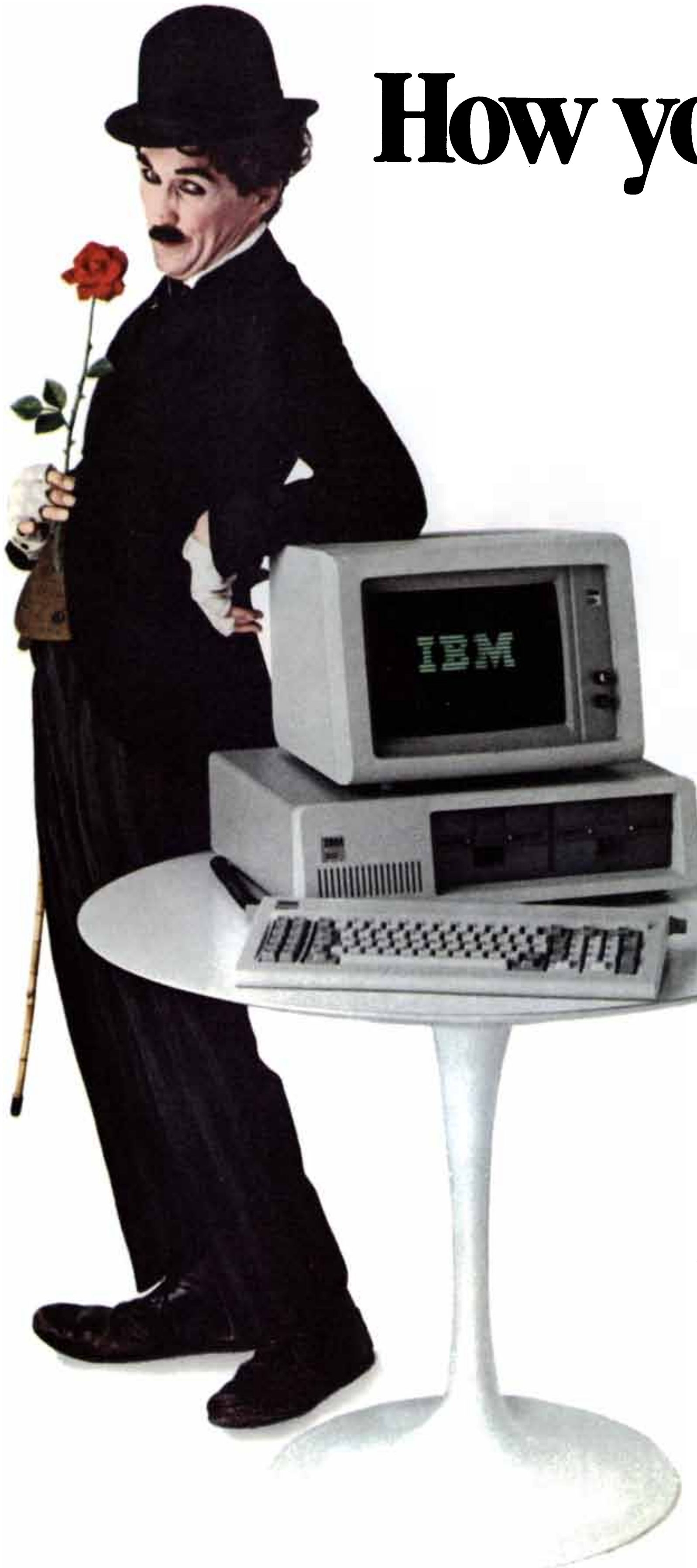
# SCIENTIFIC AMERICAN

THE SLIDE FASTENER

$2.50

June 1983

# How you can tell




They're both IBM Personal Computers. And they're the same. Only different.

The system that's on the left is the ideal solution for a person who wants to be creative, efficient and improve his or her personal productivity.

So is the other one.

The system on the left incorporates the quality, the reliability and the technological excellence that have made IBM a computer leader for over 30 years.

So does the other one.

So which is which? Simple.

On the left is the IBM Personal Computer, starting with 64KB of user memory (expandable to 640KB) and two optional 5¼″ diskette drives. It can easily satisfy your computing needs at the office, at home or in school. With 5 expansion slots, it gives you room to grow. (You can even make it function like the computer shown on the right by adding an expansion unit that houses one or two 10-million-character fixed disk drives.)

This system can run most of the same software and accept most of the same IBM hardware as the computer on the right. And its price/performance is nothing less than remarkable.

**The IBM Personal Computer**


TOYOTA
WOW!

Established 1845 **SCIENTIFIC AMERICAN** June 1983 Volume 248 Number 6

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 75 cents 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 84121. 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 $2 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 0036–8733/83. $2.00 + 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 __________

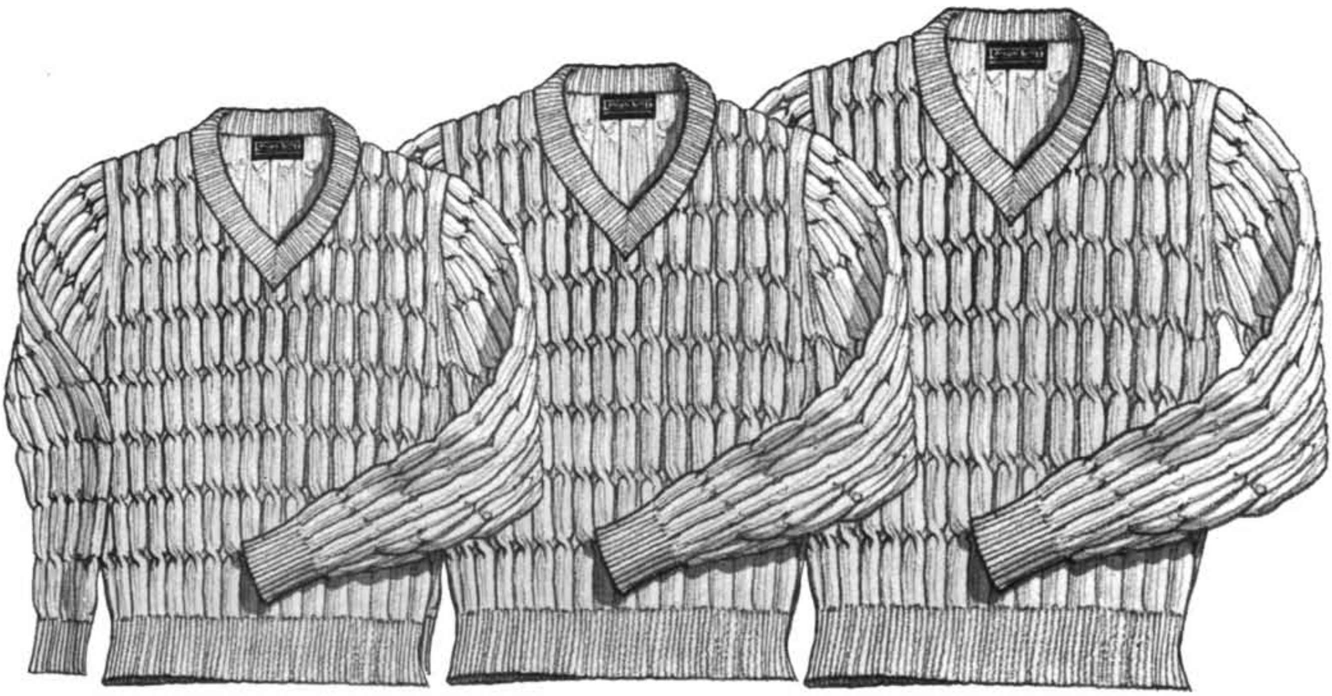
### THE COVER

The painting on the cover portrays a slide fastener, or zipper, as though it were being seen through a fairly powerful magnifying glass. The zipper is shown without its slider in order to depict the method of opening and closing, which take place inside the slider. This zipper has metal teeth, as all zippers did until improvements in plastics after World War II made it possible to manufacture zippers with nylon or polyester teeth. Each metal tooth has a projection on the top of one end and a corresponding pocket on the bottom. As the slider is raised its *V*-shaped internal neck holds the teeth at an angle that allows the projection of one tooth to slip under the pocket of the tooth on the opposite side, and then the *Y*-shaped flanges of the slider push the teeth together. The projection-pocket geometry is what holds the two sides of the zipper (the stringers) together when the zipper is closed. In opening the procedure is reversed; the slider's flanges hold the teeth at an angle that allows them to slip apart, and the neck separates them (see "The Slide Fastener," by Lewis Weiner, page 132). Each tooth is fastened to the fabric tape by legs clamped around the bead of the tape. The tape also serves for sewing the zipper to a garment.

### THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

| Page | Source |
|---|---|
| 14–19 | Edward Bell |
| 40 | Ezra Stoller, © ESTO |
| 42–49 | Albert E. Miller |
| 50 | Robert and B. Gillian Turgeon, Cornell University (*top*); Mary-Dell Chilton (*bottom*) |
| 52–57 | Bunji Tagawa |
| 58 | Jacques Tempé |
| 59 | Bunji Tagawa |
| 61 | Natural Environment Research Council, Swindon, England, and Open University |
| 62–63 | Peter Francis |
| 64–68 | Tom Prentiss |
| 70 | Peter Francis |
| 87 | Carl Mydans |
| 88–92 | Patricia J. Wynne |
| 93 | Wu Rukang and Lin Shenglong, Institute of Vertebrate Paleontology and Paleoanthropology, Peking |
| 94 | Patricia J. Wynne |
| 96–97 | National Geographic Society–Palomar Observatory Sky Survey (*top left, middle center, bottom left and bottom center*), European Southern Observatory/ Science Research Council Southern Sky Atlas (*top center and top right*), Mount Wilson and Las Campanas Observatories, courtesy Allan R. Sandage (*middle right and bottom right*), Palomar Observatory, courtesy David Burstein (*middle left*) |
| 98 | Gabor Kiss |
| 99–100 | Vera C. Rubin, Carnegie Institution of Washington (*top*); Gabor Kiss (*bottom*) |
| 101–105 | Gabor Kiss |
| 106 | Vera C. Rubin, Carnegie Institution of Washington (*left*); Gabor Kiss (*right*) |
| 108 | Gabor Kiss |
| 110 | Sam H. Ridgway |
| 112 | Alan D. Iselin |
| 114 | U.S. Navy (*top*), Sam H. Ridgway (*bottom*) |
| 115–119 | Alan D. Iselin |
| 123 | Egbert Schneider |
| 124–131 | Jerome Kuhl |
| 133–143 | Ian Worpole |
| 147 | Frank R. Seufert |
| 148–152 | Michael Goodman |







LANDS' END

# LETTERS

Sirs:

In the informative and entertaining article "The Footprints of Extinct Animals," by David J. Mossman and William A. S. Sarjeant [SCIENTIFIC AMERICAN, January], the authors cite as an example of paleoichnological misinterpretation of facts the 1834 discovery of Triassic-age tracks in the central German state of Thuringia that were originally attributed to an apelike or bearlike creature called *Chirotherium*. Later these tracks were shown to be those of a predominantly bipedal pseudosuchian reptile such as *Ticinosuchus*.

Mossman and Sarjeant also refer to sauropod tracks in the Cretaceous deposits at Paluxy Creek, Tex., discovered by the paleontologist Roland T. Bird in 1944. (Actually Bird published his first article on the Paluxy tracks in 1939.) They make no mention, however, of the alleged footprints of hominids alongside those of dinosaurs also found at Paluxy Creek in the 1930's. In view of current public debates between creationists and evolutionists it may be appropriate to set the record straight on this more famous misinterpretation.

Creationists would like to demonstrate that certain fossils are in the wrong order according to evolution theory so that they could prove the geological time scale is in error. One of their prime pieces of "evidence" in this regard has been the alleged existence of human footprints and the tracks of dinosaurs side by side in the Cretaceous limestone deposits of the Paluxy River near Glen Rose, Tex.

During the Great Depression of the 1930's some residents of Glen Rose had excavated dinosaur tracks and had fabricated other specimens, some "manlike," to sell to tourists. The fakery was disclosed in a 1939 article in *Natural History* by Bird. Subsequently other paleontologists who visited the Paluxy River site found tridactyl (three-toed) and sauropod dinosaur tracks and other Mesozoic plant and animal remains but no evidence of any Cenozoic fauna. Some of the dinosaur footprints were of such poor quality and/or so extensively eroded that they could be mistaken by lay people for "giant man prints." Some of these "man prints" have distinct claw marks protruding from what creationists call their "heel." Wann Langston, Jr., a Texas paleontologist, noted that these so-called human footprints have their "instep" along the outside edge rather than in the middle.

Some 10 years ago a film titled *Footprints in Stone* was made about the Paluxy River tracks by Films for Christ Association of Elmwood, Ill. Since the quality of the "human prints" was poor, the filmmakers decided to highlight the less obvious features such as "toes" and "sides" with shellac or oil. By stopping the film's motion Laurie R. Godfrey, a physical anthropologist at the University of Massachusetts at Amherst, has shown that the "man prints" all but disappear when the superposed outlines are eliminated. She observes that in other frames the "man print" was only a part of a larger impression or that the shellac seemed to connect erosional depressions. The film has been widely shown over the years with relatively little corresponding notice given to analyses of it from the scientific community.

According to Frederick Edwords, editor of the journal *Creation/Evolution:* "Last summer Dr. Godfrey led a research team to the site and discovered that the alleged human trackways showed no consistency. Many of the creationists' 'best' trackways featured irregular stride lengths and 'prints' that changed size and even direction with every step. The measurements she made will soon be published. Geologist Steven Schafersman discovered that worm burrows were largely responsible for the 'toes' on some of the prints."

Since *Scientific American* is widely read by both scientists (biological or otherwise) and nonscientists, it seems desirable to add this footnote to the article by Mossman and Sarjeant. Otherwise this particular myth may be perpetuated for yet another generation.

WILLIAM D. STANSFIELD

Professor, Biological Sciences Department
California Polytechnic State University
San Luis Obispo

EDITOR'S NOTE

In "The Law of the Sea," by Elisabeth Mann Borgese [SCIENTIFIC AMERICAN, March], it is stated: "Out of opposition to the Seabed Authority... the U.S., the U.S.S.R. and 15 other major industrial nations withheld their signatures from the Convention. These nations did, however, sign the Final Act of the Conference." The passage "the U.S., the U.S.S.R. and 15 other major industrial nations" should read "the U.S., the U.K. and certain other industrial nations." The error is not the responsibility of the author.

# BREAKTHROUGH: FLOAT AN EYE TO STEADY ITS GAZE.

No matter how well protected mobile cameras or telescopes may be, the motion of the vehicle carrying them causes the lenses to jitter and shake. Images are blurred and imprecise.

We discovered a way to make a moving lens sit still. By using high-energy, rare earth magnets, we suspend an electro-optical system in a magnetic field—in effect, make it float. Untouched, it remains steady while its source of transportation vibrates. This "soft ride" assures crisp, unblurred vision for an observer or a sensor.

Army scouts call it the first major advance in mobile observation since binoculars were issued to the cavalry.

We're creating breakthroughs not only in optics but also in aerospace, health care, energy and information processing.

We're McDonnell Douglas.

# THE ALL-NEW 1984 FORD TEMPO

## STYLE AND TECHNOLOGY IN TOTAL HARMONY

Available in 2-Door (above) and 4-Door (at right).

### FORM FOLLOWS FUNCTION

In its creation of Tempo, Ford set out to design a car that would prove that a spacious interior, comfort and performance could be compatible with fuel efficiency. Ford engineers began with a sound design principle: Form follows function. First came complete new functional ideas. And then Tempo's pleasing form followed—naturally.

**Designed by a world leader in aerodynamics.**

Ford is an acknowledged world leader in aerodynamic design. And Ford's aerodynamic expertise is well expressed in the '84 Tempo.

Aerodynamic design can have a profound effect on the performance of any car.

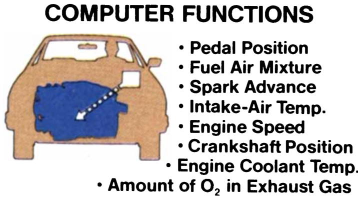
No more than six horsepower is needed for Tempo to overcome air resistance at 50 mph. This results in less fuel consumed.

And impressive fuel economy figures are not the only result. The airflow is directed to reduce lift for improved stability and directional control.

**The world's most advanced automotive computer.**

If you've ever had your car's engine buck and stall on a cold morning . . . or choke up on a hot day . . . or had *any* irregular power output due to weather, temperature or altitude—you'll appreciate the capabilities of Tempo's new EEC-IV computer.

**COMPUTER FUNCTIONS**




EEC-IV (Electronic Engine Control) monitors and controls engine operation *precisely* and *instantly* under *any* conditions for optimum power output and fuel efficiency.

**New Powertrain.**

Developed specifically for Tempo, the 2300 HSC (High Swirl Combustion) four-cylinder engine works in concert with the EEC-IV computer.

This high-compression (9-to-1) engine generates 84 horsepower at 4400 rpm.† There's a surprisingly quick power response to even slight throttle pressure. Available with 4-speed, 5-speed or automatic transaxles.

**High mileage.***

Although EPA mileage ratings were not available at the time of publication, Ford Engineering tests project 42 estimated highway and 28 estimated mpg.*

**Ride and handling.**

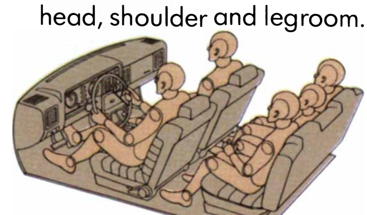
Tempo's front-wheel-drive design not only delivers added tire-to-road traction, but provides precise handling and tight cornering.

Tempo's new-design 4-wheel independent suspension permits

each wheel to absorb road shocks independently. This results in better handling and a smooth, quiet ride.

**Computer refined interior.**

Computer technology and designer ingenuity have joined forces to ensure the best use of Tempo's interior space. Its five-seat configuration allows for a comfortable combination of head, shoulder and legroom.



In fact, there is more rear head and legroom than in a Mercedes 300D.

**Commitment to quality.**

An example of Ford's commitment to quality is the closeness and uniformity of the body panel fits. You can compare Tempo's stringent standards to those set by any European Sports Sedan.

What's under Tempo's glistening paint? Three more full coats of tough acrylic enamel and two coats of primer. That's almost 10 pounds of protection. The final result: a tough, brilliant finish that is highly resistant to weather, chipping, cracking and peeling.

**Pick up the Tempo of your life.**

To obtain a brochure with more information on Ford's all-new sports sedan see your Ford Dealer or call 1-800-772-2100.

Get it together—Buckle up.

†Based on SAE standard J1349

*The above estimates are projected Ford ratings based on Ford Engineering's test data, and are expected to be very close to official EPA ratings. Use for comparison. Your mileage may differ depending on speed, distance and weather. Actual highway mileage and California ratings will probably be lower. See your Ford Dealer for a copy of the Gas Mileage Guide when available.

**HAVE YOU DRIVEN A FORD...LATELY?**

# IF PERSONAL COMPUTERS ARE FOR EVERYBODY, HOW COME THEY'RE PRICED FOR NOBODY?

**$1395*** APPLE® IIe 64K

**$999*** TRS-80® III 16K

**$1355*** IBM® PC 64K

A personal computer is supposed to be a computer for persons. Not just wealthy persons. Or whiz-kid persons. Or privileged persons.

But person persons.

In other words, all the persons whom Apple, IBM, and Radio Shack seem to have forgotten about (including, most likely, you).

But that's okay. Because now you can get a high-powered home computer without taking out a second mortgage on your home.

It's the Commodore 64. We're not talking about a low-priced computer that can barely retain a phone number. We're talking about a memory of 64K. Which means it can perform tasks most other home computers can't. Including some of those that cost a lot more. (Take another look at the three computers above.)

By itself, the Commodore 64 is all the computer you'll ever need. Yet, if you do want to expand its capabilities some day, you can do so by adding a full complement of Commodore peripherals. Such as disk drives. Modems. And printers.

You can also play terrific games on the Commodore 64. Many of which will be far more challenging than those you could ever play on a game machine alone.

And as great as all this sounds, what's even greater-sounding is the price. It's hundreds of dollars less than that of our nearest competitor.

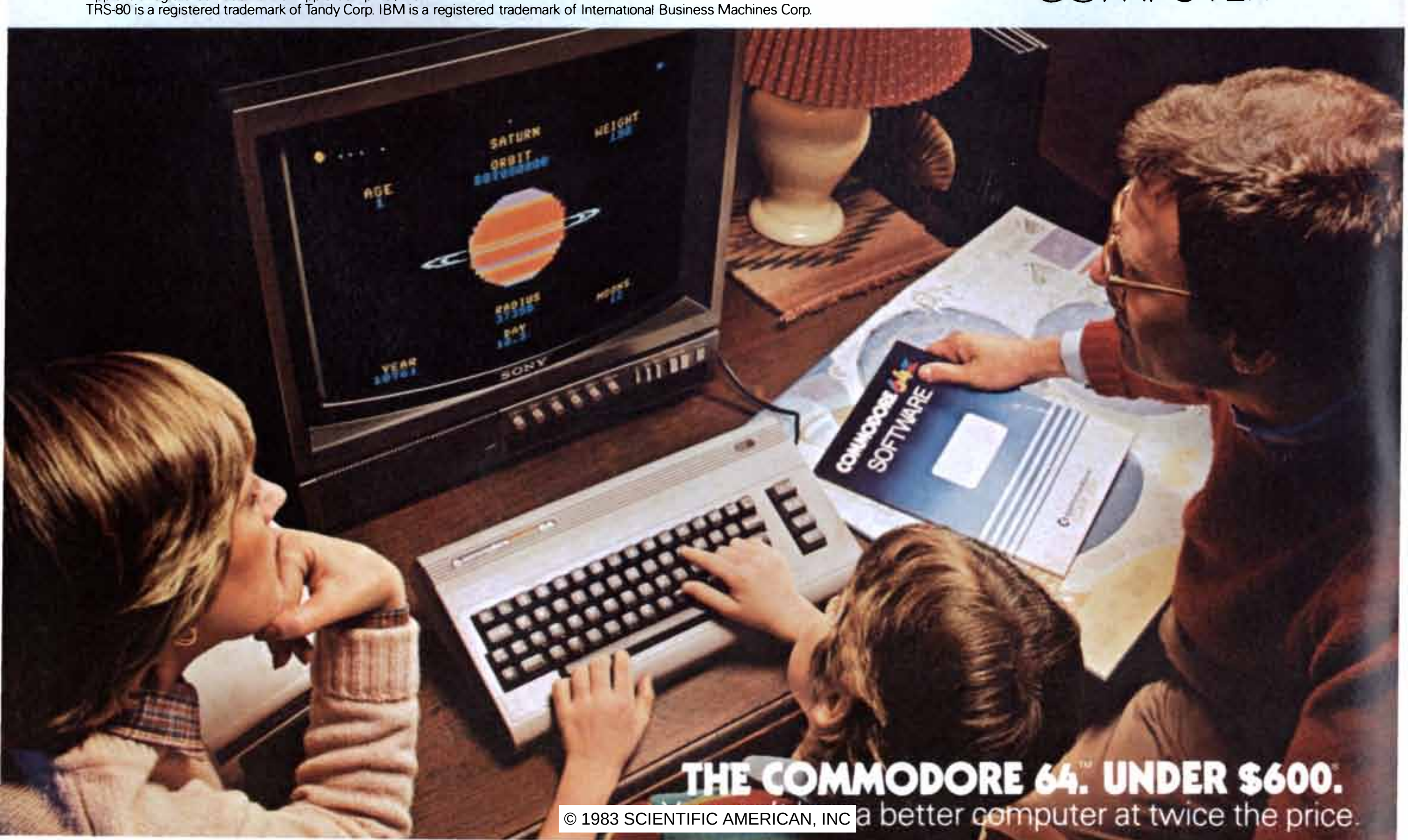
So while other companies are trying to take advantage of the computer revolution, it seems to us they're really taking advantage of something else:

Their customers.

*Manufacturers' suggested list prices as of March 20, 1983. Monitor included with TRS-80 III only. Commodore Business Machines, P.O. Box 500R, Conshohocken, PA 19428; Canada–3370 Pharmacy Avenue, Agincourt, Ont., Can. M1W 2K4.

commodore COMPUTER

Apple is a registered trademark of Apple Computer, Inc.
TRS-80 is a registered trademark of Tandy Corp. IBM is a registered trademark of International Business Machines Corp.



# 50 AND 100 YEARS AGO

## SCIENTIFIC AMERICAN

JUNE, 1933: "The disaster to the airship *Akron,* with its irreparable loss of brave and highly trained men, has shocked the entire country, and a vigorous but ill-advised clamor has arisen in the press and Congress for the discontinuance of all rigid-airship effort by the United States. It is to be hoped that not too hasty a decision will be made by Congress. The *Macon,* about to be completed, should certainly be put into service. Where long-range war operations are conducted by a fleet, the airship, with its enormous range and its ability to carry a number of fast scouting airplanes aloft, remains the most admirable source of information. Nothing can replace an airship for long-range scouting, and the Battle of Jutland would have had a far different termination had the British possessed a rigid airship of similar caliber to the one employed by the German navy."

"When Lt. Comdr. Frank M. Hawks wrote the specifications for the Texaco *Sky Chief,* he thought he was giving the manufacturer a knotty problem. He asked for an airplane that would carry fuel enough for at least 2,000 miles at well over 200 miles per hour, and he was insistent that the airplane have a landing speed not to exceed 70 m.p.h. Heretofore all fast airplanes had a landing speed in a ratio from $2\frac{1}{2}$ or 3 to 1, and Commander Hawks was asking the manufacturer to improve that ratio to something like 4 to 1. That the problem had been solved by the Northrop Corporation was immediately obvious after the new *Sky Chief* had been taken up for a few flights. Tentative performance figures showed a top speed of 250 m.p.h. and a cruising speed of more than 200 m.p.h. Performance figures showed a landing speed of from 40 to 50 m.p.h., or a ratio of 5 to 1. The improvement of the ratio was obtained by the use of large wing flaps or air 'brakes.' The lower side of the trailing edge of the wing is split from the upper side for about 20 per cent of the wing chord. Attached to the wing proper by hinges, this flap is raised or lowered by a geared shaft actuated from the cockpit."

"The remarkable growth of contraceptive clinics in all the larger cities is a very clear indication that control of human reproduction by means of contraception has reached a degree of public approbation from which it will not recede in spite of all efforts to the contrary. One of the most frequently employed of these efforts is the announcement that promiscuity and an uncontrolled orgy of sex would follow the general dissemination of contraceptive information. Another measure is deliberately to foster confusion between contraception and abortion. Criminal abortion can more correctly be considered the direct result of the *absence* of contraception. Those who attempt to prevent the spread of contraceptive information are encouraging the continuation of a social system that makes abortion not only possible but also very common."

## SCIENTIFIC AMERICAN

JUNE, 1883: "The great bridge that connects the two cities of New York and Brooklyn has been formally opened to public travel. The President of the United States, the Governor of New York and the mayors of the two cities took part in the ceremonies. On the New York side was a procession of the famous Seventh Regiment of 14 companies, preceded by a military band of 90 pieces, policemen mounted and on foot, acting as escort to the President, members of his Cabinet, Governor Cleveland and other notables, occupying 25 carriages. The President and guests walked across the bridge and were met by the Mayor of Brooklyn and other officials, and were escorted to the Sands Street bridge station by the Twenty-third Regiment. The salutes from the five naval vessels in the harbor and from Governors Island were an impressive feature in the 'pomp and circumstance' of the occasion. But the culmination of the display was reserved for the evening. It is doubtful whether there ever was, in this country at least, so magnificent an exhibition of pyrotechnics as that from the bridge. A levee held by the President and the Governor later in the evening at the Brooklyn Academy closed the public ceremonies, after which the bridge was thrown open to the public and thousands passed over from one city to the other, making a continuous procession during the entire night."

"The U.S. Board of Engineers for Fortifications has submitted a valuable report upon the practice in Europe with the heavy Armstrong, Woolwich and Krupp rifled guns. The conclusions of this report are that hammered steel in the late Spezia trials proved superior to any other material hitherto tested for armor plate. It seems probable that a hammered-steel plate, if equal in thickness to the belt armor of the *Inflexible* (22 inches), would stop the shot of the 100-ton gun (Armstrong) fired with its greatest practicable velocity. While the 12-inch rifled gun may prove a sufficient armament for the barbette batteries of our seacoast defenses against the lighter ironclads of foreign navies, iron turrets, armed with guns of 100 tons' weight at least, will be needed to meet the attack of armored ships of the latest construction. The 100-ton chambered Armstrong gun, throwing a projectile of a ton weight and fired with a charge of more than 700 pounds of powder, may be taken as a sample of the monstrous requirements of modern war."

"In the Brayton petroleum engine 10 gallons of unrefined petroleum are said to give a constant power equal to five horse power for 10 hours. Crude petroleum costs about six or possibly eight cents a gallon, making that the cost of five horse power per hour. The engine is run by the combustion of the vapor of petroleum united with atmospheric air under pressure. The motor, it is stated, has been fairly tried and appears to be constructed on reasonable principles."

"The growth of the open-hearth, or what is known as the Siemens-Martin, process of making steel has been no less remarkable than that of the Bessemer process. At present we have furnaces of a capacity from 15 to 25 tons, and by combining several furnaces single ingots weighing from 120 to 125 tons have been produced at Le Creusot in France. The world's production of open-hearth steel ingots for ship and boiler plates, propeller shafts, ordnance, wheels and axles, wire billets, armor plate, castings of various kinds and a multiplicity of other articles cannot have been less than 800,000 tons last year. The process has followed two somewhat dissimilar lines. In England iron ores of a pure quality are dissolved in a bath of pig iron, with the addition of only small quantities of scrap steel and iron. At Le Creusot large quantities of wrought iron are melted in the bath. The iron is of exceptional purity, containing less than .01 per cent of phosphorus and sulphur."

"There are not a great many points in commercial or trade matters as to which the average British manufacturer would be willing to admit the decided superiority of American concerns. But in one department those on the British side concede the unmistakable and pronounced superiority of the gentlemen on the other side. That department is advertising, which in the United States has been developed and systematized to an extent almost unknown in Britain. Advertising on a bold scale is looked upon as a *necessity* in the United States, whereas not a few British firms believe they confer a favor on the journals they patronize, if they advertise at all."

# THE AUTHORS

GERBEN DEJONG and RAYMOND LIFCHEZ ("Physical Disability and Public Policy") bring complementary perspectives to the subject of their article. DeJong is senior research associate and assistant professor of rehabilitation medicine at the Tufts University School of Medicine. He attended Calvin College, receiving his B.A. in 1968. He went on to earn two master's degrees at the University of Michigan: one in public administration (1970) and one in economics and social policy (1972). His Ph.D. in public-policy studies (1980) is from Brandeis University. From 1972 to 1975 he was director of the social-services evaluation and analysis division of the Michigan Department of Social Services. In 1976 and 1977 he was project director of the Levinson Policy Institute at Brandeis. DeJong moved to Tufts in 1978. Lifchez is associate professor of architecture at the University of California at Berkeley and an architect in private practice. He was graduated from the University of Florida in 1955 with a bachelor's degree. He holds three master's degrees: in architecture (1957) and art history (1967) from Columbia University and in city planning from Berkeley (1972). From 1964 to 1969 he was a member of the faculty at Columbia, moving to Berkeley in 1971. He has worked for several architectural firms, including that of Eero Saarinen. He is the coauthor with Barbara S. Winslow of *Design for Independent Living: The Environment and Physically Disabled People* (University of California Press, 1982).

MARY-DELL CHILTON ("A Vector for Introducing New Genes into Plants") writes: "My undergraduate studies at the University of Illinois at Urbana-Champaign were in the chemistry curriculum. My graduate studies (also at Illinois) started in organic chemistry, but in taking courses for my two minors (biochemistry and microbiology) I found out about DNA and have had an unwavering interest in it ever since. As a postdoctoral student at the University of Washington I continued to work on the DNA of bacteria, mice and maize. I became a faculty member at the University of Washington in microbiology in 1970. In 1979 I accepted a position as associate professor of biology at Washington University in St. Louis, where interaction with scientists at Monsanto led my interests in the direction of the genetic engineering of higher plants. Last month my career took a new turn: I became executive director of agricultural biotechnology for CIBA-Geigy. I shall be responsible for the development of a new research unit in North Carolina, one of whose goals will be to exploit the gene-vector system described in this article."

PETER FRANCIS ("Giant Volcanic Calderas") was born in Zambia, which was then called Northern Rhodesia. He went to England to continue his education, attending the Imperial College of Science and Technology as an undergraduate and a graduate student. His Ph.D. in geology was granted by Imperial College in 1969. Since 1971 he has been a member of the faculty of the department of earth sciences at the Open University. Since 1981 he has served as senior visiting scientist at the Lunar and Planetary Institute in Houston, where he works on remote-sensing techniques for the investigation of volcanic terrain. Francis has led expeditions into volcanic territory in central America, South America and Africa.

WU RUKANG and LIN SHENGLONG ("Peking Man") are paleoanthropologists. Both are natives of China, and the anthropology of their country is their main scientific interest. Wu was graduated from the Central University of Chongqing with a B.S. in 1940. He moved to the U.S. in 1946, getting his M.S. (1947) and Ph.D. (1949) from Washington University in St. Louis. After earning his doctorate he returned to China, becoming director of the anatomy department at the Dalian Medical College. In 1956 he went to the Institute of Vertebrate Paleontology and Paleoanthropology of the Academy of Sciences of China as research fellow; he is currently vice-director of the institute. Wu is a member of the Academy of Sciences and president of the Association of Anatomical Sciences. Lin studied archaeology at Beijing University. He joined the faculty of the University of Inner Mongolia in 1959, leaving in 1962 to continue his graduate studies at the Institute of Vertebrate Paleontology and Paleoanthropology. He is currently assistant research fellow there.

VERA C. RUBIN ("Dark Matter in Spiral Galaxies") writes: "For the past 18 years I have been a staff member of the Department of Terrestrial Magnetism of the Carnegie Institution of Washington, where I work on the dynamics of galaxies. I am also adjunct staff member of the Mount Wilson and Las Campanas Observatories. I obtained degrees in astronomy from Vassar College (a B.A. in 1948), Cornell University (an M.A. in 1951) and Georgetown University (a Ph.D. in 1954). For my M.A. thesis I searched for a large-scale systemic motion of all the galaxies whose radial velocities were then known. For my Ph.D. thesis (under George Gamow) I studied the spatial distribution of galaxies. Both subjects, which are closely related to the work I currently do, were not widely studied again until the 1970's, when more data and large computers became available. My life is busy with travel to observatories, particularly Kitt Peak near Tucson and Cerro Tololo and Las Campanas in Chile."

JOHN W. KANWISHER and SAM H. RIDGWAY ("The Physiological Ecology of Whales and Porpoises") are physiologists who are interested in diving both by human beings and by marine mammals. Kanwisher was trained in biophysics. His Ph.D. in the subject was granted in 1951 by the University of Rochester. In 1952 at a program on Bimini Island he met the comparative physiologist P. F. Scholander. He writes that as a result of the meeting "my formal education in physics and engineering was instantly put aside for physiology." He has taught biological oceanography at Harvard University and the Massachusetts Institute of Technology. Kanwisher spends much of his time as a consulting engineer, and he has patented several devices to enable divers to remain longer below the surface. Ridgway is head of the Biomedical Division of the Naval Undersea Center. His B.S. (1958) and D.V.M. degrees are from Texas A&M University. He went on to get a doctorate in neurobiology from the University of Cambridge in 1973. He has served as research veterinarian at several institutions, including the U.S. Naval Missile Center, the University of Southern California and the University of California at Santa Barbara.

JOSEPH W. DAUBEN ("Georg Cantor and the Origins of Transfinite Set Theory") is professor of history and of the history of science at the Herbert H. Lehman College of the City University of New York. He got his bachelor's degree at Claremont College in 1966. His Ph.D., awarded in 1972, is from Harvard University. In 1973 and 1974 he was at the American Academy in Rome, where he studied mathematics, perspective and art in the Italian Renaissance. In 1977 and 1978 he was at the Institute for Advanced Study. In 1980 he was Mead-Swing Visiting Scholar at Oberlin College. In 1981 he was a visiting scholar at Harvard.

LEWIS WEINER ("The Slide Fastener") is a consulting engineer much of whose work is related to the subject of his article. Born in Czechoslovakia, he was educated in Prague at the Charles University. He came to the U.S. in 1941. He holds several patents for devices employed in the manufacture of slide fasteners.

# METAMAGICAL THEMAS

*The calculus of cooperation is tested through a lottery*

by Douglas R. Hofstadter

And then one fine day out of the blue you get a letter from S. N. Platonia, the well-known Oklahoma oil trillionaire, mentioning that 20 leading rational thinkers have been selected to participate in a little game. "You are one of them," it says. "Each of you has a chance at winning a billion dollars, put up by the Platonia Institute for the Study of Human Irrationality. Here's how. If you wish, you may send a telegram with just your name on it to the Platonia Institute in Frogville, Okla. You may reverse the charges. If you reply within 48 hours, the billion is yours—unless there are two or more replies, in which case the prize is awarded to no one. And if no one replies, nothing will be awarded to anyone."

You have no way of knowing who the other 19 participants are; indeed, Platonia's letter states that the entire offer will be rescinded if it is detected that an attempt has been made by any participant to learn the identity of, or to establish contact with, any other participant. Moreover, it is a condition that the winner (if there is one) must agree in writing not to share the prize money with any other participant at any time in the future. This is to squelch any thoughts of cooperation, either before or after the prize is given out.

The brutal fact is that no one will know what anyone else is doing. Clearly everyone will want that billion. Clearly all will realize that if their name is *not* submitted, they will have no chance at all. Does this mean that 20 telegrams will arrive in Frogville, showing that even having transcendent levels of rationality—as you of course do—is of no help in such an excruciating situation?

This is the Platonia Dilemma, a little scenario I thought up recently in trying to get a better handle on the Prisoner's Dilemma, of which I wrote last month. The Prisoner's Dilemma can be formulated in terms resembling this dilemma. Imagine you receive a letter from the Platonia Institute telling you that you and just *one* other anonymous leading rational thinker have been selected for a modest cash giveaway. As before, both of you are asked to reply by telegram within 48 hours to the Platonia Institute, charges reversed. Your telegram is to include, apart from your name, just the word "cooperate" or the word "defect." If both of you wire "cooperate," you will each get $3. If both of you wire "defect," you will each get $1. If one of each word is received, the cooperator gets nothing and the defector gets $5.

What choice would you make? It would be nice if you both cooperated, so that you would each get $3, but does it not seem a bit unlikely? After all, who wants to get suckered by a nasty defector who gets $5 for being sneaky? Certainly not *you!* Therefore you would probably decide not to cooperate. It seems a regrettable but necessary choice. Of course, both of you, reasoning alike, come to the same conclusion. Hence you will both defect, and so you will get a mere dollar apiece. If, however, both of you had just been willing to take a chance, you could have got $3 apiece. What a pity!

It was my discomfort with this seemingly logical analysis of the "One-Round Prisoner's Dilemma" that led me to formulate the following letter, which I sent to 20 friends:

"Dear ——:

I am sending this letter by Special Delivery to 20 of you (namely various friends of mine around the country). I am proposing to all of you a one-round Prisoner's Dilemma game, the payoffs to be monetary (provided by *Scientific American*). It is very simple. Here is how it goes.

"Each of you is to give me a single letter: *C* or *D*, standing for 'cooperate' or 'defect.' This will be used as your move in a Prisoner's Dilemma with *each* of the 19 other players. The payoff matrix I am using for the Prisoner's Dilemma is the one enclosed numbered *1* [*see illustration on this page*].

"Thus if everyone sends in *C*, everyone will get $57, whereas if everyone sends in *D*, everyone will get $19. You can't lose! And of course anyone who sends in *D* will get at least as much as everyone else. If, for example, 11 people send in *C* and nine send in *D*, then the 11 *C*-ers will get $3 apiece from each of the other *C*-ers (making $30) and will get nothing from the *D*-ers. Therefore *C*-ers will get $30 each. The *D*-ers, in contrast, will pick up $5 apiece from each of the *C*-ers (making $55) and will get $1 from each of the other *D*-ers (making $8), for a grand total of $63. No matter what the distribution is, *D*-ers always do better than *C*-ers. Of course, the more *C*-ers there are, the better *everyone* will do!

"By the way, I should make it clear that in making your choice you should not aim to be the *winner* but simply to get as much *money* for yourself as possible. Thus you should be happier to get $30 (say as a result of saying *C* along with 10 others, even though the nine *D*-sayers get more than you) than to get $19 (by saying *D* along with everyone else, so that nobody 'beat' you). Furthermore, you are not supposed to think that at some later time you will meet with and be able to share the goods with your coparticipants. You are not aiming at maximizing the total number of dollars *Scientific American* shells out, only at maximizing the number of dollars that come to *you!*

"Of course, your hope is to be the *unique* defector, thereby really cleaning up: with 19 *C*-ers, you will get $95 and they will each get 18 times $3, namely $54. But why am I doing the multiplication or any of this figuring for you? You are very bright. So are the others. All about equally bright, I would say. Therefore all you need to do is tell me your choice. I want all the answers by telephone (call collect, please) *the day you receive this letter*.

"It is to be understood (it *almost* goes without saying, but not quite) that you are not to try to consult with others who

| | | Player *B* | |
|---|---|---|---|
| | | Cooperates | Defects |
| Player *A* | Cooperates | (3,3) | (0,5) |
| | Defects | (5,0) | (1,1) |

Where (*x*, *y*) means Player *A* gets *x* dollars and Player *B* gets *y* dollars.

*A payoff matrix for the Prisoner's Dilemma*

# INTRODUCING THE PERFORMANCE SEDAN WITH A GENETIC ADVANTAGE.

Fifteen years ago, when big was beautiful and the auto industry was churning out ungainly V-8's, BMW introduced the 2002. A performance car.

Nine years later, big was out and the future was conceded to the tiny economy car. BMW introduced the 320i. A performance car.

Today the auto industry has embraced performance as the wave of the present. And BMW has introduced the 318i. A performance car—but one with an important difference.

It arrives with a heritage, as opposed to a history of recurring identity crises.

WHY A CAR'S PAST IS ITS FUTURE.

All cars, especially those aspiring to high performance, are products of engineering, technology—and genealogy.

The BMW 318i enjoys a substantial head start on all counts.

Its engine anticipated by years the current preference for fuel injection. The BMW 3-Series power plant serves as the basis for the BMW Formula One race engine—modified on the 318i by a new L-Jetronic system that generates even greater quickness and torque.

Its gearbox has been refined to match the greater torque, while retaining the characteristic precision that has inspired comparison to the linkage of fine timepieces.

Its new independent suspension is a virtual syllabus of lessons learned on racecourses, delivering handling and road holding Car and Driver judged "magnificent."

It offers a new graduated power-assist steering system, larger disc brakes, and such technological innovations as an electronic system that actually calculates when routine servicing is needed.

It is also among the automotive leaders in incongruity, having managed to boost fuel efficiency* at the same time it was improving its performance.

"THE FINEST QUALITY CAR AVAILABLE."

Inside, the 318i rebuts the myth that a noisy, bone-rattling ride under cramped conditions somehow certifies a vehicle as a true performance car.

It is roomier than its predecessors and well equipped with noise- and vibration-dampening measures. It also offers orthopedically-designed bucket seats and other functional amenities bearing the imprint of its racing lineage.

Its "overall ergonomics," summed up AutoWeek, "demonstrate why many consider the BMW the finest quality car available on the market today."

In short, the 318i continues and advances a 55-year commitment to performance. As today's fledgling performance cars will discover, there is no available shortcut to that experience.

If you see no need to discover the point along with them, we invite you to visit your nearest BMW dealer and test drive the 318i.

**THE ULTIMATE DRIVING MACHINE.**

*EPA-estimated 27 mpg, 38 highway. Figures are for comparison purposes only. Your actual mileage may vary, depending on speed, weather and trip length; actual highway mileage will most likely be lower.  European Tourist Delivery can be arranged through your authorized U.S. BMW dealer.

# ONLY

## ATARI GIVES YOU SUCH EASILY ACCESSIBLE GRAPHICS AND SOUND CAPABILITIES IN A 64K COMPUTER.

***ONLY*** *the ATARI 1200XL Home Computer combines 64K RAM computing power with outstanding graphics and sound capabilities that are so easy to access. You can create sophisticated graphics and sound effects using quick and simple commands. This allows you to write complex programs in much less time than with many other 64K computers.*

***ONLY*** *the ATARI 1200XL Home Computer has a specially designed keyboard which is fully redefinable and features 8 programmable function keys controlling 16 functions. That's twice as many as the Commodore 64. In addition, the ATARI 1200XL keyboard locks and unlocks electronically to protect your programs.*

***ONLY*** *the ATARI 1200XL Home Computer provides a revolutionary "help" key that performs two important functions. First, it can provide the user with helpful instructions on selected programs used with the computer. Second, the "help"*

# THE NEW

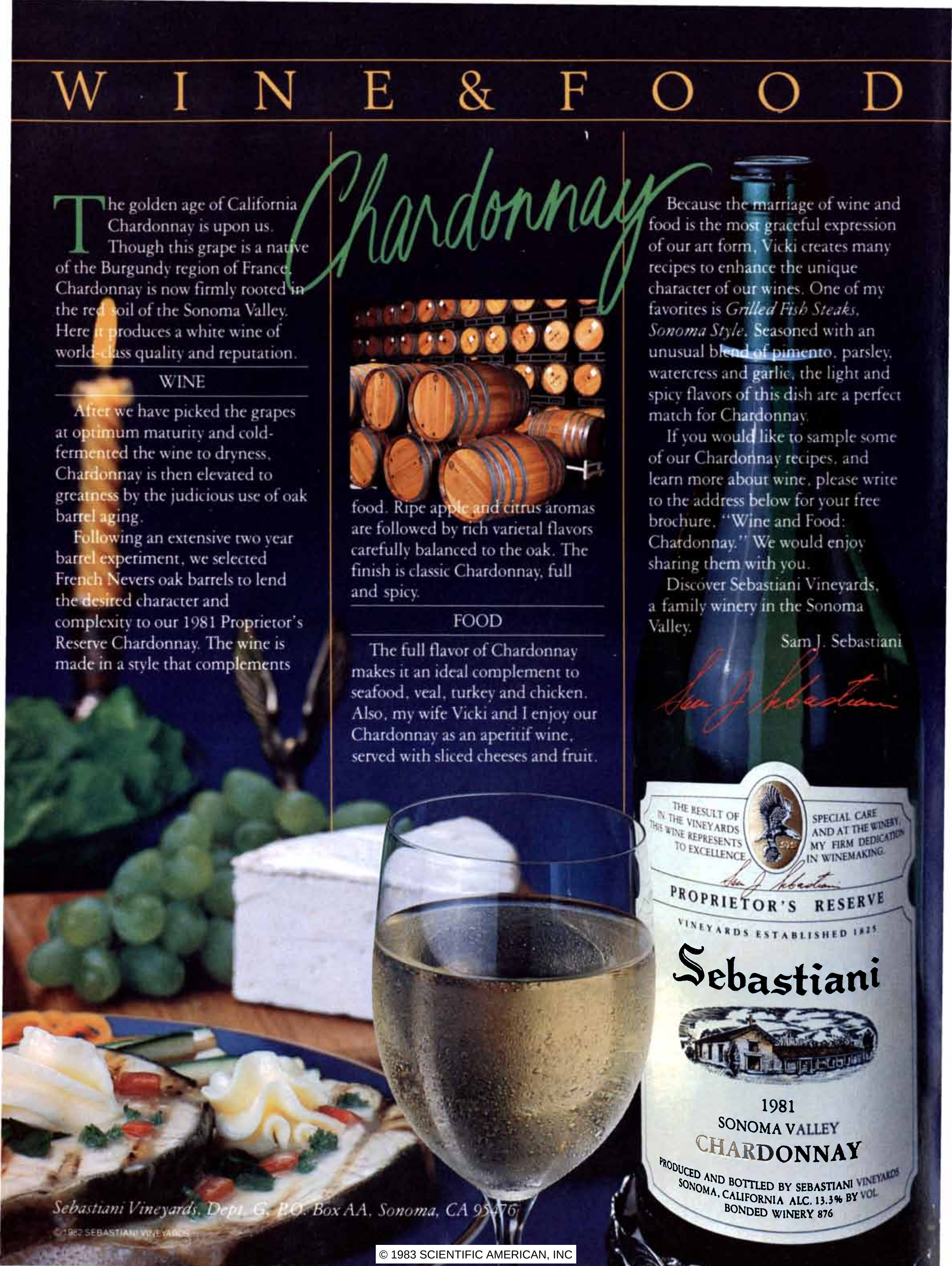
*key performs self-diagnosis—as requested by the user on the memory, audio visual circuitry and keyboard—to ensure that all components are in peak operating condition. And if you need additional help only Atari gives you a toll-free "help line" to call (800) 538-8543; in California 1-(800) 672-1404.*

***ONLY*** *the ATARI 1200XL offers you a home computer compatible with virtually all ATARI Computer peripherals and software (compatibility that other new computers like the Commodore 64 don't offer). There are over 2,000 programs currently available for the ATARI 1200XL, including the new AtariWriter™ word processing system and VisiCalc.®**

***ONLY*** *Atari puts so much more in the new 1200XL Home Computer so you get so much more out of it.*


# ATARI® 1200XL™

## HOME COMPUTER



WINE & FOOD
Chardonnay
The golden age of California Chardonnay is upon us. Though this grape is a native of the Burgundy region of France, Chardonnay is now firmly rooted in the red soil of the Sonoma Valley. Here it produces a white wine of world-class quality and reputation.
WINE
After we have picked the grapes at optimum maturity and cold-fermented the wine to dryness, Chardonnay is then elevated to greatness by the judicious use of oak barrel aging.
Following an extensive two year barrel experiment, we selected French Nevers oak barrels to lend the desired character and complexity to our 1981 Proprietor's Reserve Chardonnay. The wine is made in a style that complements food. Ripe apple and citrus aromas are followed by rich varietal flavors carefully balanced to the oak. The finish is classic Chardonnay, full and spicy.
FOOD
The full flavor of Chardonnay makes it an ideal complement to seafood, veal, turkey and chicken. Also, my wife Vicki and I enjoy our Chardonnay as an aperitif wine, served with sliced cheeses and fruit.
Because the marriage of wine and food is the most graceful expression of our art form, Vicki creates many recipes to enhance the unique character of our wines. One of my favorites is *Grilled Fish Steaks, Sonoma Style*. Seasoned with an unusual blend of pimento, parsley, watercress and garlic, the light and spicy flavors of this dish are a perfect match for Chardonnay.
If you would like to sample some of our Chardonnay recipes, and learn more about wine, please write to the address below for your free brochure, "Wine and Food: Chardonnay." We would enjoy sharing them with you.
Discover Sebastiani Vineyards, a family winery in the Sonoma Valley.
Sam J. Sebastiani
THE RESULT OF IN THE VINEYARDS THIS WINE REPRESENTS TO EXCELLENCE
SPECIAL CARE AND AT THE WINERY MY FIRM DEDICATION IN WINEMAKING.
PROPRIETOR'S RESERVE
VINEYARDS ESTABLISHED 1825
Sebastiani
1981
SONOMA VALLEY
CHARDONNAY
PRODUCED AND BOTTLED BY SEBASTIANI VINEYARDS SONOMA, CALIFORNIA ALC. 13.3% BY VOL.
BONDED WINERY 876
*Sebastiani Vineyards, Dept. G, P.O. Box AA, Sonoma, CA 95476*
©1982 SEBASTIANI VINEYARDS

you guess have been asked to participate. In fact, please consult with no one at all. The purpose is to see what people will do on their own, in isolation. Finally, I would appreciate a short statement to go along with your choice, telling me *why* you made this particular one.

Yours,
Doug H.

"P.S. By the way, it may be helpful for you to imagine a related situation, the same as the present one except that you are told all the other players have *already* submitted their choice (say a week ago) and so you are the last. Now what do you do? Do you submit *D*, knowing full well that *their* answers are already committed to paper? Now suppose immediately after having submitted your *D* (or your *C*) in that circumstance, you are told that actually the others really *haven't* submitted their answers yet but that they are all doing it today. Would you retract your answer? Or what if you knew (or at least were told) that you were the *first* person being asked for an answer?

"And, one last thing to ponder, what would you do if the payoff matrix looked like the one enclosed labeled *2* [*see illustration on this page*]?"

I want to stress that this situation is not an *iterated* Prisoner's Dilemma (discussed in last month's column). It is a one-shot, multiperson Prisoner's Dilemma. There is no possibility of learning over a period of time anything about how the others are inclined to play. Therefore all the lessons described last month are not applicable here, since they depend on the situation's being iterated. All that each recipient of my letter could go on was this thought: There are 19 people out there, somewhat like me, all in the same boat, all grappling with the same issues as I am. In other words, there was nothing to rely on except pure reason.

I had much fun preparing the letter, deciding whom to send it to, anticipating the responses and then receiving them. Before I reveal the results I invite you to think about how you would play in such a contest. I should particularly like you to take seriously the assertion that everyone is very bright. In fact, let me expand on this idea, since I thought people perhaps did not really understand what I meant by it. Thus please consider the letter to contain the following clarifying paragraph:

"All of you are very rational people. Therefore I hardly need tell you that you are to make what you consider to be your maximally rational choice. In particular, feelings of morality, guilt, vague malaise and so on are to be disregarded. Reasoning alone (of course including reasoning about the others' reasoning) should be the basis of your decision. And please remember that everyone is being told this (including *this!*)!"

I was hoping for—and expecting—a particular outcome to the experiment. As I received the replies by telephone over the next few days I jotted down notes so that I had a record of what impelled various people to choose as they did. The result was not what I had expected; in fact, my friends "faked me out" considerably. We got into heated arguments about the "rational" thing to do, and everyone expressed much interest in the entire question.

I should like to quote some of the feelings expressed by my friends caught in this deliciously tricky situation. David Policansky opened his call tersely by saying, "O.K., Hofstadter, give me the $19!" Then he presented this argument for defecting: "What you're asking us to do, in effect, is to press one of two buttons, knowing nothing except that if we press button *D*, we'll get more than if we press button *C*. Therefore *D* is better. That is the essence of my argument. I defect."

