, has provided the lion's share of the funds necessary for these conservation programs.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

THE ART OF COMPUTER PROGRAMMING, VOL. 2: SEMINUMERICAL ALGORITHMS. Addison-Wesley Publishing Company, 1969.

SYSTEM SIMULATION WITH DIGITAL COMPUTER. Narsingh Deo. Prentice-Hall, Inc., 1983.

THE GROWTH OF CORE REGIONS IN THE THIRD WORLD

URBAN CONCENTRATION, AGGLOMERATION ECONOMIES, AND THE LEVEL OF ECONOMIC DEVELOPMENT. William C. Wheaton and Hisanobu Shishido in *Economic Development and Cultural Change,* Vol. 30, No. 1, pages 17–30; October, 1981.

IMPROVED INTERNATIONAL COMPARISONS OF REAL PRODUCT AND ITS COMPOSITION: 1950–1980. Robert Summers and Alan Heston in *The Review of Income and Wealth,* Series 30, No. 2, pages 207–262; June, 1982.

URBANIZATION AND THIRD WORLD STABILITY. Henry Bienen in *World Development,* Vol. 12, No. 7, pages 661–691; July, 1984.

POPULATION REDISTRIBUTION TOWARDS LDC CORE AREAS, 1950–1980. Daniel R. Vining, Jr., in *International Regional Science Review,* Vol. 10, No. 1, in press.

THE EARTH'S HOT SPOTS

THE ICELAND MANTLE PLUME: STATUS OF THE HYPOTHESIS AFTER A DECADE OF NEW WORK. Peter R. Vogt in *Structure and Development of the Greenland-Scotland Ridge,* edited by Martin H. Bott and Svend Saxov. Plenum Publishing Corp., 1983.

HOTSPOT TRACKS AND THE EARLY RIFTING OF THE ATLANTIC. W. Jason Morgan in *Tectonophysics,* Vol. 94, No. 1–4, pages 123–139; May 1, 1983.

A HOTSPOT MODEL FOR ICELAND AND THE VØRING PLATEAU. Gregory E. Vink in *Journal of Geophysical Research,* Vol. 89, No. B12, pages 9949–9959; November 10, 1984.

THE RELEASE OF ACETYLCHOLINE

THE PRESENT STATUS OF THE VESICULAR HYPOTHESIS. M. Israël, Y. Dunant and R. Manaranche in *Progress in Neurobiology,* Vol. 13, No. 3, pages 237–275; 1979.

CONTINUOUS DETERMINATION BY A CHEMILUMINESCENT METHOD OF ACETYLCHOLINE RELEASE AND COMPARTMENTATION IN *TORPEDO* ELECTRIC ORGAN SYNAPTOSOMES. Maurice Israël and Bernard Lesbats in *Journal of Neurochemistry,* Vol. 37, No. 7, pages 1475–1483; December, 1981.

ACETYLCHOLINE MEASURED AT SHORT TIME INTERVALS DURING TRANSMISSION OF NERVE IMPULSES IN THE ELECTRIC ORGAN OF *TORPEDO.* Y. Dunant, G. J. Jones and F. Loctin in *The Journal of Physiology,* Vol. 325, pages 441–460; April, 1982.

RECONSTITUTION OF A FUNCTIONAL SYNAPTOSOMAL MEMBRANE POSSESSING THE PROTEIN CONSTITUENTS INVOLVED IN ACETYLCHOLINE TRANSLOCATION. M. Israël, B. Lesbats, N. Morel, R. Manaranche, T. Gulik-Krzywicki and J. C. Dedieu in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 81, No. 1, pages 277–281; January, 1984.

ELEMENTARY PARTICLES AND FORCES

THE NATURE OF MATTER. Edited by J. H. Mulvey. Oxford University Press, 1981.

THE DISCOVERY OF SUBATOMIC PARTICLES. Steven Weinberg. W. H. Freeman and Company, 1983.

GAUGE THEORIES OF THE STRONG, WEAK, AND ELECTROMAGNETIC INTERACTIONS. C. Quigg, Benjamin-Cummings Publishing Co., 1983.

SUPERCOLLIDER PHYSICS. E. Eichten, I. Hinchliffe, K. Lane and C. Quigg in *Reviews of Modern Physics,* Vol. 56, No. 4, pages 579–707; October, 1984.

COMPARTMENTALIZATION OF DECAY IN TREES

COMPARTMENTALIZATION OF DECAY IN TREES. Alex L. Shigo and Harold G. Marx. Agricultural Information Bulletin No. 405, U.S. Department of Agriculture, Forest Service, 1977.

TREE DECAY: AN EXPANDED CONCEPT. Alex L. Shigo. Agricultural Information Bulletin No. 419, U.S. Department of Agriculture, Forest Service, U.S. Government Printing Office, 1979.

AN EXPANDED CONCEPT OF TREE DECAY. W. Merrill and A. L. Shigo in *Phytopathology,* Vol. 69, No. 10, pages 1158–1160; October, 1979.

MECHANISMS OF COMPARTMENTALIZATION OF DECAY IN LIVING TREES. Walter C. Shortle in *Phytopathology,* Vol. 69, No. 10, pages 1147–1151; October, 1979.

THE CLOUDS OF MAGELLAN

MAGELLANIC CLOUDS. B. J. Bok in *Annual Review of Astronomy and Astrophysics,* Vol. 4, pages 95–144; 1966.

THE MAGELLANIC CLOUDS. Edited by Andre B. Muller. D. Reidel Publishing Co., 1971.

THE MAGELLANIC STREAM. D. S. Mathewson, M. N. Cleary and J. D. Murray in *The Astrophysical Journal,* Vol. 190, No. 2, Part 1, pages 291–296; June 1, 1974.

ANESTHESIOLOGY

THE PRESSURE REVERSAL OF GENERAL ANESTHESIA AND THE CRITICAL VOLUME HYPOTHESIS. K. W. Miller, W. D. M. Paton, R. A. Smith and E. B. Smith in *Molecular Pharmacology,* Vol. 10, No. 2, pages 131–143; March, 1973.

OPIATE RECEPTORS AND INTERNAL OPIATES. Solomon H. Snyder in *Scientific American,* Vol. 236, No. 3, pages 44–56; March, 1977.

HISTORY AND PRINCIPLES OF ANESTHESIOLOGY. Theodore C. Smith, Lee H. Cooperman and Harry Wollman in *The Pharmacological Basis of Therapeutics,* sixth edition, edited by Alfred Goodman Gilman, Louis S. Goodman and Alfred Gilman. Macmillan Publishing Co., Inc., 1980.

PROBING THE LIMITS OF HUMAN DEEP DIVING. P. B. Bennett and Melissa McLeod in *Philosophical Transactions of the Royal Society of London, Series B,* Vol. 304, pages 105–117; 1984.

THE PARSONS STEAM TURBINE

CHARLES PARSONS. Rollo Appleyard. Constable & Co., Ltd., 1933.

STEAM TURBINE THEORY AND PRACTICE. William J. Kearton. Pitman Publishing Co., 1958.

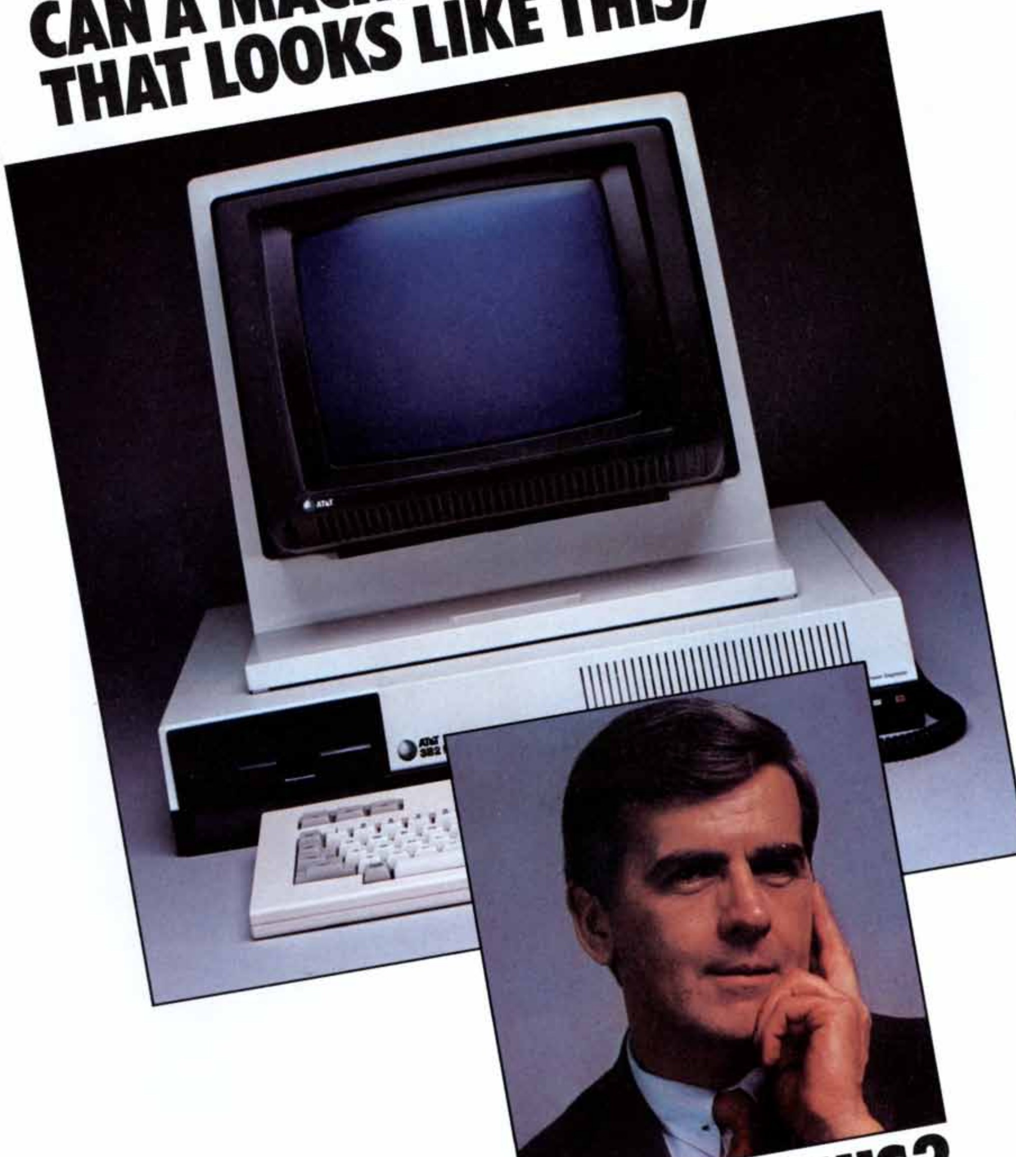
THE PARSONS CENTENARY—A HUNDRED YEARS OF STEAM TURBINES. F. R. Harris in *Proceedings of the Institution of Mechanical Engineers, Part A,* Vol. 198, No. 9, pages 183–224; 1984.