Martin Gardner (yes, I asked him to participate) vividly expressed the mental turmoil he and many others went through. "Horrible dilemma," he said, "I really don't know what to do about it. If I wanted to maximize my money, I would choose *D* and expect that others would also; to maximize my satisfaction I'd choose *C* and hope other people would do the same (by the Kantian imperative). I don't know, though, how one should behave *rationally*. You get into endless regresses: 'If they all do *X*, then I should do *Y*, but then they'll anticipate that and do *Z*, and so....' You get trapped in an endless whirlpool. It's like Newcomb's paradox." So saying, Martin defected, with a sigh of regret.

In a way echoing Martin's feelings of confusion, Chris Morgan said: "More by intuition than by anything else, I'm coming to the conclusion that there's no way to deal with the paradoxes inherent in this situation. So I've decided to flip a coin, because I can't anticipate what the others are going to do. I think—but can't know—that they're all going to negate one another." Therefore while Chris was on the telephone he flipped a coin and "chose" to cooperate.

Sidney Nagel was very displeased with his conclusion. "I actually couldn't sleep last night because I was thinking about it. I *wanted* to be a cooperator, but I couldn't find any way of justifying it. The way I figured it, what I do isn't going to affect what anyone else does. I might as well consider that everything else is already fixed, in which case the best I can do for myself is to play a *D*."

Robert Axelrod, whose work proves the superiority of cooperative strategies in the *iterated* Prisoner's Dilemma,

| | | Player *B* | |
|---|---|---|---|
| | | Cooperates | Defects |
| Player *A* | Cooperates | (3,3) | (0,50) |
| | Defects | (50,0) | (.01,.01) |

*A second payoff matrix for the Dilemma*

saw no reason whatever to cooperate in a one-shot game and defected without compunction.

Dorothy Denning was brief: "I figure that if I defect, then I always do at least as well as I would have done if I had cooperated. And so I defect." She was one of the people who faked me out. Her husband, Peter, cooperated. I had predicted the reverse.

By now you have probably been counting. So far I have mentioned five defectors and two cooperators. Suppose you had been me and had got roughly a third of the calls, and they were five to two in favor of defection. Would you dare to extrapolate these statistics to roughly 14 to six? How in the world can seven individuals' choices have anything to do with 13 *other* individuals' choices? As Nagel said, certainly one choice cannot influence another (unless you believe in some kind of telepathic transmission, a possibility we shall discount here). What justification might there be for extrapolating these results?

Clearly any such justification would rely on the idea that people are in some sense "like" one another. It would rely on the idea that in complex and tricky decisions such as this one people will resort to a cluster of reasons, images, prejudices and vague notions, some of which will tend to push them one way, others of which the other way but whose overall impact will be to push a certain percentage of people toward one alternative and another percentage toward the other. In advance you cannot hope to predict what those percentages will be, but given a sample of people in the situation you can hope that their decisions will be "typical." Thus the notion that early returns running five to two in favor of defection can be extrapolated to a final result of 14 to six (or so) would be based on assuming that the seven people are acting "typically" for people confronted with these conflicting mental pressures.

The snag is that the mental pressures are not completely explicit; they are evoked by, but not totally spelled out by, the wording of the letter. Each person brings a unique set of images and associations to each word and concept, and it is the set of those images and associations that will collectively create in that

*Bell System people are ready for new directions.*

# "WE'RE NOT JUST FACING CHANGE. WE'RE MAKING IT."

*—Ian M. Ross, President, Bell Laboratories.*

Ever since AT&T announced its planned divestiture of the 22 Bell telephone companies, I've been asked how the coming change will affect Bell Labs. Will we remain as creative and innovative? Will we remain intact? Will we retain those unique research and development capabilities responsible for our technological leadership and contributions to the nation?

My answer to each of those questions is yes.

I'm not suggesting that the challenges ahead are easy ones. But divestiture does give us more freedom to innovate, to apply our resources in new ways and in new markets. We'll be able to follow our technology wherever it leads. We intend to make the most of those opportunities.

Bell Labs as an institution will remain intact, although some of our employees have been reassigned to AT&T's American Bell subsidiary and others will be transferred to serve the divested companies. The major portion of our resources will continue to be devoted to developing new and improved systems and services for local telephone companies and their customers. We will also continue to provide the innovations to assure that AT&T's interexchange network remains the most flexible and intelligent in the world.

Change is not new to us. We're an organization that creates change. We look forward to it. We thrive on it.

But we also face the future with the strengths of our past. We've earned an international reputation as a result of the depth and breadth of our work, our ability to attract and nurture outstanding scientific and engineering talent and our ability to inspire innovation. The wealth of creativity at Bell Labs never ceases to amaze me.

These same qualities are also responsible for the design and development of the world's finest telecommunications system. I'm proud of our role in delivering that system. I'm also proud that much of the technology of the Information Age is rooted in innovations conceived and developed by Bell Labs.

We're committed to making America's telephone system a total communications and information system. Our innovations will help AT&T maintain its leadership in these high technology fields. They will also provide the divested companies with the technologies needed to make them the true gateways to the Information Age.

We have a commitment to quality and reliability. In a world scrambling for quick markets and short-term profits, Bell Labs will continue to set standards of excellence. We believe that the increased competition that comes with divestiture will prove that quality and reliability will ultimately determine success.

As we have in the past, Bell Labs will continue to explore and expand the frontiers of research. To discover and develop new knowledge and new technology. To contribute to America's strength, economic growth and ability to meet any challenges from anyone.

At Bell Labs, we're not just facing change. We're making it.

Bell System

person's mind a set of mental stresses like the set of stresses in the earth's crust in an earthquake zone. When people make their decision, you find out how all those stresses working in different directions add up, like a set of force vectors pushing in various directions with their strengths influenced by private or unmeasurable factors. The assumption that it is valid to extrapolate has to be based on the idea that everyone is alike inside, only with somewhat different weights attached to certain notions.

This way each person's decision can be likened to a geophysical experiment whose goal is to predict when and where an earthquake will occur. You set up a model of the earth's crust and you put in data representing your best understanding of the internal stresses. You know there are large uncertainties in your knowledge, and so you just have to choose what seem to be "reasonable" values for the variables. Therefore no single run of your simulation will have strong predictive power, but that's all right. You run it and you get a fault line telling you where the simulated earth's crust shifts. Then you go back and choose other values in the ranges of those variables and rerun the entire thing. If you rerun it repeatedly, eventually a pattern will emerge revealing where and how the crust is likely to shift and where it is rock solid.

This kind of simulation depends on an essential principle of statistics: the idea that when you let variables take on a few sample random values in their ranges, the overall outcome determined by a cluster of such variables will begin to emerge after a few trials and soon will give you an accurate model. You do not need to run your simulation millions of times to see valid trends emerging.

That is clearly the kind of assumption television networks make when they predict national election results on the basis of early returns from a few select towns in the East. Certainly they do not think that free will is any "freer" in the East than it is in the West, that whatever the East chooses to do the West will follow. It is just that the cluster of emotional and intellectual pressures on voters is much the same all over the nation. Obviously no one individual can be taken as representing the entire nation, but a well-selected group of residents of the East Coast can be assumed to be representative of the entire nation in terms of how much they are "pushed" by the various pressures of the election, so that their choices are likely to show general trends of the larger electorate.

Suppose it turned out that New Hampshire's Belknap County and California's Modoc County had produced, over many national elections, very similar results. Would it follow that one of the two counties had been exerting some kind of causal influence on the other? Would they have had to be in some kind of eerie cosmic resonance mediated by "sympathetic magic" for this to happen? Certainly not. All it takes is for the electorates of the two counties to be similar; then the pressures that determine how people vote will take over and automatically make the results come out similar. It is no more mysterious than the observation that a Belknap County schoolgirl and a Modoc County schoolboy will get the same answer when they are asked to divide 507 by 13: the laws of arithmetic are the same the world over, and they operate in the same way in remote minds without any need for "sympathetic magic."

All of this is elementary common sense; it should be the kind of thing any well-educated person understands automatically. And yet emotionally it cannot help but seem a little peculiar, since it flies in the face of the concept of free will and regards people's decisions as being simply the outcome of combinations of pressures with unknown values. On the other hand, perhaps it is better to look at decisions that way than it is to attribute them to free will, a philosophically murky notion at best.

That may have seemed like a digression about statistics and the question of individual actions v. the predictability of group behavior, but as a matter of fact it has plenty to do with the "correct action" to take in the dilemma of my letter. The question we were considering is: To what extent can what a *few* people do be taken as an indication of what *all* people will do? We can sharpen it: To what extent can what *one* person does be taken as an indication of what *all* people will do? The ultimate version of this question, stated in the first person, has a funny twist to it: To what extent does *my* choice tell me about the choices of the other participants?

You might think that each person is completely unique and therefore that no one can be relied on as a predictor of how other people will act, particularly in a situation involving an intense dilemma. There is more to the story, however. I tried to engineer the situation so that everyone would have the same image of it. In the dead center of the image was supposed to be the notion that everyone in the situation was using *reasoning alone*—including reasoning about the reasoning—to come to an answer.

Now, if reasoning dictates an answer, then everyone should independently come to that answer (just as the Belknap County schoolgirl and the Modoc County schoolboy would independently get 39 as their answer to the division problem). Seeing this fact is itself the critical step in the reasoning toward the correct answer, but unfortunately it eluded nearly everyone to whom I sent the letter. (That is why I came to wish I had included in the letter a paragraph stressing the rationality of the players.) Once you realize this fact it dawns on you that *either* all rational players will choose *D or* all rational players will choose *C*. That is the crux.

Any number of ideal rational thinkers faced with the same situation and suffering through similar throes of reasoning agony will necessarily come up with the same answer eventually, as long as reasoning alone is the ultimate justification for their conclusion. Otherwise reasoning would be subjective, not objective as arithmetic is. A conclusion reached by reasoning would be a matter of preference, not one of necessity. Now *some* people may believe this of reasoning, but rational thinkers understand that a valid argument must be *universally* compelling, otherwise it is simply not a valid argument.

If you grant this, then you are 90 percent of the way. All you need ask now is: "Since we are all going to submit the same letter, which one would be more logical? That is, which world is better for the *individual* rational thinker: one with all cooperators or one with all defectors?" The answer is immediate: "I get $57 if we all cooperate, $19 if we all defect. Clearly I prefer $57, hence cooperating is preferred by this particular rational thinker. Since I am typical, cooperating must be preferred by *all* rational thinkers. And so I'll cooperate." Another way of stating it, making it sound weirder, is this: "If I choose *C*, then everyone will choose *C*, so that I'll get $57. If I choose *D*, then everyone will choose *D*, so that I'll get $19. I'd rather have $57 than $19, and so I'll choose *C*. Then everyone will, and I'll get $57."

To many people this sounds like a belief in voodoo or sympathetic magic, a vision of a universe permeated by tenuous threads of synchronicity, conveying thoughts from mind to mind like pneumatic tubes carrying messages across Paris, and making people resonate to a secret harmony. Nothing could be further from the truth. This solution depends in no way on telepathy or bizarre forms of causality. It is just that the statement "I'll choose *C* and then everyone will," although entirely correct, is somewhat misleadingly phrased. It involves the word "choice," which is incompatible with the compelling quality of logic. Schoolchildren do not *choose* what 507 divided by 13 is; they figure it out. Analogously, my letter really did not allow choice; it demanded reasoning. Thus a better way to phrase the "voodoo" statement would be this: "If reasoning guides *me* to say *C*, then, since I am no different from anyone else as far as rational thinking is con-

# THE TECHNOLOGY OF THE CENTURY.

## THE 1983 BUICK CENTURY.

With today's technology, almost anything is possible. Computers that talk. Computers that teach. Electronics that put Wall Street right in your living room. And the Buick Century. A car so technologically advanced, it makes driving a pleasure instead of a chore.

The way Century accomplishes this is quite remarkable. There's an available electronic digital instrument panel, for example, that's a lot more than needles and dials. It tells you mph, km/h, fuel level, even how far you've traveled. All in computerized digital display.

Century's standard engine is electronically fuel injected. And connected to front-wheel drive. Its ride and handling are computer tuned. And all of this is surrounded by an aerodynamic shape that was refined in the wind tunnel, and further refined by computer.

Then there are, of course, the little things. Like an electronic tone for the ignition key and seat belt, instead of a buzzer. Why, Century is even built with robotic assembly techniques.

Happily, there remains one human quality that technology, even the technology of the Century, can never replace. The sheer pleasure of driving a Buick. And that's something you can experience at your Buick dealer now.

EST. HWY. 39 EPA EST. MPG 24

Use estimated MPG for comparison. Your mileage may differ depending on speed, distance, weather. Actual highway mileage lower. Some Buicks are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your Buick dealer for details.

*Official Car of the XXIIIrd Olympiad Los Angeles 1984*

BUICK

**Wouldn't you really rather have a Buick?**




cerned, it will guide everyone to say *C*."

The corresponding foray into the opposite world ("If I choose *D*, then everyone will choose *D*") can be understood more clearly by likening it to a musing by the Belknap County schoolgirl before she divides: H'm, I'd guess that 13 into 507 is about 49—maybe 39. I see I'll have to calculate it out. But I know in advance that if I find it is 49, then sure as shooting that Modoc County kid will write 49 on his paper as well, and if I get 39 as my answer, then so will he. No secret transmissions are involved; all that is needed is the universality and uniformity of arithmetic. Similarly, the argument "Whatever I do everyone else will do" is simply a statement of faith that reasoning is universal, at least among rational thinkers, not an endorsement of any mystical kind of causality.

That analysis shows why you should cooperate even when the opaque envelopes containing the other players' answers are there on the table in front of you. Faced so concretely with this unalterable set of *C*'s and *D*'s, you might think: Whatever they have done, I am better off playing *D* than playing *C*, because certainly what I *now* choose can have no retroactive effect on what they chose. And so I defect. Such a thought, however, assumes that the logic now driving you to playing *D* has no connection or relation to the logic that earlier drove the other players to their decisions. But if you accept what was stated in the letter, then you must conclude that the decision you now make will be mirrored by the plays in the envelopes in front of you. If logic now coerces *you* to play *D*, it has *already* coerced the others to do the same, and for the same reasons; conversely, if logic coerces you to play *C*, it has also already coerced the others to do so.

Imagine there is a pile of envelopes on your desk, all containing other people's answers to the arithmetic problem "What is 507 divided by 13?" Having hurriedly calculated your answer, you are about to seal a sheet bearing "49" in your envelope when at the last minute you decide to check it. You discover your error and change the 4 to a 3. Do you at that moment visualize all the answers in the other envelopes suddenly pivoting on their heels and switching from 49 to 39? Of course not. You simply recognize that what is changed is your *image* of the contents of those envelopes, not the contents themselves. You first thought there were many 49's. You now think there are many 39's. It does not follow, however, that there was an instant when you thought: "They're all switching from 49 to 39." In fact, you would be crazy to think it.

It is similar with *D*'s and *C*'s. If at first you are inclined to play one way but on careful consideration you switch to the other way, the other players will not retroactively or synchronistically follow you. However, if you just give them credit for being able to see the logic you have seen, you must assume that their answers are the same as yours. In short, you are not going to be able to undercut them; you are simply in cahoots with them, like it or not! Either all *D*'s or all *C*'s. Take your pick.

Actually saying "Take your pick" is 100 percent misleading. It is *not* as if you could merely "pick" and then other people—even in the past—would magically follow suit! The point is that since you are going to be "choosing" by using what you believe is compelling *logic*, if you truly respect your logic's compelling quality, you would have to believe others would buy it as well, which means you are certainly *not* "just picking." In fact, the more convinced you are of what you are playing, the more certain you should be that others will also play (or have already played) the same way, and for the same reasons. This holds whether you play *C* or *D*, and it is the real core of the solution. Instead of being a paradox it is a self-reinforcing solution: a benign circle of logic.

If that still sounds like retrograde causality to you, consider this little tale, which may help to make it all make more sense. Suppose you and Jane are lovers of classical music. Over the years you have found that you have incredibly similar tastes in music—a remarkable coincidence. One day you find that two concerts are being given simultaneously in the city where you live. Both of them sound excellent to you, but Concert *A* simply cannot be missed, whereas Concert *B* is a strong temptation but one you will have to resist. Still, you are extremely curious about Concert *B*, because it features Zilenko Buznani, a violinist you have heard amazing things about.

At first you are disappointed, but then a thought flashes through your mind: Maybe I can at least get a firsthand report about Buznani's playing from Jane. Since she and I hear music through virtually the same ears, it would be almost as good as my going if *she* were to go. This is comforting for a moment, until it occurs to you that something is wrong. For the same reasons you would, Jane will insist on hearing Concert *A*. After all, she loves music in the same way you do—that's precisely why you want her to tell you about Concert *B*. The more you think Jane's taste is the same as yours, the more you want her to go to the other concert. But the more her taste is like yours, the less she will want to go to it!

The two of you are linked by a bond of common taste. And if it turns out that you differ enough in taste to disagree about which concert is better, then that

# Over 300 manufacturers sell their microcomputers with Microsoft software.

# There's a reason.

**Powerful simplicity.** That's the concept behind every software program we write. Powerful programs that let you spend more time thinking about the problem . . . and less time thinking about the computer. That concept has earned us the confidence of over 300 microcomputer manufacturers. Starting with the very first microcomputer you could buy. And today, Microsoft® languages, operating systems and applications software are running on well over a million microcomputers. Worldwide.

**Made for each other.** Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

**User-oriented.** Microsoft is the only software supplier to offer a full range of compatible operating systems, languages, utilities and applications programs. If you're not a computer expert, here's what that means to you: Better programs. Programs that are not only more powerful, but easier to learn and use.

**Better tools.** We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard,™ products that materially increase the capabilities of Apple ® and IBM PC computers. Plus, business and management software such as Multiplan,™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

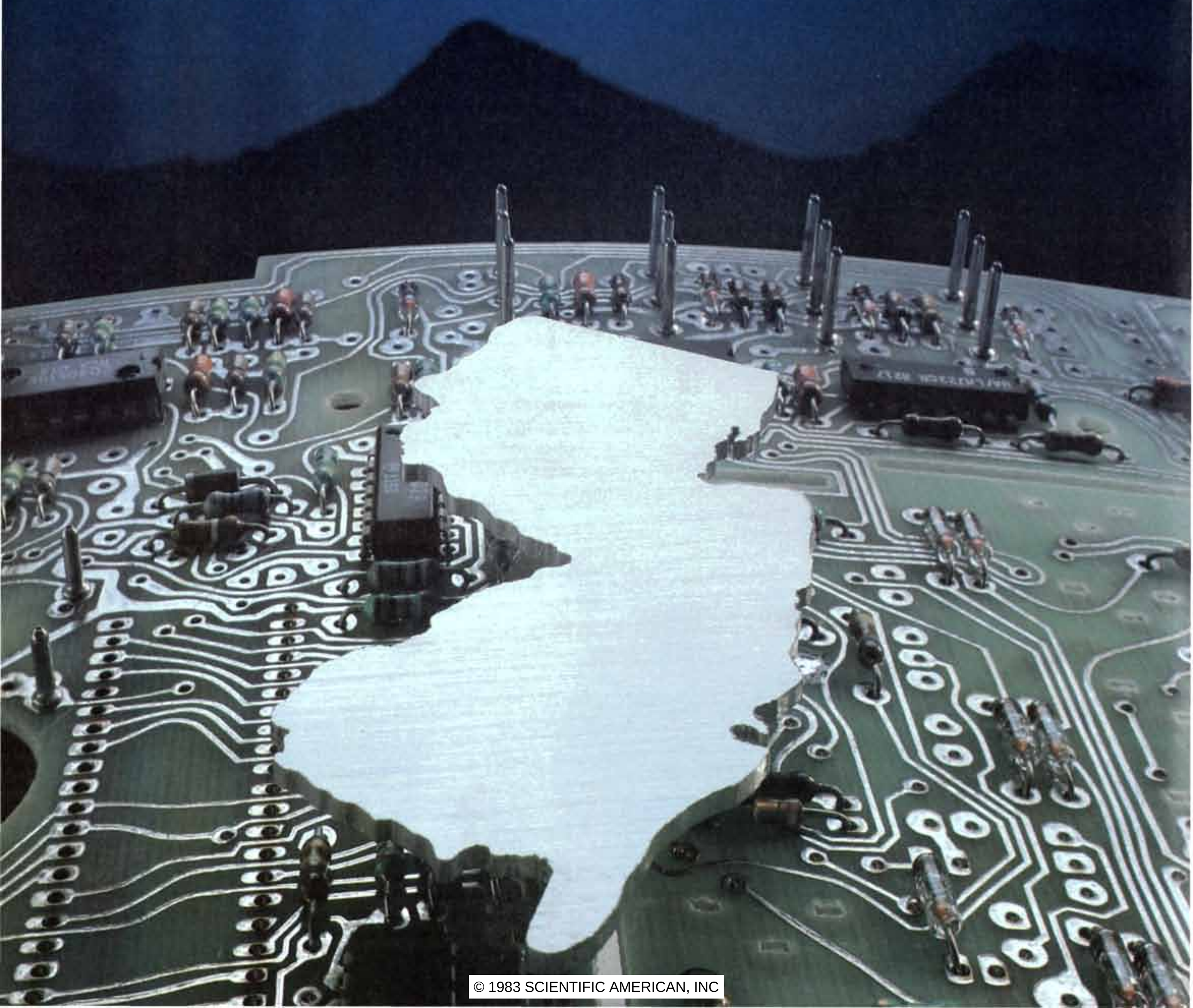
**Ask your Microsoft dealer.** Most microcomputer manufacturers offer their systems with some Microsoft software, but you'll undoubtedly want more. Programs and languages that solve your specific problems. Programs for your industry, business or home. Programs that do more, yet ask less. Ask your computer software dealer for a demonstration. You'll see why more than 300 microcomputer manufacturers offer their systems with Microsoft software. The reason is powerful simplicity.

**BETTER TOOLS FOR MICROCOMPUTERS**

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004


Microsoft is a registered trademark, and Multiplan, SoftCard, RAMCard and the Microsoft logo are trademarks of Microsoft Corporation. Apple and the Apple logo are registered trademarks of Apple Computer, Inc. IBM and the IBM logo are registered trademarks of International Business Machines Corporation. The Radio Shack logo is a registered trademark of Radio Shack, a Division of Tandy Corporation. The Burroughs logo is a registered trademark of Burroughs Corporation. The Victor logo is a registered trademark of Victor Business Products. The Wang logo is a registered trademark of Wang Laboratories, Inc.



will tend to make you lose interest in what Jane might report, since you no longer can trust her opinion as that of someone who hears music "through your ears." In other words, hoping she will choose Concert *B* is pointless, since it undermines your reasons for *caring* which concert she chooses!

The analogy is clear, I trust. Choosing *D* undermines your reasons for choosing it. To the extent that all of you in the game really *are* rational thinkers, you really will think along the same lines. And my letter was supposed to establish beyond doubt the notion that you are all "in sync," that is, to ensure you can depend on the others' thoughts to be rational, which is all you need.

Well, not quite. You need to depend not only on their being rational but also on their depending on everyone else's being rational, *and* on their depending on everyone to depend on everyone's being rational, and so on. A group of reasoners in this relationship to one another I call *superrational.* Superrational thinkers, by recursive definition, include in their calculations the fact that they are in a group of superrational thinkers. In this way they resemble elementary particles that are "renormalized."

A renormalized electron's way of interacting with, say, a renormalized photon takes into account that the photon's quantum-mechanical structure includes "virtual electrons" and that the electron's quantum-mechanical structure includes "virtual photons"; moreover, it takes into account that all these virtual particles (themselves renormalized) also interact with one another. An infinite cascade of possibilities ensues but is taken into account in one fell swoop by nature. Similarly, superrationality, or renormalized reasoning, involves seeing all the consequences of the fact that other renormalized reasoners are involved in the same situation—and doing so in a finite swoop rather than succumbing to an infinite regress of reasoning about reasoning about reasoning....

The answer I was hoping to receive from everyone was *C.* I was not so optimistic as to believe literally everyone would arrive at this conclusion, but I did expect a majority would, hence my dismay when the early returns strongly favored defection.

As more telephone calls came in I did get some *C*'s, but for the wrong reasons. Dan Dennett cooperated, saying "I would rather be the person who bought the Brooklyn Bridge than the person who sold it. Similarly, I'd feel better spending $3 gained by cooperating than $10 gained by defecting."

Charles Brenner, whom I had figured to be a surefire *D,* took me by surprise and *C*'d. When I asked him why, he candidly replied, "Because I don't want to go on record in a magazine of world-wide circulation as a defector." Very well. Know, world, that Charles Brenner is a cooperator!

Many people flirted with the idea that everyone would think "about the same" but did not take it seriously enough. Scott Buresh confided to me: "It was not an easy choice. I found myself in an oscillation mode, back and forth. I made an assumption, namely that everyone went through the same mental processes I did. I found myself wanting to cooperate roughly a third of the time. On the basis of that figure and the assumption that I was typical I figured about a third of the people would cooperate. Therefore I computed how much I stood to make in a field where six or seven people cooperate. It came out that if I were a *D,* I'd get about three times as much as I would if I were a *C.* So that I'd have to defect. Water seeks its own level, and I sank to the lower right-hand corner of the matrix." At this point I told Scott that so far a substantial majority had defected. He reacted swiftly: "Those rats—how can they all defect? It makes me so mad! I'm really disappointed in your friends, Doug."

So was I, when the final results were in: 14 people had defected and six had cooperated—exactly what the networks would have predicted. Defectors thus received $43 and cooperators got $15. I wonder what Dorothy is saying to Peter about now? I bet she is chuckling and remarking, "I told you I'd do better this way, didn't I?" Ah me, what can you do with people like that?

A striking aspect of Scott Buresh's answer is that in effect he treated his own brain as a simulation of other people's brains, and he ran the simulation long enough to get a sense of what a "typical person" would do. This is very much in the spirit of my letter. Having assessed what the statistics are likely to be, Scott then did a coolheaded calculation to maximize his profit, based on the assumption of six or seven cooperators. Of course, it came out in favor of defection. In fact, it would have done so no matter what the number of cooperators was. Any such calculation will always come out in favor of defection. As long as you feel your decision is *independent* of others' decisions, you should defect. What Scott failed to take into account was that coolheaded calculating people should take into account that coolheaded calculating people should take into account that coolheaded calculating people should take into account that....

That sounds awfully hard to take into account in a finite way, but actually it is the easiest thing in the world. All it means is that all these heavy-duty rational thinkers are going to see they are in a symmetrical situation, so that whatever reason dictates to one it will dictate to all. From this point on the process is simple. Which is better for an individual if it is a universal choice: cooperation or defection? That's all.

Actually it is not quite all, because I have swept one possibility under the rug, namely that maybe throwing a die could be better than making a deterministic choice. Like Chris Morgan, you might think the best thing to do is to choose *C* with probability $p$ and *D* with probability 1 minus $p$. Chris arbitrarily let $p$ be 1/2, but it could be any number between 0 and 1, where the two extremes represent defection and cooperation respectively. What value of $p$ would be chosen by superrational players? It is easy to figure out in a two-person Prisoner's Dilemma, where you assume that both players use the same value of $p$. The expected earnings for each, as a function of $p$, come out to be $1 + 3p - p^2$, which grows monotonically as $p$ increases from 0 to 1. Therefore the optimum value of $p$ is 1, meaning certain cooperation. Where you have more players the computations get more complex, but the answer does not change: the expectation is always maximal when $p$ is equal to 1. Hence this approach confirms the earlier one, which did not entertain probabilistic strategies.

Rolling a die to determine what you will do did not add anything new to the standard Prisoner's Dilemma, but what about the modified-matrix version I gave in the postscript to my letter? I shall let you figure that one out for yourself. And what about the Platonia Dilemma? There two things are clear: (1) if you decide not to send a telegram, your chances of winning are zero; (2) if everyone sends a telegram, your chances of winning are zero. If you believe what you choose will be the same as what everyone else chooses because you are all superrational, then neither of the alternatives is very appealing. With dice, however, a new option presents itself: roll a die with probability $p$ of coming up "good" and send in your name if and only if "good" comes up.

Now imagine 20 people all doing this and figure out what value of $p$ maximizes the likelihood of exactly *one* person getting the go-ahead. It turns out that it is $p = 1/20$ or, more generally, $p = 1/N$, where $N$ is the number of participants. In the limit where $N$ approaches infinity the chance that exactly one person will get the go-ahead is $1/e$, which is just under 37 percent. With 20 superrational players all throwing icosahedral dice the chance that *you* will come up the big winner is quite close to $1/(20e)$, which is a little less than 2 percent. That's not at all bad! Certainly it's a *lot* better than zero percent.

The objection many people raise is: "What if my roll comes up bad? Then why shouldn't I send in my name any-way? After all, if I fail to do so, I'll have no chance whatever of winning. I'll be no better off than if I had never rolled my die and had just voluntarily withdrawn." This objection seems convincing at first, but actually it is fallacious, being based on a misrepresentation of the meaning of "making a decision." A genuine decision to abide by the throw of a die means that you *must* abide by the throw of the die; if under certain circumstances you ignore the die and do something else, then you never made the decision you said you had made. Your decision is revealed by your actions, not by your words before acting.

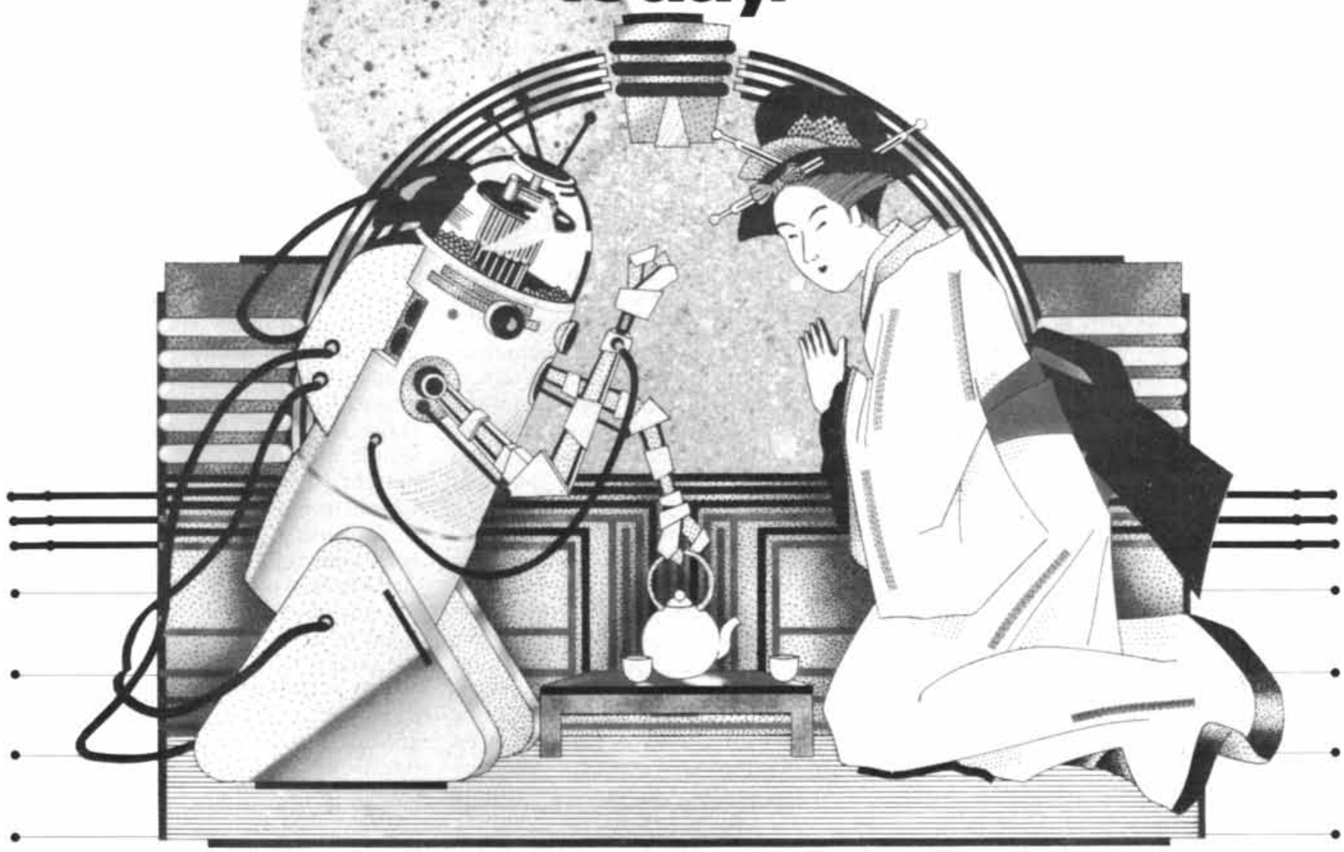
If you like the idea of rolling a die but fear your willpower may not be up to resisting the temptation to defect, imagine a third "Policansky button": this one says *R* (for "roll"), and if you press it, it rolls a die (perhaps simulated) and then irrevocably either sends your name or does not, depending on which way the die came up. This way you are never allowed to go back on your decision after the die is cast. Pushing *that* button is making a *genuine* decision to abide by the roll of a die. It would be easier on any ordinary human being to be thus shielded from the temptation, but any superrational player would have no trouble holding back after a bad roll.

This talk of holding back in the face of strong temptation brings me to the climax of this column: the announcement of a Luring Lottery open to all readers of *Scientific American,* and nonreaders too. The prize of the lottery is $1,000,000/*N*, where *N* is the number of entries submitted. Just think, if you are the only entrant (and if you submit only one entry), a cool million is yours! Perhaps, though, you doubt this will come about. It does seem a trifle iffy. If you would like to increase your chances of winning, you are encouraged to send in multiple entries without limit. Just send in one postcard per entry. If you send in 100 entries, you will have 100 times the chance of some poor soul who sends in only one entry. Come to think of it, why should you have to send in multiple entries separately? Just send *one* postcard with your name and address and a positive integer (telling how many entries you are making) to:

Luring Lottery
*Scientific American*
415 Madison Avenue
New York, N.Y. 10017

You will have the same chance of winning as if you had sent in that number of postcards. Illegible, incoherent, ill-specified or incomprehensible entries will be disqualified. Only entries received by 5:00 P.M. on June 30, 1983, will be considered. Good luck to you (but certainly not to any other reader of this column)!

# Technology Is Creating Tomorrow. You Can Invest In Tomorrow... Today.



Modern technology has encircled the globe, and nowhere is it more apparent than in Japan. Now, ancient customs and traditions share eminence with amazing scientific accomplishments.

The Japanese excel in such fields as robotics, fiber optics, semiconductors, fermentation techniques, ceramics, factory and office automation. Great technological triumphs are also being recorded in other parts of the Far East, as well as in America and Europe.

These exciting developments call for new ways to invest, and World of Technology, Inc., a global high-technology Mutual Fund, makes it easy to participate in tomorrow's technological triumphs... today!

At least 50% of the Fund's portfolio is invested in U. S. companies, while the balance includes common stocks traded on major foreign exchanges. The Fund uses management talents of investment experts in London and Denver, selecting companies large and small... but all having strong potential for growth.

Today, send for more information concerning worldwide high-technology investment opportunities.

In continental U. S., Call Free
**1·800·525·9274**

In Colorado, Call Free
**1·800·525·9769**

A mutual fund investing in high technology worldwide.

P. O. Box 1900
Denver, Colorado
80201

**World of Technology, Inc.**

Name
Address

City
State
Zip
Phone

A prospectus containing more complete information, including all charges and expenses, will be sent upon request. Read it carefully before you invest or send money.

TSAO63G

# BOOKS

## *The human primate, the land from the air, Hideki Yukawa's story, biological minerals*

by Philip Morrison

THE HUMAN PRIMATE, by R. T. Passingham. W. H. Freeman and Company ($19.95). "If ever apes learn to read, I hope that they will find the secret" of how to become human in this book. So writes the author, an Oxford neurologist and psychologist whose research is on the monkeys, in an admirably judicious yet lighthearted review of the vexed question of just what it is that makes us extraordinary among our fellow primates. The book is scrupulously documented, its arguments set as far as is now possible on experimental results, not on mere plausible conjecture. Many issues are still moot.

One graph suffices to make the case that as between distinct species (certainly not by mere modest sexual dimorphism) biology is destiny. The graph shows the development of the brain in the human being and in the chimpanzee. The two curves grow steeply at about equal rates after conception; both reach a plateau. The growth of the chimpanzee brain, however, slows markedly just before birth; the growth of the human brain continues at the fetal rate for about two years after birth. The result is the sovereign difference: the human brain ends up about three times the size of the chimpanzee's.

The author takes pains to examine the crude measure of absolute brain weight, because it lies at the core of his own experimental researches. It is of course proportioned to the weight of the individual animal, but we share with the great apes brains that are large for our build compared with those of the monkeys, just as monkeys have brains large for their size compared with those of other mammals. The primates are a brainy bunch. So too are the sea mammals; we do not quite know why, although plainly their bodies are built on a quite different plan.

The human brain is set firmly in the primate pattern; a single curve fits the data well for the volume of neocortex (those little gray cells) as a function of the brain volume, one neat power law all the way from the lemurs to proud *Homo sapiens*. We have more of that powerful hardware, but only because the brain we have is bigger, and bigger primate brains regularly have relatively more neocortical volume. The number of cells in a cylinder of cortex of a given area is the same in all mammals; the big brains have a lower cell density to accommodate the more numerous interconnections but have thicker cortical layers. Even a couple of million years ago our australopithecine ancestors had a mean brain volume relatively larger than that of today's chimpanzees, although the brainiest individual chimpanzees exceed the mean of the fossil hominids in this measure. Only the rather subtle issue around the dominance of one cerebral hemisphere over language suggests that the chimpanzee brain may differ from ours in function more than it appears. In monkeys the removal of one cerebral hemisphere has little effect on the learning even of complex tasks.

The rest of the comparative anatomy of apes and human beings tells no very different story. The world appears much the same to chimpanzee and man. We share a poor sense of smell, own very sensitive pads on the fingers, see color by day, use stereoscopic clues for depth and are deaf to high audio frequencies. Like us, the chimpanzees can make sense of mirrors and photographs and can classify their world. Without words, chimpanzees and orangutans can pick out by feel at the first trial that one of two objects which matches a visible sample. The chimpanzee has good manual skills; all Old World monkeys and apes can in fact use the light precision grip, and chimpanzees can manipulate our tools well, to make a pot of tea or flick a switch. There is no doubt that we are better at it; no one expects the chimp to play Chopin presto. Walking is again in our favor; it is an old knack in the family, three or more million years old. A photograph shows a chimp upright on two feet with its arms full of bananas; chimpanzees know the usefulness of this posture, although a 50-yard run erect is as much as has been seen among them.

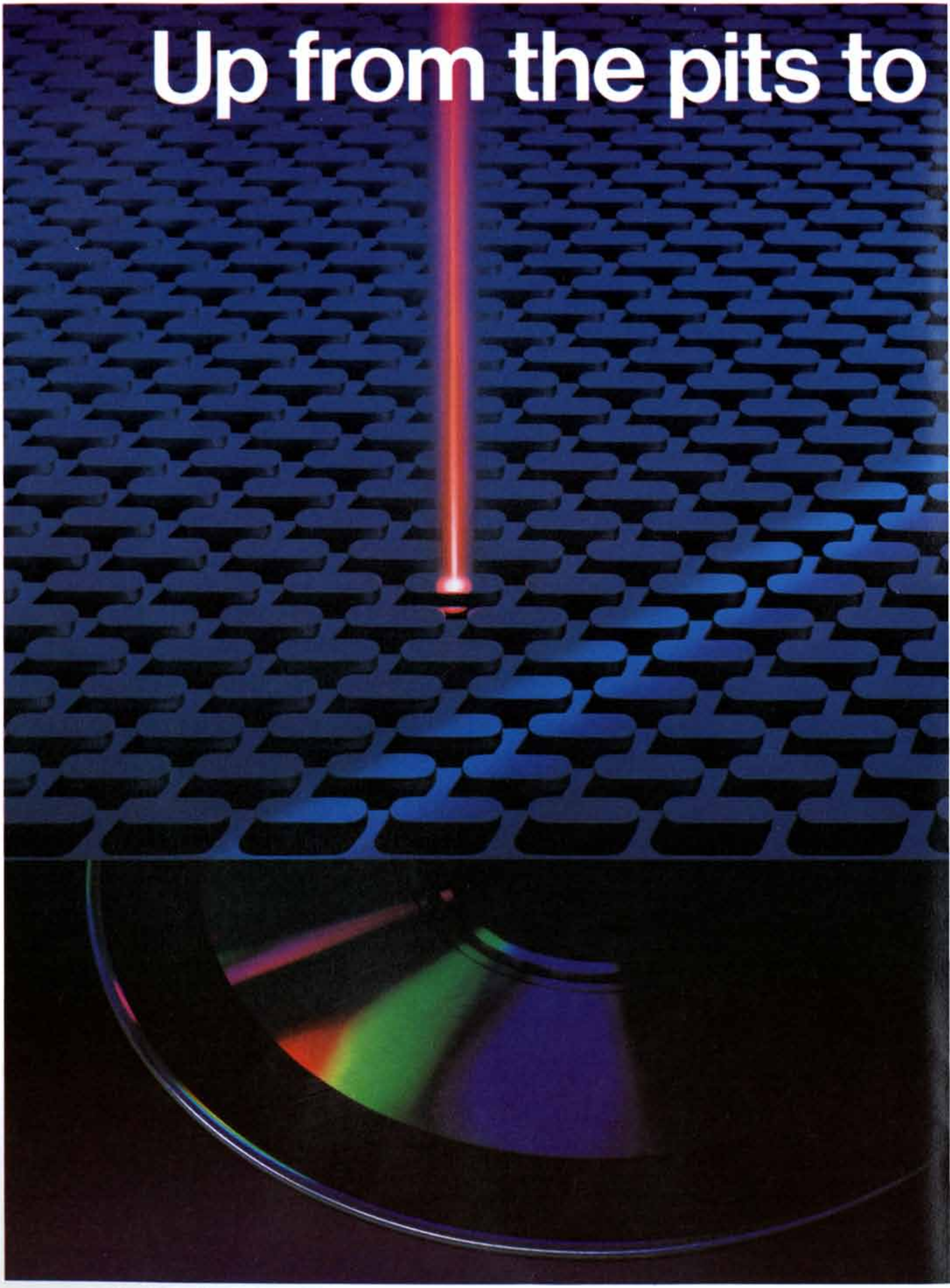
Such up-to-date comparative anatomy occupies about a third of the text; the rest is split between a similar discussion of comparative abilities and one of comparative social order. Matching and discrimination tests have been done with many mammals; monkeys and apes excel, even allowing for the fact that the testers have a rather evident primate bias. (The dolphin is doubtless a clever species, but there is no basis for saying more for it than that it may compare with some of the clever simians.) A small orangutan named Abang learned to strike a flake from a flint core and then use the flake to cut a string to release food from a box. No claim is made that Abang is like the first Paleolithic craftsman; the animal simply followed a careful demonstration.

Wild chimpanzees have been seen to use simple tools—the sticks, stones, twigs and leaves around them—to dig, pound, probe, sponge up, rake in and reach out, to throw and club and prop up what they would. Other apes and monkeys have also been seen to use many of these ready tools. The modification of tools—ape toolmaking, not simple use—is clear in captivity and is present, if rudimentary, in the wild. Chimpanzees have an entire repertory of tool using, but even a blue jay has been known to tear and crumple paper strips to help move food pellets closer to the side of its cage. Perhaps more to the point is social imitation, famous among the macaques of Japan and even commoner in chimpanzees. Again birds can show it. Social custom—particular grooming postures customary among one group of chimps and never seen in a group of matching habitat nearby—strikes a note of recognition.

For more than a century language has been seen as the Rubicon of human nature. A table here lists 13 identified vocalizations of chimpanzees in the wild: laughter, a pant-hoot for general approval, the call *wraa*, sounded when a human being or another predator is detected or when a dead chimpanzee is found. The vocabulary is touching, but it is no language. Passingham is inclined to credit the success of recent teaching of sign or symbol languages to the chimpanzees, although plainly the matter is not finally resolved. Children "have a thirst for language," but chimpanzees need careful tuition to pick up the simplest rules step by step, if indeed they can do it at all.

Of course, the seat of human culture is society, not the individual. The human family does not seem to hold a clear key; 90 percent of bird species are monogamous, and many primates too are tightly pair-bonded, in a rather random distribution among the groups of primates. The titi-monkey pair shown in a loving winding of furry tails are better as a symbol of marriage than any robin's nest; titis do seem to share plant food easily between adults and young. Chimpanzees share the meat they occasional-




Up from the pits to

# sensational sound

**Because records get stuck, get stuck, get stuck...**

Anyone who has heard a worn or scratched disc played on good audio equipment knows the problem of analog recordings. Record wear and damage cannot be avoided. Even brand-new discs are subject to wow, flutter, and distortion—the inherent drawbacks of conventional recording methods. To eliminate these difficulties, an entirely new approach to sound reproduction was developed: Compact Disc (CD) Digital Audio.

**Hitachi introduces the future of sound**

Drawing on a wealth of accumulated experience in laser, semiconductor, and audio technologies, Hitachi was able to come up with one of the world's first, completely digital sound systems for home use. Mirror-like CDs, just one fourth as large as standard LPs, are imprinted with coded series of reflective pits, each one just 0.9 × 0.5 microns in size, analogous to a speck of dust on the floor of an amphitheatre. In all, some 8 billion of these pits make up a single, one-hour disc side. During playback, they are scanned by an infrared laser beam, and the beam reading is reconverted into sound, with no wow, no flutter, and just 1/200 of the distortion heard on ordinary discs. And there's no wear at all because only laser light comes into contact with the specially coated CD surface.

**Top performance from a compact, complex package**

Within the Hitachi CD player itself are nine different large-scale integrated circuits (LSIs), each equivalent to some 32,000 transistors, yet taking up just a few square centimeters of space. Perhaps the most important of these is the 16-bit linear digital/analog converter which translates digital pickup signals into analog signals at a rate of 44,100 sound segments per second. It allows output virtually identical to the original sound, dynamic range and S/N ratios of over 90 dB, through standard amplifier/speaker systems.

Hitachi's DA-1000 digital audio disc player

**Technical excellence is embodied in all Hitachi products**

The development of a compact disc digital audio system is just one case demonstrating Hitachi's technological strength. You'll find other examples in all of Hitachi's products, from audio amplifiers, tuners and tape decks to a whole host of high-quality electronic appliances. Our comprehensive technical expertise is your guarantee of convenience, easy operation and high reliability in every product that bears the Hitachi brand.

**HITACHI**

A World Leader in Technology

*Inquiries to:* **Hitachi Sales Corporation of America,** 401 West Artesia Blvd., Compton, Calif. 90220 Tel: (213) 537-8383

ly take, but they form sex-divided bands rather than family groups; their transfer of meat is more of a "tolerated scrounging" than an example of customary sharing. There is evidence that some 1.5 million years ago hominid bands already used butchering sites; "whatever the basic social unit at that time, we must suppose that the two sexes were more dependent on each other in their economic life than is true of chimpanzees today."

Cooperation and indeed competition play their roles in the social groupings of animals. Monkey groups sometimes fight battles, but only human beings sustain war. A careful account makes it plain that human warfare is not inbuilt aggression, as the ethologist Konrad Lorenz seemed to argue, or an analogue of territoriality or of dominance hierarchies. It is not the "beast within" we need to blame. Man is the only warring primate because he alone is able to make war; he is too clever for his own good. The author considers both the genetic and the cultural aspects of the human heritage at some length; he concludes that the current controversies are premature in the absence of adequate evidence. Both nature and nurture affect behavior. These days more people might be persuaded by a trendier phrase: the need for both hardware and software. The profound software changes of human prehistory in the absence of demonstrable genetic change at any such breakneck rate, however, place a heavy burden on sociobiological explanations.

The future evolutionary task of the chimpanzees is clear. "They must expand their brains according to the primate pattern" until they can fluently form and transmit propositions about the physical and the social world, even in the sensory absence of the topical stimulus. A novel social order will follow, a growing culture, and the apes will be on their way. What we can say today about that path may be elementary, but it does not seem to present an utter mystery. Over some 10 million years of change a threshold was crossed, by brain expansion rather than by reconstruction, and a rich and ready code arose. "Of course, it is possible that, when faced with the option of becoming a man, a rational chimpanzee would take one look at human society and turn down the offer."

BRITAIN FROM THE AIR, by Bernard Stonehouse. Crown Publishers, Inc. ($30). THE EXTRAORDINARY LANDSCAPE: AERIAL PHOTOGRAPHS OF AMERICA, by William Garnett. New York Graphic Society, Little, Brown and Company ($60). The cameramen of Aerofilms, Ltd., began in about 1919 (there is a takeoff picture of a stick-and-wire biplane to prove it) to assemble their unmatched collection of commercial air photographs of the island of Britain (and a little outside it). Stonehouse, an experienced pilot and a well-known ecologist, has lovingly compiled from that overflowing archive a selection of its photographs in color, mainly from flights over the past 10 years. He has prepared a brief description of each photograph, provided a map of the implied wanderings from the island of Jersey to the far Scottish Highlands and displayed the photographs on the large pages of this album. About seven of every eight photographs focus, speaking very roughly, on the work of British hands 1,000 feet below; the rest show a landscape that is mainly nature's.

Here, for instance, are the famous bridges of Britain, monuments in the history of modern technology. They are shown of course as they lie in the countryside, an essential view not so familiar to students of engineering from the books. First see Coalbrookdale, that delicate masterpiece of Abraham Darby's cast-iron foundry, spanning the narrow blue Severn for 200 years now. The first modern suspension bridge, designed by Thomas Telford, soars over the Menai Strait in northern Wales, modern steel cables bearing the load carried until quite recent times by the original huge links of wrought-iron chain. Close by is little Conwy, where the same granite headland that bears a fine 13th-century castle also receives three early Victorian wonders: Telford's suspension bridge for the coach road and the twin rectangular iron tubes designed by Robert Stephenson that still enclose and support the railway, one track within each.

Here too is a span of Isambard Kingdom Brunel's, a chain-link flung high above the sluggish Avon below Bristol, "completed with second-hand chains." The cantilevers of the Firth of Forth hold out their powerful hands to bear the trains, and behind them the elegant new cable-suspension highway bridge is mistily visible. The Severn cable bridge spans wide, windy waters near the sea, and the half-dozen busy crossings of central London, low arches and high towers, appear repeatedly.

An older tale is told in other structures. The export industry of Norfolk about 4,000 years back was flint tools. At Grime's Graves the field is pocked with long-overgrown spoil heaps and long-filled pits, where once Neolithic miners toiled with reindeer-antler picks to win the nodules of flint 20 feet down in the chalk. Here are the Avebury rings, near Stonehenge on Salisbury Plain, best seen as a concentric whole from the air, except that the viewer can then only envy the people whose parked cars testify to their pleasant hour at the country pub near the center. Scarred to the white chalk just below the roots of rich green pasture on Berkshire Downs is the giant silhouette of a "wild horse-goddess," with an Iron Age fort close by. Roman occupation is marked by a stone garrison fort at Housesteads along Hadrian's Wall, where a battalion of legionaries once served.

Here are the centers of learning, the University of Oxford lying grandly around the rotunda of the Radcliffe Camera, Cambridge seen from across the Cam, with the chapel of King's College at the center, and Eton College, framed a little too tightly to display more than a corner of the aphoristic playing fields. There are steel mills and mining towns, and one enormous nonnuclear power station, magically transformed almost into a Turner by billows of mist and the vapor columns rising from the eight drums of its cooling towers. British foibles are not avoided: endearing architectural follies, cricket and the crowd leaving a Cup Final at Wembley Stadium. There is a multiple pileup on the M1 road, a mortal disaster among dozens of cars and lorries in the fog somehow more devastating in Bedfordshire than in Los Angeles.

Among the images of nature the cliffs of Albion are strikingly shown, the green downs edged by 500 feet of white chalk, the sea below completely lost in mist. This is a striking set of documents, craftsmanlike rather than artful, but one image crosses that line. It shows the Pool of London at sunset, ships, barges and bridges silhouetted strongly against a winding golden Thames, the city an almost unmarked black. It would evocatively illustrate an edition of Conrad.

William Garnett, the author of the other book of aerial photographs, is no company but an engaged artist, one man with a million miles of solo flight in a light plane behind him, camera in one hand, controls in the other. His quarry is the beauty of the land. A few people appear in boats, as the lone fisherman adds melancholy scale to the landscape of the Chinese painters. The only works of humankind seen often here are the boldly contoured furrows of the tractor-driving plowman, along with field boundaries and windrows.

Here also there are two classes of images, both beautiful. They are superbly reproduced in this book, a triumph of American bookmakers from coast to coast. One class is a landscape where scale is easy to read: the inverted cone of a volcanic gas vent newly melted into glacier ice and snowpack high on Mount Baker or 100 miles of Mississippi oxbows glowing in the low sun. (Garnett flies and shoots at altitudes from 100 feet up to three or four miles.) In the other class the scale is covert, enigmatic. The picture is largely texture, a play of form and color that may remind you of the land but might be the abstract canvas of a painter. The tufts of sunstruck brown kelp float in the blue sea. The eroded butte is a sculptured surface, the blue sky mirrored in the branching ravines filled with snow, the steeper slopes lacing the ravines with a chiseled gray-gold smoothness. Sometimes it is the clouds that bear the tale, sometimes it is the surf breaking on an unseen slope. The flooded fields of California rice glow deep sky blue, and across them the low dikes trace a contour map in sun gold.

Ansel Adams writes of these photographs as revelation; they disclose patterns we ordinarily miss. Garnett recalls 36 years of piloting, often intensely hard, dangerous work. He praises his tools: "Deep appreciation goes to all those unknown people who carefully engineered my plane,... crafted the parts,... gave my craft its rated strength and the integrity to serve... through unrelenting, pounding turbulence, and yet to respond with... ballerinalike movement in smooth air. Yes, to all you people at Cessna Aircraft, McCauley Propeller and Continental Engine—I love you for your care. Thank you for three decades of safe, beautiful, privileged flight." The opening photograph, by the artist's son, shows the silvery 1955 Cessna 170-B in flight; the closing image displays "the shadow of my plane on a cloud, completely encircled by a rainbow." It is a token for this happy union of technology and sensibility, in a life of devotion to an antique and spacious vision that became real only in our times.