THE AMATEUR SCIENTIST

THE STIRLING ENGINE. Graham Walker in *Scientific American,* Vol. 229, No. 2, pages 80–87; August, 1973.

LIQUID PISTON STIRLING ENGINES. Colin D. West. Van Nostrand Reinhold Company, 1983.

CAN A MACHINE
THAT LOOKS LIKE THIS,



THINK LIKE THIS?

Not yet, but we're working on it.

Computers only look smart. Actually, they can't tell the players, or anything else, without a program.

But, until recently, even the most sophisticated programs only allowed computers to operate in "yes" or "no" terms. Judgment was still an exclusive domain of the human brain.

Now, however, the rapidly developing science of artificial intelligence is moving in on the mind.

At AT&T Bell Laboratories, using the new approaches to computer programming of knowledge engineering and rule-based programming, we have developed software that can cope—like a human—with incomplete and uncertain information, such as incorrect spelling and improper abbreviations.

The Expert with an ACE in the Hole

Called expert systems, this ambiguity-tolerant software is a new approach to artificial intelligence in industry. And, like the subject matter experts it mimics, our first expert system, ACE, works with equivocal data, but produces expert judgments.

ACE, for "Automated Cable Expertise," is a software system that contains the distilled knowledge—in the form of "if-then" rules—of the people who know cables best: telephone company cable maintenance experts. ACE differs from other expert systems in two ways: it manipulates massive amounts of data, and driven by the UNIX™ Operating System on a 3B2 computer, it obtains this information automatically from the data bases of other computers.

Developmental ACE software has been working as an "assistant" for over two years now to the cable maintenance force of the Southwestern Bell Telephone Company. Every night it monitors and analyzes the performance of cable systems serving over half-a-million customers in several metropolitan areas.

But ACE does more than analyze; it makes recommendations.

The Expert at Work

Unlike a conventional computer system, ACE isn't programmed with all logical answers to all possible problems. Instead, it's given a set of about 500 rules to follow.

ACE can run through the cable records of a city the size of Fort Worth overnight, a job that would take a human up to a week. By collecting its information from other computer programs, detecting recurring patterns, requesting additional data, and testing these data against its expert-derived rules, ACE can often isolate problems much earlier than its human counterparts. It provides information on both specific trouble types and locations—such as a break in cable insulation at the corner of 3rd and Elm. And when ACE has a recommendation to make, rather than generating a mound of paper, it communicates via a CRT.

ACE frees humans to work on the causes of problems, not the symptoms.

The Experts Behind the Experts

AT&T's skill in software development rests on a solid base of accomplishments in the computer field. Forty million lines of our software direct the world's largest computer system, the U.S. telephone network. And that experience is helping to make us a leader in new software research and design.

Today, software systems are being developed to help computers write their own software. While others make it possible to type data base requests in plain English. Our Rex expert system, for Regression Expert, aids in statistical analysis. And we've got silicon compilers that speed computer-aided design. There's even a program that allows a computer to derive new rules based on the rules already in its software!