HIDEKI YUKAWA: "TABIBITO" (THE TRAVELER), translated by L. M. Brown and R. Yoshida. World Scientific Publishing Co., Singapore. Distributed in the U.S. by Heyden & Son, Inc., 247 South 41 Street, Philadelphia, Pa. 19104 ($33). "Hideki Ogawa was born in 1907 at old Tokyo's Ichibei-cho Azabu. The house smelled of plum blossoms each spring.... I cannot remember the house of my birth, not even the plum blossoms there. I know them only through my mother, but in my mind I know they are extremely beautiful. (Perhaps my subconscious tries to beautify my birth.)" So began the physicist Yukawa when at 50, a decade after his name was the first from the Far East to enter the long list of Nobel prizewinners, he wrote this candid, gentle, reflective autobiography of his youth. (Like his father before him he had adopted his wife's family name on marriage, a custom in Japan when there is no son to continue the name.) The memoir ends in November, 1934, with his famous paper in English, the birth of the physics of the hidden transient particles of nature. (Enrico Fermi's neutrino theory of the same year probably shares that proud claim.) Yukawa was then a young lecturer in theoretical physics at Osaka Imperial University. He finished the paper quickly because "my wife kept urging me, 'Please write the English paper and show it to the world.'"

The flood of post-Meiji modernization in Japan runs brimful through the book. Hideki's paternal grandfather taught Chinese learning to the children of the feudal clan of Tanabē. Hideki's father, inclined to a bookish life, was drawn by a chance at the New Knowledge, and made his way to the university in Tokyo. The terrible earthquake of 1891 near Nagoya, not far from his home countryside, gripped his imagination. He visited the scene and came away in awe. "It was then that his life began to center around geology." He became a well-known field geologist. After a long government-sponsored trip to China to examine its minerals on the eve of the Japanese expansion, he went to Kyoto University, eventually to become dean. The infant Hideki grew up in Kyoto with two sisters and four brothers. Three of the sons became university professors; the other was killed in World War II.

Hideki was a shy, bookish, silent, somewhat lonely child. He was distant from his parents. His father did not intend to send this enigmatic boy to university; a technical school seemed good enough. But his mother's pleas and the intervention of the insightful principal of his school ("A boy with Hideki's potential is very rare") opened the path. Hideki knew he was mathematically gifted. An older brother brought home from the middle school the wonderful formula for summing a long arithmetic progression all at once. Younger brother had already worked out the method on his own, innocent of algebra and its notation. Mother realized it and "began to smile in delight."

The narrative is familiar; worldwide human unity outweighs the powerful old culture. This is a Japan as vibrant as it is constrained. Einstein visits, and all Kyoto—except Hideki—turns out to see and hear him. Dostoevski and de Mille are in everyone's mind; Hideki, still in his wide pleated trousers, has a nickname borrowed from Tolstoi, perhaps meaning Ivan the Fool. He is active in baseball, but he walks apart from even his sisters. He is taught at home the Chinese characters from the classics by rote without understanding, he swims in the traditional way (endurance for many hours with no account of distance) and the high school play is storytelling of the third shogun, a performance full of swordsmen and dependent on elegant female impersonators. (Hideki refused such a part.) His family was a strict one; playfulness was rare.

He was aimed at mathematics, pretty surely, until he hit a martinet of a high school teacher. The man marked as wrong a correct proof that differed from the one given in class; Hideki could not "dedicate his life" to a subject like that. A bad draftsman (although an able calligrapher), he could not hope for engineering. His geologist father made an effort to lead the boy toward geology, but Hideki's mind was a logical one, with little power for observation or memorization. He could not finish the big textbook in English his father produced out of the library at home. Physics it was to be; two German books on the old quantum theory (Reiche and Planck) made up his young mind.

Here is his undergraduate physics class photograph at Kyoto; visiting professor Otto Laporte of Michigan is there, and so is Hideki's close friend Sin-itiro Tomonaga, who was to win a Nobel prize a few years after him. A young graduate student, Hideki is attracted by the "spirit of Copenhagen" in the lectures of a paternal Yoshio Nishina on quantum mechanics. That attitude is hard to describe, but "it is certain that it had much in common with the spirit of generosity."

An arranged marriage is charmingly described; it was a lifelong success. The new Ph.D. finds his way, at first only part-time and then in a real teaching job, to the new, lively University of Osaka. There is a physicist-president there, and Yukawa is recruited by Hidetsugu Yagi, inventor of the antenna named after him. ("I developed a deep trust in this man.") No gifted teacher, Yukawa spoke to the blackboard, and his soft unemphatic speech invited sleep.

All the time he was seized with the refractory question of the nuclear forces. His first original research paper treated the magnetic interaction between the atomic electron and the nucleus, but Fermi beats him to the spectroscopic answer and goes further. Once again Yukawa tries for the nuclear forces; this time it is an effort to use the electron and the neutrino, where once more Fermi far off in Rome has blazed a path only dimly seen in Osaka. Finally, in a new house in the hills, a 10-mile commute to the dusty and busy streets of industrial Osaka, he is forced to spend a few days at home. Sumi, his wife, is confined with the birth of their second child, and stormy fall weather makes it hard to travel among the fallen trees.

One night, unable to sleep, the new infant a few days old, the crucial point comes to him. The short range of nuclear forces demands a carrier of the force field with a mass of 200 electrons. No one could easily observe such a particle. It is present but hidden, if his calculations are right. That day in the meeting of their lively nuclear research group he

# IBM 2½ minutes HP 30 seconds


## Hewlett-Packard's new personal computer is up to 5 times faster than IBM's. And half the size.

Take two personal computers: Hewlett-Packard's new Series 200 Model 16 and the IBM PC. Run the same software—Context/MBA™—and solve the same problem.

You'll see one big difference. It takes the IBM minutes to do what the HP 200 can do in seconds.* And that's just the beginning.

With Context/MBA™ integrated software, the HP 200 can do word processing, spreadsheet projections, indexing, filing and graphics all at one time. Change one number in a projection... and the graphs change, automatically. It's that easy. And with the HP 200, it's just about that fast.

Because, with its super-powerful MC68000 micro-processor, the HP 200 can address 32 pieces of information at one time. (Twice as much as the IBM.)

And it's got the memory to handle big, complicated problems. (You can have as much as 768Kb.)

But for all its power, the HP 200 is surprisingly small. The whole computer, not including the keyboard, measures just 12″ across by 13″ deep. So, putting a personal computer on your desk doesn't mean taking everything else off.

If you'd like to see how the HP 200 measures up, call (800) 547-3400. (In Alaska, Hawaii and Oregon call (503) 758-1010 or TTY (503) 758-5566.) Ask operator #411 for more information and the name of the HP dealer or sales office nearest you.

HEWLETT PACKARD

TC-02310

*Recalculating and refreshing a screen with words, numbers and graphics. Both ran equivalent Context/MBA™ programs.

Context/MBA™ is a trademark of Context Management Inc. IBM is a trademark of International Business Machines.

tells of it. "If there is such a charged particle, it should become visible in the Wilson cloud chamber, should it not?" Kibuchi asks. "I answered, 'Yes, the particle can be found in the cosmic rays.'"

There the pi meson was duly found, after a few years of unraveling some puzzles. Today the simple Yukawa equation has given way to a grand scenario of particles found and conjectured in entire families, much extending Yukawa's pioneer path. His penetrating self-analysis is nicely translated. L. M. Brown of Northwestern has given a knowing introduction to Japanese physics in the times of the narrative, an account that makes it quite clear that Yukawa, although he was a self-contained worker, was by no means alone. The book is published now in English by a new Singapore firm, determined to make available good books (most of them original advanced treatises) at low prices for the students, many in Asia, from whose work will flow the worldwide physics of the future.

A book so personal may justify the self-indulgence of the physicist-reviewer, who first met Hideki and Sumi Yukawa in that year of grim resolution 1945 at their Kyoto house, over a ceremonial cup of tea. The gauche young visitor, in the uniform of the occupier, came half-triumphant, half-guilty, to receive a graceful welcome at their hands. They gave what they had in war-exhausted Japan, quiet, beauty and the celebration of peace, and it was treasure.

BIOMINERALIZATION AND BIOLOGICAL METAL ACCUMULATION: BIOLOGICAL AND GEOLOGICAL PERSPECTIVES, edited by P. Westbroek and E. W. De Jong. D. Reidel Publishing Company. Distributed in the U.S. and Canada by Kluwer Boston, Inc., 190 Old Derby Street, Hingham, Mass. 02043 ($69.50). Hard-shelled or spiny, or sharp-toothed and stout of bone, many a member of the animal kingdom has manifested control over and investment in the austere kingdom of minerals. Add examples such as eggshell and the pathological kidney stone, to say nothing of the tiny otoliths within the orienting instrumentation of animals from sharks to human beings, and the case is stronger. About half of the space in this varied and prompt report of an international congress held in the Netherlands only a year ago is given over to papers touching many aspects of calcification. Plainly the examples just listed are signs of the liminess of the present world of life, organisms adept at mobilizing the ions of calcium up to the crystalline level.

The topics are not new; indeed, this was the fourth interdisciplinary congress of a series, a gathering of biologists, geologists and chemists. This time their brief was widened. A third of the papers treat the accumulation of minerals other than those of calcium, and the rest of the 500-page volume, its attitude closer to plate tectonics than to electron micrographs and polysaccharide chemistry, inquires into the global implications of the ancient alliance between minerals and organisms.

There are now known about 40 minerals normally formed by extant organisms, plus nearly 20 others that appear as pathologies (including dolomite, a bladder stone of the Dalmatian dog, whose spots are not its only claim to oddity). Our bone crystallites are mainly a mineral of calcium phosphate, hydroxyapatite; the shelly stuff of mollusks is calcium carbonate, in one or another of its crystal forms; also on the list are pyrite, galena, fluorite, barite, gypsum and magnetite. It is true that some of these relations may be less than intimate: the organism induces the growth of the mineral in a pseudoinorganic manner, so to speak on the side.

Even more commonly crystal growth is more or less controlled by a little-understood assembly of specific organic macropolymers, which act to determine the mineral phase, its crystallographic orientation and its overall form. Tiny opal-like silica spheres somehow are assembled into the specific and lovely forms of the diatoms and the radiolarians. An organic framework can be seen to precede the opaline deposit and may act as a template. Life is good at constructing surface layers, a strategic device for the control of crystal growth, and there are reports here of the strong inhibition of crystal growth, even in vitro, by specific organic products that disturb nucleation. Electron micrographs at almost the resolution of the atomic layers disclose that in the tiny calcite particles of the inner ear of a rat the crystallites are oriented to yield a single-crystal diffraction pattern, even though actually the structure is an ordered overlap of several tiny crystals. The full terms of the alliance between the two species of order remain undisclosed.

The tidal flats of our times are not much different from the world of the early Precambrian, before life could leave the wet sea edge. Nowadays the evolved competitors are keen, but even so some colonial blue-green algae are found on salty shores worldwide from the Bahamas to western Australia. One remarkable study examined these matted structures over time; the living forms lithify to dominate the intertidal surfaces. These crusts of calcium carbonate form through an elaborate process of seasonal change; they resemble in detail the stromatolites of a billion or two years ago and more. The process that preserves the structure destroys the very organisms that built it; the alliance was intimate even then, although not yet easy.

The sovereign cycles of the earth are plainly two: at depth gravitational and radioactive heat drives a slow, majestic flow, to raise mountains, open and close seas and outgas deep-lying materials within the huge volume of the planet. On the surface there is a thin shell of life, sun-nourished, catalytically remarkably active, demonstrably evolving in complexity and perhaps in gross activity over the entire span of planetary history. One molecule, carbon dioxide, plays two roles. It is the nutrient that feeds the photosynthetic elaboration of the carbon skein of life. It is also the single atmospheric component most important in the regulation of the temperature of the earth's surface by its greenhouse control over outgoing infrared.

Life demands enough of carbon dioxide, and yet life cannot easily survive too much of it. Most of the carbon atoms now near the surface of the earth are in fact buried deep in the calcite jetsam of the waters of all geologic time. Otherwise our planet might be an oven-hot second Venus. For four billion years the right balance has somehow been struck. Was it a subtle feedback loop set up within the fabric of life itself? Or was the balance maintained by blind geology? The terrestrial thermostat has maintained a living planet even though the sun has slowly grown more luminous, on its way to the swollen red giant of the future.

That might be reflected in the uptake of carbon dioxide by calcium during the weathering of the exposed feldspars in crystalline rock. Hot climates mean faster weathering, less free carbon dioxide and a cooling climate, a feedback loop entirely independent of life. Such an earth cannot freeze over, either; the gas from the depths is fed out even through a glaciated surface, but weathering would not persist under an ice blanket. Or is it that weathering rates depend heavily on the transport of carbon dioxide to the rocky surfaces by the vigorous life in the soil of a hot and sunny climate? Evolution or good fortune? Gaia or Pluto?

This big book holds at least half a dozen readable review papers over the entire set of issues mentioned. Another dozen or so papers are straightforward enough to instruct and even amaze the general reader. The remaining score, terse and technical, can reward the browser. One report by a young Boston University biologist indicates by direct simulation with living cyanobacteria (a sample she took from the Charles River) that the very old gold ores of the Witwatersrand may indeed have been formed through direct microbial involvement. Is glittering gold another biomineral of the microbial mats?

Established 1845 SCIENTIFIC AMERICAN June 1983 Volume 248 Number 6

# Physical Disability and Public Policy

*The civil rights of disabled Americans require a more accessible environment. The present Administration, however, hesitates to enforce laws calling for the removal of architectural barriers*

by Gerben DeJong and Raymond Lifchez

In the past 15 years or so there has been a major change in the policy of the U.S. with respect to citizens with disabilities. Between 1968 and 1980 Congress passed a series of laws aimed at enhancing the quality of life for disabled Americans through various measures, including the removal of architectural barriers, the improvement of educational opportunities for disabled children and the encouragement of self-help organizations for people with disabilities. The centerpiece of the new legislation was the Rehabilitation Act of 1973. Section 504 of the act, which bars discrimination on the basis of handicap in any program or activity benefiting either directly or indirectly from Federal financial aid, has emerged as the most controversial and most litigated piece of the legislation. In essence the Rehabilitation Act extends to disabled Americans a previously uncodified form of civil-rights protection. Accordingly the 1973 law is sometimes called the Civil Rights Act of the Handicapped.

In spite of the clear intent of Congress in the matter the present status of disabled people in American society is uncertain, in large part because the Reagan Administration has been reluctant to enforce the new laws. Meanwhile there are signs of a public backlash: after an initial period of little opposition to these measures the nation seems to have entered a period of doubt, particularly over the cost of making the built environment more accessible to people with disabilities. At a time of widespread retrenchment in Government expenditures for social programs Americans with disabilities are finding themselves increasingly on the defensive.

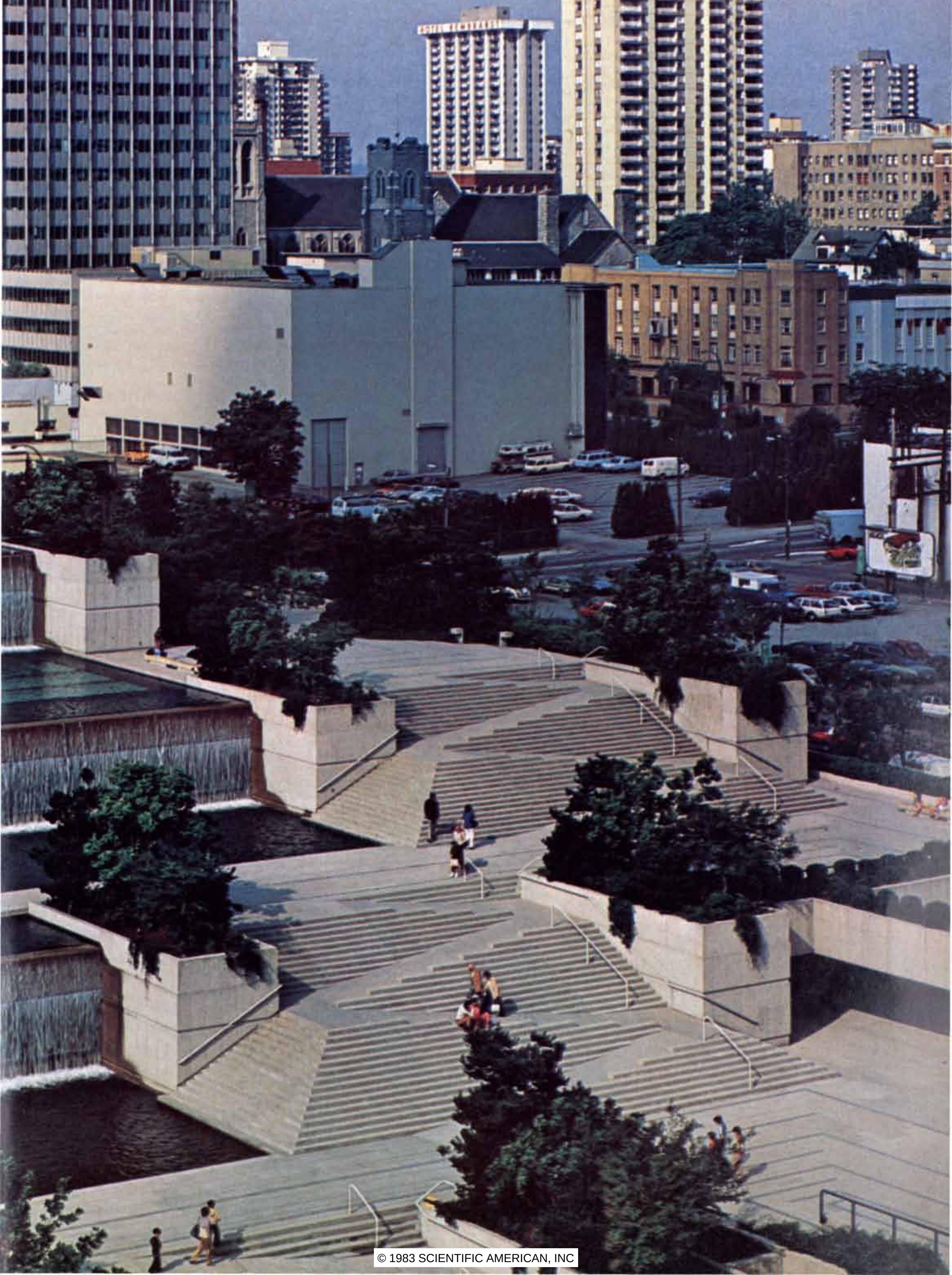
In what follows we shall review the background of this issue, summarize what is known about the dimensions of disability in the U.S., look into the causes of the current stalemate in the implementation of the existing disability laws and try to identify some of the ingredients needed to make the nation's disability policy a viable one in the face of current political and economic realities. In particular we shall focus on one highly visible aspect of the problem, namely the removal of architectural barriers to the progress and well-being of people with disabilities.

The disability legislation of the past 15 years is based on familiar civil-rights principles such as equal opportunity, nondiscrimination, integration, free choice and self-help, to name a few. The thread linking all these principles is the concept of access—access to education, to employment, to public facilities and services, to transportation, to housing and to other resources needed by disabled people to more fully realize their rights as citizens. Architectural barriers compromise access. They are seen in this context as obstacles that limit opportunity, promote discrimination, prevent integration, restrict choice and frustrate self-help.

In a sense the new disability legislation also has an affinity with the traditional noninterventionist values of the political right. The legislation emphasizes self-reliance and the right of disabled individuals to shape their own destiny. The aim of the legislation is merely to remove the barriers that prevent disabled people from achieving that goal. Although the legislation does in fact impose new burdens on certain segments of American society, the "bootstrap" philosophy implicit in the laws is unmistakable.

Much of what is known about disability in the U.S. comes from the Health Interview Survey, a continuing nationwide study sponsored by the National Center for Health Statistics. Each year the center, working in collaboration with the Bureau of the Census, selects a sample of about 40,000 households, representing approximately 120,000 persons. One respondent from each household is interviewed in depth to ascertain the state of health of each resident, including any functional limitations attributable to a chronic health condition. According to the Health Interview Survey for 1979 (the latest year for which fully tabulated results are available), some 31.5 million Americans, or 14.6 percent of the noninstitutionalized population, are limited in some way by a chronic health condition. An estimated 7.9 million (or 3.7 percent) are considered severely disabled, that is, they are unable to carry on some major activity such as attending school, working or housekeeping. As one might expect, disability increases significantly with age: 46 percent of those who are 65

**OUTSTANDING EXAMPLE of a public place designed to be accessible to people with disabilities is seen in the photograph on the opposite page. The view is of a rooftop "vertical park" adjacent to the new Provincial Government Offices and Law Courts Building in Vancouver, B.C. The stairway, which provides access to the terraced, low-rise building at several levels, incorporates a series of diagonal ramps suitable for use by people in wheelchairs. The ramps have a grade of 8 percent (that is, they rise eight feet for every 100 feet of their length), in compliance with the local building code. The $100 million civic center was designed by Arthur Erickson Architects of Toronto. The client was the British Columbia Building Corporation.**


HOTEL REMBRANDT

years old or older have a chronic activity-limiting disability, and 16.9 percent are unable to carry on a major activity.

For more detailed statistics on the prevalence of disability in the U.S. one must turn to another source: the Survey of Income and Education, which was conducted in 1976 by the Census Bureau. The results of this survey are based on interviews with people representing 158,500 households. For the purpose of the survey people were considered disabled if they had a chronic health condition that prevented them from participating in a major activity appropriate to their age group.

The Survey of Income and Education indicates that blacks are significantly more likely to be disabled than whites: 17.6 percent of the black population surveyed were found to be disabled, compared with 13.7 percent of the white population. A more striking contrast can be found in the 18-to-64 age group, in which 19.4 percent of the black population and 12.6 percent of the white population were found to be disabled. People classified as Hispanic were found to have the lowest rate of disability: 10.6 percent.

According to the same survey, the prevalence of disability varies inversely with education. In the 18-to-64 age group 38.5 percent of those with less than an eighth-grade education, 10.3 percent of those with a high school education and 5.3 percent of those with a college degree were disabled. Because the ability to work is central to the definition of disability for this age group, a person's level of education can significantly affect his or her ability to obtain work with or without a long-term health condition.

Adopting the Federal poverty line as a benchmark, the Survey of Income and Education found that poor people in the 18-to-64 age group were much more likely to be disabled than nonpoor working-age people (28.7 percent as against 11.8 percent). Another way of stating the relation between disability and poverty for this age group is to point out that disabled people were more than twice as likely to be poor than nondisabled people (18.7 percent as against 7.1 percent).

The direction of causation between poverty and disability is not well understood. Does poverty cause disability or does disability cause poverty? Based on national survey data from the 1960's Howard Luft of Stanford University estimates that at least 30 percent of the Americans with disabilities are poor because of their health problems. He reports that among white males the figure rises to 75 percent. Luft argues that the causal sequence from poverty to disability is less clear and cannot be easily disentangled from the effects of other factors such as age, sex, race and education.

The prevalence of disability varies by as much as 45 percent from one region of the country to another. Data from the Survey of Income and Education indicate that for people three years old and older New England has the lowest rate of disability (12.6 percent), whereas the East South Central region has the highest (18.3 percent). Among the contiguous 48 states Nebraska has the lowest disability rate (11.1 percent); West Virginia has the highest (21.9 percent).

Most of these statistics reveal little about the demands that disability places on the nation's system of health and human services. The two largest Federal programs for disabled people in the U.S. are the Social Security Disability Insurance program and the Supplemental Security Income program, both of which are managed by the Social Security Administration. In 1980 these two income-transfer programs made

| YEAR | PUBLIC LAW NO. | TITLE OF LAW | KEY PROVISIONS |
|---|---|---|---|
| 1968 | 90–480 | ARCHITECTURAL BARRIERS ACT | Requires that buildings built with Federal funds or leased by the Federal Government be made accessible |
| 1970 | 91–453 | URBAN MASS TRANSPORTATION ACT | Requires eligible local jurisdictions to plan and design accessible mass-transportation facilities and services |
| 1973 | 93–87 | FEDERAL AID HIGHWAY ACT | Requires that transportation facilities receiving Federal assistance under the act be made accessible; allows highway funds to be used to make pedestrian crosswalks accessible |
| 1973 | 93–112 | REHABILITATION ACT | Prohibits discrimination against qualified handicapped people in programs, services and benefits that are Federally funded; creates Architectural and Transportation Barriers Compliance Board |
| 1975 | 93–391 | DEPARTMENT OF TRANSPORTATION APPROPRIATIONS ACT | Prohibits purchase of mass-transit equipment or construction of facilities unless they are accessible to elderly and handicapped people |
| 1975 | 94–103 | DEVELOPMENTAL DISABILITIES ASSISTANCE AND BILL OF RIGHTS ACT | Establishes protection and advocacy systems for developmentally disabled people; establishes representative councils in each state for developmentally disabled people |
| 1975 | 94–142 | EDUCATION FOR ALL HANDICAPPED CHILDREN ACT | Provides for a free appropriate education for handicapped children in the least restrictive setting possible |
| 1975 | 94–173 | NATIONAL HOUSING ACT AMENDMENTS | Provides for the removal of barriers in Federally supported housing; establishes Office of Independent Living for disabled people in U.S. Department of Housing and Urban Development |
| 1978 | 95–602 | REHABILITATION COMPREHENSIVE SERVICES AND DEVELOPMENTAL DISABILITY AMENDMENTS | Establishes independent living as a priority for state vocational rehabilitation programs; provides Federal funding for independent-living centers |
| 1980 | 96–265 | SOCIAL SECURITY DISABILITY AMENDMENTS | Removes certain disincentives to work by allowing disabled people to deduct independent-living expenses in computing income benefits |

**FEDERAL DISABILITY LAWS passed by Congress between 1968 and 1980 marked a shift in the public policy of the U.S. from one that focused primarily on providing personal-adjustment services to disabled individuals to one that concentrated more on making the built environment accessible to all Americans. In addition to the major pieces of legislation listed here Congress enacted numerous other measures aimed toward the same objective. Much of this legislation has not been fully implemented by the Reagan Administration.**

$20.6 billion in cash payments to more than four million working-age people with disabilities.

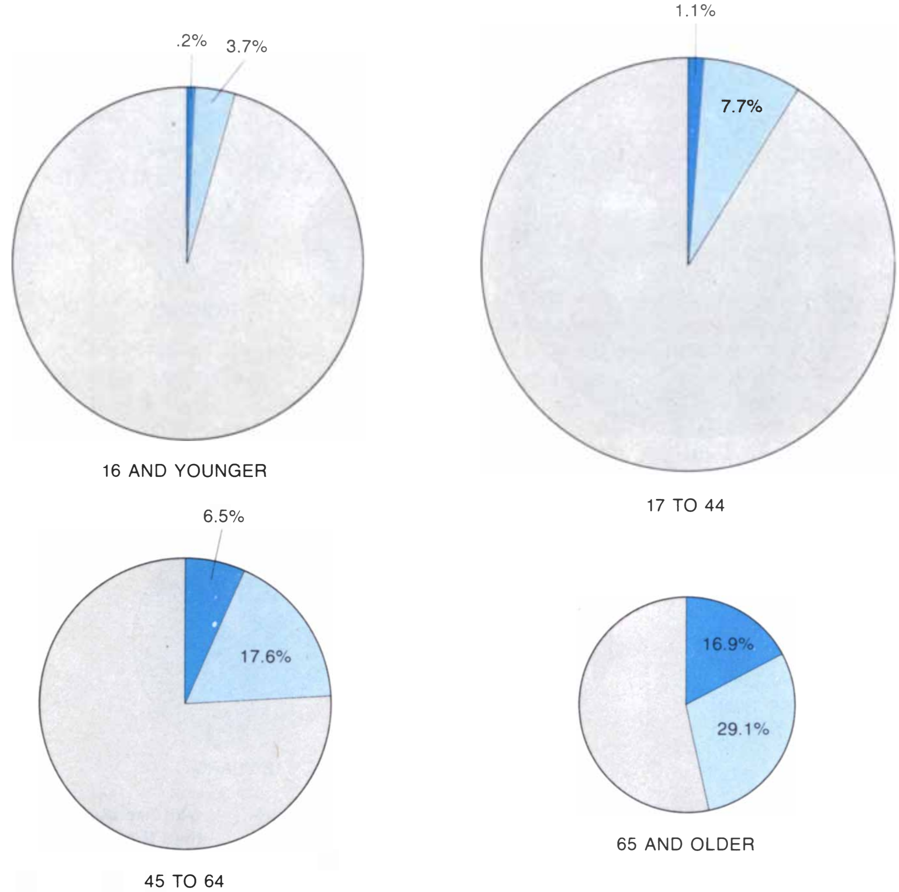
Monroe Berkowitz of Rutgers University estimates that disability-related expenditures of $63.5 billion were made on behalf of working-age people in 1977, the latest year for which aggregate cost data are available. Of this figure $47.6 billion went to public and private income-transfer payments, $12.9 billion went to medical care and $3 billion went to direct services and various labor-market interventions such as job programs and vocational rehabilitation. None of these figures includes transfer payments or services on behalf of children or people of retirement age.

The number of people who receive disability-related income-transfer payments has stabilized in recent years. Nevertheless, the problem of chronic disability continues to have a disproportionate and growing impact on the health-care system. Data from the 1979 Health Interview Survey indicate that although people with a chronic activity limitation constitute less than 15 percent of the population, they account for 29.3 percent of all visits to physicians, 40.1 percent of discharges from short-stay hospitals and 58 percent of days spent in short-stay hospitals. These percentages all represent increases from previous years.

In spite of the prominence of the issue of environmental accessibility in recent discussions of disability policy, little is known about the size of the disabled population for whom environmental accommodations are required. An important statistic to consider in this context is the number of "mobility-impaired" people in the population.

One indicator of mobility impairment is the employment of various devices to aid mobility. According to the 1977 Health Interview Survey, about five million people in the noninstitutionalized civilian population rely on one or more devices of this kind. Another million or so people in nursing homes use mobility aids or are confined to bed. It is fair to say that the problem of physical accessibility is most acute for these six million people. In addition there are thousands of Americans who do not use mobility aids but are nonetheless limited in their mobility because of a chronic health condition. A common example would be the person with coronary heart disease who cannot climb stairs without inducing a painful attack of angina.

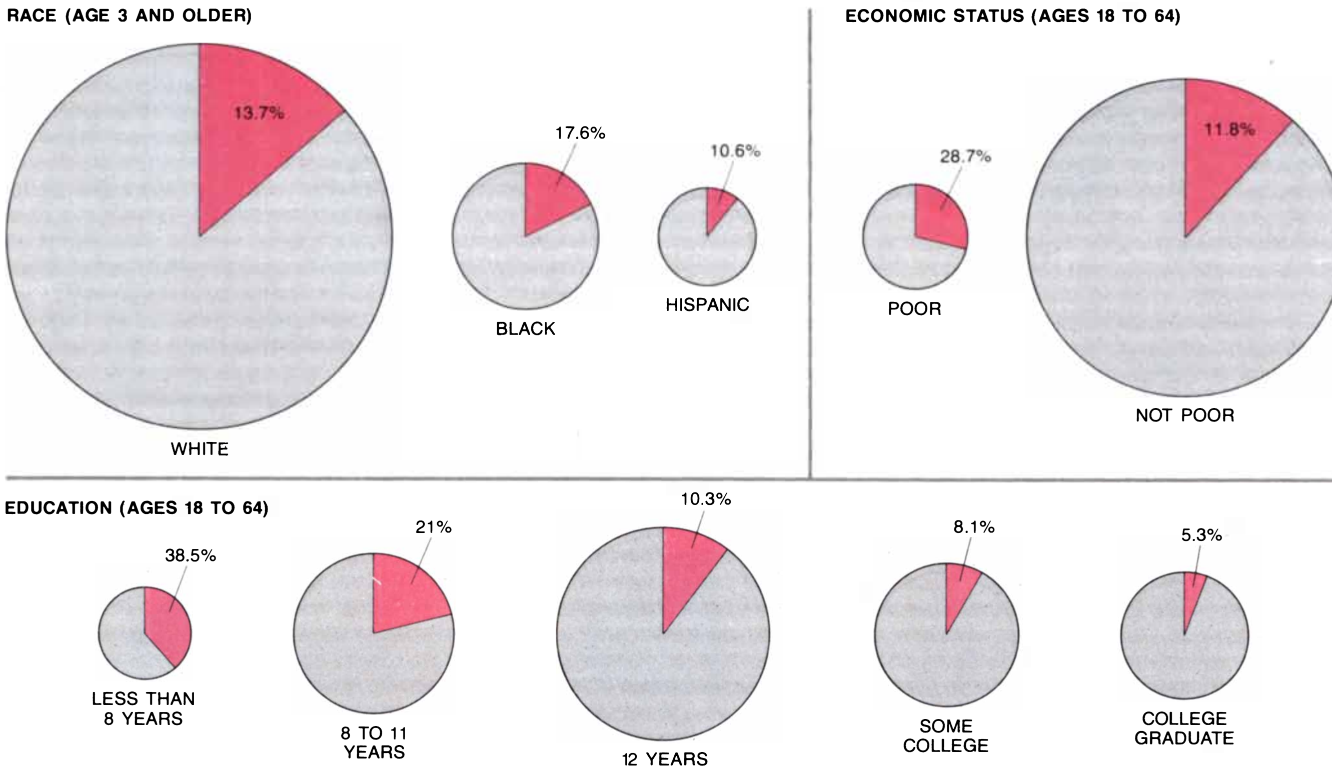
Most striking, we believe, is the recent growth in the use of mobility aids. Comparative data from the 1969 and the 1977 Health Interview Survey suggest that there has been a 40 percent growth in the use of such aids in this period. The same source indicates that there are some 645,000 people in the noninstitutionalized civilian population who use wheelchairs. Based on our knowledge of selected disabling conditions, we believe the results reported by the Health Interview Survey are somewhat low. Our impression is supported by the Social Security Administration's 1978 Survey of Disability and Work among people in the 18-to-64 age group. The Social Security Administration reports that there are 907,000 people of working age alone who use wheelchairs. No estimate is given for children or older people. As for the institutionalized population, the National Center for Health Statistics, based on its 1977 National Nursing Home Survey, reports that there are 517,000 people in nursing homes who use wheelchairs.




**PREVALENCE OF DISABILITY** **increases significantly with age. The area of each of the circles in this chart is proportional to the number of Americans in the associated age group. The colored sections show the relative size of the disabled population in each age group. The light-colored sections correspond to the percentage of people with a limited disability, the dark-colored sections to the percentage of people with a severe disability. The chart is based on the 1979 Health Interview Survey, conducted by the National Center for Health Statistics. For this purpose people were classified as severely disabled if they were unable to carry on a major activity appropriate to their age group, such as attending school, working or housekeeping.**

Most surveys of health and disability provide only an instantaneous view of disability in the U.S. More interesting from the standpoint of disability policy is the fact that the prevalence and composition of disability has changed over time. One reason for the change is the shifting age structure of the population. Since the turn of the century there has been a 27-year increase in life expectancy at birth: from 47 years in 1900 to 74 years in 1980. As a result there has been an eightfold increase in the size of the population 65 years old and older: from three million (representing 3 percent of the population) in 1900 to 25 million (representing 11 percent of the population) in 1980. As the life span has increased, so has the probability that an individual will acquire a major disability-inducing disease, and as the population of the elderly has increased, so has the prevalence of disease and disability.

The 1930's marked a major turning point in the medical management of chronic disease and disability. Until then people seldom died because of the disease or disability; usually they died because of secondary infections such as pneumonia or urinary-tract disease. The introduction of sulfonamide drugs in the late 1930's and penicillin and other antibiotics in the 1940's mitigated such terminal infections. The net effect was that the prevalence of certain diseases and disabilities increased because their average duration was prolonged.




**DEMOGRAPHIC ASPECTS of disability in the U.S. are represented here in terms of race, economic status and education. The statistics, which are derived from the Survey of Income and Education, conducted in 1976 by the Bureau of the Census, show that people who are black, poor or less educated are more likely to be disabled than the rest of the population. Again the area of each circle is proportional to the number of Americans in each category; the colored section in this case represents the total percentage of disabled people.**

Although it has now been more than four decades since the introduction of sulfonamides, health surveys still reveal large increases in the prevalence of disability. According to Alain Colvez and Madeleine Blanchett of the Quebec Ministry of Social Affairs, the results of the Health Interview Survey indicate that the number of Americans with a chronic activity limitation increased by 8.2 million, or 37.3 percent, from 1966 to 1976; meanwhile the general population increased by only 10 percent.

More impressive was the large increase in severe disability. Colvez and Blanchett report that the number of Americans unable to carry on their major activity because of a chronic health condition increased by 3.4 million, or 83.2 percent, during the 10-year period. In terms of prevalence rates, the rate of severe disability increased from 213 people per 10,000 in 1966 to 355 per 10,000 in 1976: an increase of 66.7 percent. More recent data from the Health Interview Survey indicate that by 1979 the prevalence rate had increased to 365 people per 10,000: an increase of 71.4 percent since 1966. The leading causes of activity restriction have also changed over this period. Arthritis and rheumatism have replaced heart disease as the leading cause of disability, and diabetes has become the fastest-growing disabling condition among the leading causes of disability.

Colvez and Blanchett argue that none of these developments can be attributed solely to factors such as methodological changes in the Health Interview Survey, changes in the respondent's perception of what constituted disability or changes in the size or overall age distribution of the U.S. population. Although we cannot offer a comprehensive reason for the findings reported by Colvez and Blanchett, we believe at least some of them can be attributed to the introduction of new medical interventions that are now saving lives at the price of longer-term and severer residual disabilities. In addition new interventions are in some cases significantly extending the lives of those already disabled.

The combination of these two trends augurs a variety of new demands facing American society in the form of new technology and highly intensive services: nonvocal communication devices, respirator management, 24-hour attendant care and a variety of community-based long-term care services. At present few of these services are routinely available; usually they can only be acquired with great difficulty from multiple providers and funding sources.

These developments present important ethical and public-policy issues that cannot be ignored in any consideration of the demands of disabled people for fuller access to society and its institutions. Many questions have already been asked about the quality of life that can be expected for those whose existence appears marginal. These questions, however, implicitly place the onus on the disabled individual rather than on society, whose services can in many cases materially affect the quality of life. What about society's obligation to provide a minimum quality of life for those who survive as the result of medical research and technological advances that have often been publicly demanded, publicly supported and publicly financed? The financial consequences of these public decisions are so overwhelming that few individuals or families are capable of assuming the financial burden for services and environmental supports needed to provide a minimum quality of life.

In spite of the growing number of people with severe physical disabilities, the total number remains small compared with the rest of the population. Moreover, the needs of such people are varied and often highly specialized, making it difficult to form an effective political constituency representing these needs. Meanwhile this segment of the population depends on the limited awareness of a larger public, which is hesitant to provide needed services and accommodations. It is not widely understood that these needs arise because society has been successful—both in saving and in prolonging lives. The new social needs are the price of that success.

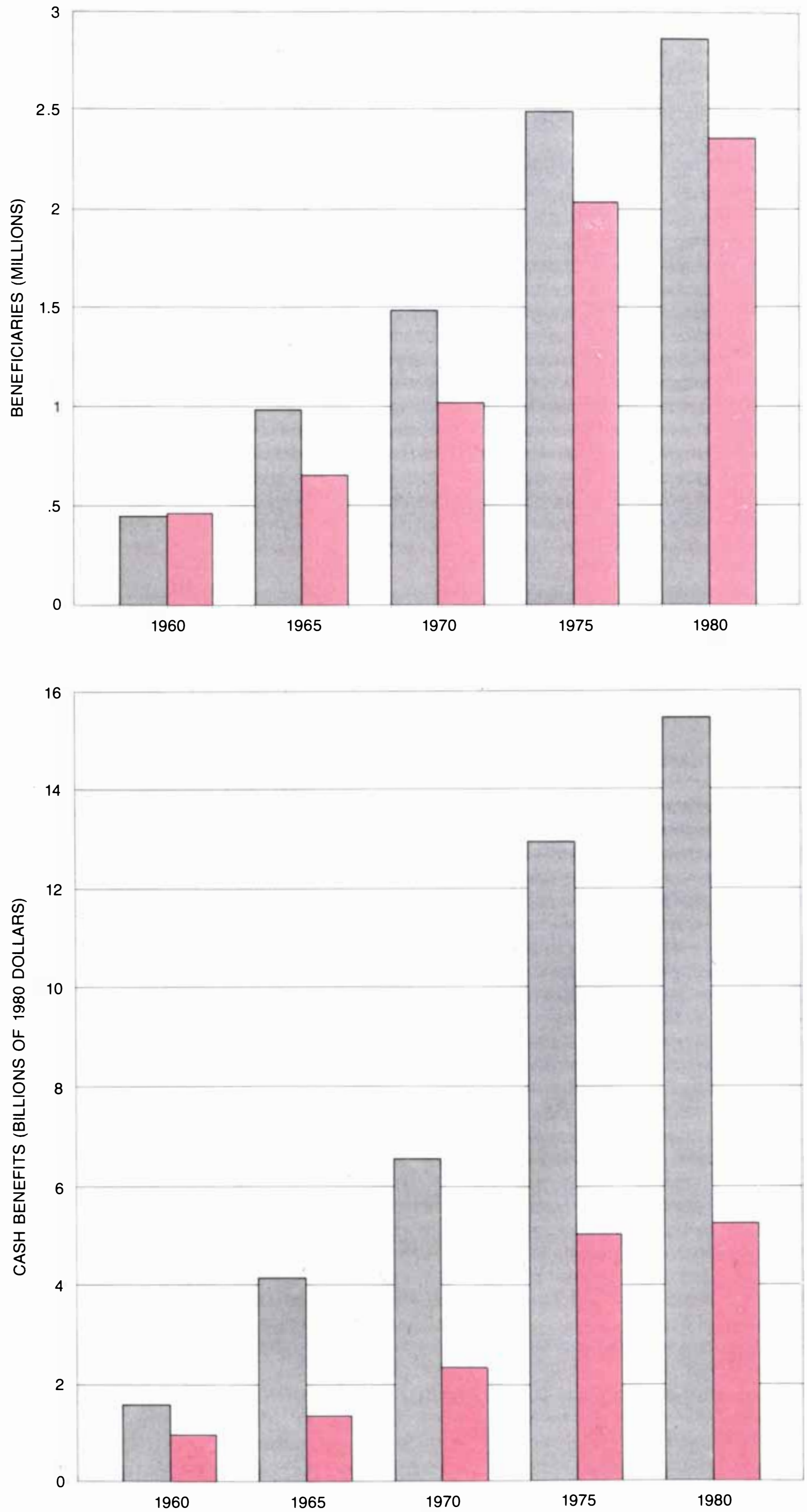
The disabled population in the U.S. has also undergone a political and social transformation in the form of a social movement commonly referred to as the independent-living movement. This movement has changed how disabled people view themselves and how they wish to be viewed in American society: no longer as passive victims deserving of charitable intervention but as self-directed individuals seeking to remove environmental barriers that preclude their full participation in society.

The independent-living movement had its origins in the early 1970's when small groups of disabled people banded together to form self-help organizations known as independent-living centers. The first centers were established in Berkeley, Calif., and Boston, Mass. Today about 100 independent-living centers across the country provide or arrange for a variety of services such as housing referral, transportation, attendant care, peer support and legal advocacy. The leaders of the independent-living movement have also been active on the political front: in promoting the passage of disability legislation, in seeking to make local transit systems more accessible, in removing architectural barriers and in changing the awareness of political leaders and other decision makers about the particular needs of disabled people.

The independent-living movement offers a radically different view of the problem of disability and its solution. According to representatives of the movement, the problem of disability is one not only of physical impairment but also of unnecessary dependence on relatives and professionals, of architectural barriers and of unprotected rights. In this view the pathology is not in the individual, as the medical model would suggest, but rather in the physical, social, political and economic environment that has up to now limited the choices available to people with disabilities. The solution to these problems is not more professional intervention but more self-help initiatives leading to the removal of barriers and to the full participation of disabled people in society.

The independent-living movement should not be viewed as an iconoclastic one but rather as one that affirms traditional American values such as self-reliance. The movement's commitment to the nation's political and economic institutions is remarkable, considering how disabled people have been excluded from the mainstream of the nation's political and economic life. Disabled people still want to be part of the system. They want access.

The barrier-free movement—a shorthand term to describe the national effort




**TWO LARGEST FEDERAL INCOME-TRANSFER PROGRAMS for disabled people of working age have grown rapidly over the past two decades, as this pair of charts shows. The two programs represented are the Social Security Disability Insurance program (*gray bars*) and the Supplemental Security Income program and its predecessors (*colored bars*). Both programs are now run by the Social Security Administration. Between 1970 and 1980 the number of disabled people receiving benefits under one or both of the programs doubled (*top chart*). Since 1977 the number of beneficiaries in both programs has leveled off. The total benefits paid under both programs also rose sharply, reaching a total of $20.6 billion in 1980 (*bottom chart*).**

to eliminate architectural barriers—actually preceded the rise of the independent-living movement. Although the exact moment of conception cannot be determined, the barrier-free movement goes back at least to 1958. That was the year the President's Committee on Employment of the Handicapped, with the help of the Veterans Administration, published a tentative guide to help make public buildings accessible to people with impaired mobility. The effort, which depended entirely on voluntary action, produced no discernible results.

A milestone year for the barrier-free movement was 1961. That was the year the American National Standards Institute (A.N.S.I.), a private architectural standard-setting organization, issued Standard A117.1. The standard specified the minimum requirements and working details for structures such as walkways, parking spaces, ramps, stairs, floor surfaces, mirrors, water fountains, public telephones, control identifications and warning signals. Work on the A.N.S.I. standard was sponsored by the President's Committee on Employment of the Handicapped and by the National Easter Seal Society for Crippled Children and Adults.

The promulgation of the A.N.S.I. standard in 1961 and the widespread endorsement it received generated interest in the architectural-barriers problem and created the expectation that the problem would be speedily addressed. On the Federal level, however, no action beyond voluntary compliance was immediately taken to adopt the standard with respect to Federal construction. In response, as part of the 1965 Amendments to the Vocational Rehabilitation Act, Congress established the National Commission on Architectural Barriers to Rehabilitation of the Handicapped in order to study the problem.

The commission's study and three concurrent surveys found compliance to be minimal. Buildings constructed and renovated with Federal funds remained inaccessible. At the local level only 3 percent of the cities surveyed had adopted accessibility codes, and roughly two-thirds of the architects surveyed were unfamiliar with accessibility standards. More than anything, these surveys demonstrated that voluntary compliance alone would not yield the intended results.

The shortcomings of voluntary compliance led Congress to pass the Architectural Barriers Act of 1968. The act's stated purpose was to "ensure that certain public buildings, financed with Federal funds, are so designed and constructed as to be accessible to the physically handicapped." The act delegated authority for setting accessibility standards to various Federal departments and agencies. There was no attempt to add to or take away from what had been defined as the minimum standards of access already accepted in principle by Federal agencies since 1961. In fact, each of the delegated Federal agencies promptly adopted the 1961 A.N.S.I. standard as the standard for their respective regulations.

The task of enforcing the 1968 act, however, remained unresolved until the enactment of the Rehabilitation Act of 1973. Section 502 of the Rehabilitation Act established the Architectural and Transportation Barriers Compliance Board, which was authorized to investigate complaints, hold public hearings, issue compliance orders and seek enforcement of its orders by the courts if necessary.

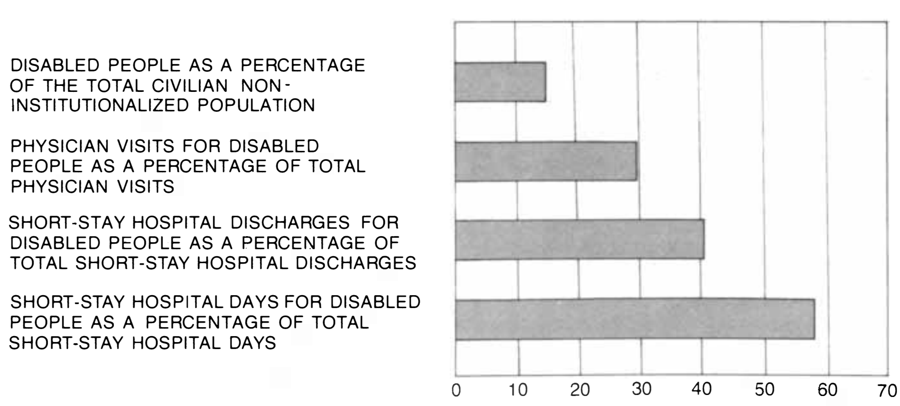
The legal machinery to address the problem of architectural barriers was strengthened in 1977 when regulations pursuant to Section 504 were issued by the Department of Health, Education, and Welfare. The regulations implementing Section 504 required that programs receiving Federal financial assistance be made accessible to, and usable by, disabled people in order to provide them with effective services. The regulations specified that new facilities constructed after June 3, 1977, would have to meet accessibility standards and that existing facilities would have to be made accessible by June 2, 1980. The regulations also specified that conformance with the A.N.S.I. standard would constitute compliance, but other methods would be allowed if they provided equivalent access. The effect of the regulations was broad, since Federal financial assistance extends to many areas such as education, transportation facilities, public housing and other activities affecting the lives of disabled people.

The Federal lead with respect to architectural accessibility has been followed by the states. All 50 states have enacted laws requiring that certain facilities be made physically accessible to disabled people. According to information compiled by the National Center for a Barrier Free Environment, approximately half of the states rely either on the original 1961 A.N.S.I. standard or on a slightly modified 1971 version. The other half have adopted their own standards or have borrowed standards from another state.

A survey of state accessibility laws in 1979 by the American Bar Association found that state legislation is not equally forceful throughout the country. The survey made a distinction between two types of state accessibility laws. Laws of the first type are classified as declaratory: they merely assert that disabled people have a right of access. According to the bar association, such laws "have consistently been held by the courts to be insufficient manifestations of legislative intent to require any environmental modifications to facilitate such use." Laws of the second type are considered more affirmative: they "vary from simple delegations of authority to promulgate rules, to lengthy acts incorporating minimum standards... and specific administrative and enforcement methodologies."

One of the most troublesome problems faced by architects in trying to eliminate architectural barriers in their designs is the application of accessibility standards to different settings. For example, there is little dispute over the specifications for an accessible bathroom, but there is no agreement on how many accessible bathrooms there should be or how they should be spaced throughout a building. It is the implementation and application of standards that have caused major difficulties for practicing architects.

Although most Federal, state and lo-




**UTILIZATION OF HEALTH-CARE SERVICES by disabled people is disproportionately high. As these bars for the year 1979 indicate, although people with a chronic activity-limiting health condition totaled less than 15 percent of the U.S. population, they accounted for much higher percentages of all visits to a physician, discharges from short-stay hospitals and days spent in short-stay hospitals. The data are from the National Center for Health Statistics.**

cal agencies have embraced the A.N.S.I. standard, many have adopted different criteria for applying it. Hence architects sometimes find themselves faced with conflicting regulations, for example when Federal, state and local agencies join on a public housing project. A similar problem can arise in private commercial construction where the public is served. Even here the architect may be confronted with state and local building codes that have different requirements.

The architect is often caught between the requirements of his client and the demands of people with disabilities. In the past most clients knew or cared little about barrier-free design. Most doubted that there was even a serious demand for accessible facilities. Today there is a much greater awareness of the scope of disability and the need for various types of environmental accessibility.

In spite of increased awareness over the past 15 years, it is not clear that clients are any more interested today in making their facilities accessible. Architects observe that clients often approach the issue of accessibility with one or more of three attitudes: (1) disabled people are not viewed as customers, clients or tenants; (2) barrier-free design is automatically assumed to be too expensive; (3) disabled people are thought likely to stigmatize the facility in the minds of nondisabled users. Few architects are equipped to cope with clients who are so disposed. Meanwhile on the other side of the issue are advocacy groups on behalf of disabled people who are often assumed to be making unreasonable demands.

Unfortunately there are few mechanisms available to reconcile such differences. The problem is not merely the accessibility standards or codes but the failure to provide the technical assistance needed to implement them. By themselves standards and codes are viewed as architectural constraints that invite resistance, if not hostility. In addition there is the constant worry about building inspectors and the constant threat of court action, real or imagined. What is needed is a technical-assistance body that can offer creative solutions meeting both the letter and the spirit of existing standards and codes. What is also needed is a decision-making body that can render these creative solutions and compromises legally binding.

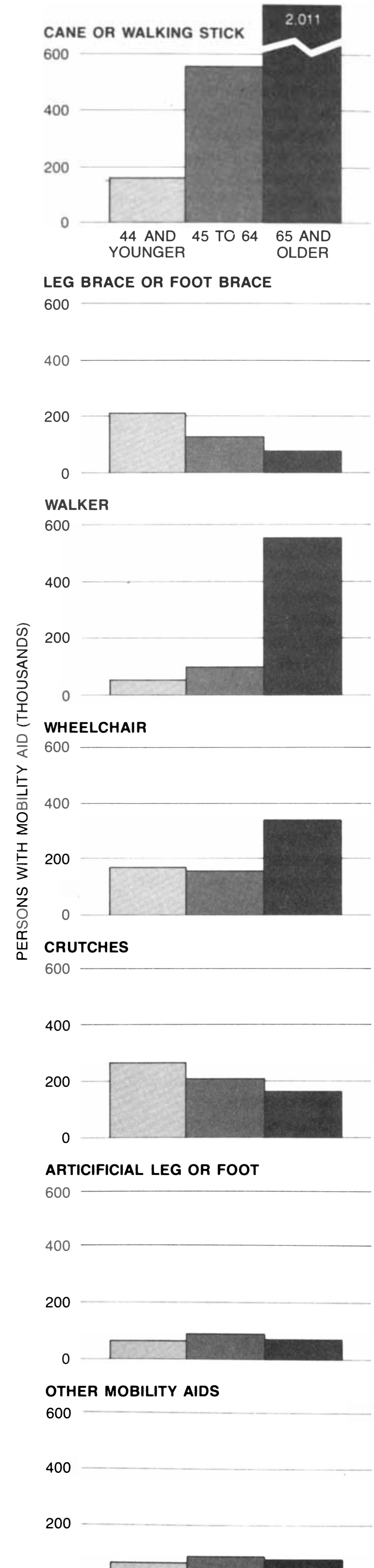
One model of technical assistance is the Special Office for the Handicapped in the North Carolina Department of Insurance. The office includes an architect who works with practicing architects and local building inspectors to administer accessibility standards. As a result the North Carolina architectural-barriers compliance board has received only two requests for waivers over a nine-year period. On the national level the National Center for a Barrier Free Environment, based in Washington, serves as a nongovernmental technical-assistance organization seeking to advise architects, clients and various regulatory boards. Since building codes are largely a state and local matter, however, the greatest need for technical assistance remains at the state and local level.

Architects are not neutral observers merely trying to accommodate the demands of clients and disabled users. Most architects are able-bodied and bring to the design process all the able-bodied attitudes and assumptions that have shaped design concepts in Western culture. Moreover, the architect, like most people, has become accustomed to the social segregation of disabled people in the environment.

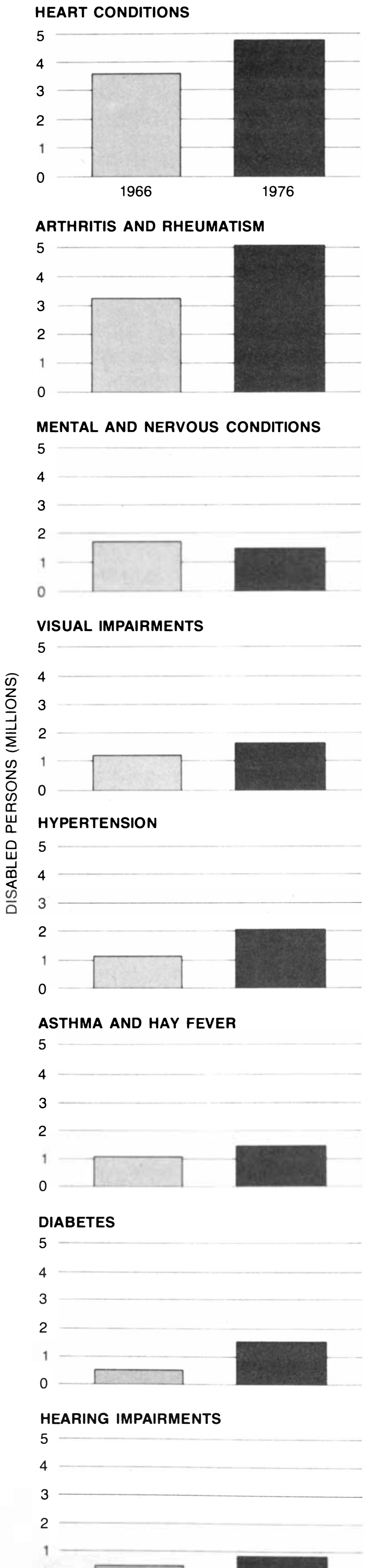
What can be done to correct for this "cultural lag" in professional architecture? We should like to suggest three steps. First, architects must come to know more about disability, just as in recent years they have come to know more about solar energy and climate and how to design with these environmental factors in mind. Architects must come to understand that disabled people are a complex collection of individuals with diverse functional limitations. An important place to start is in the curriculum of both graduate and undergraduate schools of architecture.

Second, architects need to seek guidance from outside their profession. For example, architects are known to enlist the help of behavioral scientists in the design of specialized environments such as housing for the elderly. Now they need to accept the idea that disabled people have something to tell them. Architects need to involve disabled people as design consultants and learn where and how such consultants can be employed.

Third, the architectural profession, through its association and publications, needs to actively support the barrier-free movement and to reward practitioners for outstanding work. So far the profession has viewed the barrier-free movement largely as an infringement on its creative freedom and has treated accessibility standards as a "cookbook" approach to design. The American Institute of Architects, the national professional organization, has never come out openly against accessibility standards, but it has done little to promote the concept of accessibility. Its barrier-free committee was never popular and was




**MOBILITY AIDS are now employed by millions of Americans with disabling health conditions, as these bars for the year 1977 indicate. The use of such devices has increased sharply in the U.S. in the past decade or so.**




**MAJOR CAUSES of disability in the U.S. are represented in this chart for the years 1966 and 1976. In the intervening period the population of the U.S. increased by 10 percent, and the number of people with a disability rose by 37.3 percent. Arthritis and rheumatism have replaced heart disease as the leading cause of disability. Meanwhile diabetes has become the fastest-growing major cause of disability.**

eventually disbanded. Even the institute's Bartlett Award, given for excellence in accessibility design, fell from favor because of the careless way winners were selected and because some winning designs did not comply with minimal national standards. Similarly, architectural magazines have done little to promote accessibility as an important design issue. Coverage of barrier-free architecture has mainly been confined to technical or business pages or has appeared in connection with institutional buildings designed for special populations.

Although technical assistance and professional awareness are indispensable, there still is a need at the national level for leadership in developing a comprehensive standard for accessibility. For two decades the 1961 A.N.S.I. standard served as the point of departure for the development of accessibility regulations. The inadequacies of the 1961 A.N.S.I. standard have allowed Federal agencies to apply it in conflicting ways. The price of varying Federal requirements has been high both in dollars and in frustration.

Recognizing the shortcomings of the 1961 A.N.S.I. standard, the Department of Health, Education, and Welfare—in cooperation with the President's Committee on Employment of the Handicapped and the National Easter Seal Society—awarded a contract in 1974 to Syracuse University's School of Architecture to revise and update the 1961 standard. What started as a two-year project became a six-year conflict ending in 1980 when a revised standard was formally adopted by the American National Standards Institute. In spite of the consensus-building efforts of those associated with the Syracuse project, several participants representing Federal agencies voted against the new A.N.S.I. standard on the final ballot.

The failure to achieve a consensus in 1980 set the stage for what has been one of the most acrimonious debates on the issue of environmental accessibility. The General Services Administration and the Department of Health, Education, and Welfare announced the development of a new accessibility standard, as did the Postal Service. In addition the Architectural and Transportation Barriers Compliance Board, as authorized under a 1978 statute, proceeded to develop a set of "minimum guidelines and requirements" that was to serve as the basis for the setting of standards by other Federal agencies. The emerging consensus that had led to the 1980 A.N.S.I. standard suddenly unraveled with the proliferation of conflicting accessibility standards. The situation was further complicated by the arrival of a new Administration that in many ways preferred no regulations at all.

For most of 1981 the conflict was centered on the compliance board. The opening salvo was the board's publication of Minimum Guidelines and Requirements for Accessible Design. The new standard appeared in the January 16 *Federal Register,* just four days before the Reagan Administration took office. The new minimum guidelines were characterized by Vice-President George Bush's Regulatory Review Task Force as overzealous and an example of unnecessary Government regulation. The compliance board soon found itself targeted for extinction by the Office of Management and Budget. The board survived the threat but not without further tumult.

The compliance board was itself divided—11 citizens (nine of whom were disabled) representing the public and 11 members representing various Federal agencies. The scales were tipped when one of the public members was persuaded by the Administration to join the Federal members in a 12-to-10 vote to rescind the minimum requirements. The Federal members argued that the new requirements were too costly. Total recision was averted late in 1981 when compromises were made to eliminate the most costly provisions. The Administration had also come to realize that uniform minimum requirements would simplify rule making for the other Federal standard-setting agencies. On August 4, 1982, the compliance board promulgated its final version of the new minimum requirements.

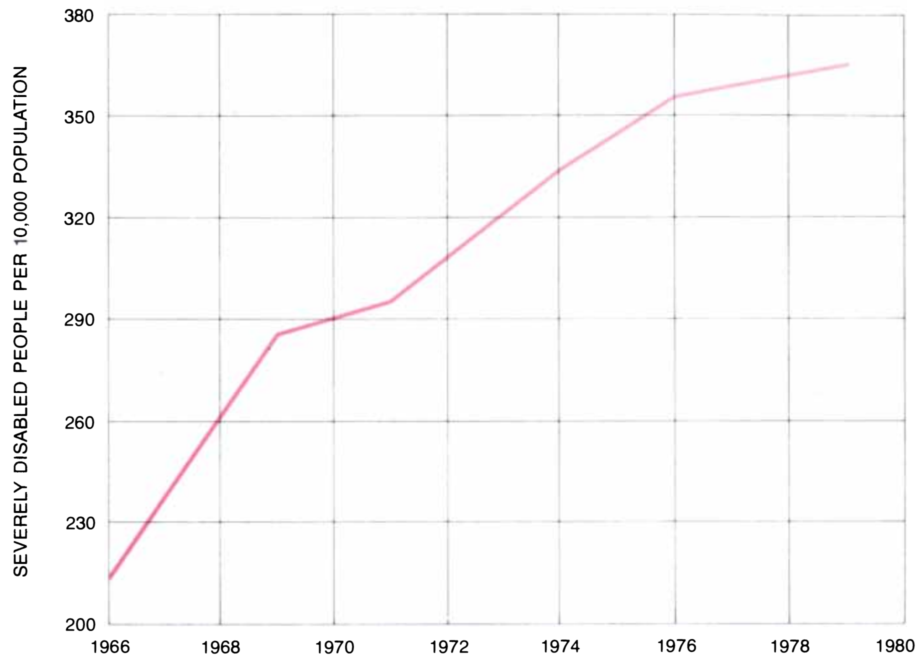
Although the new guidelines do not go as far as many barrier-free advocates would like, there is at least a consensus on what the minimum requirements should be. The minimum requirements issued by the compliance board are similar to the 1980 A.N.S.I. standard and differ mainly with regard to "scoping provisions," or application criteria, an area largely unaddressed by the 1980 A.N.S.I. standard. Even under the board's guidelines most issues of this kind are left to the Federal standard-setting agencies. Among these agencies only the Postal Service appears to be recalcitrant, as was demonstrated by its vaguely worded proposed regulations published in October, 1982.

In spite of the Reagan Administration's attempt to put the brakes on the barrier-free movement, there is in one sense no turning back. The groundwork—accessibility standards, laws, regulations, technical know-how, public consciousness and, to a degree, professional awareness—on which the movement is based is substantial. Many of the gains are here to stay. In short, the efforts of the past two decades to develop accessibility standards and enforcement mechanisms have paid off. The built environment, in spite of many shortcomings, is far more accessible to mobility-impaired persons than it was a few years ago.

In considering the future of environmental accessibility for physically disabled Americans one must consider the full range of policy tools that can be used to stimulate public and private efforts toward achieving a barrier-free environment. The central weakness in the current policy of environmental accessibility is its overreliance on just one approach to policy implementation: the regulatory approach. When this approach is not accompanied by other incentives and inducements, it becomes inherently adversarial: it eventually pits one group against another. The adversarial nature of the regulatory approach has also helped to polarize the national debate on environmental accessibility and to reduce the chances for mutual accommodation.

Any critique of the regulatory approach, however, must take into account the fact that the development of accessibility regulations arose in direct response to the failure of voluntary compliance. To repeal existing regulations would be to return to an earlier era of neglect. The champions of voluntarism would do well to consider this earlier period in the history of the barrier-free movement. Although the development of the first accessibility standards had considerable private sponsorship, the standards themselves had little private endorsement in the form of active implementation.

Seldom can a public-policy goal such as environmental accessibility be obtained with the benefit of just one policy instrument. Other approaches must be tried. Two are noteworthy here: technical assistance and economic incentives.




**INCIDENCE OF SEVERE DISABILITY** has increased by more than 70 percent over the 13-year period from 1966 to 1979, rising from 213 to 365 Americans per 10,000 population.

In our review of the architectural-barriers problem we observed how mediating bodies providing technical assistance to designers, builders, clients and users can help to minimize unnecessary conflict. Inherent in the technical-assistance approach to policy implementation is a large measure of mutual problem solving and cooperation. The need for technical assistance is recognized in the statutory provisions governing the Architectural and Transportation Barriers Compliance Board. So far, however, funding for technical assistance at the Federal level has been quite modest.

One of the most powerful tools of policy implementation is the variety of economic incentives that government offers through the tax system and through government expenditures. The economic-incentives approach to policy implementation has not been widely tried with respect to architectural barriers. Some states have introduced modest tax deductions for the removal of such barriers. Until recently the Federal income-tax code allowed businesses a modest tax deduction of $25,000 in any given year. The statutory authority for the Federal tax deduction, however, expired last year. Although the size of the Federal tax write-off was far too small to be noticed by most large businesses, such cost-sharing schemes can help to foster compliance with governmentally sponsored accessibility standards.

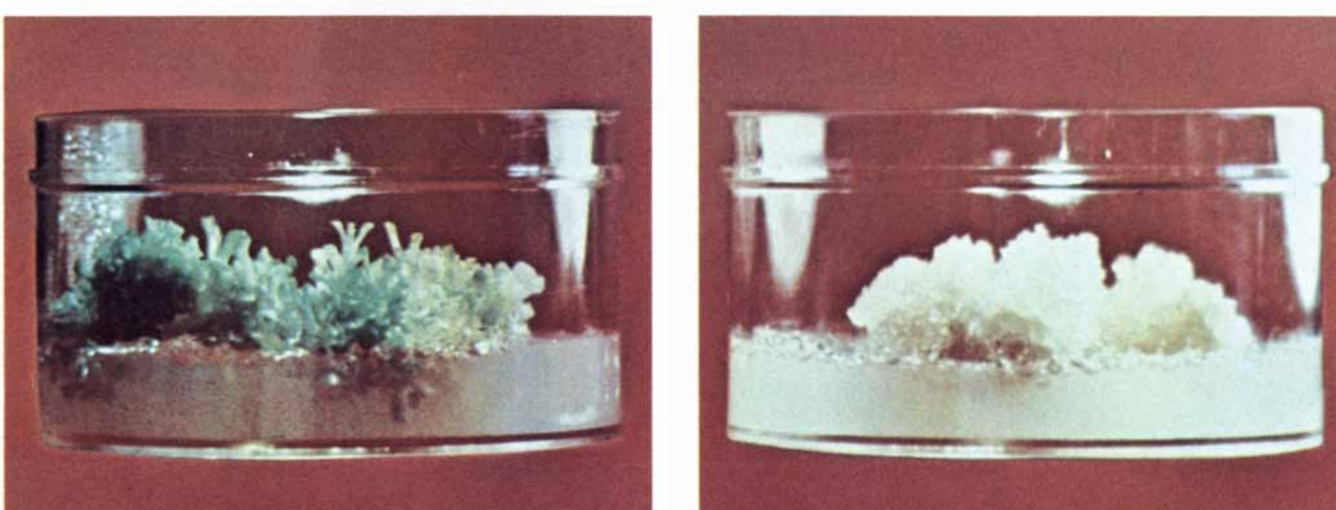
Pervading the public debate on environmental accessibility is a legal standard commonly referred to as the "reasonable accommodation" rule. This rule was affirmed in 1979 by the United States Supreme Court in the case of *Southeastern Community College* v. *Davis,* in which the court ruled that "substantial modifications" resulting in "undue administrative or financial burdens" were not required under Section 504 of the Rehabilitation Act of 1973.

Implicit in the reasonable-accommodation rule is a cost-benefit consideration: Are overall costs reasonable in the light of anticipated benefits? One should be cautious, however, about shaping disability policy solely on the basis of cost-benefit analysis. In the extreme there are instances where taking cost-benefit criteria to their logical conclusion could entail withholding essential life-support systems. We believe in many cases a more appropriate decision rule would be the cost-effectiveness criterion: How, in the face of limited resources, can a particular right be honored or societal responsibility be met in the least costly or most cost-effective way? Phrasing the issue in these terms does not challenge the rights of disabled persons or absolve society of its responsibilities, but it does face up to the unavoidable economic consequences.

The increasing preoccupation with cost-benefit analysis also betrays the degree to which society hesitates to make financial outlays on behalf of disabled people. The current outlook on disability issues would change materially if it were clearly understood, as we have argued, that much of the cost is the product of society's success over the years in saving and prolonging the lives of millions of Americans, often at the cost of a long-term residual disability. This success has its price, and it is a price that must be paid. For disabled people environmental accessibility is fundamental to their quality of life and their ability to participate in the life of the community.

The ultimate and most pervasive of environmental barriers are the attitudinal ones, particularly the view that disabled people are helpless, pathetic victims deserving of charitable intervention. There is now more than enough experience to indicate that disabled people can, with appropriate environmental supports, lead full and independent lives. Without the removal of attitudinal barriers the disability legislation of the past decade will not realize its full promise.

**CROWN GALL** on a tobacco plant has developed where the top of the plant was cut off and the cut stem was inoculated with *Agrobacterium tumefaciens*. The infecting bacteria carry a circular molecule of accessory DNA: the *Ti* (tumor-inducing) plasmid. A segment of the plasmid called *T*-DNA (transferred DNA) is inserted into plant cells. It gives rise to the tumor, which in this case is a teratoma: a tumor mass from which shoots and leaves grow in an unorganized way. Geneticists hope to exploit the plasmid as a vector for introducing genes into plants, giving rise to plants with improved traits. The photograph was made by Robert and B. Gillian Turgeon of Cornell University.



**TOBACCO TUMOR TISSUE** grows in culture in the author's laboratory at Washington University in St. Louis. The bacteria have been eliminated from these cell lines derived from infected plants, but their influence remains as a genetic change in the transformed plant cells. The cells grow in culture without any added auxin or cytokinin, hormones needed to enable normal plant cells to grow in culture. Here two kinds of tumorous growth are shown: a teratoma at the left and a callus (a mass of undifferentiated cells) at the right.

# A Vector for Introducing New Genes into Plants

*The induction of a plant tumor by a bacterium is a natural form of genetic engineering. The piece of DNA the bacterium injects may serve as a tool for the genetic modification of crop plants*

by Mary-Dell Chilton

For a long time, perhaps millions of years, the common soil bacterium *Agrobacterium tumefaciens* has been doing what molecular biologists are now striving to do. It has been inserting foreign genes into plants and getting the plants to express those genes in the form of proteins. In the process the bacterium causes the plant cells to proliferate and form a gall, or tumor, most commonly at or near the junction of the root and the stem (the crown).

Crown gall disease can afflict a wide range of dicotyledonous (broad-leaved) plants, and it causes considerable loss in certain crops, notably grapes, stone fruits and ornamental plants. Since the turn of the century it has been studied by plant pathologists because of its economic impact. In the past few years, as it has become clear that the infective process is a natural form of genetic engineering, *A. tumefaciens* and the tumor it engenders have become the subject of fiercely competitive investigation by molecular geneticists in both academic and industrial laboratories. What has long appeared to be simply the agent of a bothersome plant disease is likely to become a major tool for the genetic manipulation of plants: for putting new genes into plants and thereby giving rise to new varieties with desired traits.

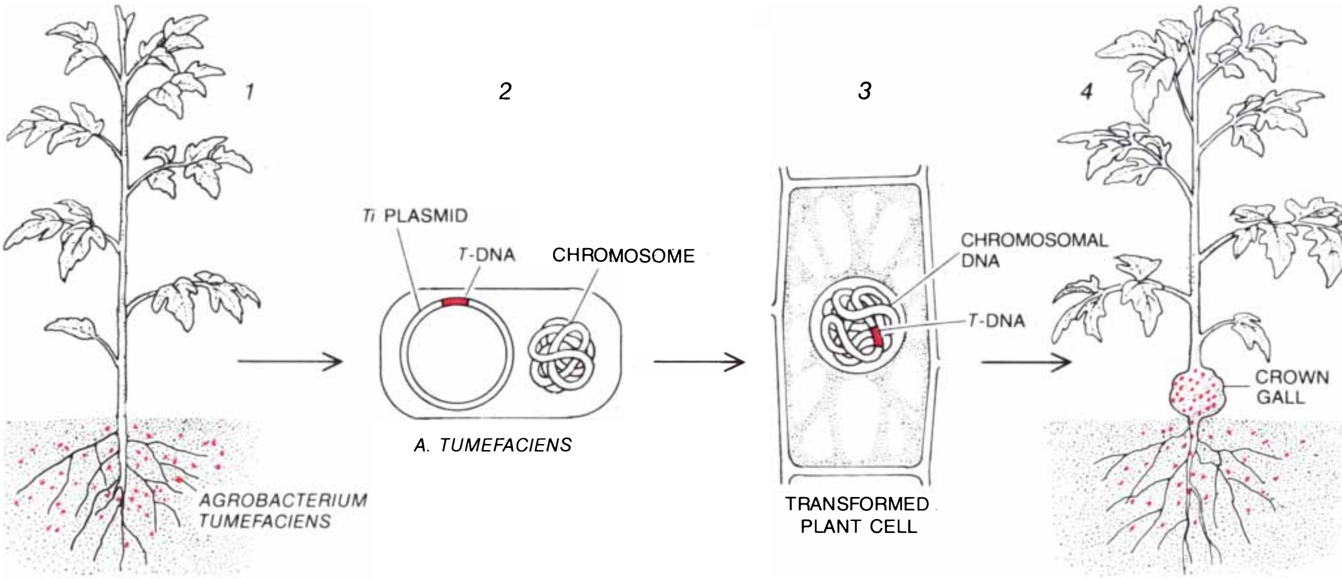
In 1907 Erwin F. Smith and C. O. Townsend of the U.S. Department of Agriculture identified *A. tumefaciens* as the causative agent of crown gall disease. They did so by isolating the bacteria from plant galls and then showing that inoculation of the bacteria into a wound in a healthy plant would generate a new gall. It has since become clear that infection takes place only in freshly wounded tissue. The wound site becomes refractory after only two days or so; indeed, washing the "wound juice" (the sap that flows out of cut tissue) away from the site of a fresh wound can block the induction of a tumor. Not all strains of *A. tumefaciens* are virulent, or capable of inducing a tumor. Virulent bacteria are known to bind to specific pectin constituents of the plant-cell wall. The bacterial cells actually enter only dead, broken plant cells, and yet they stimulate adjacent living plant cells to form a tumor; the bacteria themselves proliferate among the tumor cells.

In 1947 Armin C. Braun of the Rockefeller Institute for Medical Research first succeeded in cultivating gall tissue, free of the inducing bacterium, in the laboratory on a completely defined medium containing only sucrose and inorganic salts. He found that the tissue had the characteristic property of a tumor: a habit of uncontrolled growth. Whereas normal plant cells or tissues grow slowly in culture, and then only when plant hormones (cytokinins and auxins) are supplied, the gall tissue grew rapidly in the absence of any exogenous hormone. It continued to grow luxuriantly when it was transferred to fresh medium time after time. Braun concluded that the plant cells had somehow been transformed—changed to tumor cells—by a hit-and-run act on the part of *A. tumefaciens*. He proposed that the bacterium introduces a "tumor-inducing principle" into the plant cell.

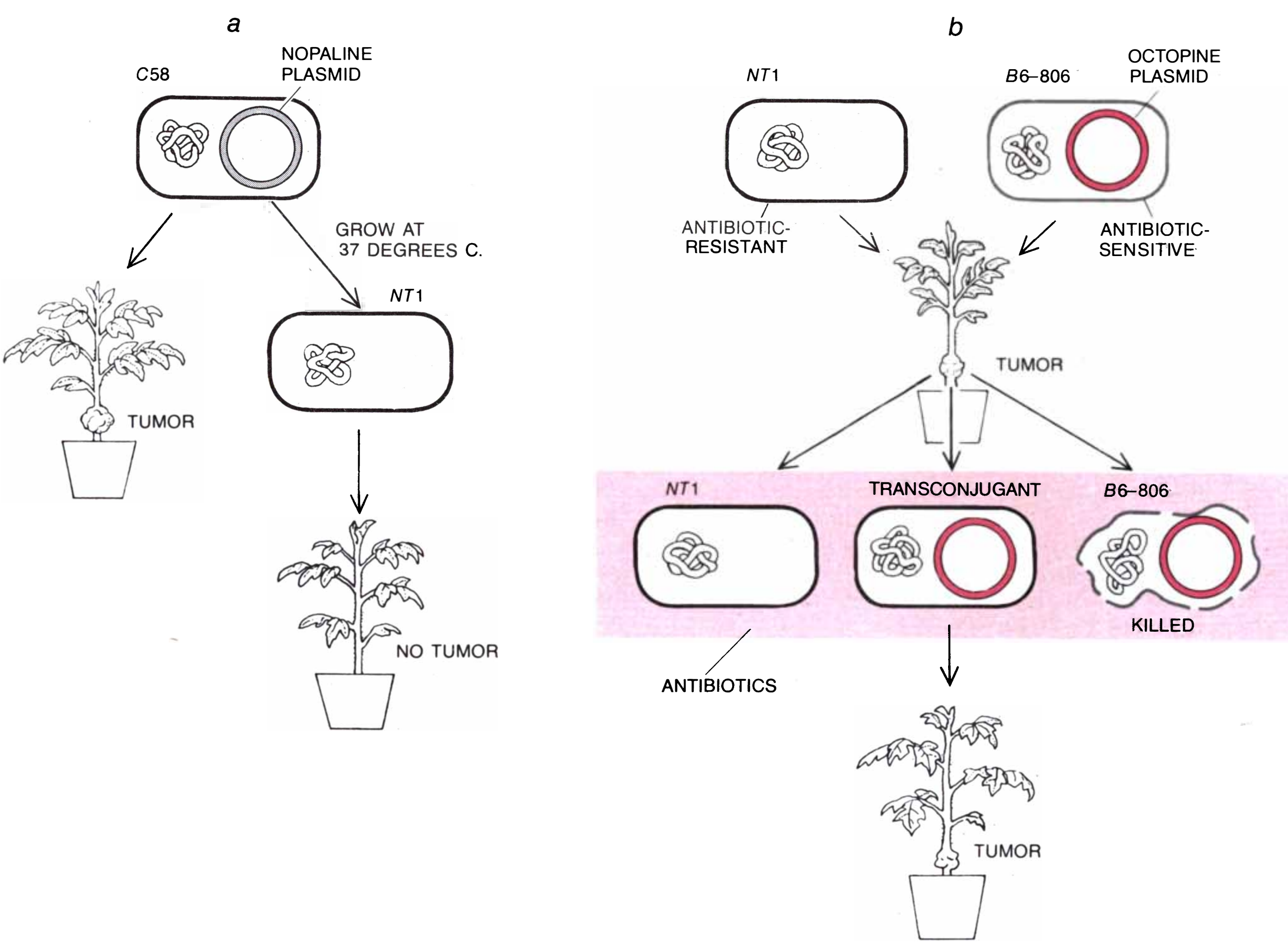
An important metabolic clue to the nature of the tumor-inducing principle was discovered in the 1960's by Georges M. Morel and his colleagues at the Institut National de la Recherche Agronomique in Versailles. They found that crown gall cells synthesize a class of novel chemical compounds (which they named opines) that are not present in normal cells of the same plant. The opines are derivatives of common metabolic intermediates, for the most part amino acids (the constituents of proteins) and various keto acids or sugars. Two opines were studied intensively: octopine, a compound of the amino acid arginine and pyruvic acid, and nopaline, a product of arginine and alpha-ketoglutaric acid. Morel's group established two significant findings. One was that whether a tumor synthesizes octopine or nopaline depends not on the species of the host plant but on the strain of the bacterium that induces the tumor. The other was that a given bacterial strain is able to grow on either octopine or nopaline but not on both; it can metabolize whichever of the compounds is synthesized by the tumor it induces. Morel made a striking prediction. He proposed that the bacterium must insert into the plant cell a gene governing the synthesis of either octopine or nopaline; the tumor-inducing principle must be DNA.

If that is the case, it soon became clear, the DNA must be on some mobile element. Allen Kerr of the University of Adelaide found that the property of virulence was transferred from a virulent strain to an avirulent one when the two strains were inoculated into the same plant. The transfer took place even between distantly related species of *Agrobacterium*, making it unlikely that the genes responsible for virulence were on the bacterial chromosome. Kerr thought the genes were probably carried by a bacterial virus or a plasmid: a small circle of nonchromosomal double-strand DNA, usually carrying ancillary genes such as those affecting pathogenicity, the metabolism of rare compounds or resistance to antibiotics.

Additional evidence that virulence in *Agrobacterium* might be determined by a virus or a plasmid came from the finding that virulence could be lost. In 1971 Robert H. Hamilton, Jr., and M. Z. Fall of Pennsylvania State University discovered a strain of *A. tumefaciens* whose virulence trait is temperature-sensitive. When strain *C*58 is grown at a moderately high temperature (37 degrees Celsius), the bacteria lose virulence: they never spontaneously regain the ability to induce a tumor. In other words, something is permanently lost from the cells. Again the implication was that the tu-




**TUMOR IS INITIATED** when bacteria enter a lesion, which is usually near the crown of the plant (the junction of root and stem), and attach themselves to cells (*1*). A virulent bacterium carries, in addition to its chromosomal DNA, a *Ti* plasmid (*2*). The plasmid's *T*-DNA is introduced into a cell and becomes integrated into the cell's chromosomal DNA (*3*). Transformed cells proliferate to form a crown gall tumor (*4*). The tumor cells synthesize compounds called opines, which serve as nutrients for *A. tumefaciens* cells inhabiting the gall. Two well-studied opines are octopine and nopaline. The *Ti* plasmid carried by a given strain of the bacterium induces synthesis of a given opine.




**PLASMID IS IDENTIFIED** as the tumorigenic agent by experiments such as these. *A. tumefaciens* strain *C*58 is normally virulent (*a*). It grows on nopaline and induces tumors that synthesize nopaline. After growth at 37 degrees Celsius, *C*58 loses virulence and the ability to grow on nopaline; the *Ti* plasmid of *C*58 is missing from the avirulent derivative, *NT*1. Virulence can be restored to *NT*1 by mating with a virulent strain (*b*). *NT*1 cells (mutants resistant to two antibiotics) are inoculated into a tomato plant with *B*6–806, a virulent strain that grows on octopine. A tumor forms. It is ground up and spread on a medium with the two antibiotics, so that only resistant *NT*1 cells form colonies. Many of the cells are transconjugants: they carry the octopine-type plasmid of strain *B*6–806, which they have acquired through conjugation, a bacterial version of mating. Such cells are virulent, grow on octopine and induce octopine-synthesizing tumors.

mor-inducing genes must be on a virus or a plasmid.

The elusive DNA was identified in 1974. Jeff Schell and Marc Van Montagu and their co-workers at the State University of Ghent found very large plasmids in all virulent strains; the plasmids could not be found in avirulent strains. Soon it was shown that the loss of virulence by strain *C*58 at 37 degrees C. is due to the loss of its plasmid. The plasmid is Kerr's mobile element; it is transferred from a virulent strain to an avirulent one by conjugation, a bacterial version of mating. Conjugation experiments done by Rob A. Schilperoort and his colleagues at the State University of Leiden revealed it is the plasmid that confers on bacteria the ability to catabolize (break down) octopine or nopaline, and so to grow on one of those compounds, and that determines whether the tumor will synthesize octopine or nopaline. Even the spread of a particular virulence plasmid through a population of *A. tumefaciens* is promoted by the presence of the opine whose synthesis and catabolism it specifies: the transfer of the plasmid in conjugation is specifically induced by a particular opine.

The *Ti* (for tumor-inducing) plasmid, in other words, is the central element of a rather wonderful, highly evolved ecological interrelation. The presence of a *Ti* plasmid enables a bacterium to subvert the plant cells' metabolic machinery to make a substance that is of no use to the plant but on which the bacteria (and hence also the symbiotic plasmid) can proliferate. The growth of a tumor makes for a large supply of that substance. In nature the tumor tissue is at soil level, and its opine presumably diffuses into the soil, where the transfer of plasmids to hitherto avirulent bacteria promotes the survival of both the bacterial species and the plasmid.

If a crown gall is generated by gene transfer and if the genes are on the *Ti* plasmid, clearly plasmid DNA should be present in crown galls. At the University of Washington in Seattle, Eugene W. Nester, Milton P. Gordon and I, together with a number of postdoctoral and graduate students, undertook to find that DNA. At the time the best way to detect a particular DNA (the *Ti* plasmid in our case) in the presence of a large amount of other DNA (the plant's own) was to measure the rate of DNA reassociation. The DNA double helix is denatured (its two complementary strands are separated) by heating to 100 degrees C. When the temperature is reduced to 68 degrees, complementary strands that "find" each other renature, or reassociate, to form a double helix again. The rate of reassociation is proportional to the square of the concentration of the DNA. Our approach was to compare the rate of reassociation of denatured *Ti* plasmid DNA alone with the rate of its reassociation in the presence of normal plant DNA or in the presence of crown gall DNA. If the tumor-cell DNA included copies of the plasmid DNA, it should add to the concentration of complementary strands, and so make the plasmid DNA (which we could detect because it was labeled with a radioactive isotope) renature faster, in a quantitatively predictable way.

Our initial results showed that the addition of denatured tumor DNA to denatured plasmid DNA made no significant difference in the reassociation rate, so that clearly there was no copy of the entire *Ti* plasmid in the tumor. Might part of it be there? We cut up the plasmid DNA with a restriction endonuclease: one of a family of enzymes, each of which recognizes a particular short sequence of nucleotides (DNA subunits) as a unique cleavage site and therefore always cuts any given DNA into a reproducible set of fragments. The endonuclease *Sma* I cuts a *Ti* plasmid into some 20 fragments of different sizes, so that digesting a quantity of plasmid DNA yields some 20 unique kinds of fragments. These can be separated according to their size by gel electrophoresis. Each group of fragments forms a narrow horizontal band on the gel; the bands can be visualized by staining them with a fluorescent dye and the DNA in each band can be purified for further experiments.

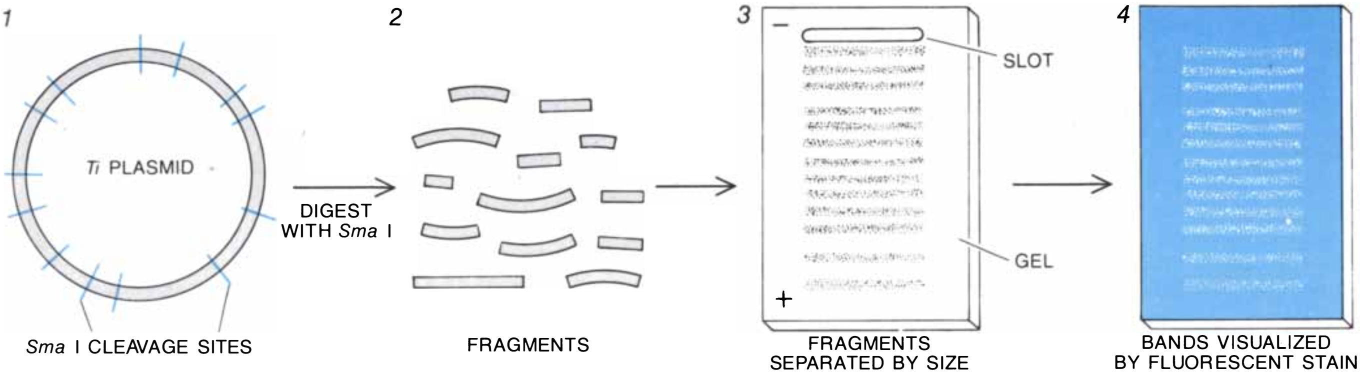
When we subjected the *Ti* plasmid DNA to electrophoresis and then did the renaturation experiment with the individual DNA fragments instead of the plasmid as a whole, we were excited to find that two of the fragments renatured faster in the presence of tumor DNA; the others did not. Those two segments of the *Ti* plasmid must therefore be present in the tumor cells. Subsequent experiments with many tumor lines have shown that specific fragments of the *Ti* plasmid that induced the tumor are always present in the crown gall cells. We call these fragments transferred DNA, or *T*-DNA.

Analysis of renaturation rates has now been supplanted by a more direct method of detecting particular segments of DNA embedded in a large amount of nonspecific DNA. Southern-blot analysis, developed by E. M. Southern of the University of Edinburgh in 1975, combines electrophoresis with renaturation, and it has become a fundamental tool of molecular genetics. To find plasmid DNA in tumor cells one begins by digesting the tumor-cell DNA with, for example, *Sma* I. Whereas such a digest yielded 20 fragments from each plasmid, however, it produces a million or so fragments of the enormously larger genome (the total gene complement) of the plant-tumor cell. When these are spread out on a gel by electrophoresis, they form what looks like a continuous smear (although it is really a set of a million or so discrete bands). Any *T*-DNA fragments in the tumor are buried in that smear; the Southern technique picks them out.

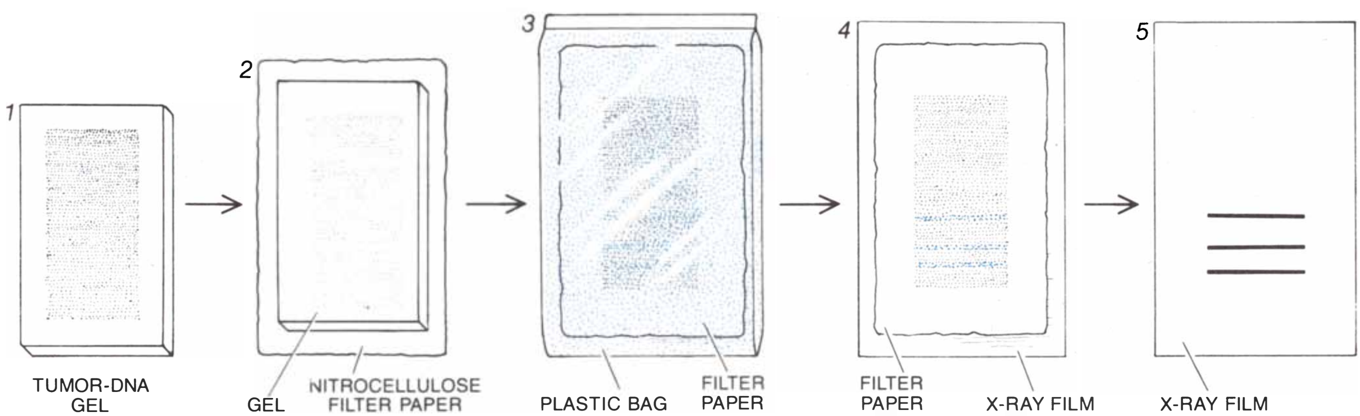
The tumor DNA fragments are denatured and then transferred by blotting to a sheet of nitrocellulose filter paper, which now carries a sharp but invisible print of the original pattern. The DNA is fixed to the paper. The paper is incubated with a solution of denatured, radioactively labeled fragments of plasmid DNA, which serve as a probe. If there is plasmid DNA in the tumor, probe DNA reassociates with it, forming hybrid double-strand DNA; the rest of the probe DNA is washed away, leaving radioactive plasmid DNA bound to its counterpart strands in the tumor DNA. The paper is put in contact with X-ray film. When the film is developed, dark bands on the autoradiogram show where *T*-DNA has been incorporated into plant DNA. Comparison of those bands with the bands on a gel displaying a *Sma* I digest of pure plasmid DNA shows which *Sma* I fragments are, in whole or in part, in *T*-DNA: fragments 3*b*, 10*c*, 16 and 17.

To understand the structure of *Ti* plasmids and their *T*-DNA in detail it was necessary first to determine how the plasmid's fragments map into a closed circle and then to locate genes for particular functions on the circular fragment map. Our group at the University of Washington deduced the first fragment map of a *Ti* plasmid (for the octopine plasmid designated *B*6–806) by digesting the plasmid with the endonuclease *Hpa* I and seeing how the resulting fragments overlapped with the *Sma* I fragments cut from the same plasmid. We prepared a large Southern blot of the *Hpa* I bands [*see illustration on page 55*] and cut it into longitudinal strips. A different labeled *Sma* I fragment was allowed to hybridize with each strip. A labeled fragment that binds to only one *Hpa* I band must map within the band; a fragment that binds to two *Hpa* I bands must span the junction between those bands, and so on. By applying criteria such as these and by testing fragments generated with several endonucleases one finally builds up a complete fragment map.

The next step was to map the plasmid's genes by associating mutations at particular sites with changes in particular functions. For this purpose several groups exploited the technique of transposon mutagenesis. Some genes for resistance to an antibiotic are carried on transposons, which are short segments of DNA that tend to jump, almost at random, to new sites on the bacterial chromosome or on a plasmid. The presence of a transposon simultaneously




**Ti PLASMID IS ANALYZED by gel electrophoresis. The plasmid is digested with a restriction endonuclease, an enzyme that cleaves DNA at specific sites. Fragments made by the endonuclease *Sma* I (actually thousands of copies of each fragment, from thousands of copies of the plasmid) are placed in a slot at one end of an agarose gel. In an electric field the negatively charged DNA fragments move toward the positive pole. The smaller they are, the faster they move, and so the fragments form a series of bands, which can be visualized with a fluorescent dye that binds to DNA. A given plasmid digested with a given endonuclease yields a distinctive pattern of bands.**




**PRESENCE OF PLASMID DNA (*T*-DNA) in gall tissue is demonstrated by Southern transfer. Tumor-DNA fragments generated by digestion with *Sma* I are separated by gel electrophoresis. The thousands of bands of fragments form a continuous smear on the gel (*1*). The DNA is denatured (the two strands are separated), transferred by blotting to filter paper and fixed to it (*2*). The paper is incubated with probe DNA (*color*): *Ti*-plasmid DNA or fragments of it, labeled with a radioactive isotope and denatured (*3*). Probe fragments reassociate with complementary DNA, that is, they "find" and bind to bands constituting *T*-DNA. Unbound probe DNA is washed away (*4*). The *T*-DNA bands are made visible by autoradiography (*5*). The fragments can be identified by comparison with the original *Sma* I gel.**

generates a mutation (because insertion of the transposon interrupts a gene on the chromosome or plasmid), identifies itself (because the bacterium is rendered resistant to an antibiotic) and flags its location (because the pattern of endonuclease fragments is changed by the insertion of new DNA into an old fragment, giving rise to new ones). At the University of Washington, Nester and David Garfinkel introduced the transposon by means of a "suicide vector": a plasmid, constructed by John Beringer of the John Innes Institute in England, that carries the kanamycin-resistance transposon *Tn*5. Introduced into *Agrobacterium,* this plasmid conveniently self-destructs, but only after the *Tn*5 transposon has had an opportunity to jump to a new site on the bacterial chromosome or on a *Ti* plasmid.

One selects kanamycin-resistant bacterial colonies in which the *Tn*5 is on the plasmid rather than on the chromosomes [*see top illustration on page 56*] and then tests them for any change in genetic properties. Many mutations are "silent," but in some cases there is a detectable change. If the bacteria cannot catabolize octopine, for example, a gene needed for that process must have undergone mutation. To map the gene one isolates *Ti* plasmid DNA from the mutant bacteria, digests it with *Sma* I and examines the resulting electrophoresis pattern. It is different from the normal pattern: one of the usual fragments (call it *X*) is missing and two new fragments are seen, and the combined size of the two new fragments equals the size of fragment *X* plus *Tn*5. The gene must be within fragment *X*. Knowing that *Sma* I cuts near the middle of *Tn*5, one can determine, from the size of the two new fragments and certain other information, the precise location of the gene.

From this kind of analysis by Nester's group, by Schell and Van Montagu and by Schilperoort's group, a genetic map of octopine and nopaline *Ti* plasmids eventually emerged. A large sector of the plasmid to the left of the *T*-DNA carries the virulence functions. Mutations in this sector presumably block the attachment of the bacterium to a plant cell, the transfer of plasmid DNA into the cell or the integration of the DNA in the plant-cell chromosome, or perhaps several of these functions. Another large sector, on the other side of the *T*-DNA, governs the uptake and catabolism of octopine or nopaline. Genes governing the plasmid's ability to transfer in conjugation have been mapped, as have genes for the catabolism of additional opines. Large sectors of the plasmid still appear to be silent, but that may be because they govern traits of which we are not yet aware, perhaps including the ability to catabolize still more nutrient substances.

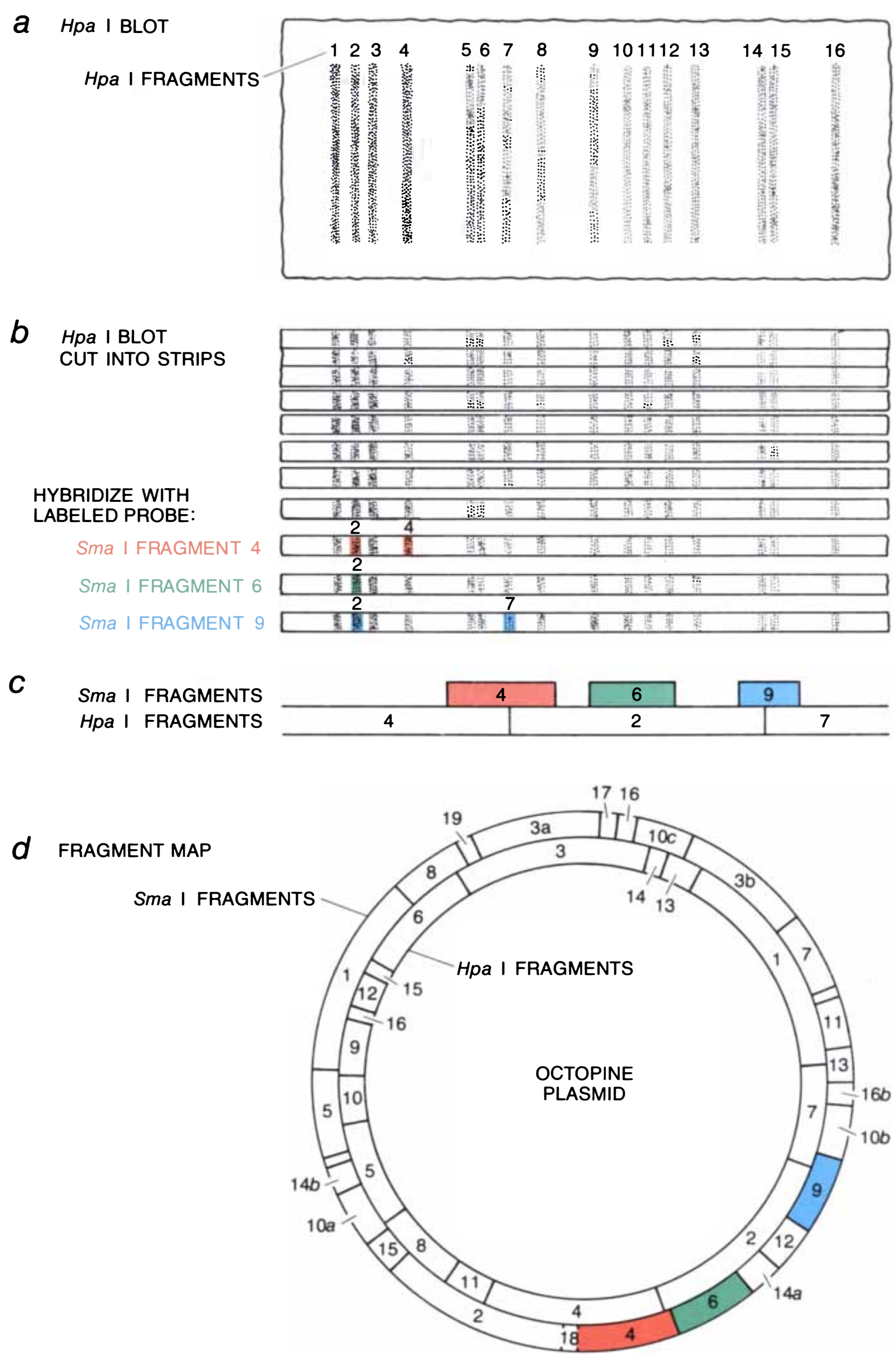
Mutations within the *T*-DNA were of particular interest in that they identified genes transferred from the bacterium to the plant. In the nopaline plasmid studied by Schell and Van Montagu one mutation eliminated nopaline synthesis in tumors, and it was situated within the *T*-DNA. Morel's prescient hypothesis was thereby confirmed; here was direct evidence, at the level of the gene, that the ability to synthesize nopaline is governed by DNA transferred from the bacterium to the plant. Cells of the tumor incited by this mutant *Ti* plasmid turned out to harbor not only the usual *T*-DNA but also a copy of the transposon that had given rise to the mutation. Now, the transposon in question, *Tn7*, is a sizable piece of DNA. Its presence in the tumor cells meant that a large segment of foreign DNA, enough to carry several genes, can be transferred along with the *T*-DNA. Clearly, then, in principle it should be possible to make *T*-DNA a vector for introducing a desirable gene into plant cells. If such a gene could be made to function in the plant cells and if the cells could be regenerated into healthy plants, a method would be available for genetic engineering in plants.

Additional transposon mutations in *T*-DNA enabled Schilperoort's group to identify sites controlling the perturbation of hormone activity that is apparently responsible for tumor growth. The mutations bring about two kinds of changes in tumor morphology. One class of mutations cause roots to sprout from the tumor and the other class give rise to shoots. Just such changes are caused in normal plant tissue by a hormone imbalance: a high ratio of auxins to cytokinins induces excess formation of roots and a low ratio of auxins to cytokinins generates more shoots. Schilperoort showed that the effect of the tumor-morphology mutations could indeed be reversed by correcting the putative hormone deficiency, that is, by adding cytokinins to the medium in the case of "rooty" mutants and by adding auxins in the case of "shooty" mutants.

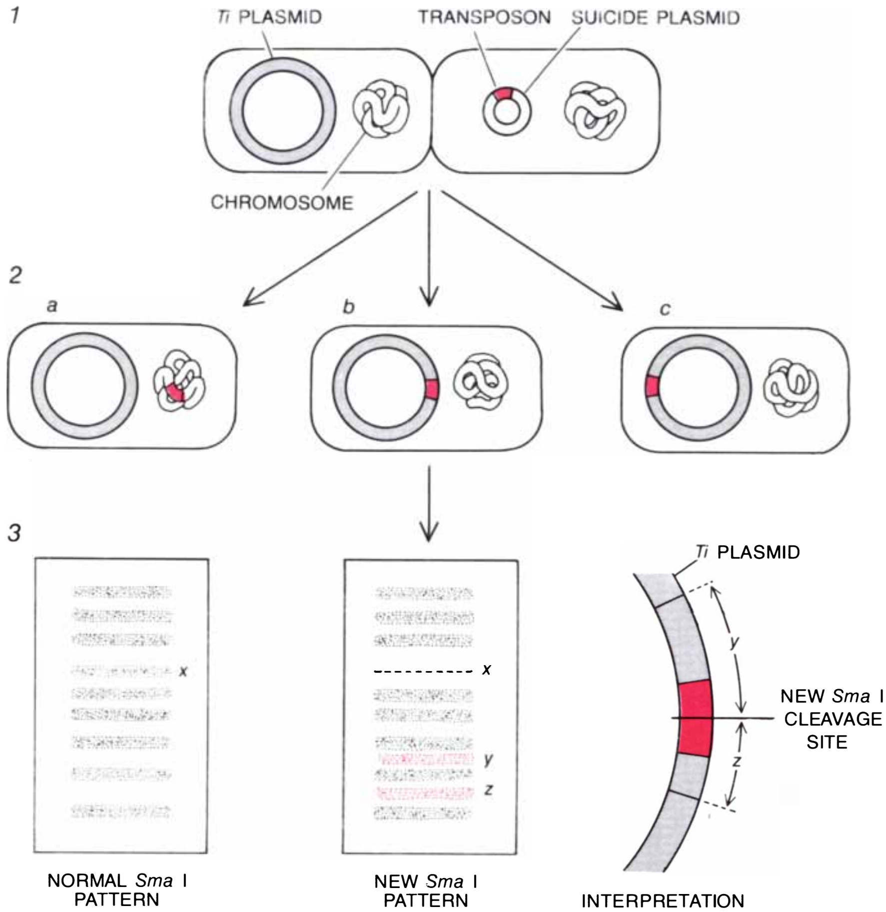
In Nester's laboratory several students worked out a detailed map of *T*-DNA, including the rooty and shooty gene loci. Again the method was transposon mutagenesis, but now that something was known about the structure of the *Ti* plasmid the mutagenesis could be "site-directed": one could choose a particular *T*-DNA fragment, clone it in a small vector plasmid, generate a mutation in it and then transfer the fragment into *A. tumefaciens* and onto the *Ti* plasmid by means of double recombination. The result was a *Ti* plasmid with a mutation in a predetermined segment.

Intensive mutagenesis of *T*-DNA in this way defined the rooty and shooty loci and also the site of another class of mutations generating unusually large tumors. Extensive regions of the *T*-DNA appeared to be silent, that is, mutations within them had no apparent effect. Curiously, none of the 75 mutant plasmids studied had completely lost its virulence. This implies that all the genes governing the transmission of *T*-DNA to a plant cell and its insertion into the chromosomal DNA must fall outside the *T*-DNA, in the virulence region to the left of it. In this respect *T*-DNA is different from both viruses and transposons, both of which carry with them the vital functions allowing them to transfer to a new host.