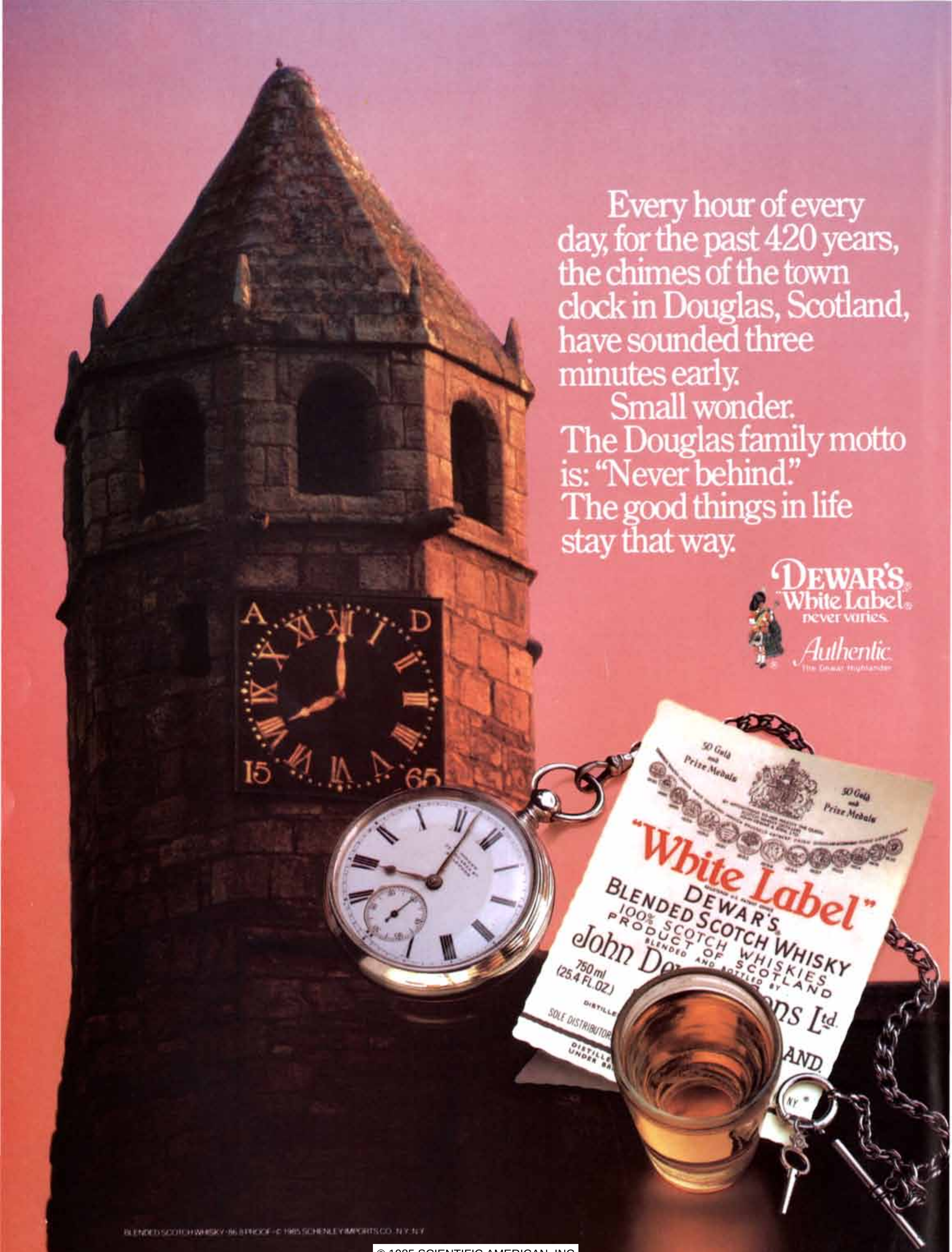
All these systems are just a few examples of software being developed today to meet our customers' needs for tomorrow.

As you cross from industry to industry, tech-talk sometimes turns into bafflegab. We can help. In our new booklet, "Tech Talk," you will find the latest terms and acronyms from the fields of microelectronics, photonics, computers and telecommunications.

To order your copy, send $2.00* to AT&T, Department A, P.O. Box 4, Hawthorne, NJ 07507.

*Please add 6% sales tax for all orders originating in New Jersey.

AT&T


Every hour of every day, for the past 420 years, the chimes of the town clock in Douglas, Scotland, have sounded three minutes early.
Small wonder. The Douglas family motto is: "Never behind." The good things in life stay that way.
DEWAR'S White Label never varies.
Authentic.
The Dewar Highlander
"White Label"
DEWAR'S
BLENDED SCOTCH WHISKY
100% SCOTCH WHISKIES
PRODUCT OF SCOTLAND
750 ml (25.4 FL.OZ.)
BLENDED SCOTCH WHISKY · 86.8 PROOF · © 1985 SCHENLEY IMPORTS CO. N.Y., N.Y.