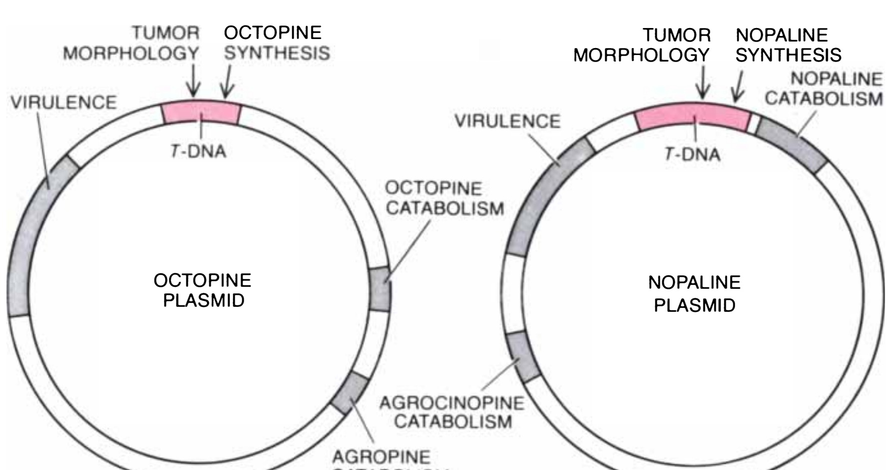
A gene is expressed when its DNA is transcribed into a messenger RNA and the RNA is translated into protein. Crown gall cells contain RNA transcribed from *T*-DNA genes. The RNA seems to have the special characteristics of messenger RNA in eukaryotic cells (the nucleated cells of all organisms higher than the bacteria) rather than those of prokaryotic (bacterial) messenger RNA. The sizes and map positions of individual molecules of the tumor RNA have been determined. Individual RNA's correspond to the rooty and the "large" tumor-morphology loci in the octopine and nopaline plasmids; two RNA's are encoded by the shooty locus. Interestingly, just about all the silent regions of *T*-DNA also encode messenger RNA's, some of which may correspond




**FRAGMENT MAP OF PLASMID is derived by overlapping two sets of endonuclease fragments. A large Southern blot is prepared displaying the fragment pattern generated by cleaving the *Ti* plasmid with the endonuclease *Hpa* I (*a*). The blot is cut into strips and each strip is incubated with a different radioactively labeled *Sma* I probe, which reassociates with its *Hpa* I counterpart (*b*). *Sma* I fragment 4, for example, hybridizes with *Hpa* I fragments 2 and 4; *Sma* I fragment 6 hybridizes only with *Hpa* I fragment 2. Such data yield a linear pattern of overlapping fragments (*c*) and eventually yield a complete map of the circular *Ti* plasmid (*d*).**




**GENETIC MAP OF PLASMID is determined by transposon mutagenesis. A gene conferring antibiotic resistance may be carried on a transposon, a piece of DNA that tends to jump to a new site. A "suicide" plasmid carrying a kanamycin-resistance transposon (*color*) is introduced into *Agrobacterium* by conjugation (*1*). The transposon may jump (*2*) to a site on the bacterial chromosome (*a*) or on the *Ti* plasmid (*b, c*). Wherever the transposon becomes integrated it causes a mutation. By screening for kanamycin resistance it is possible to select cells in which the *Ti* plasmid has thus been subjected to mutation. Each cell is tested for a change in genetic function. The gene for that function can be determined by comparing the fragment pattern of the mutant plasmid with the normal pattern (*3*). Fragment *x* has disappeared from the plasmid of cell *b* and new fragments *y* and *z* are seen. They are the result of *Sma* I cleavage at a new site, in the transposon. The gene must be in missing fragment *x,* whose identity is confirmed by the fact that its size is equal to the sum of *y* and *z* minus the size of the transposon.**



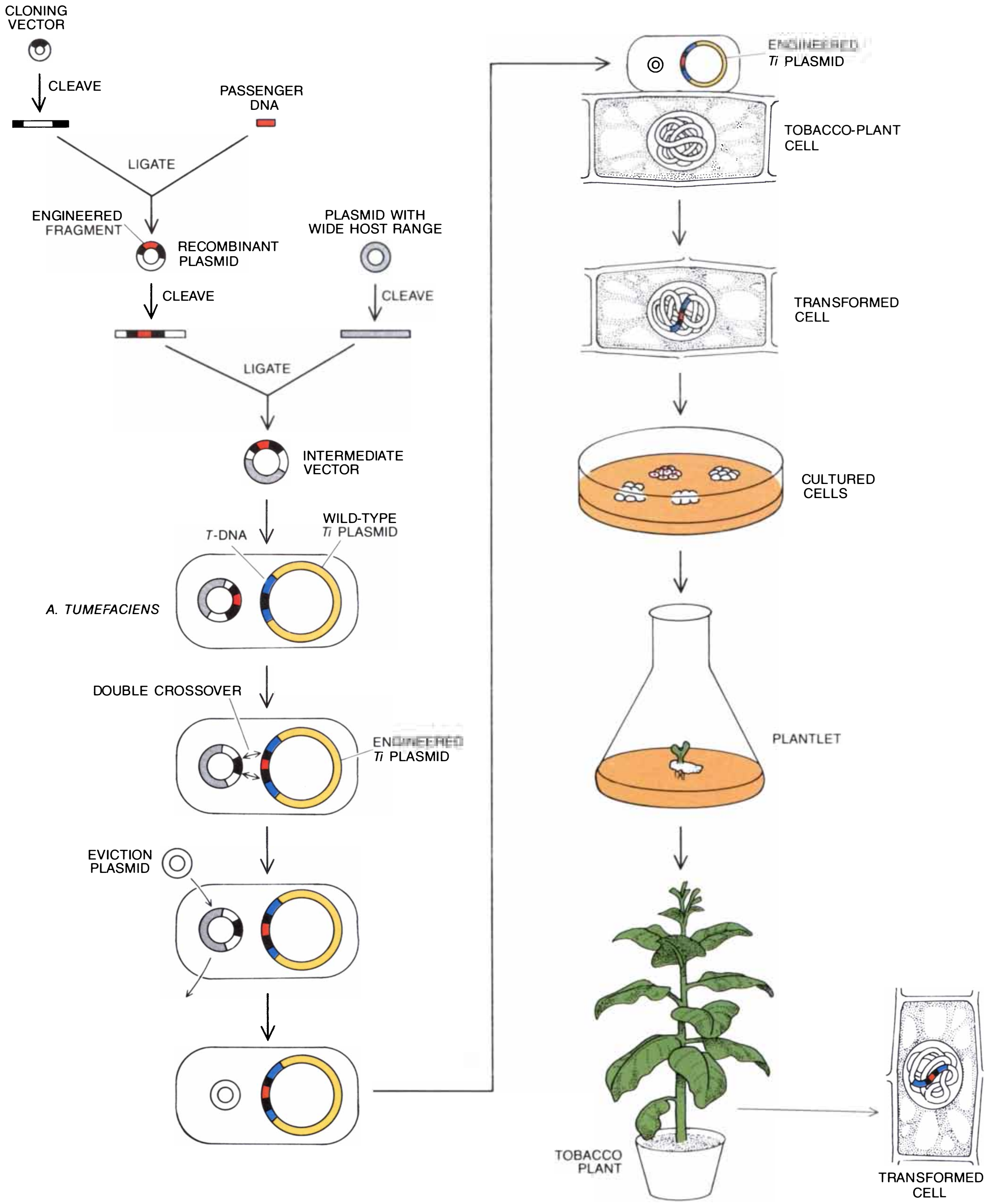

**GENETIC MAPS are beginning to be built up for an octopine *Ti* plasmid (*left*) and a nopaline one (*right*). The *T*-DNA has been defined. Within it gene loci governing synthesis of octopine or nopaline have been mapped. So have genes affecting tumor morphology, which are revealed by the "rooty," "shooty" and "large" mutations. In both plasmids a region to the left of the *T*-DNA governs the virulence functions that enable the plasmid to induce a tumor. Regions outside the *T*-DNA also govern the catabolism of octopine or nopaline and other opines.**

to genes for opines that have not yet been discovered.

The unique biological characteristics of the *Ti* plasmid make it a natural agent for gene transfer. It brings about the insertion into the plant's chromosomal DNA of *T*-DNA, which is thereupon maintained by the host plant, replicated along with the host plant's own genes, transcribed into RNA and (at least in large part) expressed as protein. How does the plasmid "decide" what part of itself is *T*-DNA? The answer may lie in specific signal sequences that have been identified in the border regions at both ends of the *T*-DNA. In these regions there is a 25-nucleotide "direct repeat" sequence of nucleotides. (Nucleotides are the four subunits of DNA whose sequence in the coding regions of a gene encodes the genetic information specifying a particular protein. The precise repetition of a short nucleotide sequence, either direct or "inverted," often serves as a signal for recombinational events.) The direct repeats may define the limits of the *T*-DNA; any foreign DNA inserted between the repeats should be handled as *T*-DNA, that is, it should be transferred to the plant cell.

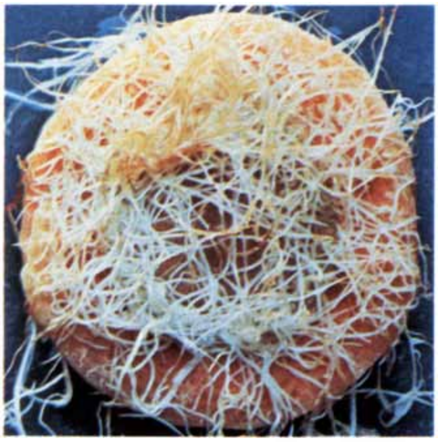
Let me describe one strategy for introducing a foreign gene into a plant by means of a *Ti* plasmid and *A. tumefaciens.* One begins with a recombinant plasmid: a small plasmid suitable for cloning in the common bacterium *Escherichia coli,* into which a known fragment of *T*-DNA has been spliced as in the site-directed mutagenesis process described above. The recombinant plasmid is cleaved open, at a site in the *T*-DNA segment, with a restriction endonuclease. A piece of "passenger" DNA is spliced into the opening. The passenger segment is bipartite: it consists of the foreign gene one wants to send into the plant on *T*-DNA and, attached to that gene, a genetic marker, say a bacterial gene for resistance to kanamycin. The combined plasmid is then recloned into a larger plasmid, which is introduced into an *Agrobacterium* strain carrying a wild-type (unmodified) *Ti* plasmid. As the bacteria proliferate a rare double-recombination event will sometimes take place, yielding a clone of bacteria whose *T*-DNA harbors an insert: the engineered passenger DNA. Such bacteria can be identified and selected by their survival on a medium containing kanamycin after the recombinant vector plasmid has been evicted. The selected bacteria are introduced into a plant. Their *T*-DNA is inserted into plant cells, and with it the passenger DNA.

A number of experiments have made it clear that this strategy or any of several similar ones can give rise to genetically modified crown gall cells growing in a culture dish. That result is still a far cry from the objective of growing genetical-




**FOREIGN GENE IS INSERTED** into a plant cell and a plant is regenerated. A small plasmid carrying a fragment of *T*-DNA (*black*) including the rooty locus is cut open with an endonuclease. Into it is inserted a piece of "passenger" DNA: a yeast gene for the enzyme alcohol dehydrogenase together with a gene for resistance to kanamycin (for screening). The recombinant plasmid is recloned in a plasmid that has a wide host range; the new intermediate vector is introduced into an *A. tumefaciens* cell carrying a wild-type (unmodified) *Ti* plasmid. The engineered *T*-DNA fragment in the intermediate vector can recombine with the same fragment (*black*) in the *Ti* plasmid. A rare double-crossover recombination brings the engineered fragment, with its passenger, into the *Ti* plasmid. The vector is evicted by introducing an "incompatible" plasmid. Bacteria carrying the engineered fragment in their *T*-DNA are identified by resistance to kanamycin. Such bacteria are used to transform tobacco-plant cells. The *T*-DNA, with its passenger, is integrated into the plant's DNA, giving rise to rooty mutant cells. The cells yield healthy plants whose transformed cells contain *T*-DNA carrying alcohol dehydrogenase gene.

ly modified healthy plants from such a transformed cell line. Few complete plants have been grown from wild-type crown gall cells, and in every case they have turned out to be mutants most or all of whose *T*-DNA has been deleted. The conclusion is that the presence of *T*-DNA genes is incompatible with healthy plant growth, presumably because they perturb auxin and cytokinin levels. How then is one to regenerate a healthy plant with the *T*-DNA and its passenger integrated into the plant's chromosomal DNA?



**CARROT CELLS transformed by *Ri* plasmids, which are carried by *Agrobacterium rhizogenes,* regenerate into healthy carrot plants. The plasmid's *T*-DNA transforms cells of a slice of carrot into rapidly growing hairy roots (*top left*). A single root placed in a culture medium proliferates (*top right*). When a root is grown on a solid culture medium to form a callus and callus is transferred to a liquid medium, it forms an embryo that then develops into a plant (*bottom*). Other kinds of plant can also be regenerated from their hairy roots. The *Ri* plasmid seems to be a naturally disarmed gene vector that can give rise to genetically modified healthy plants. These photographs were made by Jacques Tempé of the University of Paris at Orsay.**

One possible way was suggested by some surprising results obtained recently by Andrew N. Binns of the University of Pennsylvania and by Kenneth A. Barton, Antonius J. M. Matzke and me in my laboratory at Washington University in St. Louis. We inserted a yeast gene encoding the enzyme alcohol dehydrogenase into the rooty locus of the nopaline plasmid and infected a tobacco plant with the resulting plasmid. The outcome was a culture of rooty mutant cells. The surprise was that the cells were not typically tumorous: they required added cytokinin for growth. Moreover, the cells (which synthesized nopaline and did not require added auxin) grew into complete tobacco plants. DNA isolated from the plants contained some 20 copies of *T*-DNA per cell—each including a copy of the inserted eukaryotic gene for alcohol dehydrogenase! In other words, far from being deletion mutants our plants had an all-time record number of copies of *T*-DNA per cell. After self-pollination the plants produced seeds that grew into healthy plants whose cells contained full-length *T*-DNA, still including the yeast gene, in multiple copies. We conclude that a rooty mutation somehow "disarms" nopaline *T*-DNA so that tobacco cells transformed by the *T*-DNA can regenerate. Whether this will work for other kinds of *Ti* plasmids and in other kinds of plants remains to be seen.

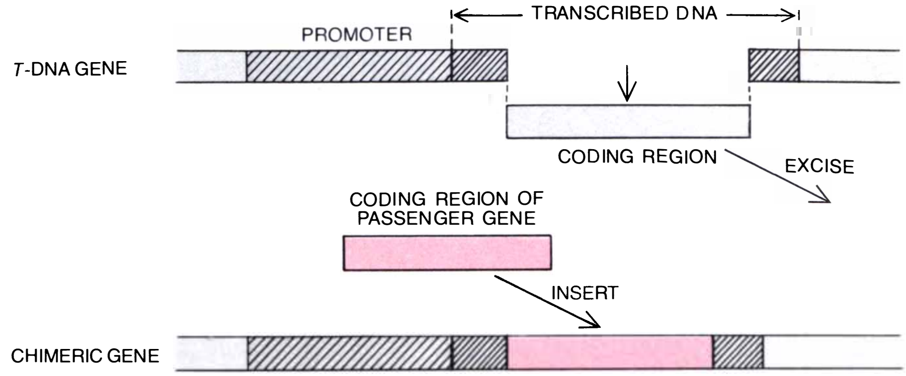
A more general way to build a disarmed *T*-DNA vector may be to delete all the *T*-DNA's own genes, leaving only the passenger DNA and the noncoding direct-repeat signal sequences. Such vectors have now been constructed in several laboratories, including mine. They may turn out to be the vectors of choice for most plants that are susceptible to infection by *A. tumefaciens.*

There is another family of *Agrobacterium* plasmids that may also serve as gene vectors yielding healthy genetically modified plants. They cause a proliferation of roots, called hairy-root disease, in plants infected by *A. rhizogenes* and are called *Ri* (for root-inducing) plasmids. The roots, like crown gall tissue, grow rapidly in culture in the absence of the infecting bacterium. Jacques Tempé and his colleagues at the Institut National de la Recherche Agronomique have shown that the root cells contain opines; they also contain several copies of a *T*-DNA. The important thing is that, like the rooty-mutant nopaline *Ti* plasmids, the *Ri* plasmids give rise to transformed plant cells that can regenerate into intact healthy, fertile plants. The *Ri* plasmids appear to be naturally disarmed vectors.

Given that foreign genes can be introduced into certain plants and that healthy genetically modified plants can be regenerated, the next question is whether those genes will be expressed—will give rise to messenger RNA and protein—and so have an effect in their new setting. A number of experiments in which plant, animal, yeast and bacterial genes have been transferred have not been successful in this respect. Our yeast gene for alcohol dehydrogenase, for example, is not expressed (the enzyme is not synthesized) in the regenerated tobacco plants. The expression of a gene is regulated by various signal sequences flanking the DNA that codes for a protein. The most important regulatory sequences are in the promoter region, upstream from the gene. The signals surrounding an inserted foreign gene must be recognizable by the plant cell if the gene is to be expressed. For several years it has been possible to get bacteria to synthesize animal proteins by inserting into the bacteria a chimeric gene: one in which regulatory information from a bacterial gene is carefully spliced to the coding region of an animal gene.

Very recently Michael W. Bevan in my laboratory, the group at Ghent and workers at the Monsanto Company in St. Louis have been able to make an analogous chimeric gene that is expressed in plant cells. In all three laboratories the procedure was to excise the coding region from a *T*-DNA gene, splice a bacterial gene for resistance to kanamycin between the *T*-DNA gene's regulatory regions and then introduce the chimeric gene into *Ti* plasmids more or less as described above. Plant cells transformed by the plasmids expressed the bacterial gene: they were resistant to a concentration of kanamycin that would have killed unprotected cells in culture. This success indicates that the problem of gene expression in a genetically modified plant can, in principle at least, be solved. As more is learned about the regulatory mechanisms of plant genes it should become possible not only to achieve expression but also to get selective expression by modifying the promoter region of genes. One might cause a gene to be expressed only in leaves or only in seeds, or only in response to a chemical signal.

So far both regeneration and expression have not been reported in the same experiment, but this will surely be accomplished soon. Then the question will be how to choose and isolate genes that are likely to improve crop plants. There are obvious objectives such as larger yields, resistance to pests and tolerance of cold and drought. The trouble is that there is no known way to find genes controlling such traits, most of which are probably governed by a number of genes and perhaps by elaborate regulatory mechanisms. Even if the appropriate genes could be identified, there is no known technique for isolating them by recombinant-DNA methods; only genes whose protein product is known have so far been isolated.




**CHIMERIC GENE must be constructed if a passenger gene is to be expressed by plant cells. The coding region of a *T*-DNA gene is excised precisely, leaving undisturbed the promoter region that precedes the gene (and regulates its transcription into RNA) and the signal sequences (*heavy hatching*) that flank the coding region (and may regulate the processing of the RNA for translation into protein). The coding region of the passenger gene is inserted between the regulatory regions so that it will be expressed as if it were an integral part of the *T*-DNA.**

On the other hand, a feasible objective for the near future is the introduction into crop plants of novel foreign genes conferring resistance to herbicides or pesticides. If a plant can be grown in the presence of powerful agricultural chemicals, it should give a higher yield because of diminished competition from weeds or diminished predation.

The *Ti* and *Ri* plasmids' promise is limited to the modification of dicotyledonous plants, which *Agrobacterium* infects. The bacterium does not infect monocotyledonous plants, including grasses and cereals, and most of the world's important food crops are monocots. The reason for their immunity to the bacterium is not understood. There is as yet little information on how *T*-DNA travels from the plasmid to the plant cell's nucleus and on how (or even whether) *T*-DNA picks its sites of insertion into plant DNA. Research is needed in this important area, because better understanding of the sequence of events in infection may give insight into what step fails in the monocots. Many dicotyledonous plant species readily regenerate intact plants from tissue-culture cells; among the monocots regeneration is rarer and more difficult to achieve, but recently advances have been made in regenerating important crop plants from cells. That should spur efforts to extend the host range of the *Ti* plasmid to monocots and to develop new vectors that will function in monocots.

*Agrobacterium* has shown that it is possible to get foreign DNA into plant chromosomes. Can biologists mimic the natural process with a plasmid of their own choice or design, or does the symbiotic combination of *A. tumefaciens* and the *Ti* plasmid have inimitable advantages shaped by evolution and natural selection? Until now counterfeit plasmids introduced into plant cells by a variety of techniques have failed to integrate their DNA into the host DNA. Even the *Ti* plasmid itself nearly fails when it is introduced artificially into tobacco protoplasts (cells denuded of their tough outer wall). The efficiency of *T*-DNA insertion is very low and the ends of the inserted DNA are abnormal, suggesting that the usual integration mechanism is not functioning.

It is a safe bet that, in contrast, *Agrobacterium* knows how to get *T*-DNA into the plant-cell nucleus with high efficiency. How? There may be a bacterial enzyme that excises *T*-DNA from the plasmid. The form of the *T*-DNA on arrival in the nucleus must be important. The DNA could be linear or circular, single-strand or double-strand, naked or complexed with bacterial proteins; it could replicate independently in the plant cell or it could become integrated immediately. The infecting bacterial cell may somehow bring the plant cell into a receptive state for DNA insertion, or the bacterium may have a special affinity for receptive cells. The signal sequences at the end of the *T*-DNA may be important for excision from the plasmid or for insertion into the plant DNA; what structure (and in which organism) reads the signals is not known.

For all its mysteries the *Ti* plasmid is today the only successful gene vector for higher plants. Quite apart from the great promise of its exploitation for crop improvement, the plasmid has become an important investigative tool. It provides a way to manipulate plant genes and thereby to study gene structure and regulation. It has revealed genes that influence plant-hormone levels, and such genes will be valuable for the study of development. Finally, *T*-DNA has demonstrated for the first time that bacterial plasmid DNA recombines with plant chromosomal DNA in nature, thus revealing a previously unknown evolutionary channel for the transmission of genes.

# Giant Volcanic Calderas

*They are craters tens of kilometers in diameter that remain after eruptions far more violent than any in the span of human history. Perhaps 10 of them have been created over the past million years*

by Peter Francis

The eruption of Mount St. Helens in southern Washington on May 18, 1980, ejected .6 cubic kilometer of magma and left a crater two kilometers in diameter; it is deservedly regarded as having been a spectacular event. How then might one describe an eruption that happened 950 kilometers to the east of Mount St. Helens some 600,000 years ago? It disgorged 1,000 cubic kilometers of pumice and ash and left an elongated caldera (the term for a large volcanic crater) 70 kilometers in its longest dimension. The camouflaging effects of vegetation and glaciation have made the traces of that eruption quite difficult to recognize today; the most obvious vestige is the Old Faithful geyser in Yellowstone National Park. Yellowstone is nonetheless a product of volcanic processes operating on the largest scale: it is a resurgent caldera, that is, a caldera whose floor has slowly domed upward in the millenniums since the eruption. Resurgent calderas are by far the largest calderas found on the earth. Thus volcanic eruptions such as the one that formed the Yellowstone caldera must be counted among the greatest catastrophes of nature, perhaps comparable to the impact of an asteroid.

Fortunately they are rare. Not one has taken place throughout the few thousand years of recorded history, and in the U.S. only three are known to have happened over the past million years. In addition to the one in Yellowstone, an eruption 700,000 years ago formed the Long Valley caldera in California and an eruption a million years ago formed the Valles caldera in New Mexico. (The Long Valley caldera has been of particular concern lately because of a series of small earthquakes in the area.) Eruptions of similar age may eventually be shown to have happened in other parts of the world. Still, it will probably turn out that in the past million years no more than 10 have happened worldwide. On the other hand, detailed mapping of the San Juan Mountains in Colorado by Thomas A. Steven and Peter W. Lipman of the U.S. Geological Survey revealed at least 18 calderas between 20 and 30 million years old, and many others of comparable age have been recognized in southern New Mexico, Arizona and Nevada. In the past decade volcanologists have made rapid progress in understanding the origins of giant, resurgent calderas and the catastrophic eruptions that form them.

The fundamental mechanism of caldera formation is clear. The sudden ejection of large volumes of magma from a magma chamber a few kilometers below the surface of the earth abruptly removes the underpinning of the roof of the chamber. Hence the roof founders downward, leaving a caldera at the surface. In a classic work on the volcanic geology of Scotland published in 1909, C. T. Clough, H. B. Maufe and E. B. Bailey offered the first account of this process, which they termed cauldron subsidence. The process takes place on a wide range of scales, leaving calderas that vary in diameter from a few kilometers to 50 or more.

Apart from sheer size the definitive feature of a resurgent caldera is the slow upheaval of its floor, which probably results from the intrusion of fresh magma into the magma chamber that created the caldera in the first place. The amount of vertical resurgence can exceed one kilometer. Unlike an ordinary volcano, therefore, a resurgent caldera is a broad depression with a central massif. The phenomenon of resurgence was first recognized in 1939 by the Dutch geologist R. W. van Bemmelen during a study of the Toba caldera in northern Sumatra. Van Bemmelen estimated that the floor of the caldera had subsided by as much as two kilometers, allowing a deep lake to form, but had later been heaved up hundreds of meters to create the island known as Samosir in the middle of the lake. Toba remains today the largest resurgent caldera known: its maximum dimension is nearly 100 kilometers. Toba, however, has been relatively little investigated, and the details of resurgence have been worked out largely by geologists of the Geological Survey who have studied the calderas of the American Southwest. Indeed, the term resurgent caldera was introduced (in 1962) by Robert L. Smith and R. S. Bailey of the Survey.

One further characteristic of resurgent calderas can be traced to the nature of volcanic processes. In any volcanic eruption the magma reaching the surface can erupt in three ways: as lava, as air-fall material or as a pyroclastic flow. Lava is simply magma that reaches the surface, spreads out as a liquid and solidifies into igneous rock that is finely crystalline or even glassy. Depending on the composition of the magma, the rock ranges from basalt, a dark gray material relatively poor in silica ($SiO_2$), through andesite and dacite to rhyolite, a light gray material rich in silica. Air-fall material consists of pumice (a frothy, glassy substance) along with finer particles of ash and dust. (The term dust is usually given to particles less than four micrometers in diameter. Particles between four and 63 micrometers are called fine ash.) The material is created when fragments of solidified magma are carried

**CERRO GALÁN CALDERA in northwestern Argentina went undetected in conventional geological surveys of the area but was discovered in this false-color image, which was made by a Landsat satellite in orbit 900 kilometers above the ground. The sheer size of the caldera had helped to hide it. The rim of the caldera, in places covered with snow, is an ellipse 34 kilometers along its greatest (north-south) diameter. The center of the caldera, also covered with snow, is an elevation whose apex is six kilometers above sea level. The elevation (a resurgent center) exists because the accumulation of fresh magma under the surface in the millenniums since the catastrophic eruption at the site has heaved the caldera floor upward. The grayish, heavily gullied terrain surrounding the northern half of the caldera is the much-eroded remnant of a great apron of ignimbrite: rock that formed from the pumice and ash ejected in the eruption.**

upward in the atmosphere by intensely hot convecting gases.

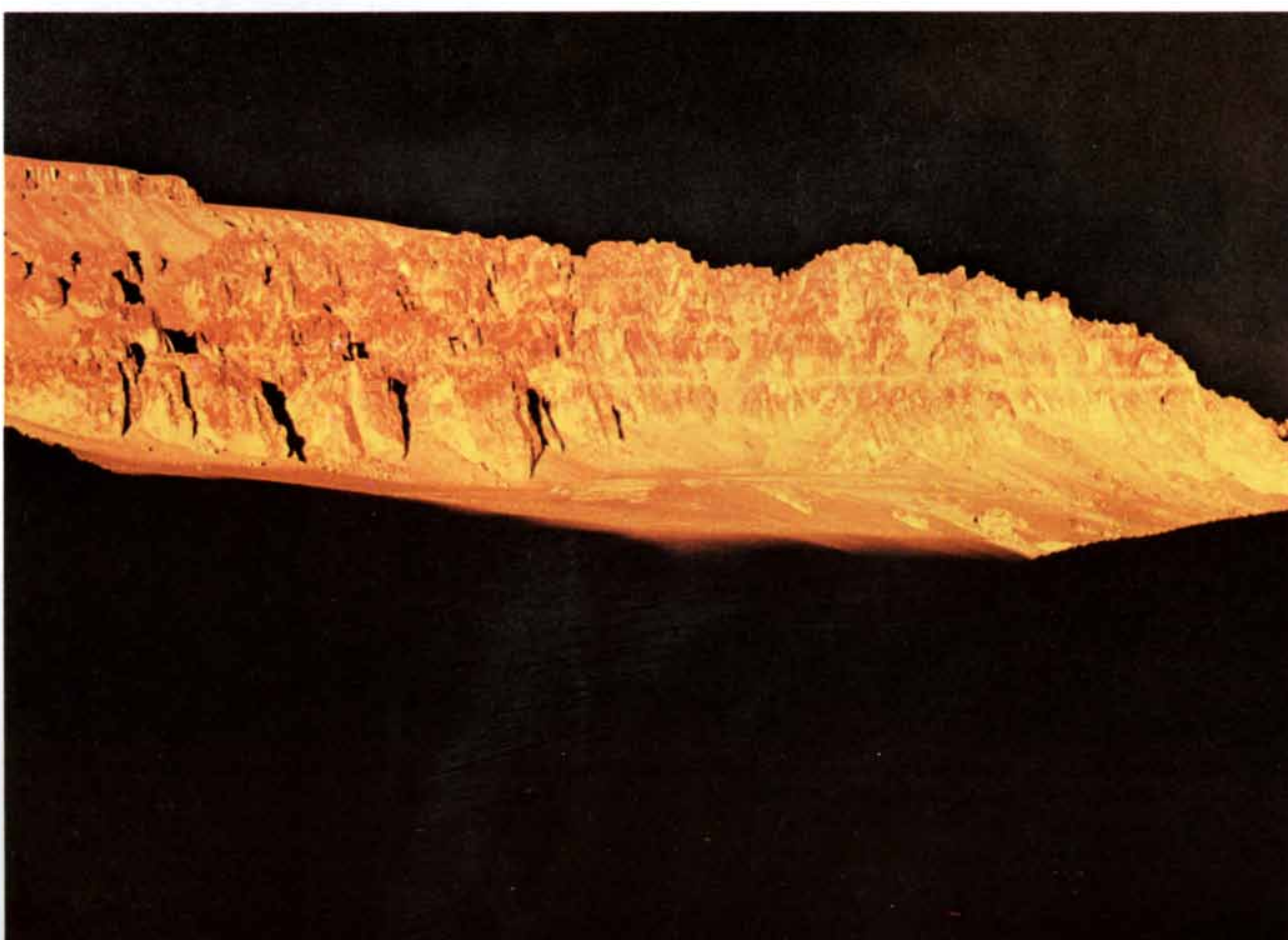
Pyroclastic flows are also composed of pumice, ash and dust. Here, however, the volcanic material forms an incandescent, ground-hugging cloud that races along the surface at speeds of up to 100 meters per second, buoyed up and fluidized by hot gas. Often the cloud deposits clasts, or fragments, of pumice many tens of kilometers from the source of the eruption.

Most volcanic eruptions produce lava and air-fall ashes; pyroclastic flows are less common. Yet in eruptions that form a resurgent caldera pyroclastic flows account for by far the greatest proportion of the ejecta. Indeed, pyroclastic flows often accumulate in a newly formed caldera to a thickness of more than a kilometer. At the bottom of such a pile the pumice clasts typically soften and weld together, producing a distinctive and unmistakable rock. In 1935 P. Marshall, studying the rock resulting from pyroclastic flows in New Zealand, named it ignimbrite: fire-cloud rock. The term is now employed for the deposits laid down by pumiceous pyroclastic flows, whether the clasts are welded or not. Large thicknesses of welded ignimbrite are excellent evidence for the location of ancient calderas.

Resurgent calderas are large and the material ejected by the eruptions that make them is extremely voluminous. Nevertheless, resurgent calderas are not particularly easy to identify. In many instances they are camouflaged by their very size. A few years ago M. C. W. Baker and I, working first at the Open University in England and later at the Lunar and Planetary Institute in Houston, undertook a search for calderas in the Andean cordillera of South America by means of images made with the Landsat satellites. The search capitalized on the greatest advantage that satellite imagery holds over conventional aerial photography: the altitude of the spacecraft (900 kilometers) gives an otherwise unobtainable synoptic view of large structures on the earth. We were able to identify two major calderas that had previously gone undetected by geological surveys.

The more impressive of the two is the Cerro Galán caldera in northwestern Argentina, which is 34 kilometers in diameter and is surrounded by a spectacular apron of ignimbrite extending as far as 70 kilometers from the caldera rim itself. We later learned that the caldera had also been noted by J. D. Friedman of the Geological Survey and Grant H. Heiken of the Johnson Space Center in photographs made from the *Skylab 4* spacecraft in 1974. The caldera is so young (2.6 million years) and its structure is so unambiguous that we had little doubt what it is.

The second caldera was less obvious. The Landsat images we examined suggested to us that the Kari Kari mountain massif in Bolivia, which is five kilometers high, probably represents the resurgent center of a large, old caldera. Earlier mapping of Kari Kari, however, had suggested the massif was a batholith: a great mass of coarsely crystalline igneous rock that solidified in the crust of the earth and was later exposed by erosion. Our field work at the site soon confirmed our identification. The welded texture of the rock of the massif showed conclusively that it is ignimbrite. The resurgent



**CLIFF OF IGNIMBRITE nearly half a kilometer high was exposed by erosion at Cerro Galán at a site on the apron of ignimbrite about 20 kilometers to the west of the caldera rim. Each of the horizontal layers in the cliff represents material deposited by an individual eruption; the uppermost layer represents material deposited by the eruption that formed the Cerro Galán caldera some 2.6 million years ago.**

center is the remaining evidence of a caldera that was originally about 36 kilometers across in its largest dimension. It is 20 million years old. We have found a number of other calderas in unexplored parts of the central Andes. Field investigation will be long and difficult.

At first one might think the most likely place for an eruption that forms a resurgent caldera would be a subduction zone, the boundary on the surface of the earth where a plate of oceanic crust slides under a continental plate and plunges into the underlying mantle. After all, subduction zones are sites of intense seismic and volcanic activity. The Toba caldera in Sumatra is in such a setting. In most instances, however, the geological setting is much less straightforward. Most of the younger calderas in the U.S., for example, are hundreds of kilometers from any modern subduction zone.

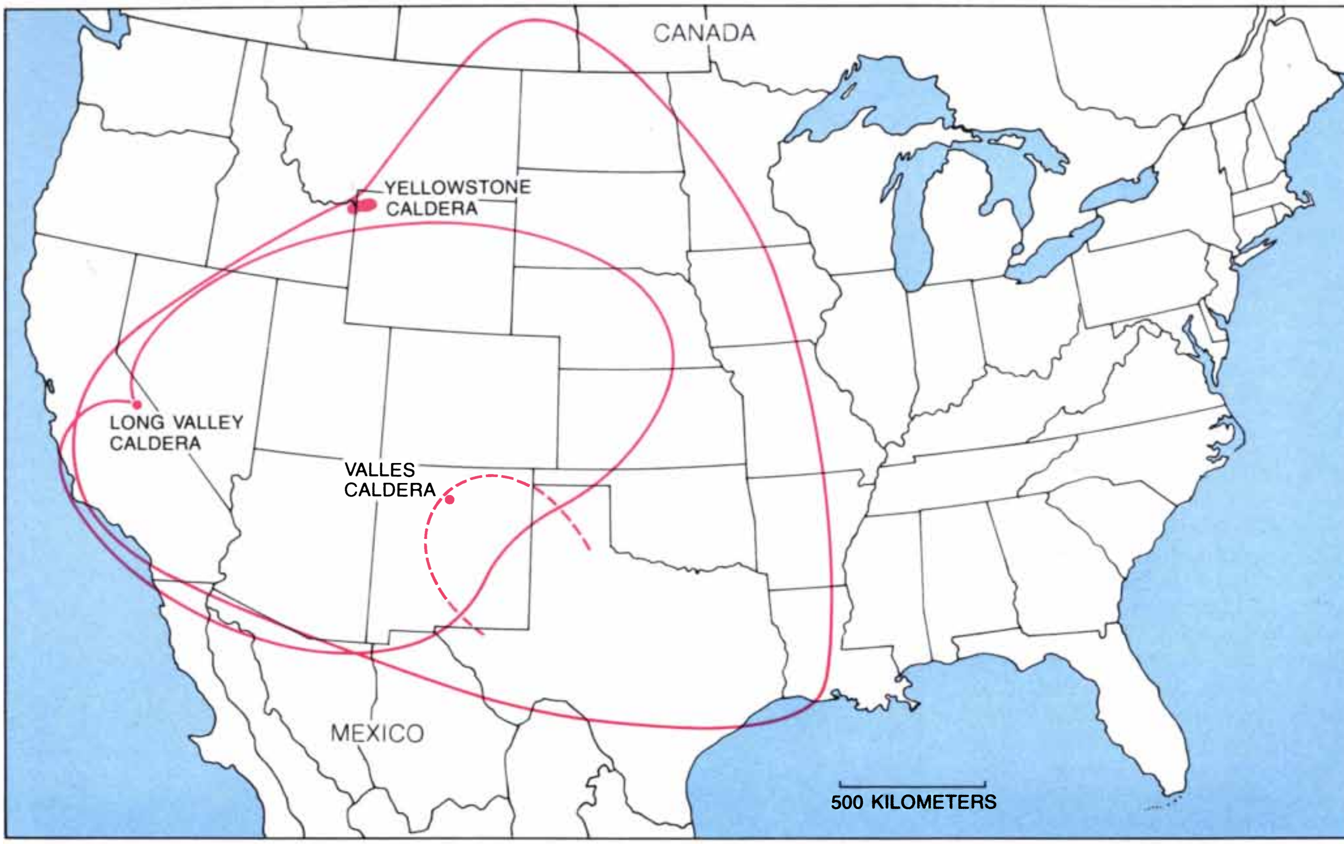
Nevertheless, resurgent calderas are not randomly distributed around the globe. The ignimbrites that characterize them result from the eruption of dacitic or rhyolitic magma, which is viscous, rich in silica and typically is produced in areas where the continental crust is thick. Resurgent calderas may form, therefore, in regions of the continental crust where a thermal plume (a "hot spot") in the mantle of the earth is large enough and long-lasting enough to melt huge volumes of rock. The plume does not melt the continental crust directly. Instead it melts part of the mantle to create a basaltic magma. The basaltic magma rises, melting the rock at shallower levels.

Consider some specific calderas. In the U.S. the Yellowstone caldera lies at the northeastern end of a trail of volcanic activity that begins in Idaho in the basaltic rock of the Snake River plain. Over the past 15 million years the focus of volcanic activity has shifted along the trail to its present position in Wyoming, perhaps in response to the movement of the plate that includes the North American continental crust over a stationary thermal plume in the mantle. A number of other calderas, no older than a few tens of millions of years, are found in a zone many hundreds of kilometers wide in Nevada, Arizona, Utah and New Mexico. The youngest caldera in the group is on the flanks of the Rio Grande Rift, which runs for hundreds of kilometers northward through New Mexico into Colorado. It is thought the continental crust at the Rio Grande Rift has somehow been thinned, producing the rift itself. A similar process is thought to have caused rifts in the oceanic crust near many of the island arcs in the Pacific. At the Rio Grande Rift the thinning may have allowed heat from the mantle to act with great effect.

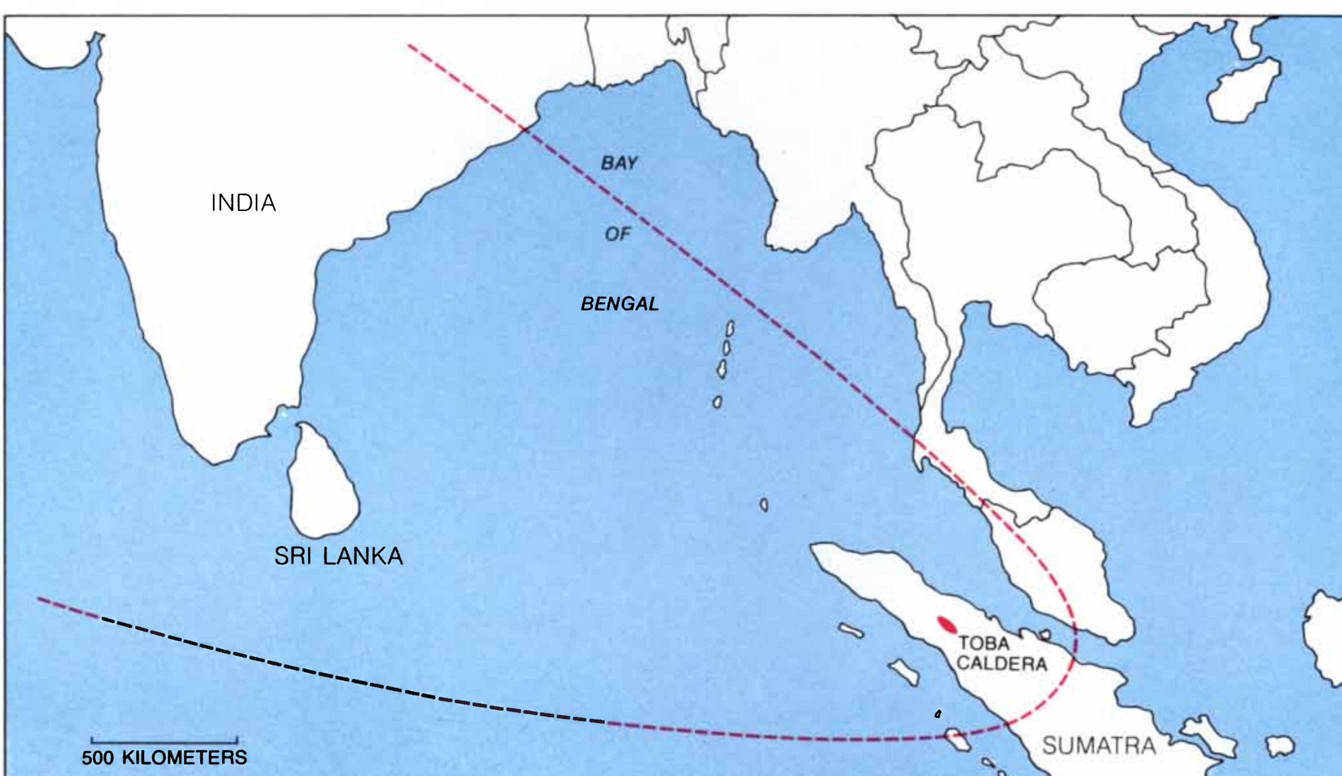
Other settings are also possible. In Ar-

**WELDED IGNIMBRITE, shown actual size, was photographed on the flanks of the Valles caldera in New Mexico. The dark streaks are compressed clasts, or fragments, of glassy pumice (a frothy volcanic rock) that softened and fused in the heat of the layer of ejecta laid down by the caldera-forming eruption. The clasts now lie in the midst of a hard, dense rock.**

**UNWELDED IGNIMBRITE was photographed on the flanks of the Cerro Galán caldera. The clasts of pumice lie in a matrix of volcanic ash and dust. They can be removed by hand.**




**RAIN OF ASH** is mapped for the three eruptions that left resurgent calderas in the U.S. over the past million years. An eruption a million years ago left the Valles caldera in New Mexico, an eruption 700,000 years ago left the Long Valley caldera in California and an eruption 600,000 years ago left the Yellowstone caldera in Wyoming. The mapping was done largely by G. A. Izett and his colleagues in the U.S. Geological Survey; the uncertain extent of the Valles caldera's ash is shown as a broken line. Ordinary volcanoes can distribute ash over large areas, but the thickness of the ash layer laid down by a caldera-forming eruption is centimeters rather than a fraction of a millimeter.



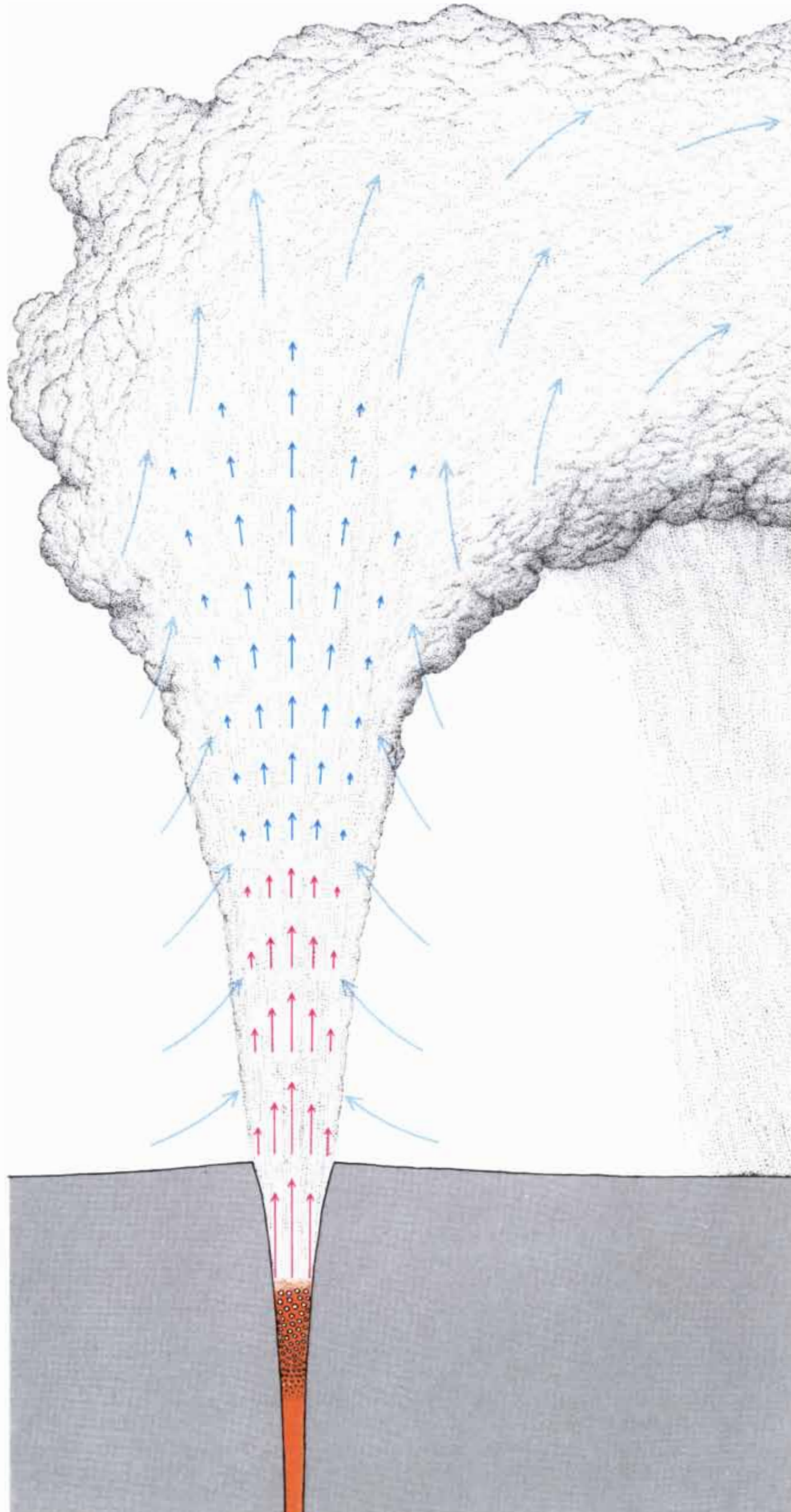

**ASH FROM THE TOBA CALDERA** in Sumatra rained down on a region including most of the Bay of Bengal and parts of Sumatra, Sri Lanka and India 75,000 years ago. The pattern of distribution was determined by a group led by Dragoslav Ninkovich of the Lamont-Doherty Geological Observatory of Columbia University; the group found an ash layer more than 10 centimeters thick in cores extracted from the sea floor as far as 2,000 kilometers from the Toba caldera itself. The Toba caldera is the youngest and the largest caldera known.

gentina and Bolivia resurgent calderas have formed not only along the main volcanic cordillera of the Andes but also in a second cordillera more than 200 kilometers farther inland. Here one finds no clear evidence of crustal thinning. On the contrary, the continental crust may be as much as 40 to 50 kilometers thick under the Cerro Galán and Kari Kari calderas. It is thought magma conduits to the surface of the inland cordillera may have developed by local extensional stress: a fracturing of the crust caused by the pressure of magma moving upward because of its buoyancy.

A number of events can be identified in the evolution of a typical caldera: precaldera doming, caldera collapse, eruption of air-fall material and pyroclastic flows, postcaldera resurgence and finally late-stage extrusions of lava. Precaldera doming is the elevation of the surface of the earth that precedes a massive eruption. It happens when a great volume of magma intrudes itself into a shallow level of the continental crust, creating a pluton, or magma chamber, whose top may lie only four or five kilometers below the surface. The doming generates a local extensional-stress field in the surface rock, and the stresses are important to the next event in the evolution: the collapse of the caldera.

An ambiguity persists, however. Does the upward pressure of the pluton cause the failure of the rock that forms the roof of the chamber, allowing magma to escape to the surface along a great ring-shaped fracture that develops above the pluton's perimeter? The escape of the magma would open a subterranean space into which the roof could then subside. Or does the roof of the pluton sink into the magma and in the process open the ring fracture along which magma erupts? T. H. Druitt and R. S. J. Sparks of the University of Cambridge have argued that whether the subsidence is a cause or an effect of the eruption, the ring fracture, which becomes the wall of the caldera, must be vertical or must bell outward with increasing depth in order to accommodate the rate of eruption and the pistonlike fall of the roof.

The magma at the top of the pluton has a temperature of 700 to 1,000 degrees Celsius and is rich in dissolved gases, mostly water vapor. The magma rises toward the surface along the newly opened ring fracture. As it rises, the pressure on it lessens, until at a depth of about a kilometer the gases come out of solution, much as they do when the cork on a bottle of champagne is popped. Dacitic or rhyolitic magma is much more viscous than champagne, however (or even basaltic magma), and so the gases do not merely bubble away. Instead they blow the magma apart. The actual proc-



**PLINIAN COLUMN** is generated in a volcanic eruption when the rate of eruption is great, when the erupting magma includes large amounts of dissolved gas and when the fissure venting the magma is narrow: perhaps 50 to 100 meters in diameter. At a depth of about a kilometer the gas comes out of solution, causing the magma to explode into incandescent pumice and ash. The particles are blasted upward ballistically (*solid red arrows*) to form the lower part of the plinian column. In the upper part of the column they are lofted many kilometers higher (*solid blue arrows*) by convection currents set up in the atmosphere. It is thought that plinian columns are often the first phase in the catastrophic eruption that forms a resurgent caldera.

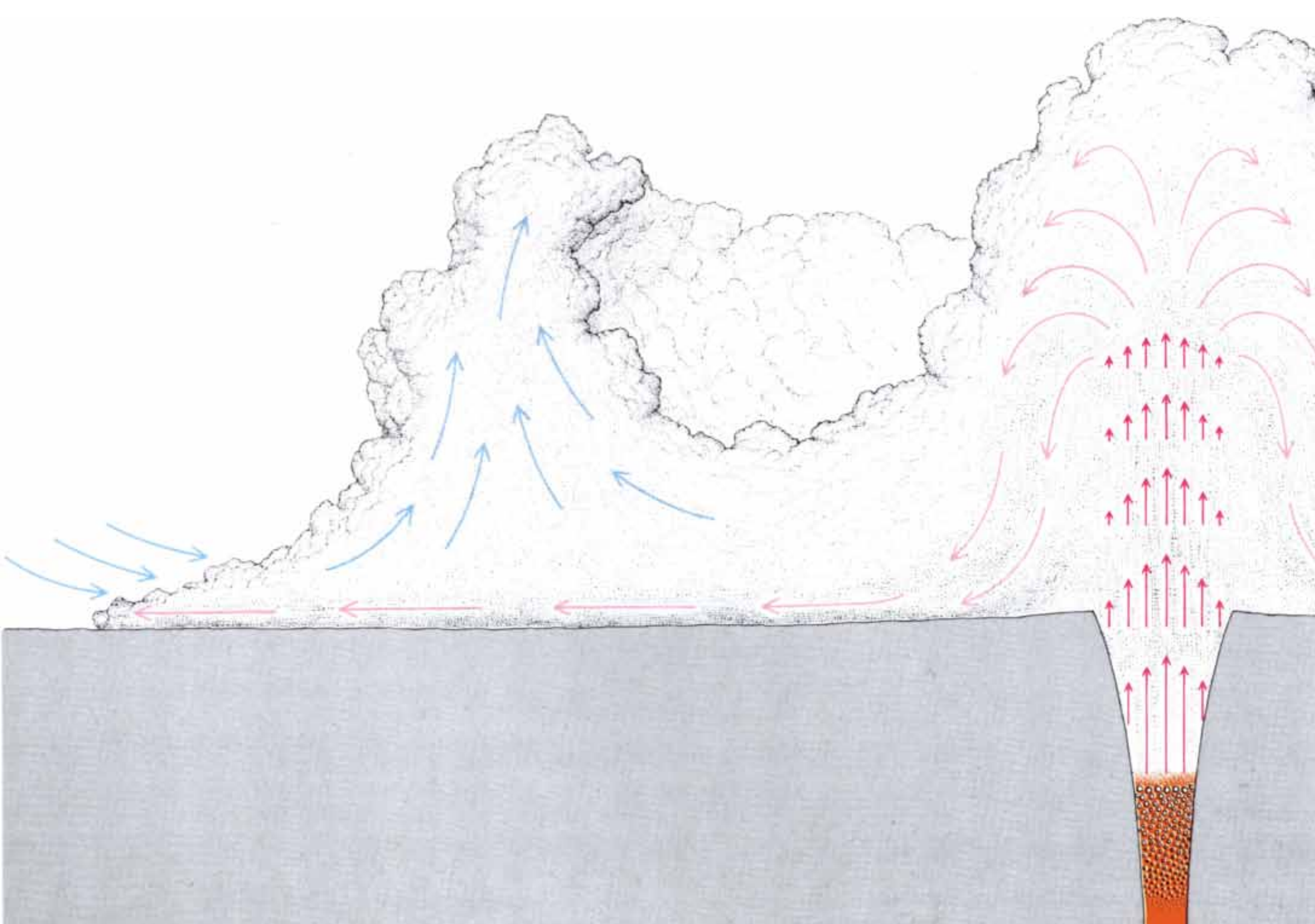
ess is complex but the result is simple: the magma escaping from the pluton toward the surface is expanded into pumice and explosively fragmented into incandescent solid particles ranging in size from micrometers to meters.

The key to understanding the effects of a great caldera-forming eruption is the action of the incandescent mixture of dust, ashes, pumice and liberated gases when it is vented at the surface. The kinetics of the venting have been analyzed by a group originally centered on G. P. L. Walker of the Imperial College of Science and Technology, notably Sparks of Cambridge, Lionel Wilson of the University of Lancaster and Colin J. N. Wilson of the University of Auckland. They have shown that when the rate of eruption is great and the vent is relatively small (perhaps 50 or 100 meters in diameter), an eruption column develops, rising tens of kilometers into the atmosphere. (The eruption column of Mount St. Helens on May 18, 1980, reached about 20 kilometers.)

It should be emphasized that the pumice in the column is not simply propelled upward from the vent like buckshot from a shotgun. Directly above the vent the energy of the column indeed is largely kinetic. The upward velocities are hundreds of meters per second. As the pumice rises, however, it rapidly decelerates; it is slowed not only by gravity but also by aerodynamic drag. Then a second process begins to contribute energy. The decelerating assemblage of incandescent pumice, ash and gas entrains and heats up air from around the column. As a result the assemblage becomes buoyant and begins to rise convectively. It may even accelerate upward again. Eruption columns rising by convection are well known; they are the essence of what is called a plinian eruption, after Pliny the Elder, whose description of Vesuvius in A.D. 79 provides the first documented example. Convection can drive a plinian column to heights as great as 50 kilometers.

Massive plinian columns may signal the start of a catastrophic collapse that creates a caldera. As the eruption proceeds, however, the plinian columns typically give way to pyroclastic flows, which account for by far the greatest fraction of the volume of the eruption. The reasons are several. The size of the vent can be increased by the blast of the initial eruption or by the collapse of the caldera floor. The gas content of the magma can decrease as lower levels of the pluton are tapped. In these circumstances it becomes more difficult for a plinian column to sustain itself. The column becomes denser than the surrounding air and soon collapses. That is, the mass of incandescent pumice and gas acquires only the first, ballistic component of upward motion. It still may reach a height of 10 kilometers, but without convective flow to sustain it, it falls back around the vent.

Now the pyroclastic flows develop. They are made up of the falling material, and they surge outward from where the material reaches the ground. Their mobility is extraordinary. It is known from the distribution of the ignimbrites they deposit that they can sweep over hills as much as a kilometer high and travel distances of up to 150 kilometers.



**PYROCLASTIC FLOW is generated in a volcanic eruption when the rate of eruption slows, when the erupting magma includes little gas or when the vent widens. Often such a flow results when a plinian column collapses, so that volcanic particles rise only ballistically (*solid***

No large-scale pyroclastic flows have ever been observed in action, but their hill-climbing abilities imply a velocity on the order of 100 meters per second. The smaller flows in the later stages of the eruption of Mount St. Helens had velocities of 30 meters per second.

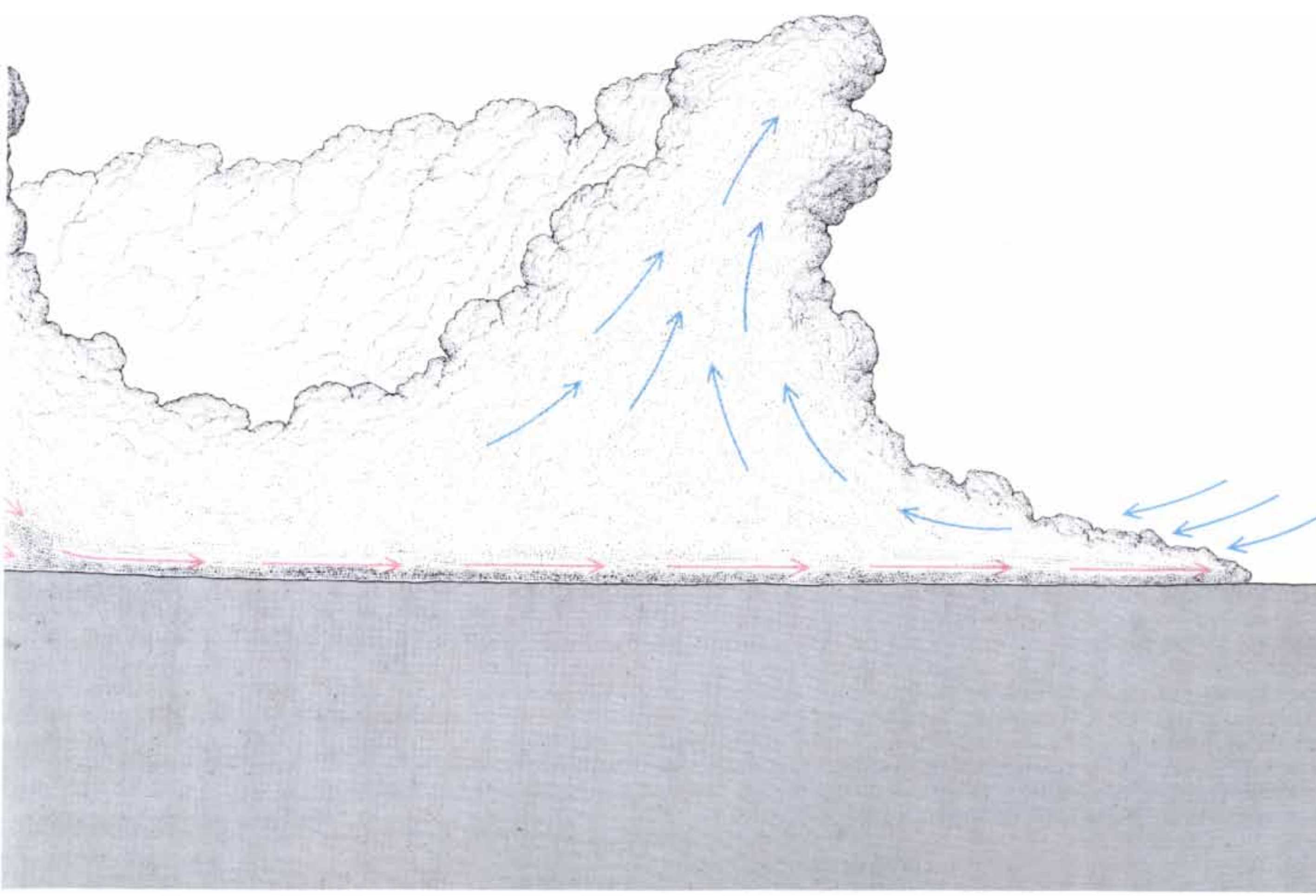
The explanation of the mobility is fairly well established as a result of recent theoretical and experimental studies such as those of Sparks and Colin Wilson. The flows travel in a partially fluidized condition. That is, the gases that continue to leave the hot pumice particles, combined with the atmospheric gas entrained at the front of the flow, stream upward among the particles, supporting the finer grains. The finer grains in turn provide a lubricating medium in which the larger clasts are dispersed. The flow thus comes to have a lower density and a much lower viscosity than an unfluidized aggregation of pumice and dust would have. It makes efficient use of its kinetic energy, which derives from its vertical fall from a height of several kilometers.

The main deposit of a massive pyroclastic flow is a blanket of pumice and smaller particles that can be many meters thick more than 50 kilometers from the vent. In addition the fine particles entrained by the flow typically form a secondary ash cloud that rises many kilometers by convection. The subsequent fall of particles from the cloud can deposit a thin layer of ash over a region much larger than the one covered by the ignimbrite of the pyroclastic flow itself. Indeed, the layer, which is called co-ignimbrite ash, can amount to as much as a third of the entire volume of the ignimbrite.

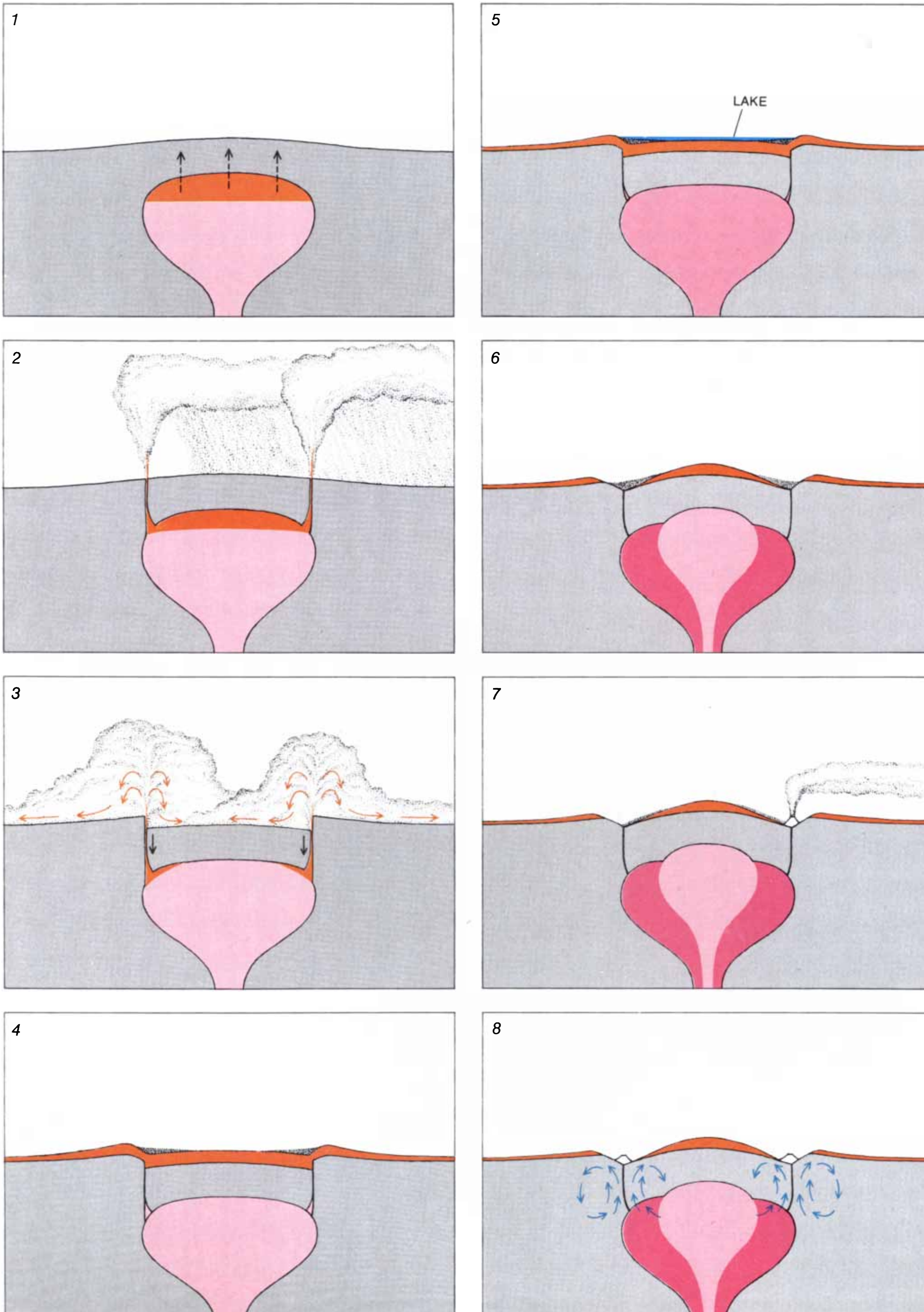
Why are the eruptions that form resurgent calderas exclusively eruptions of dacitic or rhyolitic magma? Two interrelated factors are responsible. First, because of the lower viscosity of a basaltic magma, gases coming out of solution in the magma can easily escape. The eruption is not catastrophically explosive, and the production of pumice and fine ash is inhibited. Second, because a basaltic eruption gives rise to scant fine ash, the transfer of heat from the eruption column to the atmosphere is lessened. (Such transfer is far more efficient for particles less than a millimeter in diameter than it is for larger ones.) Hence it is difficult for an eruption of basaltic magma to generate convective currents in the atmosphere and become a plinian column.

The "fire fountains" seen on the active volcanoes of the Hawaiian Islands provide an excellent example. In a fire fountain great volumes of lava are sprayed high into the air, but the lava, which is basaltic, emerges as large liquid gobbets, sometimes a meter or more across. Hardly any fine ash erupts. The gobbets do not readily transfer heat to the atmosphere, and so there is virtually no chance that a convective column might develop. Instead the eruption leaves lava spattered on the ground around the vent. Since the lava retains almost all its original heat, secondary lava flows are common.

One of the most surprising aspects of the catastrophic eruptions that form a resurgent caldera is how short they are. Michael T. Ledbetter of the



**_red arrows_) and then fall back to the ground. There they form the flow: a ground-hugging cloud of particles buoyed up and fluidized by hot gases. The flow can climb hills a kilometer high and can travel distances of 150 kilometers at velocities of 100 meters per second.**


1
5
LAKE
2
6
3
7
4
8

University of Georgia and Sparks have studied the rate at which ash particles settle in seawater. The rate depends on the density and the radius of each particle; therefore fine ash among the ejecta from the beginning of an eruption and coarse ash among the ejecta from the end of the eruption may reach the sea floor simultaneously. Ledbetter and Sparks deduce from the data they have collected that the 300 cubic kilometers of ash erupted from the Atitlán caldera in Guatemala 84,000 years ago fell over a period of only 20 to 27 days. They think it unlikely that the eruption was continuous. Instead a plinian column probably collapsed a number of times, causing the successive emplacement of distinct ignimbrite sheets. (The Atitlán caldera is 28 kilometers across, but its center is not resurgent.) The same approach suggests the eruption that formed the Toba caldera deposited more than 1,000 cubic kilometers of ash in as little as nine days.

After the eruption several things happen. Typically a lake fills the new caldera. Sediment erodes from the caldera wall and accumulates at the bottom of the lake. On a slower scale the resurgence of the caldera floor begins. In general it is not a simple pistonlike rise of the floor (and of the ignimbrite that now blankets it) along the ring fracture. Mapping of the ignimbrite in Cerro Galán and other calderas shows that it is tilted away from a well-defined central point, often at an angle exceeding 45 degrees. This suggests resurgence is a local phenomenon. Probably it occurs above a small pluton newly intruded into the one that caused the eruption. In some cases, including Yellowstone, two separate centers of resurgence are present in one caldera.

It was the recognition that young lake sediments had been raised hundreds of meters that enabled van Bemmelen to demonstrate that the resurgence of the floor of the Toba caldera had formed the 640-square-kilometer island of Samosir. At Cerro Galán resurgence of more than one kilometer has raised the center of the caldera to an altitude of more than six kilometers above sea level, making it one of the highest mountains in Argentina. Little is known about the rate of resurgence, but in the case of the Long Valley caldera, the one at which the dating of the geological events is least equivocal, the process may have continued for up to 200,000 years. Since the Toba caldera is only about 75,000 years old, the resurgence of Samosir may not yet be complete.

After resurgence come the final episodes in the evolution of a caldera: the relatively quiet effusion of dacitic or rhyolitic lava from a necklace of vents along the ring fracture. Typically the volume of material released is small, but the effusions continue intermittently long after the catastrophic caldera-forming eruption. At Long Valley distinct episodes of effusion took place 500,000, 300,000 and 100,000 years ago. It is fair to conclude that the volcanic events associated with the formation of a caldera may continue with little violence for up to a million years. Hot springs and geysers, representing geothermally heated water that finds its way to the surface, may be present for much longer than that.

The hazards to human life that would be presented by a caldera-forming eruption have never been analyzed in detail. They would surely be far-reaching. Much would depend on the total volume of the eruption and on the proportions of air-fall ash, ignimbrite and co-ignimbrite ash. Consider a 1,000-cubic-kilometer eruption. A tract of land with a surface area of perhaps 500 square kilometers would sink, and the resulting caldera would be filled entirely with ignimbrite. A surrounding area of up to 30,000 square kilometers would also be covered by ignimbrite. The depth of the cover would range from more than 100 meters on the caldera rim to a few meters at the farthest extent of the ignimbrite. Throughout the covered area plants and animals would be killed. Buildings would be demolished or buried.

In the short term the most serious consequence of the eruption would probably be the fall of co-ignimbrite ash, which would have a pattern dictated by the prevailing wind and the size distribution of the particles. Particles much larger than about 80 micrometers stay in the atmosphere for a few hours at most. Smaller particles stay aloft for much longer. Particles less than 10 micrometers in diameter have settling velocities in the atmosphere no greater than the velocities typical of atmospheric turbulence. They may thus stay airborne for times of up to 100 days and so can be transported great distances. Co-ignimbrite ashes tend to be extremely fine; they would probably be dispersed over several million square kilometers.

Imagine the eruption of Mount St. Helens multiplied more than a hundredfold and taking place at Yellowstone today. Apart from the immediate effects, which would include the immobilization of ground and air traffic by co-ignimbrite ash, the most severe effect would probably be on agriculture, since millions of square kilometers of grain-growing farmland would be covered by ash to a depth of several centimeters. The result is easy to forecast: the disruption of at least one year's harvest, and possibly several. Over a longer term the climate of the earth would be affected by the injection of dust into the atmosphere together with an extremely long-lasting aerosol of particles that arise from acidic volcanic gases. The resulting worldwide disruption of agriculture could persist for a number of years.

Could such an eruption be predicted? The most likely indication of a future caldera-forming eruption might be the "leakage" of dacitic or rhyolitic magma to the surface as a ring fracture develops. In addition seismic signals might indicate the movement of magma into a chamber a few kilometers below the surface. (Of the two types of seismic waves, *S* and *P*, the *S* waves are the most revealing, because they cannot pass through a liquid.) The site of the chamber would be confirmed by a local anomaly in the earth's gravitational field. (The density of magma is less than the density of solid crustal rock.) A further important indicator would be the tumescence, or elevation, of the terrain, and particularly an increase in the rate of tumescence. Conventional surveying techniques can readily reveal such changes.

The difficulty lies in interpreting those various signs as the precursors of a particular event and in placing them on a timetable in the absence of a known case history. Seismic activity and tumescence have recently been noted at the Long Valley caldera and have been taken to indicate the emplacement of fresh magma at a depth of several kilometers. Will a large eruption follow? Or do the events presage merely the quiet extrusion of a new lava dome at the surface? The immediate precursors of a caldera-forming eruption are largely unknown. Moreover, the precursor activity may

**DISTINCT STAGES mark the evolution of a resurgent caldera. In the first stage magma at depths of as little as four or five kilometers below the surface forms a pluton, or magma chamber, which slightly domes the surface (*1*). In the pluton there are zones that differ in composition, with viscous magma rich in silica ($SiO_2$) and dissolved gases at the top (*orange*). Eventually an eruption begins. Plinian columns develop above a great ring-shaped fracture (*2*). Then within minutes or hours the roof of the pluton collapses along the fracture, leaving a caldera, and the plinian columns give way to pyroclastic flows (*3*). The eruption lasts for only a few days. It fills the caldera with ignimbrite and also blankets the surrounding area (*4*). Meanwhile the caldera wall begins to erode (*black stipple*). A lake may also form (*5*). Still later, over a period of a few hundred thousand years, the intrusion of fresh magma into the pluton causes parts of the caldera floor to move upward again (*6*). Minor volcanic activity persists along the ring fracture (*7*). The heat of the pluton may power convection currents of water rich in minerals (*8*) and so give rise to hot springs and geysers at the surface for millions of years.**

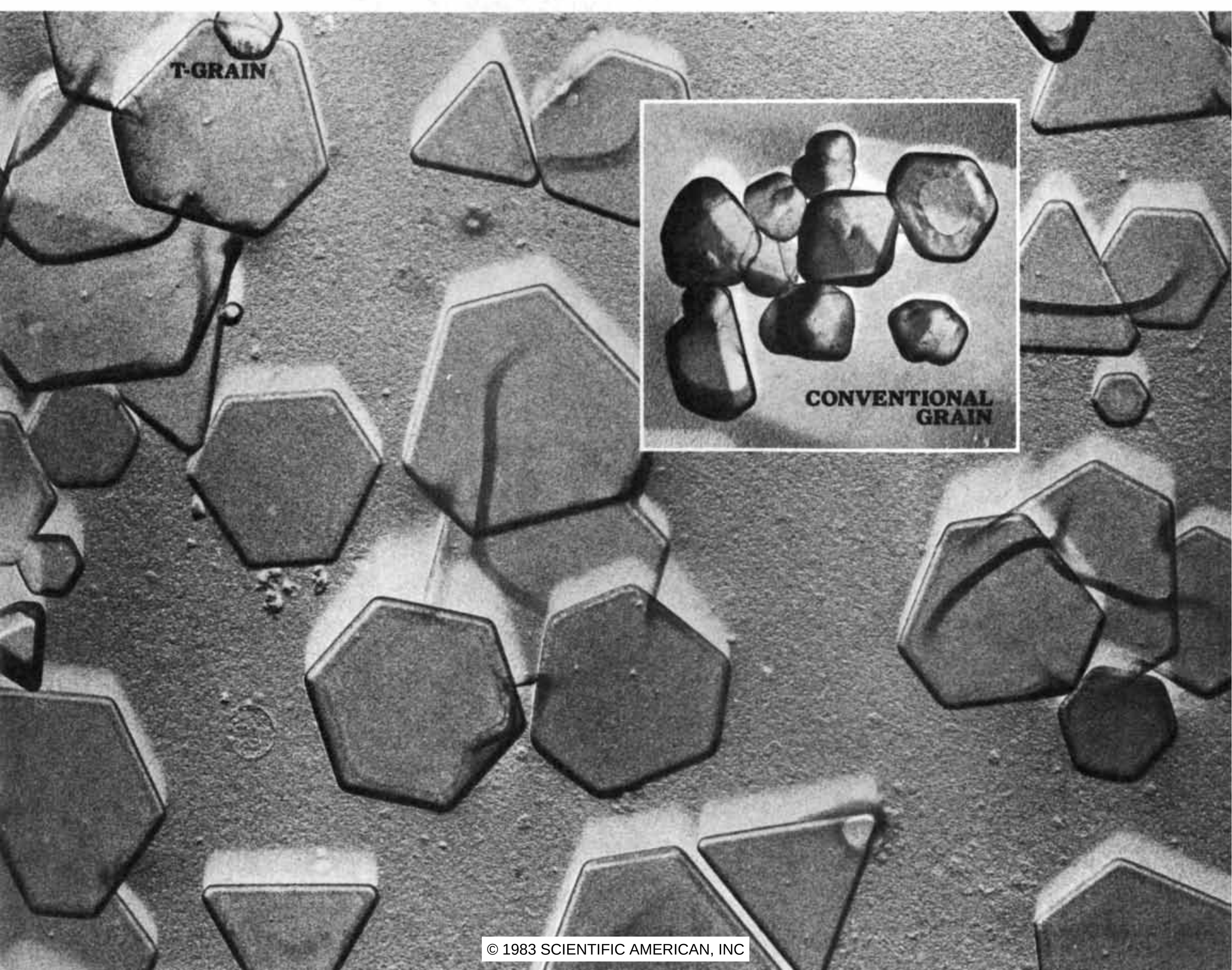
continue on a time scale spanning many human generations. Hence it may not be possible to predict an eruption correctly. All the same, signs of renewed activity at any known caldera should not be ignored.

The eruptions that form resurgent calderas are so rare that on balance they are probably advantageous to man. A caldera overlying a pluton marks a thermal anomaly in the crust that may continue to manifest itself in hot springs for two or even three million years after the eruption. Apart from the splendid spectacle afforded by the geysers, the hot water and steam are potential sources of useful energy, and so is the hot rock overlying the pluton itself. An important experiment under way in the U.S. is the Hot Dry Rock Project of the Los Alamos National Laboratory. On the flanks of the Valles caldera water is being pumped down a deep borehole into a zone of hot, fractured rock. Heated water returns to the surface through a second borehole. The hot water can be employed for the generation of electricity.

Some less apparent benefits of calderas lie in the useful minerals that are deposited in resurgent calderas or around them over a period of time. The Kari Kari caldera provides an example that is geologically typical and also rich in historical associations. On the flanks of the caldera are the city of Potosí and also the Cerro Rico, a peak whose summit is 4,900 meters above sea level. The peak is composed of rhyolite that was intruded into the caldera's ring fracture about 13 million years ago. As it cooled it was chemically reworked by hydrothermal solutions: hot, acidic liquids ascending after the reaction of water with deep-lying hot rock. In the 17th century Potosí was the largest city in the Western Hemisphere, solely because of the wealth of silver mined from the veins laid down by the solutions. The importance of the mining to Spain is hard to overstate: it gave rise to a well-known Spanish expression, "worth a potosí." Ultimately the flood of silver from Potosí and other mines in the New World disastrously inflated the Spanish currency. Mining has been continuous at Cerro Rico since 1544, but the silver easily won from the upper levels of the deposit has long been exhausted. Tin is now sought at deeper levels.

Clearly the deep-seated sources of heat that create calderas are capable of driving hydrothermal circulation over long periods, and clearly the hydrothermal activity can work complex chemical reactions that allow mineral deposits to precipitate. It may be a while before the details are understood, but the association of calderas with geothermal energy and with valuable mineral deposits ought to inspire gratitude. Without this association resurgent calderas might appear to be unalloyed catastrophes.

**TWO LAYERS OF ASH at Crater Lake in Oregon result from an eruption 7,000 years ago that tapped the viscous magma at the top of a pluton and then the less viscous magma under it. The layers of ash reverse that relation. The lower layer, light in color, was deposited first. It is rich in silica and poor in iron and magnesium. The upper, darker layer was deposited immediately afterward. It is poorer in silica and richer in iron and magnesium. Crater Lake's diameter of only nine kilometers makes it too small to be considered a major caldera.**



T-GRAIN
CONVENTIONAL GRAIN

### 4. Gastroenterology

**Gary M. Gray, M.D.,** Stanford University School of Medicine

**Peter B. Gregory, M.D.,** Stanford University School of Medicine

**John Austin Collins, M.D.,** Stanford University School of Medicine

**Douglas Wilmore, M.D.,** Harvard Medical School and Brigham and Women's Hospital

### 5. Hematology

**Stanley L. Schrier, M.D.,** Stanford University School of Medicine

### 6. Immunology

**John David, M.D.,** Harvard Medical School and Harvard School of Public Health

### 7. Infectious Disease

**Thomas C. Merigan, M.D.,** Stanford University School of Medicine

**Morton N. Swartz, M.D., F.A.C.P.,** Harvard Medical School and Massachusetts General Hospital

**Cyrus C. Hopkins, M.D.,** Harvard Medical School and Massachusetts General Hospital

**Adolf W. Karchmer, M.D., F.A.C.P.,** Harvard Medical School and Massachusetts General Hospital

**Robert H. Rubin, M.D., F.A.C.P.,** Harvard Medical School and Massachusetts General Hospital

**Harvey B. Simon, M.D., F.A.C.P.,** Harvard Medical School and Massachusetts General Hospital

**Peter F. Weller, M.D.,** Harvard Medical School; Beth Israel Hospital, Boston; and Brigham and Women's Hospital

### 8. Intensive and Emergency Care

**Edward Rubenstein, M.D., F.A.C.P.,** Stanford University School of Medicine

### 9. Metabolism

**George F. Cahill, Jr., M.D.,** Harvard Medical School, Howard Hughes Medical Institute; and Brigham and Women's Hospital

### 10. Nephrology

**Roy H. Maffly, M.D.,** Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center

**William Bennett, M.D.,** Health Sciences University, Portland, Oregon

### 11. Neurology

**Robert W. P. Cutler, M.D.,** Stanford University School of Medicine

### 12. Oncology

**Saul A. Rosenberg, M.D., F.A.C.P.,** Stanford University School of Medicine

### 13. Psychiatry

**Ned H. Cassem, M.D.,** Harvard Medical School and Massachusetts General Hospital

### 14. Respiratory Medicine

**Eugene D. Robin, M.D., F.A.C.P.,** Stanford University School of Medicine

### 15. Rheumatology

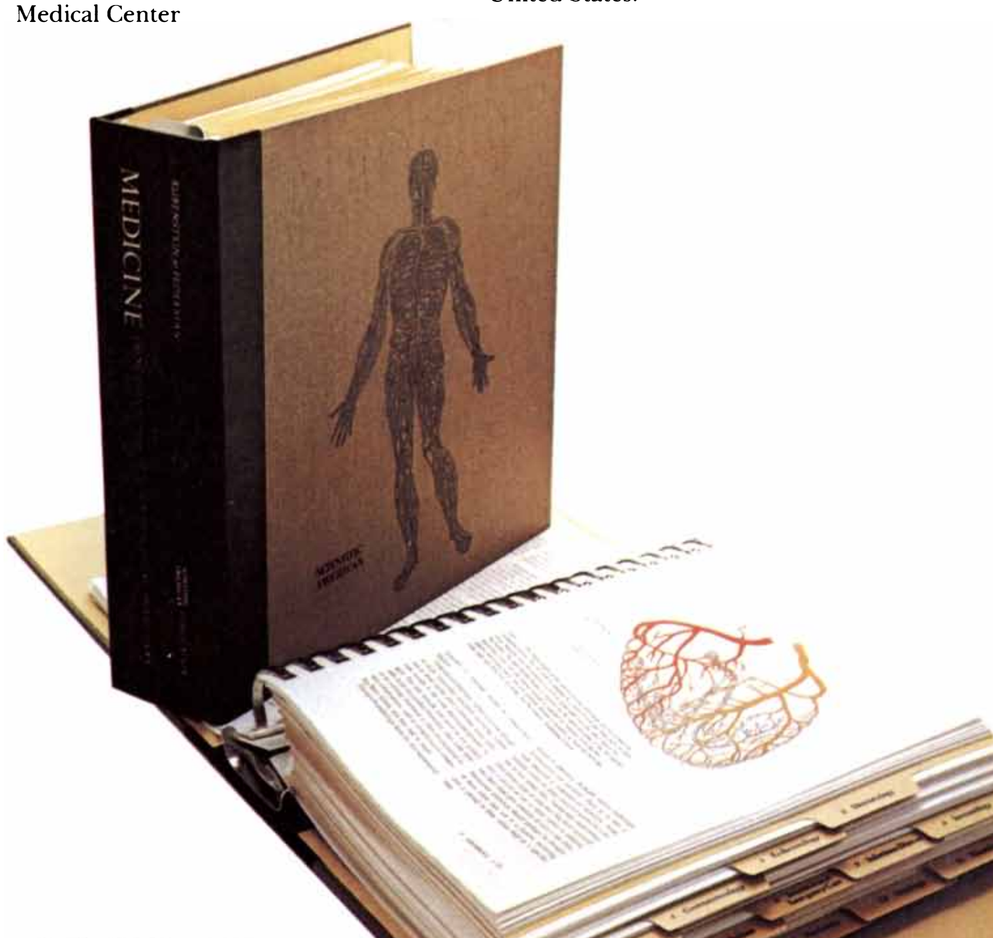
**Stephen M. Krane, M.D.,** Harvard Medical School and Massachusetts General Hospital

**Dwight R. Robinson, M.D.,** Harvard Medical School and Massachusetts General Hospital

**Andrei Calin, M.D., M.A., M.R.C.P.,** Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center

## Order by Phone

You can order SCIENTIFIC AMERICAN *Medicine* by telephone. **Please call this toll-free number: 1-800-345-8112 (in Pennsylvania, call 1-800-662-2444);** you will be billed after your subscription begins. Toll-free calls are acceptable only for orders placed in the continental United States.



# SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, N.Y. 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the price of $220 for first year of service is tax deductible, as is current renewal at $170. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*

☐ I shall also enroll in the CME program

☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*

☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date________ Account Number________

*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name________

MD Specialty________

Address________

City________ State________ Zip________

Signature________

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its pos used unless airmail delivery is prepaid.

# SCIENCE AND THE CITIZEN

## *Ill-starred Wars*

President Reagan's televised appeal to "those who gave us nuclear weapons" to join in a new long-term effort to develop a nationwide anti-ballistic-missile (ABM) system has met with a largely negative response from the scientific community. The main arguments advanced in opposition to such an undertaking are much the same as they were in the 1960's, when the feasibility of ballistic-missile defense first emerged as a major public issue.

At that time the proposed method of dealing with a long-range missile attack on the U.S. by the U.S.S.R. called for the deployment of an extensive array of ground-based radars and nuclear-armed interceptor missiles. Critics of the idea pointed out that any system of this kind, no matter what the cost, was unlikely to be fully effective against a determined adversary and hence of dubious value given the great destructiveness of even a small fraction of the possible incoming nuclear warheads. The ABM system could not provide an impenetrable shield over the entire country because of various countermeasures available to the attacker. For example, the attacking country could change its offensive tactics, black out the system's long-range radars with high-altitude nuclear explosions, fool the defense with a large number of comparatively cheap decoys and other "penetration aids" or simply overwhelm it by adding enough offensive warheads to the attacking force to exhaust the supply of antimissile missiles. Such an action-reaction cycle in military preparations, it was argued, would lead inevitably to an intensification of the strategic-arms race [see "Anti-Ballistic-Missile Systems," by Richard L. Garwin and Hans A. Bethe; SCIENTIFIC AMERICAN, March, 1968].

The force of these arguments was formally recognized in the anti-ballistic-missile treaty of 1972, in which each superpower agreed "not to deploy ABM systems for a defense of the territory of its country and not to provide a base for such a defense." The treaty limited each side to local ABM systems at no more than two sites (later reduced to one).

Although President Reagan did not specify what kind of ABM system he had in mind when he spoke of his new "vision of the future," a spokesman for the Administration acknowledged that the systems currently under consideration include both particle-beam weapons and laser weapons, based either on the ground or in space. According to independent studies carried out at the Massachusetts Institute of Technology, these systems face insurmountable technological obstacles; moreover, they would be just as vulnerable to offensive countermeasures as the ABM systems of the 1960's were [see "Particle-Beam Weapons," by John Parmentola and Kosta Tsipis, SCIENTIFIC AMERICAN, April, 1979, and "Laser Weapons," by Kosta Tsipis, SCIENTIFIC AMERICAN, December, 1981].

The one thing that does seem certain about the president's initiative is how it will be perceived by military analysts on the other side of the nuclear "balance of terror." According to a statement signed by 244 scientists in the U.S.S.R., "there are no effective defensive means in nuclear war, and their creation is practically impossible." Nevertheless, the statement adds, the proposed "attempt to create so-called 'defensive weapons' to counter the strategic nuclear force of the other side ... will inevitably lead to the emergence of a new element strengthening the American 'first strike' potential [by] striving to lessen the power of a retaliatory strike.... By his statement the President is creating a most dangerous illusion, which may cause an even more threatening spiral of the arms race. We are firmly convinced that this act will result in a sharp lessening of international security, including the security of the U.S."

## *Underground Science*

Cosmic rays were once the principal source of particles for experiments in high-energy physics, but lately physicists have been going to considerable lengths—or considerable depths—to escape them. Experiments have been set up in mine shafts and tunnels, where they are shielded from most cosmic rays by the overburden of earth and rock. Indeed, in the past few years there has been a sudden proliferation of underground experiments, due in large part to interest in a search for the decay of the proton. The number of suitable deep sites, however, is quite limited. It has now been proposed to build a National Underground Science Facility, which would provide a permanent home and supporting staff for such work.

The need for cosmic peace and quiet is most acute in experiments that attempt to measure some exceedingly rare process whose signal might be lost in a noisy background of extraneous events. The conjectured decay of the proton is a notable example. The proton was long considered an immortal form of matter, but theoretical work in the past several years has given reason to believe a proton might occasionally decay. The average lifetime would have to be at least $10^{30}$ years, or some $10^{20}$ times the age of the universe, and a detector made up of thousands of tons of material could be expected to register no more than a few decays per year. They would have to be distinguished from the much more numerous events induced by the passage of cosmic-ray particles through the detector volume. Putting the apparatus deep underground minimizes this distracting and possibly misleading background.

Of the proton-decay detectors now in operation the largest is a tank filled with 10,000 tons of water installed 600 meters below the surface in a salt mine near Cleveland. It is operated by investigators from the University of California at Irvine, the University of Michigan and the Brookhaven National Laboratory. In January the group announced its first result: no decay events had been identified, and a new lower limit of $5 \times 10^{31}$ years could be set on the average proton lifetime. For theorists the result suggests that certain simple and appealing models of the structure of matter may have to be excluded from consideration. For experimental physicists the result implies that a still larger and more sensitive detector may be needed to gather evidence of proton decay. Even if one of the existing experiments detects the disintegration of a proton, apparatus with finer spatial resolution will be needed to study the events in detail. An experiment of this kind would be a leading candidate for installation in a national underground science laboratory.

In the search for proton decay the most troublesome background events are those induced by neutrinos, particles that cannot be screened out by any feasible thickness of overburden; neutrinos readily pass through the entire earth. Several other proposed underground experiments would examine the neutrinos themselves. For example, the energy spectrum of cosmic-ray neutrinos is only poorly known and might be measured by an underground detector acting as a neutrino "telescope." The conjectured phenomenon called neutrino oscillation, in which one type of neutrino is converted spontaneously into another type, might also be detected.

Another neutrino experiment, proposed by Raymond Davis, Jr., of Brookhaven, would yield a refined measurement of the flux of neutrinos emitted by the sun. Among mine-shaft physicists Davis is a pioneer: for almost 15 years he has been counting solar neutrinos with a detector in the Homestake Mine in South Dakota. The active element in the detector is chlorine, which can be converted by a neutrino into argon. The results have been consistently puzzling (the measured flux of neutrinos is smaller than expected), and their interpreta-

tion might be clarified by a new experiment based on gallium.

Several more possible uses of an underground science facility have been discussed. The search for massive magnetic monopoles, for example, might benefit from a reduced background of cosmic rays. Attempts to detect gravitational radiation, or gravity waves, might exploit another characteristic of a deep underground site: freedom from vibration. In the stillness of the subterranean laboratory an instrument sensitive to low-frequency gravity waves (from one to 100 hertz) might be operated.

The chief proponents of the National Underground Science Facility have been Alfred K. Mann of the University of Pennsylvania and R. R. Sharp, Jr., and Leonard M. Simmons, Jr., of the Los Alamos National Laboratory. At a meeting held last September at Los Alamos, Mann reviewed the work that might be done in the facility and the technical issues that must be addressed in building it. The depth of the laboratory, he points out, should be sufficient to provide shielding from cosmic rays equivalent to at least 2,500 meters of water; because rock is denser than water, the actual depth could be appreciably less. Large rooms would be needed; one of them should be at least 50 by 200 by 50 feet. The main access shaft would have to be at least 12 feet in diameter.

Mann and Sharp have made a survey of the available underground sites in the western U.S. They identified 13 existing mines and tunnels that might be suitable for the laboratory. None of the sites is ideal, however, and some of the best ones are in working mines, where commercial operations would be likely to interfere with scientific activities. Attention has therefore focused on a plan to build an underground facility in a corner of the Nevada Test Site, the desert tract near Yucca Flat where the Department of Energy conducts underground tests of nuclear weapons. The area was set aside some years ago for nonclassified work. The proposed laboratory would be at a depth of 3,600 feet, equivalent to 2,800 meters of water.

Some physicists have objected to the idea of working at the Nevada Test Site. Nevertheless, Mann, Sharp and Simmons point out that the area has important advantages; indeed, it may have been chosen for weapons testing for some of the same reasons that make it suitable for an underground laboratory. The geology and hydrology of the site have been thoroughly explored and are exceptionally stable. The staff of the test site itself has expertise in deep excavation; more deep, large-diameter shafts have been drilled there, Mann says, than in any comparable area elsewhere in the U.S. Construction costs (estimated at $42.3 million) would be paid by the Department of Energy, but it is possible that not all the cost would come out of the department's basic-science budget.

A committee chaired by Norman F. Ramsey, Jr., of Harvard University has been formed to advise the director of the Los Alamos laboratory on matters related to science underground and the proposed facility. In a preliminary report the committee stated that a dedicated underground facility is urgently needed to provide space and technical support for important experiments in particle physics and astrophysics. The committee also expressed approval of the Los Alamos proposal to construct the facility at the Nevada Test Site.

### *Messenger with Forked Tongue*

In the three decades since James D. Watson and Francis H. C. Crick elucidated the structure of DNA the picture of how the genetic apparatus works has become steadily more complex. In place of the original simple dictum that "DNA makes RNA and RNA makes protein" is an intricate mechanism in which many intermediary operations have an importance almost equal to that of the fundamental genetic events. Throughout this period of intensive revision in molecular genetics one assumption has remained intact, namely the assumption that DNA and RNA molecules are linear polymers. Now even that presupposition has been contradicted. In *Proceedings of the National Academy of Sciences* John C. Wallace and Mary Edmonds of the University of Pittsburgh report finding an RNA molecule whose chain is split into two branches.

The finding of Wallace and Edmonds is a by-product of the considerable body of work that has been done on the processing of RNA in the living cell. The components of the RNA polymer are the nucleotide bases adenine, guanine, cytosine and uracil. Each nucleotide includes a five-carbon sugar unit whose carbons are generally designated by numbers. The monomeric units of the polymer are connected by a phosphate group interposed between the 3′ carbon of one sugar and the 5′ carbon of the adjacent sugar. The sequence of the nucleotides in the chain constitutes the information for constructing a protein in the process called translation. The RNA molecule is assembled on a DNA template in the cell nucleus in the process called transcription; the RNA is then transported to the cytoplasm, where the translation is done.

Intervening between transcription and translation are several significant operations. These processing operations transform the raw RNA transcript, known as heterogeneous nuclear RNA (hnRNA) in animal cells, into mature messenger RNA (mRNA), the form employed in translation. Among the processing functions are the addition of different structures to the ends of the hnRNA. A sequence of from 150 to 200 adenine nucleotides is attached to one end; the resulting molecule is designated poly(A)$^{+}$hnRNA. A structure called a cap, consisting of a modified guanine nucleotide, is attached to the other end of the molecule by a linkage that includes three phosphate units.

The attachment of the poly(A) sequence and the cap is thought to have a role in preparing the RNA for translation. The specific contribution made by each of these modifications, however, is not well understood. Wallace and Edmonds were investigating the detailed structure of the cap when they made their recent finding. The cap can readily be separated from the rest of the hnRNA by means of the enzymes called ribonucleases, which sever the covalent bonds linking adjacent nucleotides and thus yield the monomers. The chemical structure of the cap protects it from degradation by ribonucleases. Hence the application of the enzymes to hnRNA leaves an assortment of individual nucleotides and the intact cap.

Both the cap and the nucleotides have a negative electric charge. The charge of the cap, however, is much larger than that of the monomers. The two kinds of molecule can therefore be separated by ion-exchange chromatography, a technique that can be employed to separate molecules on the basis of differences in charge. Wallace and Edmonds applied the procedure to ribonuclease-treated hnRNA taken from the cancer cells designated HeLa cells. As expected, the procedure revealed the presence of the single nucleotides and the cap.

There was also a second molecular structure with a negative charge roughly equal to that of the cap. At first Wallace and Edmonds assumed the new substance was a kind of cap that had not previously been observed. Further chemical tests showed, however, that it was not a cap. Instead the sequence of tests demonstrated that the newly discovered segment of RNA includes a junction resembling the crotch of a tree

with two branches extending from it. The central member of the junction is a single nucleotide. As in linear RNA, strings of nucleotides extend from the 3′ and 5′ carbons of the sugar unit. In addition an anomalous third string of nucleotides extends from the 2′ carbon.

Wallace and Edmonds note that this is the first report of a branched RNA molecule. Since branching is found only in the hnRNA and not in the mature mRNA, the branches could play a role in RNA processing. Indeed, Wallace and Edmonds go on to speculate that the branching could contribute to one of the major steps in processing. Unprocessed hnRNA includes two kinds of base sequence. One kind corresponds to the parts of the gene called exons, which are ultimately translated into protein. The other kind corresponds to the parts of the gene called introns, which are not translated. In the assembly of the finished mRNA the introns are removed and the ends of adjacent exons are spliced. Consider a piece of hnRNA with two exons separated by an intron. The authors note that if the RNA were cut at one end of the intron, the free end of the adjacent exon could be brought to the other end of the intron, where it joins the second exon. The free end could then be attached to the 2′ carbon of the first nucleotide of the second exon. Such an operation would yield the branched structure observed by Wallace and Edmonds. Later the dangling intron attached to the 3′ carbon could be removed and the spliced exon could be transferred from the 2′ carbon to the usual 3′ site.

## *Clouded Ball*

Cold clouds of neutral hydrogen atoms have a special place in the conduct of modern astronomy. Detected with radio telescopes by means of their radiation at a wavelength of 21 centimeters, they have proved to compose as much as a fifth of the mass in a typical spiral galaxy. Thus they are excellent markers of a galaxy's spiral structure. In recent months Yervant Terzian, Stephen Schneider, George Helov and Edwin Salpeter of Cornell University have employed the radio telescope at Arecibo in Puerto Rico to detect clouds of neutral hydrogen atoms in galaxies other than our own galaxy. In January they aimed the telescope toward a point between two small groups of galaxies in the constellation Leo. They intended only to calibrate their instruments by measuring the "background" at 21 centimeters when the telescope was pointed at "blank sky." Instead they found an immense cloud of neutral hydrogen in intergalactic space.

The intergalactic cloud is 10 megaparsecs from the earth (one megaparsec is 3.26 million light-years). It is at least 100 kiloparsecs across, or about three times the diameter of our galaxy. Its content of hydrogen atoms is at least $10^9$ times the mass of the sun. Optical images of the sky show no sign of stars between the two groups of galaxies, but then the density of the hydrogen in the intergalactic cloud is only from $10^{-3}$ to $10^{-4}$ atoms per cubic centimeter. The spontaneous formation of stars by the gravitational collapse of the matter in a cloud would require a density of $10^4$ or even $10^6$ atoms per cubic centimeter.

On the other hand, the western edge of the cloud is receding from the solar system (as part of the expansion of the universe) faster than the eastern edge. The difference in velocity, which is about 80 kilometers per second, can be taken as evidence that the cloud is rotating as it recedes, and the rotation in turn implies that the cloud is kept from flying apart by the gravitational effect of undetected matter amounting to as much as $10^{11}$ solar masses, or 100 times the mass of the detected hydrogen atoms. The cloud would then be equal in mass to a galaxy.

## *Picolife*

Exceedingly small plant life is proving to be surprisingly important in the ecology of the oceans. At two sites in the tropical Pacific a group from the Marine Ecology Laboratory of the Bedford Institute of Oceanography in Nova Scotia sampled the ocean water each day for six days and passed the samples through various filters. The intent was to trap picoplankton: organisms whose wet weight is on the order of a picogram, or $10^{-12}$ gram. W. K. W. Li and his colleagues defined picoplankton as cells that pass through a filter whose pores are one micrometer in diameter but do not pass through a filter with .2-micrometer pores.

The cells that met the criteria were found to contain as much as 90 percent of the chlorophyll in the water samples. Thus many such cells are plants. It had earlier been supposed that picoplankton must be predominantly heterotrophic bacteria (bacteria whose mode of nutrition is not photosynthetic). A technique in which chlorophyll is made to fluoresce red or orange under the microscope revealed as many as 1.5 million picoplanktonic plant cells in each milliliter of seawater in the uppermost 50 meters of the oceanic water column.

Evidence that the picoplanktonic plants contribute to the photosynthetic fixation of carbon at the base of the ocean's food chain came from further experiments, which Li and his colleagues describe in *Science*. In one experiment "more than 70 percent of the total measurable activity of the enzyme ribulose biphosphate carboxylase was found in the picoplankton." The enzyme is present in organisms that derive organic carbon photosynthetically from inorganic sources such as carbon dioxide. In another experiment cells that had passed through the one-micrometer filter were given inorganic carbon labeled with the radioactive isotope carbon 14. The rate at which radioactive carbon accumulated in the cells turned out to be substantially less in the dark than it was in the light.

The experiments "support a view of the structure of the pelagic ecosystem in which most of the activity (production and metabolism) is carried on by very small organisms." It is thought that the oceans are inhabited by forms of life even smaller than picoplankton, weighing about as much as a virus. Essentially nothing is known about them.

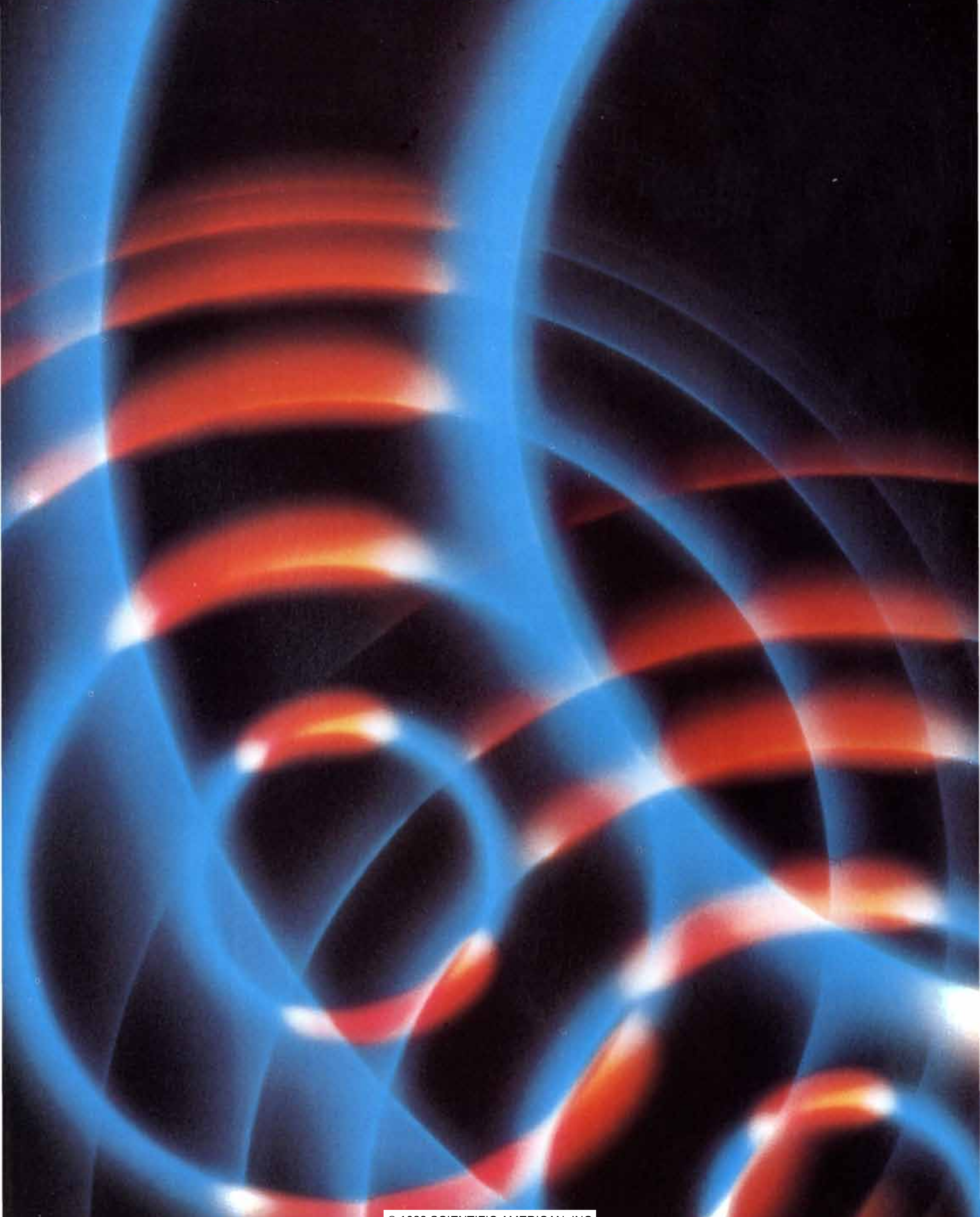
## *Spinning Still*

A standard industrial distillation column operates with no moving parts to separate two or more components of a liquid on the basis of their different boiling temperatures. The process is well understood and reliable, but the capital cost of the column and its associated equipment is high. Engineers at Imperial Chemical Industries in the U.K. are proposing a new distillation process in which the separating unit is a spinning drum. They think the technique may offer several advantages over the standard system, including lower capital costs.

In the conventional process for distilling a substance such as petroleum a stream of heated fluid is pumped into the lower part of a fractionating column as a mixture of liquid and vapor. The column, which may be as much as 150 feet high, contains perhaps 40 fractionating trays spaced regularly through its height. The vapors rise through the column and at the top are condensed by cooling. Part of the condensed liquid (representing the component with the lowest boiling temperature) is taken off. The rest goes back down through the fractionating trays, coming in contact with rising vapor at each level. At each stage the liquid absorbs heavier constituents from the vapor and loses lighter ones. When a steady state has been reached, the column is separating all the constituents, the ones with a lower boiling point at the top and the ones with a higher boiling point at the bottom.

In the new Higee (for high gravity) system the key unit is a flat, round drum filled with the kind of porous packing material installed in many distillation towers. The drum is spun at about 1,800 revolutions per minute, generating a centrifugal force of about 900 *g* (900 times the gravitational force at the surface of the earth). The heated liquid is pumped into the center of the drum and flows outward by centrifugal force. A countercurrent of vapor enters through holes in the rim and flows inward owing

# The Acoustic Detection



# The Acoustic Detection

*Scientists have studied microstructural discontinuities in high-carbon steel since the early 1920s. By monitoring acoustic emissions, a materials research engineer at the General Motors Research Laboratories has arrived at a more detailed understanding of how one type of discontinuity occurs.*

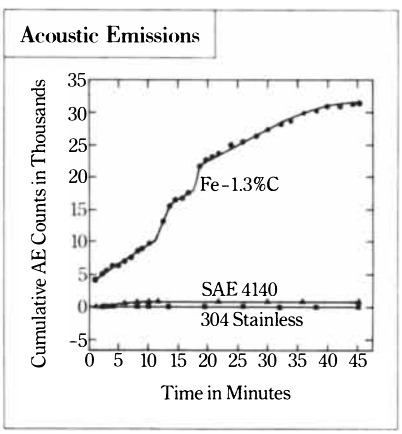



*Figure 1: Cumulative acoustic emission counts for Fe-1.3%C steel, and control specimens of SAE 4140 and 304 stainless steel.*

*Figure 2: Artist's rendering of two proposed sources of microcracking: (A) impingement of the plates during the formation of martensite and (B) carbon atom rearrangement during the aging of martensite.*

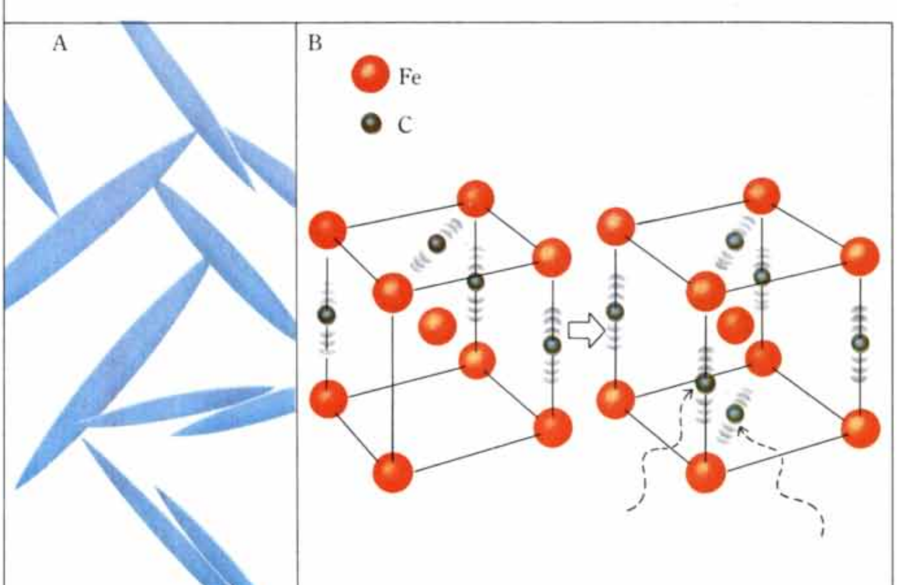
MARTENSITE is a hard microconstituent of steel which forms when austenite, iron containing carbon in solid solution, is quenched from a high temperature. The martensitic transformation produces steel that is hard and strong, but non-ductile. Through heat treatment, the steel can be tailored to applications requiring different degrees of ductility. High-carbon martensite—a highly stressed microstructure with a plate-like morphology—contains microscopic ruptures or separations 10 to 20 microns in length. These structural discontinuities, termed "microcracks," influence the mechanical properties of steel.

Although aspects of the microcracking phenomenon have been understood by metallurgists for more than fifty years, there is still no definitive explanation for when or how it occurs. An engineer at the General Motors Research Laboratories has devised an experiment that detects the microcracks as they occur.

The elastic energy released when microcracks form should produce a stress wave and associated high-frequency acoustic emission (AE). Using a piezoelectric transducer as the monitoring device, Dr. Michael Shea set out to determine what could be learned about the microcracking process by measuring AE.

The more widely accepted of two current hypotheses—the "impingement model"—asserts that microcracking is transformation-induced, taking place due to the collision of martensite plates during the quench. The other model maintains that microcracking occurs during the aging of martensite after the plates have already formed. The "aging model" suggests that thermal activation enables carbon atoms to rearrange themselves, producing localized stresses high enough to cause microcracking. Dr. Shea's ongoing research into high-carbon martensite led him to believe that the aging hypothesis was important. He proceeded to determine if AE is produced during aging.

For his study, Dr. Shea chose Fe-1.3%C steel, which undergoes martensitic transformation during quenching and is known to form



microcracks. To provide baseline data, control specimens of 304 stainless steel and SAE 4140 steel were put through the same procedures as the test composition. When quenched, 304 stainless steel produces no martensite, and SAE 4140 forms a low-carbon martensite which has a lath-type morphology, and generally does not microcrack.

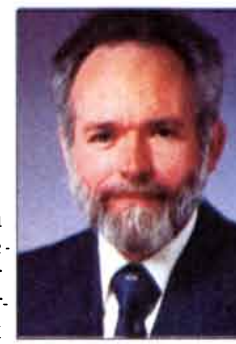
SPECIMENS of the three compositions were quenched to -196°C and then slowly heated to room temperature. Acoustic measurements were made beginning at 0°C, at which point carbon atom mobility is sufficient to allow rearrangement processes to take place, and continued for 45 minutes after the specimens had reached room temperature. No AE was recorded for 304 stainless steel, and only a slight amount for SAE 4140. Significant emission, however, was measured for the Fe-1.3%C steel specimen during the entire testing period (see Figure 1). Since martensite had already formed during the quench, these results support the hypothesis that microcracking is produced during aging of the freshly-formed plates. Dr. Shea ruled out both slip and twinning as sources of AE since the literature indicates that neither factor is significant during aging of martensite below 40°C. The possibility that the AE resulted from isothermal transformation of austenite to martensite could also be excluded because this process does not take place in the composition studied.

"These results demonstrate conclusively," says Dr. Shea, "that microcracking occurs during the aging of high-carbon martensite, thereby providing support for the less accepted of the two models.

"The next challenge," he continues, "will be to quantify the relative contributions of both models—impingement and aging—in an effort to determine which, in fact, is the more important mechanism, thus furthering our understanding of microcrack formation. Then, perhaps, we can more systematically explore ways to minimize microcracking."



## THE MAN BEHIND THE WORK



Dr. Michael Shea is a Staff Research Engineer in the Metallurgy Department at the General Motors Research Laboratories.

Dr. Shea received his undergraduate and graduate degrees in metallurgical engineering from Michigan Technological University, and his Ph.D. in materials engineering from Rensselaer Polytechnic Institute. His thesis concerned deformation and fracture of cesium chloride type superlattices. He joined General Motors in 1971.

The areas of metallurgical research pursued by Dr. Shea at General Motors include the mechanical properties of high-carbon steels, mechanically-induced transformation of austenite, and structure/property relationships in nodular cast iron. His exploration of the microcracking phenomenon in martensite was conducted with the help of instrumentation developed by GM colleague Dr. Douglas Harvey.

to a higher pressure at the outside. In the porous packing the liquid and the vapor flow past each other in thin layers, making for a rapid separation.

A demonstration unit set up by Imperial Chemical Industries in 1981 operates with two rotors, each rotor about 30 inches in diameter and six inches thick. It has successfully separated ethanol and isopropanol over extended periods of time. These liquids boil respectively at 78 and 82 degrees Celsius and would require a conventional distillation column about 100 feet high and three feet in diameter for a separation of comparable quality. The developers say their system can be cleaned, repaired or moved easily. What remains to be seen if it is scaled up to industrial size is whether the maintenance costs can be kept within reasonable bounds.

## *Sharing the Schlepping*

Why do car pools tend to fall apart? Four co-workers agree that in principle it makes sense to share the driving, the fuel and the space on the highway for the trip to work and back. On Monday, however, Jones cannot take her turn because she must be out of town. On Tuesday, Smith must drive to work because he needs his car at lunchtime, even though his turn does not come until Thursday. Jones trades with Brown, who is scheduled to drive on Wednesday, but Roberts, who should drive on Tuesday, cannot trade with Smith and drive on Thursday. Jones is delayed returning from her trip, and so she cannot drive on Wednesday either. Smith refuses to drive three days in a row, so that either Brown or Roberts must drive on Wednesday. By the following week no one can remember who is to drive when.

Ronald Fagin and John H. Williams of the San Jose Research Laboratory of the International Business Machines Corporation have now devised an ingenious system for determining who is to drive on any given day. The system allows the members of the car pool to miss a turn, to drive out of turn or to trade driving days with one another at will. More important, it sets up a simple bookkeeping method that fairly accounts for the contribution each person makes to the car pool. Since the system is fair and flexible and leaves no doubt about whose turn is next, it can be applied whenever the people in a group agree to take turns doing some regular task, which might well be cooking rather than driving.

Fagin and Williams describe their car-pooling algorithm in the *IBM Journal of Research and Development.* Suppose the maximum number of people who can ride together in one car is $N$, and $U$ is the least common multiple of all the numbers less than or equal to $N$. Thus if there are three people in the car pool, $U$ is equal to six; if there are four people, $U$ is 12. The bookkeeping system worked out by Fagin and Williams maintains a running score for each member of the car pool. Points are added to a member's score for driving and subtracted for riding; the score remains unchanged when a member misses a day or drives alone. If the number of participants on a given day is $K$ (which can be less than $N$), the driver on that day is given $U(K-1)/K$ points, and $U/K$ points are deducted from the running total of each rider. Hence if three out of four people ride together, the driver receives $12 \times 2/3$, or eight, points and the two riders each lose $12/3$, or four, points. Each day the participant with the fewest points is selected as the driver.

Fagin and Williams point out that simpler token systems, such as the system in which a driver always receives one ride token from each rider, are inherently unfair. A person whose turn to drive came up repeatedly on sparsely attended days would receive less credit than someone who drove when the car was full. A better system would require, say, Jones to drive a third of the time she shares a ride with two other people and half of the time she rides with only one other person. The new system is fair in an even more flexible sense. The total number of times Jones must drive is equal to the sum of half of her driving share for two-person car pools, a third of her share for three-person car pools and a fourth of her share for four-person car pools. Thus she need not drive half of the time she shares a ride with one other person because the system shows how she can make up the deficit in the larger car pools. Moreover, the car pool need not be confined to one car; each person's contribution to car pooling could be fairly credited, say, throughout an institution.

## *The Big Picture*

It was air travelers in the 1940's who called attention to the remarkable large drawings on the *pampa colorada* of Peru, a gravel-covered region of coastal red desert near Nazca some 300 kilometers south of Lima. Ever since, the who, when and how of the drawings' origin have fired the imagination of various commentators and tested the ingenuity of soberer scholars. Thanks largely to the patient field work of Maria Reiche, a German-born mathematician who lives in Peru, some of the wilder proposals have been laid to rest. These ranged from attributing the drawings to the work of extraterrestrials marking their desert spaceports to the claim by archaeoastronomers that the drawings provided sight lines for the observation of celestial phenomena.

Archaeological studies of the *pampa colorada* have identified both the who and the when of the Nazca figures. Pottery fragments remaining in the ancient work camps of the people who made the drawings and the carbon-isotope age of charcoal from the same workers' hearths unequivocally indicate that the vast desert displays were made by Nazca Indians sometime before A.D. 600.

The question of how is easily answered in part: the drawings were formed by the painstaking removal of gravel from the desert surface, exposing the lighter-colored underlying soil. The troublesome point has been how the drawings were laid out; many of them are too large to be seen in their entirety from ground level. Even the knowledgeable Reiche has suggested that the Nazca must have possessed surveying "instruments and equipment ... buried and hidden from the eyes of the [conquistadors] as the one treasure that was not to be surrendered."

Writing in *The Skeptical Inquirer,* Joe Nickell of the University of Kentucky reports his doubt that the Nazca needed much surveying skill or special instrumentation to construct the desert drawings. He goes on to recount his success in replicating a 440-foot-long portrait of a condor using only limited and simple equipment. Nickell and his colleagues prepared a scale drawing of the Nazca original (Reiche had long ago found that the Nazca figure-makers had made such small mock-ups for guidance), determined a center line and plotted the coordinates of various points on the figure. Any given number of units on the small drawing was then translated into the same number of larger units (about 12.7 inches long) on the full-scale figure. Taking advantage of a fresh landfill area in West Liberty, Ky., the survey group used a rope subdivided into 100 Nazca "feet" to locate various points on the condor figure. The only surveying instrument they required was a wood cross to ensure that measurements were made at a right angle to the center line.

The survey group staked out 165 points and connected them with twine to duplicate the main outlines of the condor. Where a circle was required (for example in the representation of the condor's head) it was scribed with a stake and rope; all other curves were done freehand. After completing the outline the group marked it by hand on the dark soil of the landfill with powdered lime and arranged to have the replica photographed from an altitude of 1,000 feet. The match with the Nazca original was remarkable in its exactness. The ground drawing may possibly have been the world's largest art reproduction (until rain washed away the lime). Summarizing his group's experience, Nickell emphasizes that the task required no materials not readily available to the Nazca.

# Peking Man

*In 230,000 years of cave dwelling Homo erectus pekinensis left an abundance of fossils and artifacts that now give information on biological, technological and perhaps even social evolution*

by Wu Rukang and Lin Shenglong

One of the sites with the longest history of habitation by man or his ancestors is a cave near the railroad station in Zhoukoudian, a town some 50 kilometers southwest of Beijing. Whereas modern cities are generally no more than a few thousand years old, the cave at Zhoukoudian was occupied almost continuously for more than 200,000 years. The multiple layers of fossil-bearing deposits indicate that early men first took shelter in the cave 460,000 years ago, and the last of them did not abandon the site until 230,000 years ago, when they were forced out by the filling of the cave with rubble and sediment. The species of mankind that lived there is classified as *Homo erectus pekinensis*, or Peking man. (The name became established before the customary English spelling of the city was changed to Beijing.)

The long record of habitation at Zhoukoudian offers an opportunity to trace the development of a single community over a period that spans a significant fraction of the evolution of the genus *Homo*. The period is long enough for progressive changes in the form of the fossils themselves to be discerned; one of the important changes in physical features is an increase in cranial capacity. Equally important, it is possible to reconstruct certain events in the cultural evolution of the species. There is evidence in the cave deposits that Peking man was able to control fire and that he employed it for cooking. The fossilized remains of animal bones indicate that the Zhoukoudian cave dwellers were effective hunters of both large and small game; fossilized seeds suggest another component of the diet. An abundance of flaked-stone implements provides information on toolmaking skills. From an analysis of the materials recovered from the cave it is even possible to speculate on the social organization of the community. For example, there is information to support conjectures about the sharing of food and the division of labor between the sexes.

In the past five years a comprehensive investigation of the Zhoukoudian site has been undertaken by more than 120 Chinese scientists, including us. The work is sponsored by the Institute of Vertebrate Paleontology and Paleoanthropology of the Chinese Academy of Sciences, and the contributors represent 17 universities and research institutions. Their fields of expertise range from early man and his artifacts to the study of ancient climate, caves and the terrains in which they form, ancient soils, pollen and the dating of archaeological materials. Here we shall present the salient findings of the group and attempt to sketch a portrait of Peking man in his native environment.

In the overall scheme of higher primate evolution Peking man and the other examples of *Homo erectus* are a comparatively late development. The first hominoid, the ancestor of both man and the anthropoid apes, had branched off from the other primates by about 35 million years ago. It walked quadrupedally and relied on seeds and fruits for its food. From 10 to eight million years ago the first hominid, the founder of a genetic lineage whose only living representative is modern man, diverged from the other hominoids such as the chimpanzee and the gorilla. The most primitive hominids apparently had a small cranium, with a capacity of perhaps 350 cubic centimeters; they could walk on their hind limbs and may have been able to carry objects while walking. It was not until about four million years ago that an advanced hominid, classified in the genus *Australopithecus*, appeared. The earliest remains of australopithecines have been found in the middle Awash River valley of Ethiopia. The skeleton was evidently capable of supporting a fully bipedal gait and the brain case had a capacity close to 500 cubic centimeters.

In the past 100 years many fossils of manlike skulls and skeletal fragments more advanced than those of *Australopithecus* have been found in Europe, Asia and Africa. For a time almost every discovery gave rise to a new taxonomic division, but virtually all these later specimens have now been subsumed in the single species *Homo erectus*. On the basis of chronological studies of the geologic context in which the fossils were found it has been ascertained that the species emerged at least 1.5 million years ago. *Homo erectus* could walk upright, as the species name suggests, but his skull was still very primitive, with a cranial capacity ranging from 850 to 1,100 cubic centimeters. He was able to make more advanced stone implements than the pebble tools of earlier hominids. Some later specimens of *Homo erectus*, including Peking man, developed a quite elaborate culture characterized by the inhabiting of caves and the hunting of game. They were also able to make use of fire to cook food, although it is not yet clear whether they were actually able to kindle a fire.

The subsequent course of development from *Homo erectus* to modern man cannot be traced in detail. Suffice it to say that about 200,000 years ago there appeared a form of man with a less heavily built face and a larger brain case than those of *Homo erectus*. The members of this species, early representatives of *Homo sapiens*, came to flourish in Europe, Asia and Africa somewhat less than 100,000 years ago. They were capable of building shelters in the open air and of starting a fire, and they made highly refined stone tools. Because of the closeness of their morphology to that of modern man and the progressiveness of their culture, they are considered to be among the direct ancestors of the modern *Homo sapiens*.

### The Cave of Zhoukoudian

The cave in Zhoukoudian where the remains of Peking man have been found is in a limestone formation called Dragon-Bone Hill. Before it was known as a site of early man it had long been a favorite haunt of dragon-bone collectors. Dragons had an important place in traditional Chinese culture, and their bones (actually the fossils of various mammals) were thought to have great medicinal value. Indeed, it was the dragon-bone collectors who found the first evidence of Peking man, although they did not recognize its significance.

At the beginning of the 20th century a number of paleontologists and anthropologists visited Zhoukoudian in search of evidence of primitive man. Notable among them were J. G. Anderson, a Swedish geologist, and Walter W. Granger, a paleontologist of the American Museum of Natural History. On a routine survey of Dragon-Bone Hill in 1921 Anderson, Granger and O. Zdansky, an Austrian paleontologist working as Anderson's assistant, were directed by local residents to a cave that was said to be "full of dragon bones." Nothing of interest was found, however, except some quartz fragments; because the fragments were far from their geologic context, Anderson considered them artifacts of early man.

Although survey excavations were undertaken in the cave, the progress of discovery was slow. In 1923 Zdansky found two anthropomorphic teeth. When a systematic excavation was started in 1927, Birger Böhlin, a Swedish paleontologist, discovered a well-preserved lower molar. Davidson Black, who was then professor of anatomy at Peking Union Medical College, assigned the fossil to the new species *Sinanthropus pekinensis.* From 1927 to 1937, when work was stopped by the Japanese invasion of China, the cave was continuously excavated. Pei Wenzhong, who was in charge of the Zhoukoudian excavation from 1928 through 1935, found the first complete cranium in the winter of 1929, thereby providing a sound basis for scientific studies of Peking man.

Since the founding of the People's Republic of China in 1949 the excavation of the cave has been resumed by the Institute of Vertebrate Paleontology and Paleoanthropology. Up to 1966 fossils of more than 40 males and females of various ages had been found associated with tens of thousands of stone artifacts and evidence of fire use. In addition the fossils of two species of plants (and the pollen of many more species) and 96 kinds of mammals were found. The abundance of the fossils and the thoroughness with which they have been studied make it possible to reconstruct elements of Peking man's 200,000-year history.

The cave, called formally Zhoukoudian Locality I, is a huge karst cave in a limestone structure formed in the Ordovician period, about half a billion years ago. Karst landscape, which takes its name from that of a region along the Dalmatian coast of Yugoslavia, is formed by the work of underground water on massive soluble limestones. The karst cave inhabited by Peking man is 140 meters long from east to west. The

**LIMESTONE CAVE was inhabited by Peking man, a variety of *Homo erectus,* beginning 460,000 years ago and continuing until 230,000 years ago. The cave was discovered in the early 1920's. Excavations over the past 50 years and more have revealed evidence of the use of fire, an assemblage of 100,000 stone artifacts, fossilized bones and teeth of more than 40 individuals and fossil remains of animal and plant species that were presumably part of Peking man's diet. The photograph shows the east entry to the cave, which was the primary entrance in the first phase of Peking man's occupation. The entry was abandoned after a massive collapse of some 300,000 years ago.**

widest part, which measures 40 meters, is at the eastern end; the narrowest part, at the western end, is only a little more than two meters wide.

The original cave, however, was not like this, nor was it habitable at the outset for Peking man. The formation of the cavern inside the hill began in the Pliocene epoch, about five million years ago, when water percolating along both horizontal and vertical cracks dissolved the limestone and carried it away in solution. The horizontal erosion created a long cave with both ends narrower than the central part; the vertical erosion followed precipitous fissures, creating funnel-shaped gullies whose bottom was deep, rough and bumpy. The original cave was a hollow in Dragon-Bone Hill without openings to the surface. Owing to the continual scouring of the hill by the ancient Zhoukou River, however, a spur on the eastern slope of the hill was cut off early in the Pleistocene epoch, exposing the eastern wall of the cave to water erosion. As a result the eastern slope developed a diagonal fissure approximately at the present-day location of the small east entry to the cave. The fissure became progressively wider and in the end established a connection between the Zhoukou River and the cave. Sand and dirt carried by the river were washed into the cave, gradually leveling the uneven bottom and creating a spacious, flat floor. It was at this point that Peking man took up residence.

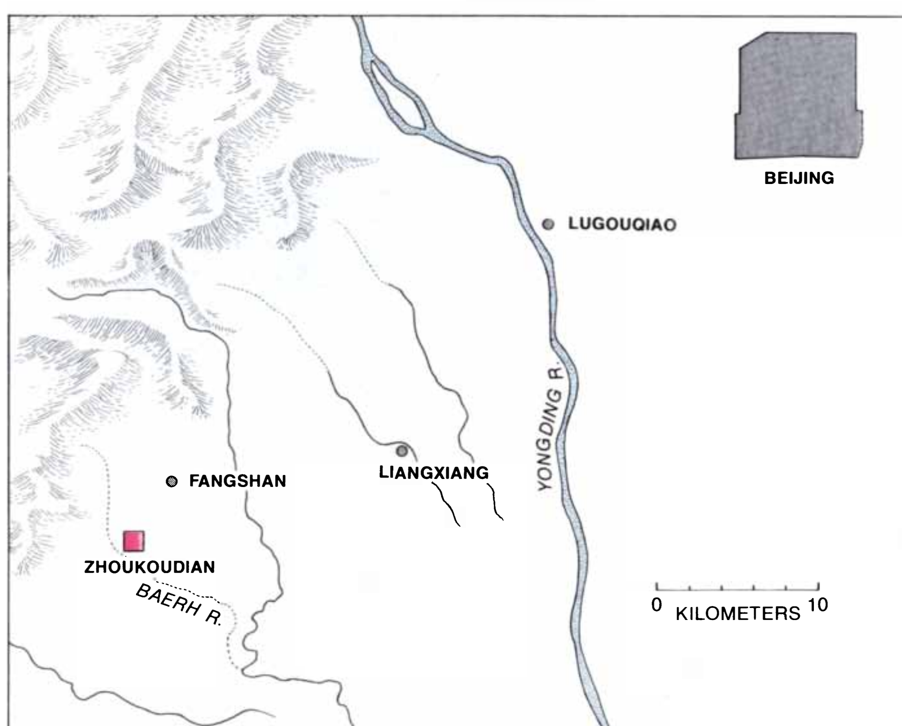
At first Peking man made his entrance from the east and lived mainly in the eastern section. About 350,000 years ago, however, there was a massive cave-in in the part of the eastern section now called the Dove Hall. An enormous amount of eroded rock fell from the roof, blocking the east entry completely. Peking man had to abandon what had been the most spacious part of the cave, but he was able to enter the cave again through a fissure in the central section. He thereafter lived in the western section until 230,000 years ago, when the cave was completely filled with the detritus of human occupation and with eroded limestone.

The cave remained filled with these deposits until its excavation. The deposits that included remains of Peking man had a thickness of about 40 meters, which was divided into 10 layers by Pierre Teilhard de Chardin and Yang Zhongjian, a Chinese paleontologist, in 1929. The Chinese archaeologist Jia Lanpo analyzed the deposits below the 10th layer in 1951 and defined three more layers (11 through 13). The various investigators have agreed to identify the deposits as those of the Zhoukoudian period. Although the period was generally recognized as belonging to the Middle Pleistocene, about half a million years ago, until recently the chronology was so sweeping and lacking in detail that it gave no idea about the evolution of Peking man himself.

Dozens of workers in five institutions have recently finished a series of chronological investigations and so for the first time have given a comparatively exact age for each layer. The 10th layer was dated by counting the tracks left in crystalline minerals by the disintegration of nuclei of uranium 238, the most abundant isotope of uranium. The rate of such fission events is constant, and so the number of tracks is a measure of the time that has elapsed since the mineral formed. It was concluded that the 10th layer was laid down 460,000 years ago.

Workers at the Institute of Geology of the Chinese Academy of Sciences have ascertained by the dating technique called the uranium-series method that the ninth and eighth layers were deposited 420,000 years ago. In the uranium-series method what is measured are the relative abundances of uranium 234 and the product of its radioactive decay, thorium 230. The decay proceeds at a constant, known rate; the ratio of the two elements therefore indicates the time since the uranium-bearing mineral was formed.

Still another method was applied to the dating of the seventh layer, which turns out to be between 370,000 and 400,000 years old. The age was established by examining the natural remanent magnetism of minerals in the layer. When a magnetic mineral crystallizes, its magnetic axis becomes aligned with the direction of the earth's mag-


**SITE OF THE CAVE is the small town of Zhoukoudian, some 50 kilometers southwest of Beijing (*right*). It is administratively subordinate to Fangshan, a suburban county of Beijing about 150 kilometers from the seacoast (*left*). In the northwestern part of the town is Dragon-Bone Hill, named for the mammalian fossils collected there by local residents, who thought they were dragon bones of great medicinal value. Zhoukoudian Locality I, which is one of four paleoanthropologic sites on the hill, is the cave that was inhabited by Peking man.**

netic field. Periodic reversals of the earth's field thus leave a record in the rocks, which identifies their time of deposition. The uranium-series method has also been applied to the deposits in the three topmost layers. The results indicate an absolute age of 230,000 years.

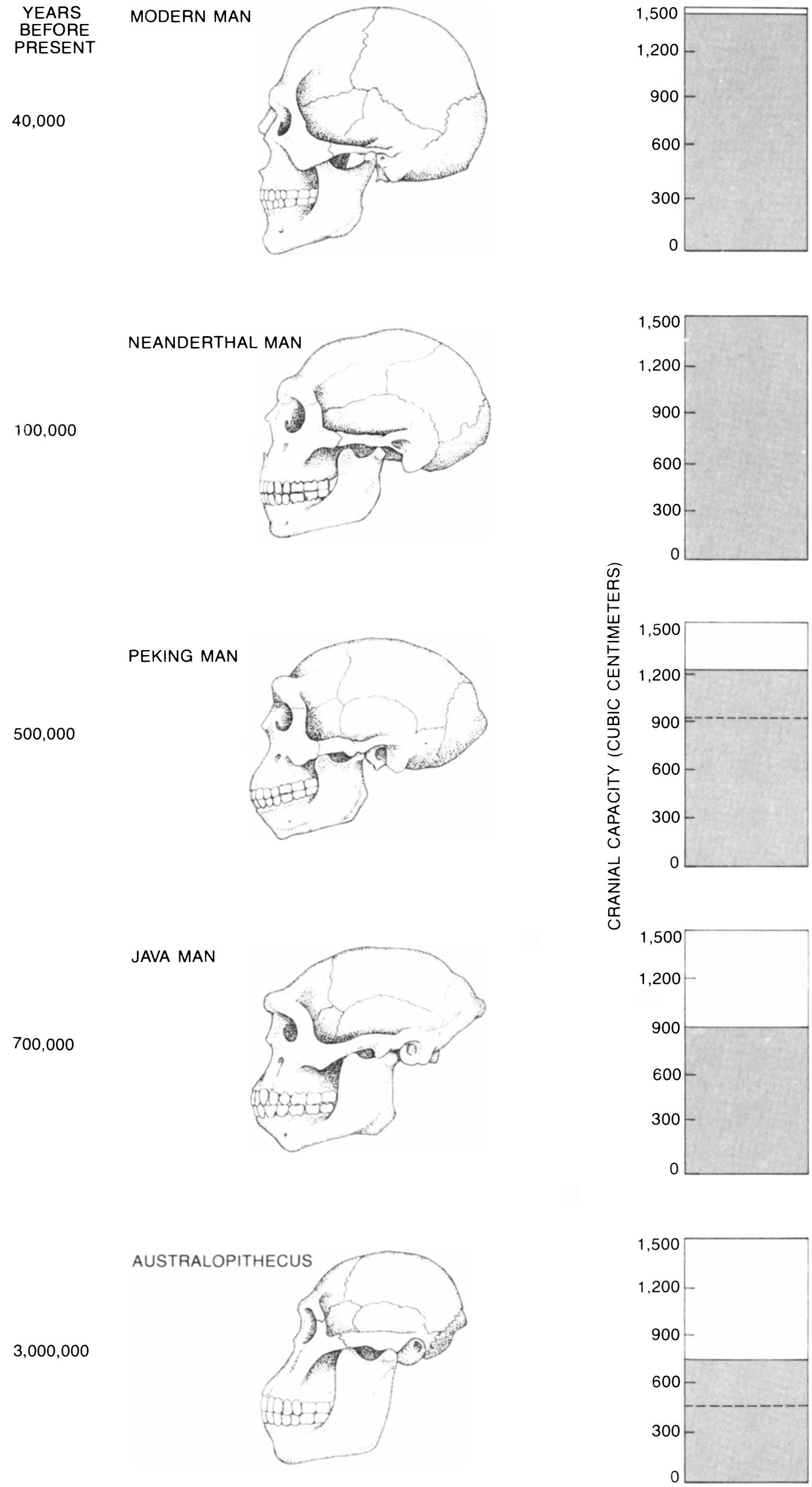
### The Fossils of Peking Man

The fossils of Peking man discovered over the past 50 years have been catalogued as follows: six complete or almost complete skulls and 12 other skull fragments, 15 pieces of mandibles, 157 teeth, three fragments of humerus (the bone of the upper arm), one clavicle (collarbone), seven fragments of femur (the thighbone), one fragment of tibia (shinbone) and one lunate bone (a wristbone shaped like a half moon). They were found at 15 widely scattered places in the cave. The fossils represent material from more than 40 male and female individuals, who died at various ages. From the fossil material it is possible to reconstruct the physical form and appearance of Peking man.

Like other examples of *Homo erectus* that emerged in the Middle Pleistocene, Peking man had a skeleton much like that of modern man. The main difference is that the limb bones had a thicker wall and a smaller marrow cavity. The skull of Peking man, however, differed substantially from the modern form. It was much thicker and flatter and had protruding brows and a marked angle at the rear. The brain case was larger than that of *Homo habilis,* a species that flourished 1.8 million years ago, and even larger than that of Java man, the somewhat older form of *Homo erectus;* the cranial capacity was still much smaller than that of modern man. The teeth were larger and more robust than those of *Homo sapiens* and had traces of an enamel collar, called a cingulum, around the crown. The cingulum is a primitive trait in prehistoric man.

All fossils of Peking man found so far have been analyzed by one of us (Wu Rukang) and Dong Xingren of the Institute of Vertebrate Paleontology and Paleoanthropology. We have given particular attention to the question of Peking man's cranial capacity. It should be stressed that brain size cannot be taken unconditionally as an indicator of human intelligence; a hominid with a bigger brain case is not necessarily smarter. Nevertheless, modern man, with an average cranial capacity of 1,450 cubic centimeters, must have evolved from an earlier form of man with a smaller cranium. The tendency is therefore for the human cranial capacity to increase, and the capacity of a given species or subspecies may provide some information about its place in human evolution.

It is generally accepted that the cranial capacity of man and his ancestors was

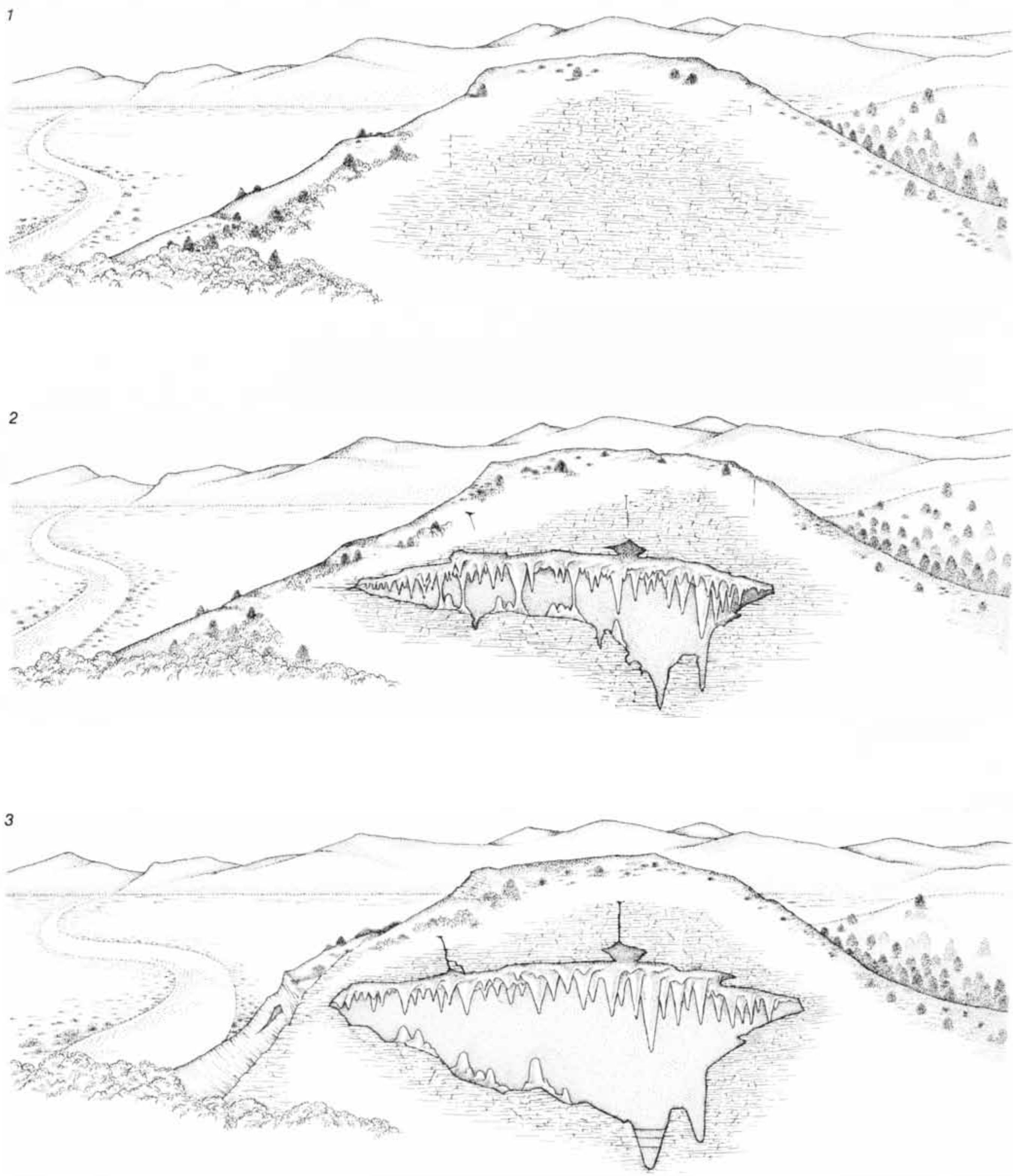



**HUMAN EVOLUTION can be traced through morphological changes in the skull (*left*) and through an increase in cranial capacity (*right*). The early hominid *Australopithecus,* which appeared some four million years ago, had a flat skull with protruding jawbones and a small brain case (between 450 and 750 cubic centimeters). About 1.5 million years later came *Homo erectus.* Two examples of *Homo erectus,* Java man and Peking man, were comparatively advanced in their facial structure and cranial capacity, which ranged from 850 cubic centimeters to more than 1,000 cubic centimeters. They differed in their age, skull shape and brain size. Not until some 100,000 years ago did Neanderthal man, the first member of the species *Homo sapiens,* come into existence; by then the brain case was much larger and the jawbones were less protuberant. Modern man, the subspecies *Homo sapiens sapiens,* emerged about 40,000 years ago.**

stable at from 500 to 800 cubic centimeters in the interval from three to 1.5 million years ago. This is the cranial size measured in fossils of *Australopithecus* and *Homo habilis.* Java man, the form of *Homo erectus* that appeared about a million years ago, had a cranial capacity of some 900 cubic centimeters. When Peking man emerged, the average capacity had reached 1,054 cubic centimeters.

It should be pointed out that this brain-case value for Peking man is an average based on the six relatively complete skulls found at Zhoukoudian. One of the six is the skull of a child who died at the age of eight or nine. A juvenile skull is obviously smaller than that of an adult; if it were not taken into account, the average cranial capacity of Peking man would be 1,088 cubic centimeters.

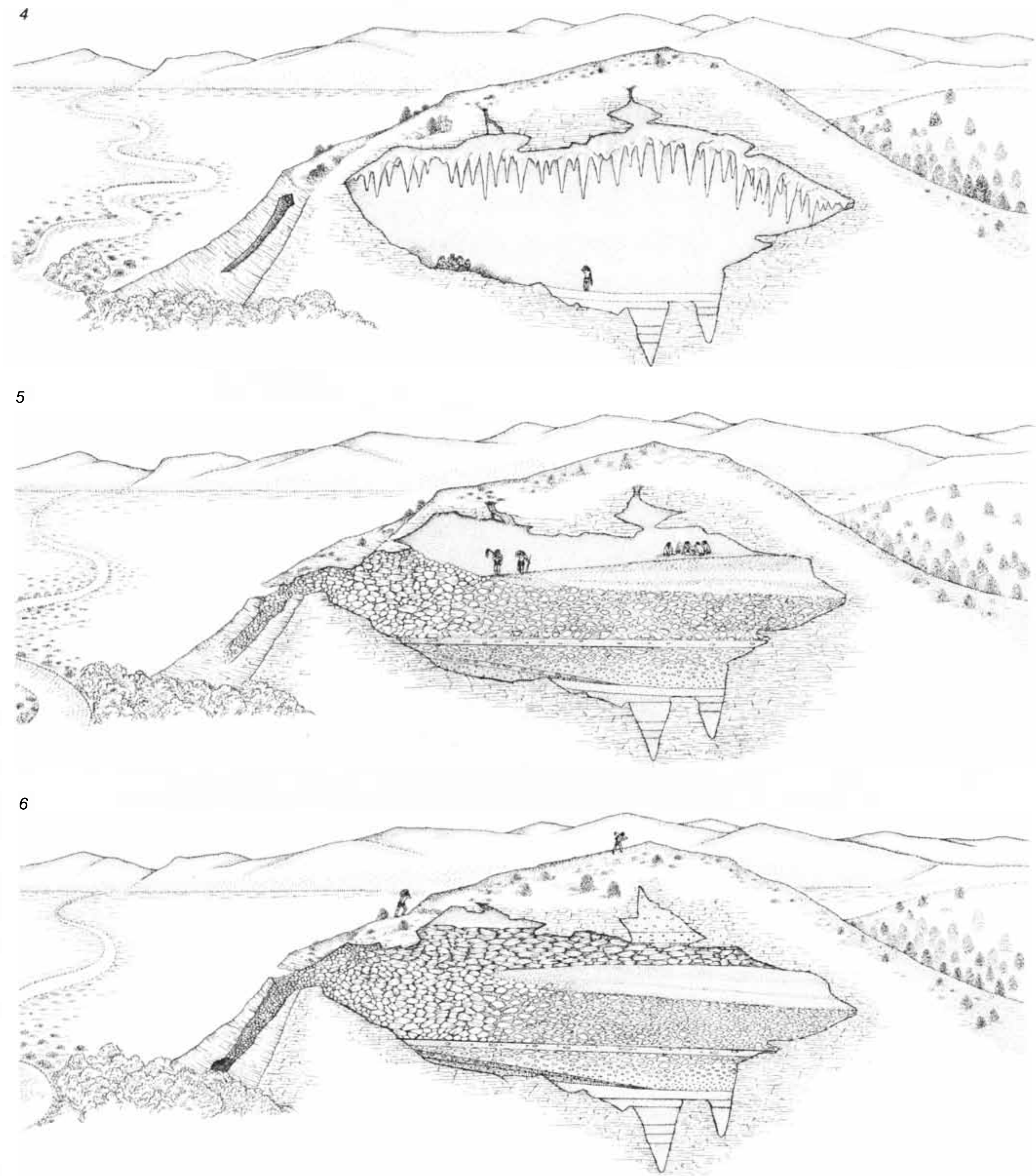
Because of the long time span covered




**EVOLUTION OF THE CAVE at Zhoukoudian Locality I is schematized here in six stages. Dragon-Bone Hill, in which the cave was created, is a limestone structure formed about 450 million years ago (*1*). As limestone was dissolved by underground water the cave began to form inside the hill about five million years ago (*2*). When the eastern spur of the hill was cut off by the scouring of the ancient Zhoukou River some three million years ago, a small entry to the cave appeared on the eastern slope and became progressively larger (*3*). Sand and**

by the fossils of Peking man, it may even be possible to trace changes in morphology, and in particular an increase in cranial capacity, over the history of the habitation of the site. The average capacity of four skulls found in the eighth and ninth layers (which are dated to more than 400,000 years ago) is 1,075 cubic centimeters. The one skull unearthed in the third layer (230,000 years old) has a capacity of 1,140 cubic centimeters. It seems that Peking man became more manlike in anatomy after 200,000 years of cave dwelling.

It is acknowledged in the anthropological community that the morphological evolution of early man was much slower and so less obvious than the transformation of behavior and life-style. Toolmaking techniques and the tools themselves are therefore an important




**dirt were carried by the river into the cave, leveling the gullies created by the erosion of the limestone. The resulting spacious shelter was adopted by Peking man about 460,000 years ago (*4*). The east entry and the eastern part of the cave were the main areas of habitation until the collapse of the roof 300,000 years ago forced a move to the western part, with access through a fissure in the central section (*5*). By the time Peking man left the cave 230,000 years ago it had been filled with fallen rock and the detritus of human occupation (*6*).**

measure of human evolution. In the case of Peking man a systematic study of the lithic culture was done three years ago by Pei Wenzhong and Zhang Shenshui of the Institute of Vertebrate Paleontology and Paleoanthropology. They came to believe Peking man's evolution is more clearly mirrored in his tools and toolmaking behavior than in his fossil remains.

### The Tools of Peking Man

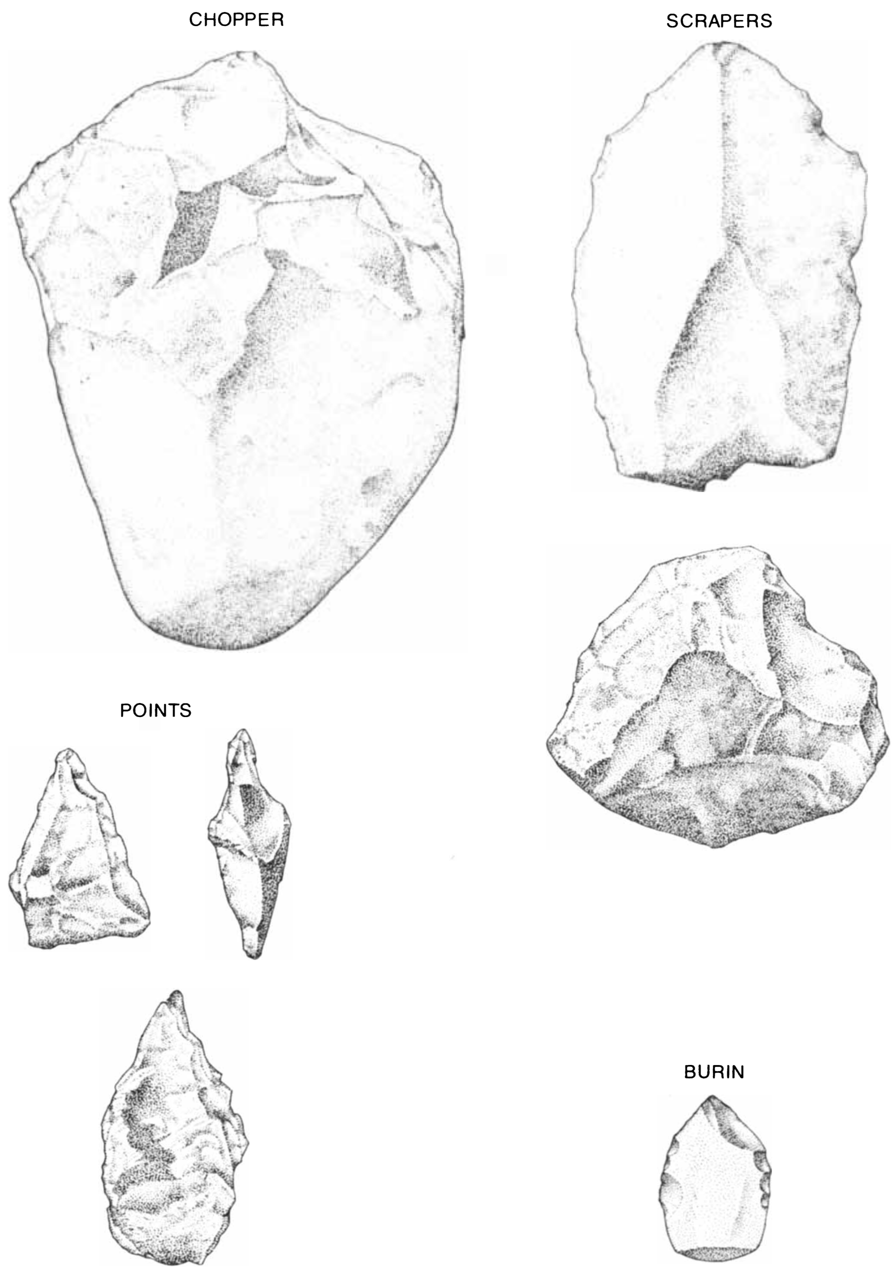
The stone industry of Peking man was an advanced one both in the selection of materials exploited and in the toolmaking techniques applied to them. The stone artifacts produced by Peking man are primarily made of vein quartz, rock crystal, flint and sandstone, suggesting that he did not rely exclusively on water-rounded pebbles for tool material. He often had to walk some distance to get vein quartz and rock crystal where they had been exposed by the weathering of granite formations.

Most of the tools are flakes of various sizes made in one of three ways. In the technique called anvil percussion a large flat stone (the anvil) was placed on the ground and forcefully struck with a piece of sandstone. Flakes chipped off the sandstone were gathered up, and those with a suitable shape and edge were selected either for direct use or for later retouching. In another method, called direct percussion, the core flint was held in one hand and flakes were detached from it by striking it with a hammerstone held in the other hand. The third flaking method is called bipolar percussion. Again a large, flat anvil was set on the ground. A piece of vein quartz was then held upright on the anvil with one hand and crushed by vertical blows with a hammerstone grasped in the other. In this way flakes could be chipped off both ends; they are called bipolar flakes.

Although some of the flakes recovered at Zhoukoudian were used as they were, it is evident that some of them had been trimmed into more specialized tools such as scrapers, points, choppers, burins and awls. The retouching was done mainly by the direct-percussion process. Most of the tools were trimmed on their back surface rather than on the one that had been worked in the initial flaking operation. Sometimes both faces were retouched.

The evolution of Peking man's lithic culture was divided into three periods by Pei and Zhang. The division was made according to changes in tool size, in tool material and in toolmaking technique. The stone tools found in the eighth through the 11th layers represent the earliest culture, which prevailed between 460,000 and 420,000 years ago. It is characterized by large tools weighing more than 50 grams and longer than 60 millimeters and by the indiscriminate application of the three toolmaking processes. At this stage tools made out of softer materials such as sandstone account for 15 to 20 percent of the artifacts. In the middle stage, from 370,000 to 350,000 years ago, the anvil-percussion process was practically abandoned and the bipolar-percussion process became the main method of flaking stone. As a result the proportion of the tools having a weight of less than 20 grams and a length of less than 40 millimeters increased to 68 percent. In contrast large tools diminished to 12 percent.

The last stage lasted from 300,000 to 230,000 years ago. It was clearly the most advanced: the tools had become smaller and the tool materials were of better quality. Among the stone tools excavated from layers one through five the small tools increased dramatically to 78 percent of the total, whereas the large tools decreased further to 5 percent. Although the tools of the third period were still made mainly out of quartz, fewer were made out of the coarse varieties of the stone, such as vein quartz, and the fraction of the tools made out of flint increased to as much as 30 percent in the uppermost layers. In the meantime sandstone tools




**STONE TOOLS made by Peking man are of four main types: the chopper (*upper left*), the scraper (*upper right*), the point (*lower left*) and the burin (*lower right*). They were formed out of vein quartz, rock crystal, flint and sandstone. Some of the flakes were put into service as tools without further trimming; others were retouched into more specialized tools such as points and burins. The stone industry was dominated by small tools with a length of less than 40 millimeters and a weight of less than 20 grams, although most of the early tools are larger.**

**USE OF FIRE** by Peking man, attested to by layers of ashes, can be traced to the beginning of his residence in the cave. Four layers of ashes have been identified in the cave deposits; the thickest layer is some six meters deep and the thinnest one is a little more than a meter deep.

had diminished to less than 1 percent of the total.

Another indication of the advanced state of Peking man's culture is his use of fire: evidently the art had already been mastered at the very beginning of his habitation of the cave. There are four large, thick layers of ashes deposited in periods ranging in age from 460,000 to 230,000 years. The thickest one, formed between 310,000 and 290,000 years ago, is six meters deep in certain places. Some of the ash deposits are in scattered piles, suggesting that Peking man had the ability to control fire and keep it burning for a long time.

How did Peking man make a fire and control it? A conclusive answer cannot be given, but an inference can be drawn. Peking man seems to have been too primitive to start a fire; instead he was probably dependent on natural fires outside the cave, generally started by lightning. Presumably the fire was captured by lighting a bunch of twigs or some other kindling and bringing it back to the cave. Given the rarity of natural fires, it must have been vitally important to avoid letting the fire go out. One way of maintaining a fire would have been to add wood to it continuously. Another would have been to cover burning charcoal with ash or soil to slow its burning. A new fire could then be made by blowing on the coals. Some charcoal found in the cave may be a remnant of the latter process.

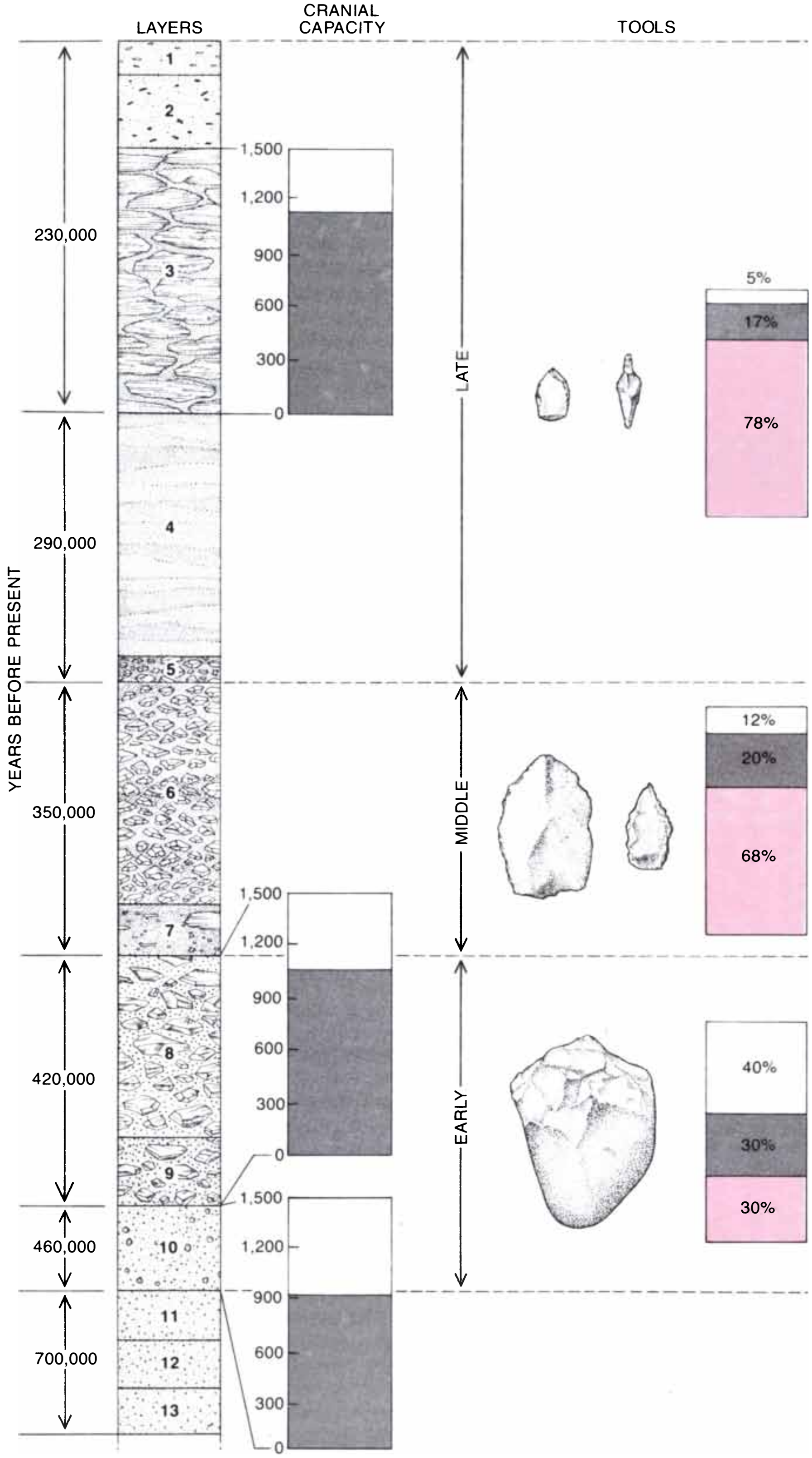
What was the nature of the environment outside the cave at the time Peking man occupied the site? On the basis of a study of pollen at Zhoukoudian Locality I, Kong Zhaochen of the Institute of Botany of the Chinese Academy of Sciences and his colleagues have suggested that Peking man lived in the climate of an interglacial period; indeed, the climate of the period was not much different from that of northern China today. The vegetation consisted of temperate deciduous forests and steppes on the plains and in the valleys and coniferous forests on the mountains.

A temperate climate entails the rigors of a cold winter every year but also provides many kinds of edible plant products. The adoption of a cave as shelter and the use of fire may have been inspired directly by the need for surviving cold weather.

A variety of plants supplied not only firewood but also fruits and seeds. Among the deposits in the cave are quantities of charred seeds of the Chinese hackberry, which were obviously gathered and roasted by Peking man. Evidently the seeds served as food. The analysis of pollen in the deposits gives evidence of other plants, such as walnut, hazelnut, pine, elm and rambler rose. The fruits and seeds of those species must also be considered candidates for inclusion in Peking man's diet.

Hunting was a valuable adaptation to the environment since meat could supply more calories and protein than a vegetarian diet. Peking man was evidently able to compete successfully with large carnivores as a hunter. An abundance of fossil bones of mammals of various sizes found in the cave indicates that Peking man not only hunted small game but also was capable of killing large animals.

Among carnivores a natural relation




**EVOLUTION OF PEKING MAN in the course of 230,000 years of cave dwelling is suggested by an increase in brain size (*left*) and the development of a more refined lithic culture (*right*). The cave deposits that include fossils of Peking man are divided into 13 layers. Skulls have been recovered from layer 10, from the boundary between layer 9 and layer 8 and from layer 3. The measured cranial capacities are 915 cubic centimeters for the earliest skull, an average of 1,075 cubic centimeters for four later skulls and 1,140 cubic centimeters for the most recent one. It seems the brain size increased by more than 100 cubic centimeters during the occupation of the cave. The cultural development of Peking man is broken down into three stages. In the earliest stage the artifacts were mainly choppers and scrapers; 40 percent were large tools (*white*), 30 percent were of intermediate size (*gray*) and 30 percent were small (*color*). In the latest stage more complex tools appeared. The proportion of large and medium tools decreased respectively to 5 and 17 percent, whereas the proportion of small ones increased to 78 percent.**

generally exists between the size of the predator and that of its prey. A fox cannot kill a zebra, but the zebra is the favorite food of lions. Owing to his use of weapons, Peking man escaped the limitations of his size. In particular he became an efficient hunter of deer, some of which were larger and faster than he was. The large quantity of fossils belonging to at least 3,000 individual deer of two species, the thick-jawed deer (*Megaceros pachyosteus*) and the sika deer (*Pseudaxis grayi*), could be regarded as an indication that deer were the commonest prey of Peking man.

## Social Adaptations

As we have shown, Peking man was a cave dweller, a fire user, a deer hunter, a seed gatherer and a maker of specialized tools. The fossil and artifactual evidence attest to his biological and technological adaptation. It is more difficult to trace his social adaptation since little evidence is available. Nevertheless, three hypotheses can be put forward on the basis of our studies.

Generally speaking, gathering is a simple labor that can be undertaken by single individuals. Hunting, in contrast, and particularly the hunting of large animals, is so complicated, difficult and hazardous that the cooperation of numerous individuals is needed. It can be inferred, therefore, that Peking man was more likely to have been living in a group than in solitude when he began to hunt deer. Furthermore, thousands of fossils of prey species found in the cave suggest that these primitive hunters may have preferred to bring prey back to the cave and share the meat with others in the community rather than to consume it where it was killed.

Second, the hunting of large and fast-moving animals such as deer may have been difficult for women because of physiological limitations (such as pregnancy and child rearing). It is thus possible to speculate that the hunting behavior of Peking man may have caused or contributed to the sexual division of labor within the group. The pattern of male hunters and female gatherers, which is common in hunting-and-gathering societies today, may have already been established.

Third, the existence of a consistently progressing lithic culture throughout a period lasting for 200,000 years or more suggests that the earliest practice of education may have taken place in Peking man's cave. It is out of the question that each generation could recapitulate the entire history of stone-tool development, from striking a pebble in order to make a crude tool to trimming a flake into a specialized scraper. Toolmaking techniques, like modern science and technology, must have been conveyed from the old to the young.

# Dark Matter in Spiral Galaxies

*It appears that much of the matter in spiral galaxies emits no light. Moreover, it is not concentrated near the center of the galaxies*

by Vera C. Rubin

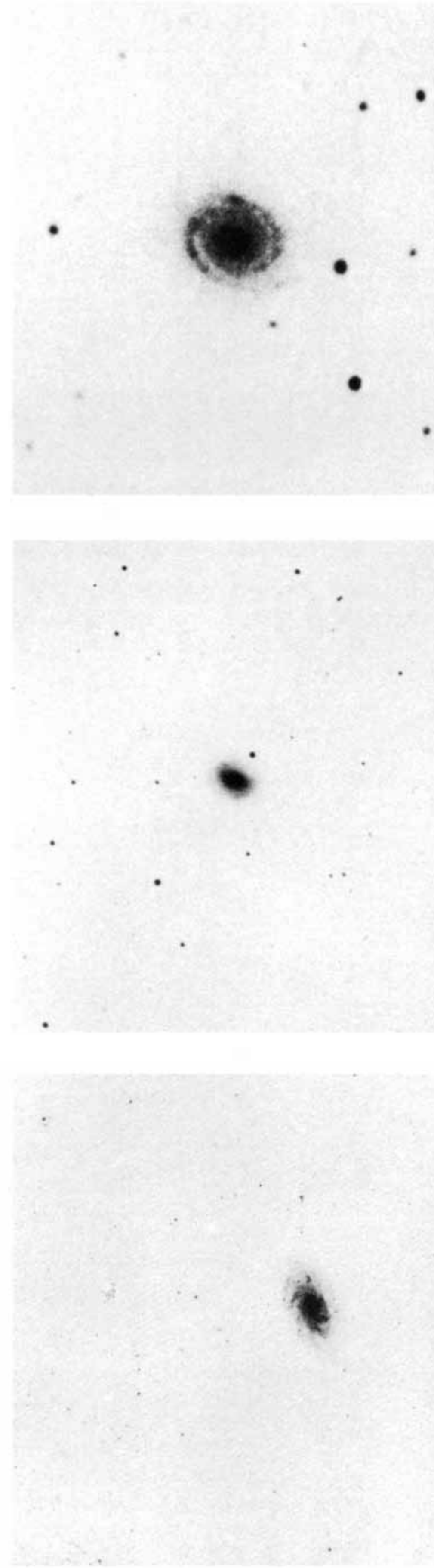
After evidence was obtained (in the 1920's) that the universe is expanding it became reasonable to ask: Will the universe continue to expand indefinitely or is there enough mass in it for the mutual attraction of its constituents to retard the expansion and finally bring it to a halt? Most cosmologists agree that the universe started in a big bang 10 to 20 billion years ago from an infinitely small and dense state and that it has been expanding ever since. It can be calculated that the critical density of matter needed to brake the expansion and "close" the universe is on the order of $5 \times 10^{-30}$ gram per cubic centimeter, which is equal to about three hydrogen atoms per cubic meter. The amount of luminous matter in the form of galaxies, however, comes to only about $7.5 \times 10^{-32}$ gram per cubic centimeter. Therefore if the expansion of the universe is to stop, the density of the invisible matter must exceed the density of the luminous matter by a factor of roughly 70.

With this factor in mind astronomers over the past half century have sought to determine the mass of the galaxies that populate the universe out to the limits of observation. From the luminosity of typical galaxies one can estimate that they have a mass ranging from a few billion to a few trillion times the mass of the sun. The actual stellar population of a galaxy is of course highly diverse. Some stars are 10,000 times more luminous than the sun per unit of mass; others are only a small fraction as luminous. Given this diversity one would like to know: Is the distribution of luminosity in galaxies a reliable indicator of the distribution of mass? And, by extrapolation, is the distribution of luminosity in galaxies a reliable indicator of the distribution of mass in the universe?

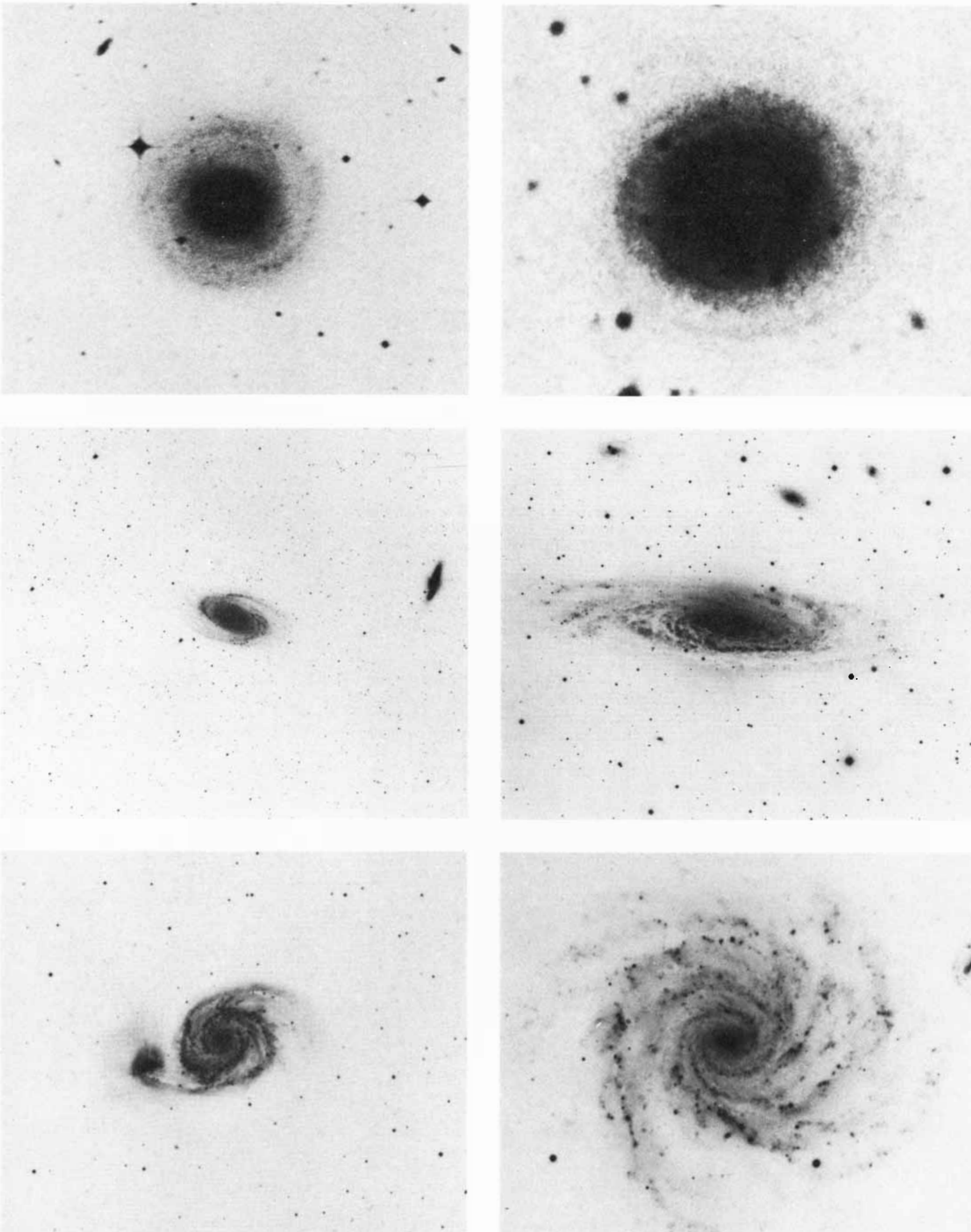
My colleagues and I in the Department of Terrestrial Magnetism of the Carnegie Institution of Washington have sought to answer these questions by measuring the rotational velocity of selected galaxies at various distances from their center of rotation. It has been known for a long time that outside the bright nucleus of a typical spiral galaxy the luminosity of the galaxy falls off rapidly with distance from the center. If luminosity were a true indicator of mass, most of the mass would be concentrated toward the center. Outside the nucleus the rotational velocity would fall off inversely as the square root of the distance, in conformity with Kepler's law for the orbital velocity of bodies in the solar system. Instead it has been found that the rotational velocity of spiral galaxies in a diverse sample either remains constant with increasing distance from the center or rises slightly out as far as it is possible to make measurements. This unexpected result indicates that the falloff in luminous mass with distance from the center is balanced by an increase in nonluminous mass.

Our results, taken together with those of many other workers who have attacked the mass question in other ways, now makes it possible to say with some confidence that the distribution of light is not a valid indicator of the distribution of mass either in galaxies or in the universe at large. As much as 90 percent of the mass of the universe is evidently not radiating at any wavelength with enough intensity to be detected on the earth. Originally astronomers described the nonluminous component as "missing matter." Today they recognize that it is not missing; it is just not visible. Such dark matter could be in the form of extremely dim stars of low mass, of large planets like Jupiter or of black holes, either small or massive. Other candidates include neutrinos (if indeed they have mass, as recent work suggests) or such hypothetical particles as magnetic monopoles or gravitinos.

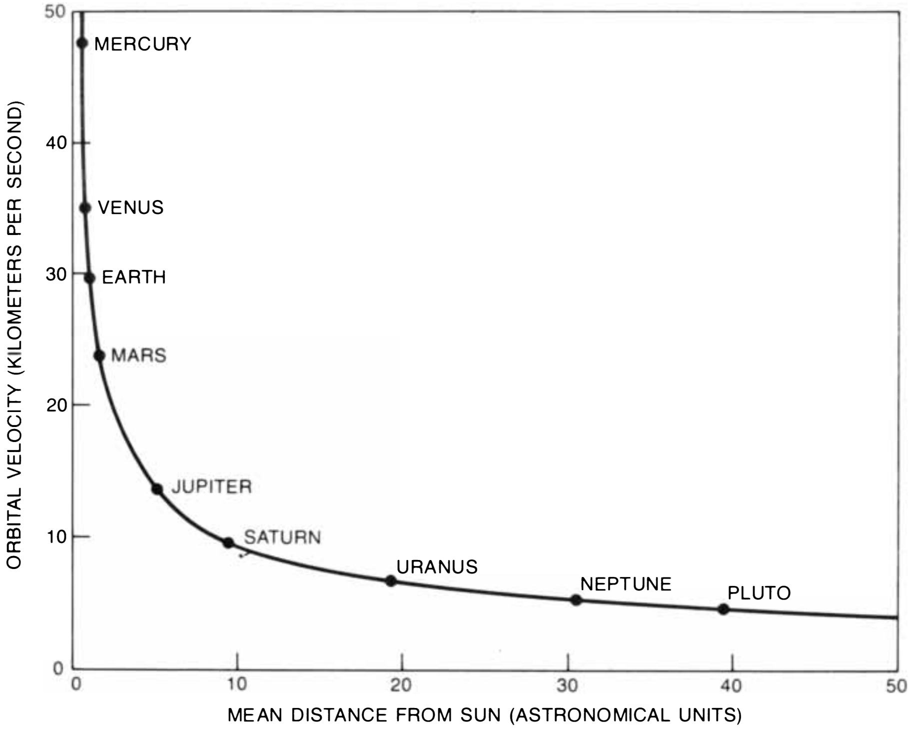
Early in this century it was reasona-



**SPIRAL GALAXIES whose unseen mass has been investigated by the author fall into three principal classes: Sa, Sb and Sc. Within each**



class the galaxies vary enormously in size and luminosity. Here nine examples are reproduced in negative images in which 1.2 centimeters equals 10 kiloparsecs (32,600 light-years). Three Sa galaxies are at the top, three Sb galaxies are in the middle and three Sc galaxies are at the bottom. In the progression from Sa to Sc the nucleus gets smaller with respect to the disk and spiral structure gets more pronounced.




**KEPLER'S LAW for the orbital velocity of planets in the solar system, in which more than 99 percent of the total mass resides in the sun, yields this plotted curve. Orbital velocity decreases inversely as the square root of *r*, the planet's mean distance from the sun. The distance is shown here in astronomical units; one A.U. equals the mean distance between the earth and the sun. Pluto, at 39.5 A.U., lies 100 times farther from the sun than Mercury, at .39 A.U. Mercury's orbital velocity is about 47.9 kilometers per second; Pluto's velocity is accordingly slower by a factor of 10, or 4.7 kilometers per second ($47.9 \times 1/\sqrt{100}$). The author's results show that the orbital velocities of stars in a spiral galaxy depart strongly from a Keplerian distribution.**

ble for astronomers to assume that the distribution of luminous matter, wherever it was found, coincided with the distribution of mass. Nearly 50 years ago, however, Sinclair Smith and Fritz Zwicky of the California Institute of Technology discovered that in some large clusters of galaxies the individual members are moving so rapidly that their mutual gravitational attraction is insufficient to keep the clusters from flying apart. Either such clusters should be dissolving or there must be enough dark matter present to hold them together. Almost all the evidence suggests that clusters of galaxies are stable configurations. Hence the early observations of Smith and Zwicky marshaled the first evidence that such clusters harbor matter both luminous and nonluminous.

Recent work by many other astronomers has strengthened this conclusion. Studies of the dynamics of individual galaxies, including our own galaxy, of pairs of galaxies, of groups of galaxies and of clusters of galaxies all point to a component of unobservable but ubiquitous mass. Such studies detect the presence of nonluminous mass solely by its gravitational effects.

For the past several years W. Kent Ford, Jr., Norbert Thonnard, David Burstein and I have sought to learn about the distribution of mass in the universe by investigating the distribution of matter within galaxies with a structure similar to that of our own galaxy, namely the general class of spiral galaxies. We have adopted this approach because spiral galaxies have a geometry favorable for the identification of mass, whether it is luminous or nonluminous, and modern large telescopes equipped with image-tube spectrographs make it possible to complete an observation of a single galaxy with an exposure of about three hours. Before I describe our observations it will be helpful if I review how celestial objects respond to the gravitational force acting on them and how that response can reveal the large-scale distribution of matter.

Toward the end of the 17th century Robert Hooke suspected that the planets were subject to a gravitational force from the sun whose intensity decreased inversely as the square of the distance. Isaac Newton then recognized that all pairs of objects in the universe have a gravitational attraction for each other that is proportional to the product of their masses and inversely proportional to the square of the distance between them. In other words, if the distance between the objects is increased by, say, a factor of two, their mutual attraction decreases by a factor of four.

For planets in orbit around the sun, which embodies essentially all the mass in the solar system, the decrease in gravitational attraction with distance is exactly paralleled by a decrease in the velocity needed to hold the planet in its orbit. Therefore Mercury, lying at .39 astronomical unit from the sun (that is, .39 of the mean distance between the sun and the earth), has an orbital velocity of about 47.9 kilometers per second. Pluto, 100 times farther away at a mean distance of 39.5 astronomical units, has an orbital velocity only a tenth that of Mercury, or 4.7 kilometers per second. Spiral galaxies rotate because they retain the angular momentum and the orbital momentum of the initial clumps of gas from which they formed.

In a spiral galaxy the gas, dust and stars in the disk of the galaxy (together with any associated planets and their satellites) are all in orbit around a common center. Like the planets in the solar system, the gas and stars move in response to the combined gravitational attraction of all the other mass. If the galaxy is visualized as a spheroid, the gravitational attraction due to the mass $M_r$ lying between the center and an object of mass $m$ in an equatorial orbit at a distance $r$ from the center is given by Newton's law $GmM_r/r^2$, where $G$ is the constant of gravitation. If the galaxy is neither contracting nor expanding, the gravitational force is exactly equal to the centrifugal force on the mass at distance $r$: $GmM_r/r^2 = mV_r^2/r$, where $V_r$ is the orbital velocity.

When this equation is solved for $V_r$, the value of $m$ drops out and the velocity of a body at distance $r$ from the center is determined only by the mass $M_r$ inward from its position. If, as in the solar system, virtually all the mass is near the center, then the velocities outward from the center decrease as $1/r^2$. Such a decrease in orbital velocity is called Keplerian after Johannes Kepler, who first stated the laws of planetary motion.

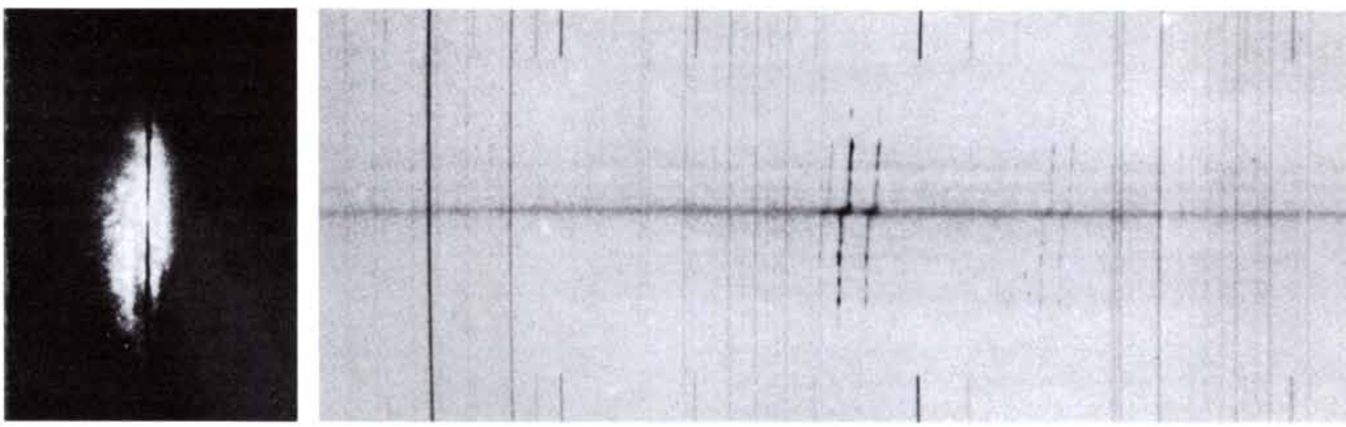
In a galaxy the brightness is strongly peaked near the center and falls off rapidly with distance. Astronomers had long assumed that the mass too decreased rapidly with distance, in accordance with the distribution of luminosity. Hence it was expected that stars at increasing distances from the center would have decreasing Keplerian orbital velocities. Until recently few velocity observations had been made in the faint outer regions of galaxies, either to confirm this expectation or refute it.

Although the forms of spiral galaxies are exceedingly diverse, astronomers are able to group them into three useful classes following a scheme proposed some 60 years ago by Edwin P. Hubble. Galaxies designated Sa have a large central bulge surrounded by tightly wound smooth arms in which "knots," or bright regions, are barely resolved. Sb galaxies have a less pronounced central bulge and more open arms with more pronounced knots. Sc galaxies have a small central bulge and well-separated arms speckled with distinct luminous segments. The progression from type Sa to type Sc is one of decreasing prominence of the central bulge and increasing prominence of the disk rotating about it. That the disk is indeed rotating is assumed on simple dynamical grounds.

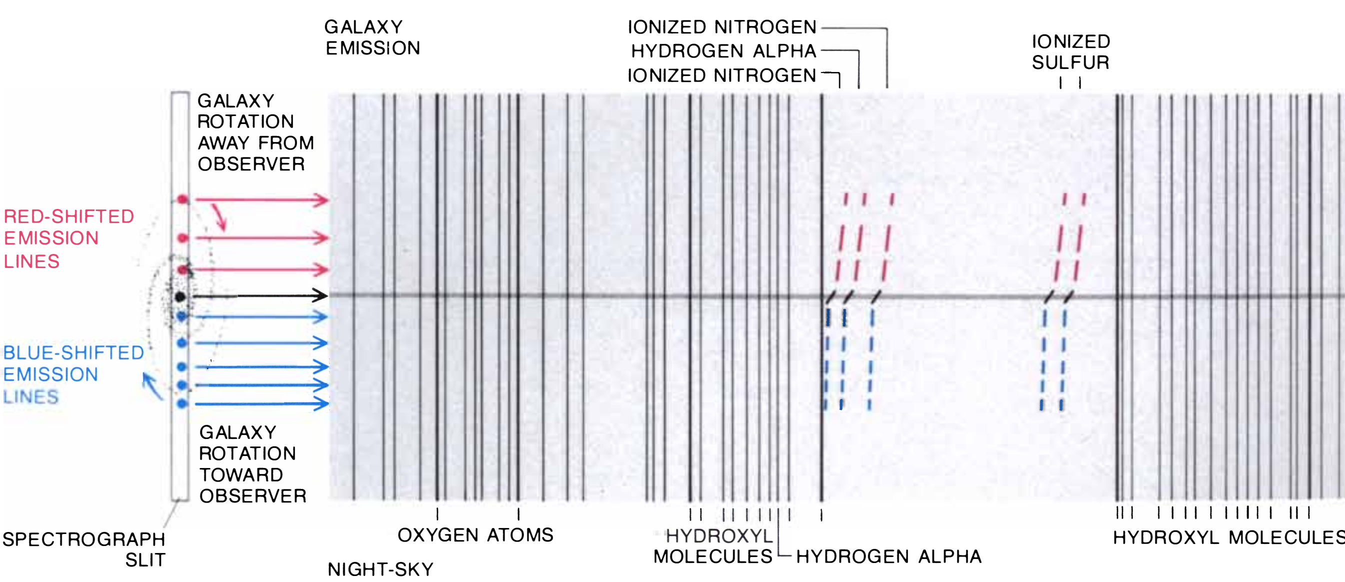
Within each type there are systematic variations in size and luminosity. For example, Sc galaxies range from small, low-luminosity, low-mass objects to galaxies of enormous luminosity and mass. For completeness, therefore, the study of the dynamics of galaxies should include not only objects with a range of morphological types but also objects with a range of luminosities.

Only for the closest stars in our own galaxy is it possible to detect motion by observing the changing position of the star against the background of more distant stars and galaxies on the celestial sphere. Even for the Andromeda galaxy, the large spiral galaxy closest to our own, it would take some 20,000 years for an orbital velocity of 200 kilometers per second (a velocity comparable to the sun's) to carry a star one second of arc across the sky. This is the minimum angular separation that can be detected optically from the earth. To study the motions in galaxies a different method is needed, one based on the phenomenon of the Doppler shift.

Doppler shifts are shifts in the frequency of waves from a source caused by the motion of the source toward or away from the observer. When the spectrum of the bright nucleus of a spiral galaxy is recorded, the absorption lines arising from the constituent stars are shifted toward the long-wavelength (red) end of the spectrum compared with the same lines in spectra made in laboratories on the earth. Such red shifts in the spectra of all but a few of the nearest galaxies, first observed in about 1915 by V. M. Slipher of the Lowell



**SPECTRUM OF SPIRAL GALAXY NGC 7541 (*right*) was recorded with the four-meter telescope at the Kitt Peak National Observatory by the author and W. Kent Ford, Jr. NGC 7541 is a type Sc spiral, 60 megaparsecs distant. (A megaparsec is 3.26 million light-years.) The exposure time was 114 minutes. The galaxy is seen at the left as it appears on a television monitor in the telescope's console room. The dark line through the galaxy shows the orientation of the spectrograph slit. Light from across the disk is sampled (*see illustration below*).**




**EMISSION LINES in the spectrogram of NGC 7541 arise from two sources: the night sky and atoms in the gas clouds surrounding bright stars in the galaxy. Most of the night-sky lines, which extend across the entire width of the spectrogram, are from hydroxyl (OH) molecules in the atmosphere of the earth. A few arise from oxygen and hydrogen atoms in the earth's atmosphere. The rotation of NGC 7541 shifts the position of the emission lines from the disk of the galaxy to either a shorter (bluer) wavelength or a longer (redder) one, depending on whether the rotation is carrying the stars and gas in the disk toward or away from the observer. Because the galaxy itself is traveling away from the observer as part of the general expansion of the universe, the hydrogen-alpha line from gas in the galaxy is red-shifted from the position of the same line in the night sky. The displacement is a measure of the galaxy's velocity of recession. The slant of the galactic emission lines shows that the orbital velocity of stars and gas in the disk is increasing with distance from the galactic center.**

Observatory, provide the evidence that the universe is expanding, carrying almost all the other isolated galaxies away from ours and away from one another. As a result of Smith and Zwicky's work it is known that in pairs, groups and clusters of galaxies the local gravitational field overcomes the general expansion, so that these denser agglomerations of matter remain bound. Although the distances between clusters of galaxies are increasing, the distances between galaxies within clusters remain about the same. Slipher also noted that the spectra of individual galaxies can yield additional information about the motions of stars and gas within the galaxy.

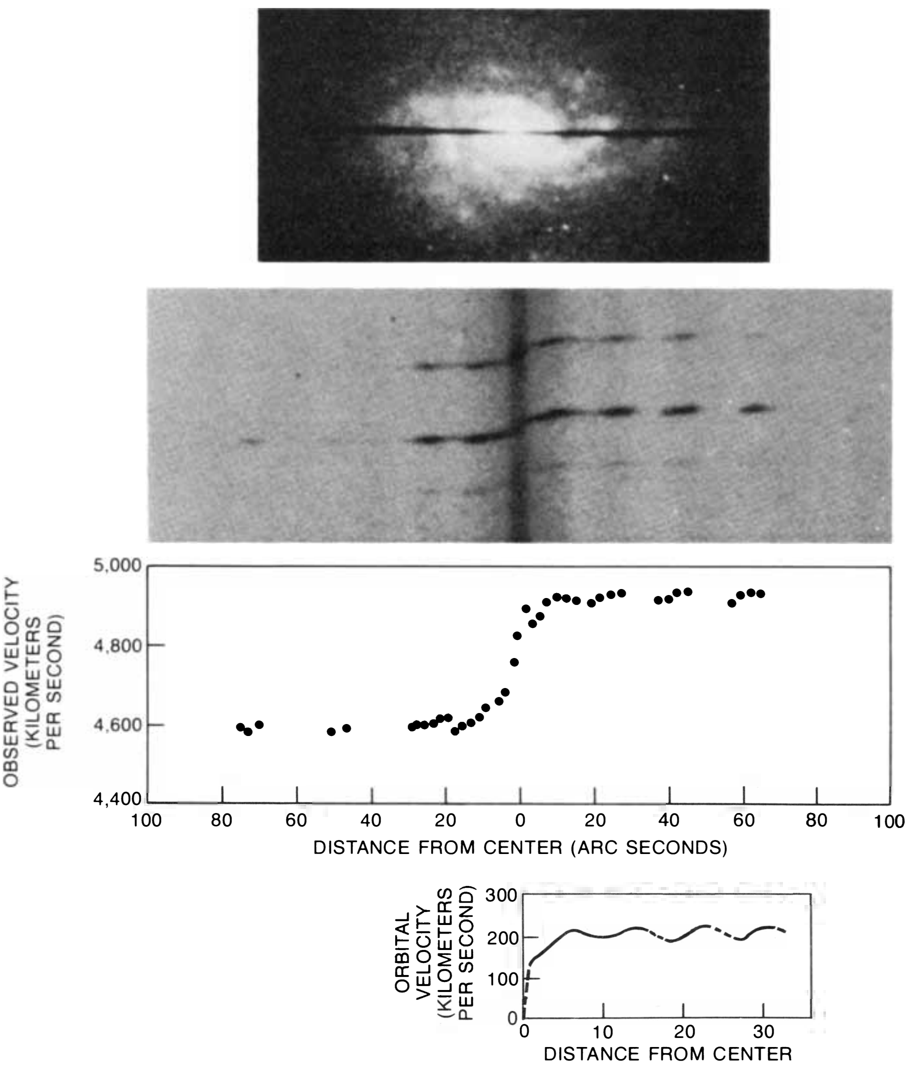
If the disk of a spiral galaxy is oriented so that its plane is sharply tilted with respect to the line of sight from the earth, the rotation of the galaxy will carry the stars and gas on one side of the galactic nucleus toward our galaxy and those on the other side away from it. The spectral lines of the approaching material will therefore be blue-shifted, or raised in frequency, and the lines of the receding material will be red-shifted, or lowered in frequency. A measurement at any point on a spectral line will therefore supply both the angular distance of that point from the galactic nucleus and the velocity along the line of sight at that distance.

It is difficult to make spectroscopic measurements of the velocities of individual stars, which are faint even in galaxies fairly close to our own. In our work, therefore, we observe not stars but the light from the clouds of gas, rich in hydrogen and helium, that surround certain hot stars. The spectra of such clouds consist of bright emission lines that arise as an electron in an excited atom drops from a higher energy state to a lower one. In addition to emission lines of hydrogen and helium there usually are bright lines from atoms of nitrogen and sulfur that are singly ionized, or stripped of one electron. These lines are called forbidden because they arise only from atoms in the near-vacuum of space; in terrestrial laboratories such singly ionized atoms are rapidly deexcited by collisions with other atoms before the forbidden transition can occur.

Until recently it was not possible to get high-resolution optical spectra of the faint outer regions of galaxies. It is the present availability of large optical telescopes, of high-resolution, long-slit spectrographs and of efficient electronic imaging devices that have made our observing program feasible. Six years ago my colleagues and I set out to measure the rotational velocities completely across the luminous disk of suitably tilted spiral galaxies. Our aim was to study the internal dynamics and distribution of mass in individual galaxies as a function of the galaxies' morphology. We have now observed 60 spiral galaxies: 20 each of the three major types Sa, Sb and Sc. We have selected galaxies that have a well-defined type, that are well inclined to the plane of the sky (yielding a large component of orbital velocity along the line of sight), that have an angular diameter no larger than the slit of the spectrograph and that span a large range of luminosities within each type.

Most of the spectra have been obtained with two four-meter telescopes, the one at the Kitt Peak National Observatory in Arizona and the one at the Cerro Tololo Inter-American Observatory in Chile. A few of the spectra were recorded with the 2.5-meter telescope at the Las Campanas Observatory in Chile.

After the photons from the galactic source pass through the slit of the spectrograph and are dispersed by a diffraction grating, they are focused on a "Carnegie" image tube (RCA C33063), where they are multiplied by a factor of 10 or more before they are recorded by the photographic emulsion. Exposures of two to three hours are recorded on Kodak IIIa-J plates whose sensitivity, matched to that of the image tube, has been much increased by having previously been baked at 65 degrees Celsius for two hours in a special "forming" gas (nitrogen with an admixture of 2 percent hydrogen) and preexposed to flashes of light. Without the image tube and




**MEASUREMENT OF THE ROTATION of NGC 2998, an Sc galaxy at a distance of 96 megaparsecs in the constellation Ursa Major, begins with the making of a spectrogram. The picture at the top shows the galaxy and superposed spectrograph slit as they appear on a television monitor at the four-meter Kitt Peak telescope. Below it is the hydrogen-alpha region of the spectrogram that resulted from an exposure of 200 minutes. The plotted points depict velocities across the galactic disk as measured from the hydrogen-alpha line. The entire galaxy is receding at 4,800 kilometers per second; the left side of the galaxy is approaching, the right side receding. The last step is to draw a rotation curve by smoothing velocities from both sides of the disk and translating angular distance on the sky into linear distance in the galaxy.**

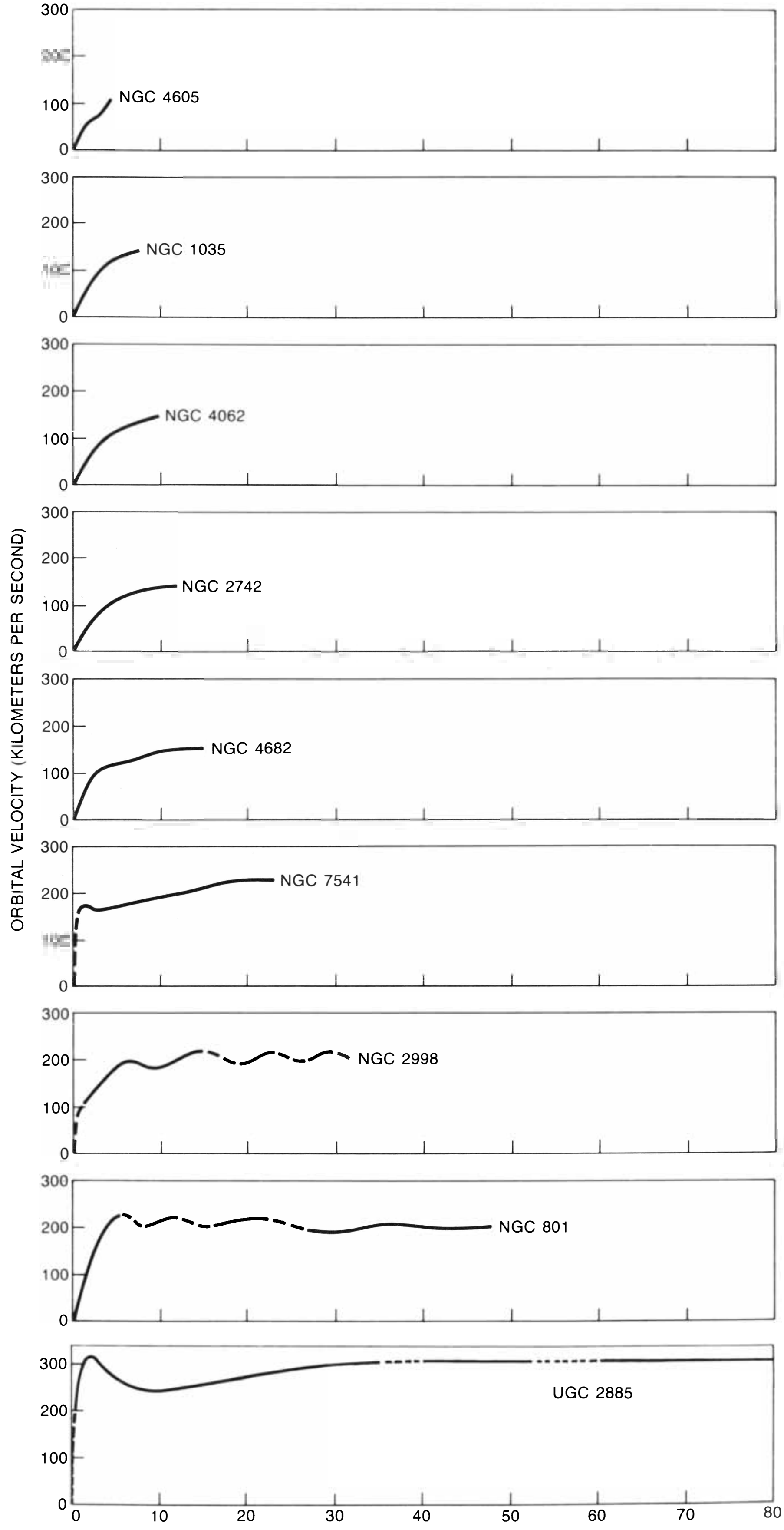
the plate-sensitizing methods exposure times would have been prohibitively long: from 20 to 60 hours.

Generally two exposures are made of each galaxy. In one exposure the spectrograph slit is made to coincide with the major (long) axis of the galaxy; each point on the spectrum arises from a single region of the galactic disk. The Doppler, or velocity, displacements of the emission lines are readily discerned in the developed image. A second exposure is made with the spectrograph slit aligned with the minor axis of the galactic disk. Since the orbital velocities are now perpendicular to the line of sight, no Doppler shifts are evident. The absence of line displacements with the slit along the minor axis is confirming evidence that the motions we study are indeed orbital ones.

In order to have a reference scale against which to measure the displacement of emission lines in galactic spectra astronomers formerly recorded neon lines from a lamp along the edges of the spectrum. We have now dispensed with this procedure. Instead we measure displacements directly from the unshifted lines on each plate that are emitted by hydroxyl (OH) molecules in the earth's atmosphere. Many astronomers have adopted sophisticated plate-scanning devices to measure line positions, particularly for faint signals. We, however, still measure the location of the emission lines with the aid of a microscope whose stage can be moved in two directions. We are able to measure positions in each coordinate to the high accuracy of one micrometer.

In our work we define the nominal radius of a galaxy as that distance at which the surface brightness of the galaxy has fallen to the threshold of detectability on plates made with the 48-inch Schmidt telescope on Palomar Mountain, a value equal to 25th magnitude per square second of arc. For establishing the distance to the objects examined, and hence their actual size, we adopt a value for the Hubble constant (which specifies the expansion rate of the universe) of 50 kilometers per second per megaparsec. (A megaparsec is 3.26 million light-years.)

From the measured velocities of the strongest emission lines we compute a smooth rotation curve by averaging together the approaching and receding velocities from the two sides of the galactic disk. Although each galaxy exhibits distinctive features in its rotational pattern, the systematic trends that emerge are impressive. With increasing luminosity galaxies are bigger, orbital velocities are higher and the velocity gradient across the nuclear bulge is steeper. Moreover, each type of galaxy displays characteristic rotational prop-

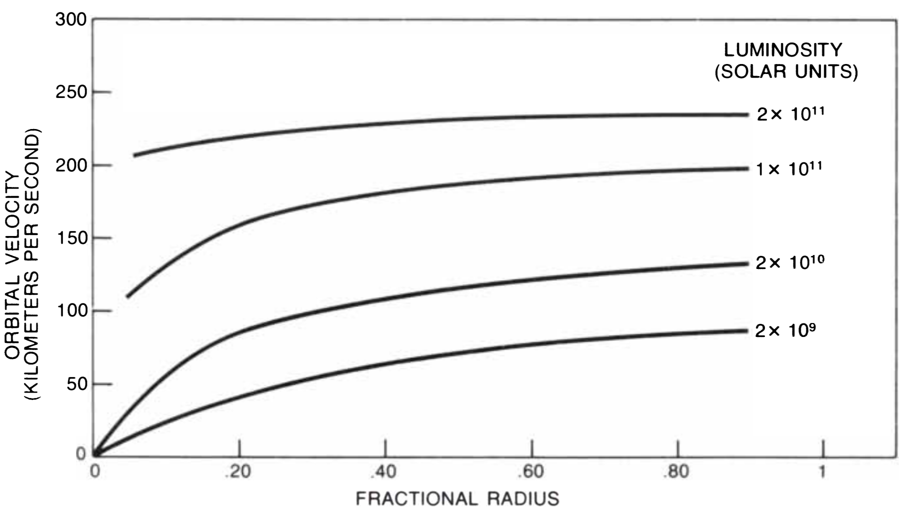



**ROTATION CURVES** show orbital velocities of nine Sc galaxies from the center outward. Galaxies increase in luminosity from top to bottom. With increasing luminosity galaxies are larger, orbital velocities are higher and velocity gradients near the galactic center are steeper.

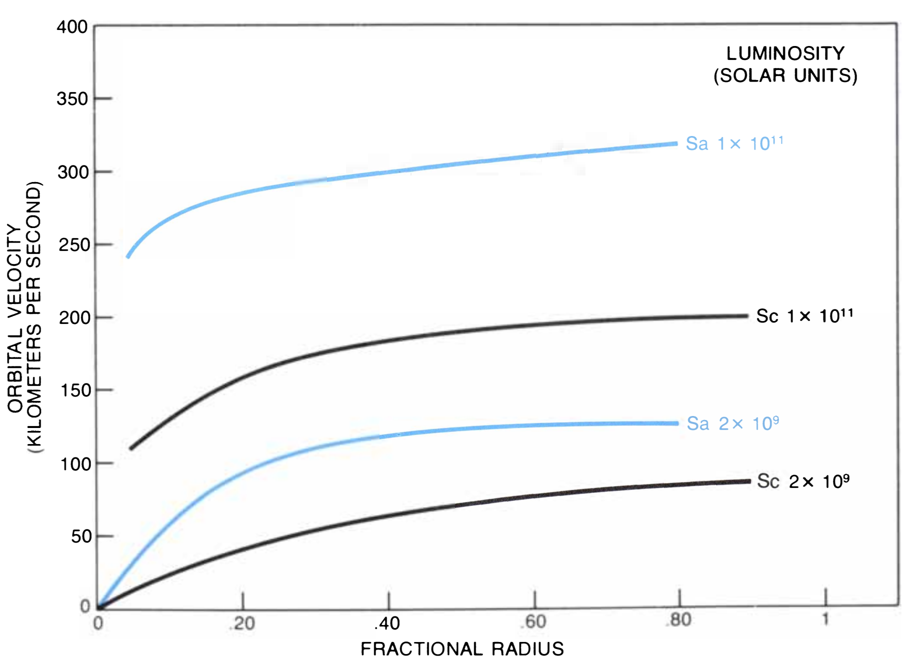
erties. For example, the most luminous Sa galaxies rotate more than 50 percent faster at the midpoint of their radius than equally luminous Sc galaxies. Among Sc galaxies the most luminous rotate more than twice as fast at comparable radial distances as Sc galaxies that are only a hundredth as luminous.

One overwhelming conclusion emerges from our observations. Virtually all the rotation curves are either flat or rising out to the visible limits of the galaxy. There are no extensive regions where the velocities fall off with distance from the center, as would be predicted if mass were centrally concentrated. The conclusion is inescapable: mass, unlike luminosity, is not concentrated near the center of spiral galaxies. Thus the light distribution in a galaxy is not at all a guide to mass distribution.

On the basis of their rotational velocities the masses of the galaxies in our study range from $6 \times 10^9$ to $2 \times 10^{12}$ times the mass of the sun inside their optical radius. We cannot yet specify the total mass of any one galaxy because we do not see any "edge" to the mass. Instead the mass inside any given radial distance is increasing linearly with distance and, contrary to what one might expect, is not converging to a limiting mass at the edge of the visible disk. The linear increase of mass with radius indicates that each successive shell of matter in the galaxy must contain just as much mass as every other shell of the same thickness. Since the volume of each successive shell increases as the square of the radius, the density of matter in successive shells must decrease as 1 over the radius squared in order for the product of the density times the volume to remain constant.




**ORBITAL VELOCITIES are depicted schematically for Sc galaxies of varying luminosity as a function of the optically visible radius of each galaxy. Luminosities in solar units differ by two orders of magnitude. At every radial distance orbital velocities increase with luminosity.**



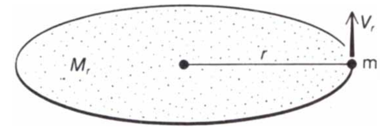

**COMPARISON OF Sa AND Sc GALAXIES shows that for equal luminosity orbital velocities are significantly higher in Sa galaxies than they are in Sc galaxies at every radial distance. This implies that Sa galaxies harbor more mass per unit of luminosity than Sc galaxies do.**

The theoretical model that least disturbs generally accepted ideas about galaxies accounts for the observed rotation curves by embedding each spiral galaxy in a spherical "halo" of matter that extends well beyond the visible limits of the galactic disk. The gravitational attraction of this unseen mass keeps the orbital velocities of the galaxies from decreasing with distance from the galactic center. It is perhaps disappointing that the observations yield almost no information on the detailed distribution of the invisible dark matter. One can nonetheless say that the dark matter is not part of the overall background density of matter in the universe but rather is strongly clumped around galaxies. This is evident because the density of nonluminous matter decreases, albeit slowly, with distance from the galactic center, and the density even at large radial distances is between 100 and 1,000 times higher than the mean density of the universe.

Although there are other models that try to account for the high orbital velocities, all are less satisfactory than a single halo of dark matter. If all the required unseen matter is put in a disk, the disk will quickly become unstable and form itself into a bar. The important finding that halos are necessary for stabilizing a disk was first elucidated by Jeremiah P. Ostriker and P. J. E. Peebles of Princeton University.

The observed dynamic effects are reproduced by models of spiral galaxies that put the mass in a nucleus, a surrounding bulge, a disk and a halo. Particularly interesting models have been developed by John N. Bahcall and Raymond M. Soneira of the Institute for Advanced Study, Maarten Schmidt of Cal Tech and S. Casertano of the Scuola Normale Superiore in Pisa. Perhaps the most radical idea for explaining the observed high rotational velocities is one advanced independently by Joel E. Tohline of Louisiana State University and M. Milgrom and J. Bekenstein of the Weizmann Institute of Science. They have proposed that at great distances the Newtonian theory of gravitation must be modified, thereby allowing rotational velocities in galaxies to remain high at such distances from the galactic center even in the absence of unseen mass.

Additional evidence on the high rotational velocities of matter in spiral galaxies is provided by the 21-centimeter radio waves emitted by the neutral (unionized) hydrogen in the galactic disk.




$$\frac{GmM_r}{r^2} = \frac{mV_r^2}{r} \rightarrow M_r = \frac{rV_r^2}{G}$$

GRAVITATIONAL FORCE — CENTRIFUGAL FORCE

| | NGC 1035 | | NGC 2998 | |
|---|---|---|---|---|
| RADIUS (KILOPARSECS) | VELOCITY $V_r$ (KILOMETERS PER SECOND) | INTERIOR MASS $M_r$ ($10^{10}$ SOLAR MASSES) | VELOCITY $V_r$ (KILOMETERS PER SECOND) | INTERIOR MASS $M_r$ ($10^{10}$ SOLAR MASSES) |
| .5 | 39 | .018 | 87 | .088 |
| 1 | 65 | .098 | 102 | .24 |
| 2 | 91 | .39 | 126 | .74 |
| 3 | 107 | .80 | 142 | 1.4 |
| 5 | 123 | 1.8 | 182 | 3.9 |
| 8 | 135 | 3.4 | 204 | 7.7 |
| 20 | | | 214 | 21 |
| 30 | | | 214 | 32 |

**MASS INSIDE A GIVEN RADIAL DISTANCE can be calculated from the equivalence of gravitational force and centrifugal force at distance $r$ from the center of the galaxy. In the equations $G$ is the constant of gravitation, $m$ is the mass at distance $r$, $M_r$ is the mass inside $r$ and $V_r$ is the orbital velocity of mass $m$. The mass inside $r$ increases linearly with distance. The table gives the mass inside $r$ for two Sc galaxies: NGC 1035, of low luminosity, and NGC 2998, of high luminosity. At every distance from the galactic center the more luminous galaxy exhibits a higher orbital velocity and therefore must have much more mass inside that distance.**

Early studies of the 21-centimeter radiation of a few spiral galaxies by Morton S. Roberts of the National Radio Astronomy Observatory showed that the rotational velocities of the hydrogen are high. With multiple radio telescopes, notably the array at Westerbork in the Netherlands and the Very Large Array at Socorro, N.M., it is possible to match and even exceed the resolving power of optical telescopes and thereby to study the distribution of hydrogen in galaxies similar to those we have observed. Albert Bosma of the State University of Leiden has shown for a wide variety of galaxy types that the orbital velocities of neutral hydrogen remain high at large distances from the galactic center.

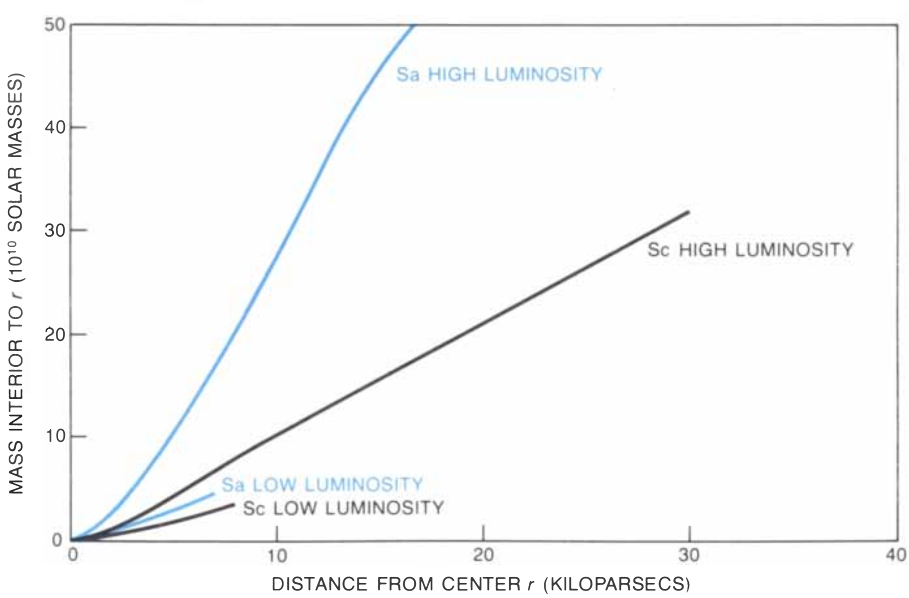
In general the apparent diameters of galaxies are similar whether they are measured by optical observations or by radio ones. For a small set of galaxies, however, hydrogen extends several times as far out from the center as the luminous stars do. For such objects it is possible to determine the gravitational potential beyond the limits of the optically visible galaxy. In several instances the hydrogen does not remain in a plane but is warped sharply near the edge of the visible disk. It is therefore not certain whether the gas velocities that have been measured at the largest distances from the center are true circular orbital velocities or whether they represent more complex motions.

Renzo Sancisi of the University of Groningen, who has studied such warped galaxies, has suggested that the orbital velocities may in fact be decreasing beyond the limits of the visible galaxy. The velocities, however, seem to decrease only slightly, perhaps by 20 kilometers per second, or about 10 percent, and then hold constant at that value at larger distances. The radio observations are continuing and should offer important information on the far outer regions of galaxies.

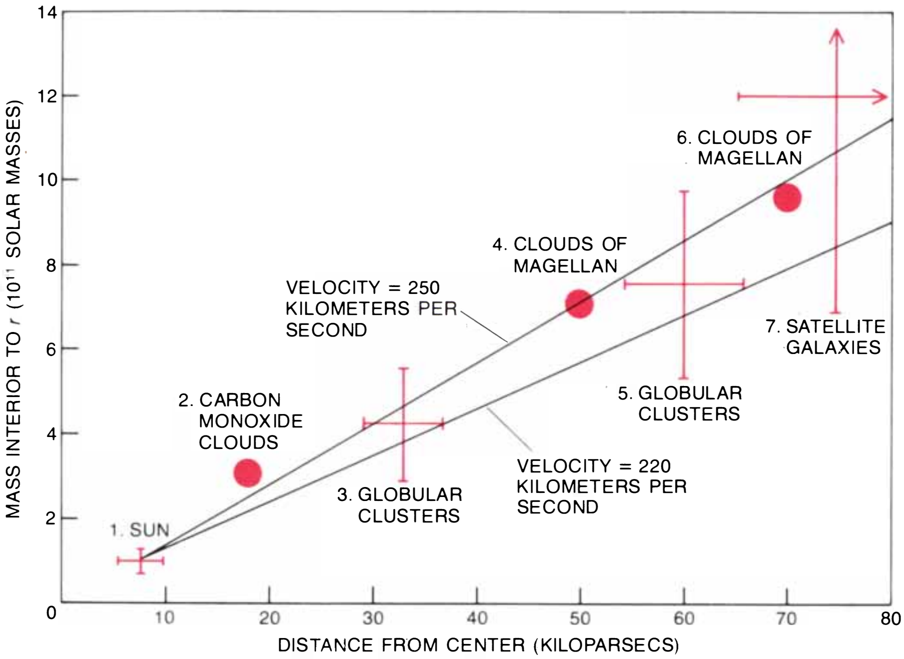
Students of galaxies are fortunate in being able to examine the properties of galaxies a long way off and then return to the galaxy where they live and ask if it exhibits the same features as other galaxies. It was not so long ago that astronomers believed the sun, about eight kiloparsecs from the center of our galaxy, was near the edge of it and that the galaxy was only of moderate size. Now all the evidence indicates that our galaxy too extends well beyond the position of the sun and that its mass continues to increase.

The velocity of the sun in its orbit around the center of the galaxy is placed at 220 kilometers per second by James E. Gunn and Gillian R. Knapp of Princeton and Scott D. Tremaine of the Massachusetts Institute of Technology. Other estimates run as high as 260 kilometers per second. At the lower value the amount of mass between the sun and the center of the galaxy is about $10^{11}$ solar masses. On the evidence that substantial mass lies beyond the sun's distance from the galactic center, the galactic mass out to 100 kiloparsecs may reach $10^{12}$ solar masses, which would place our galaxy in a class with the largest galaxies of its type.

Some 30 years ago Jan H. Oort of the Leiden Observatory demonstrated that the observable mass of stars and gas in the galactic disk in the vicinity of the sun is too low by almost a factor of two to




**COMPARISONS OF INTERIOR MASS for both low- and high-luminosity galaxies show that the mass rises with approximate linearity with distance $r$ from the center and gives no sign of approaching a limit at the edge of the optically visible galaxy. At every radial distance Sa galaxies exhibit higher mass and therefore higher density than Sc galaxies of equal luminosity.**




**STUDIES OF OUR OWN GALAXY yield estimates of the mass inside *r* measured in kiloparsecs from the galactic center. The estimates are calculated from the orbital velocities and mean distances of a variety of objects. The value adopted for the orbital velocity of the sun at a distance of eight kiloparsecs is 220 kilometers per second. The second point is obtained from the mean velocity of carbon monoxide clouds at a mean distance of 18 kiloparsecs, measured by Leo Blitz of the University of Maryland at College Park. The third and fifth points are derived from the velocity of globular clusters of stars in the "halo" of our galaxy at two different average distances from the nucleus. The velocities of the nearer clusters were analyzed by Carlos Frenk, who was then working at the University of Cambridge, and Simon White of the University of California at Berkeley. The velocities of the more distant clusters were studied by F. D. A. Hartwick of the University of Victoria and Wallace L. W. Sargent of the California Institute of Technology. The fourth point was obtained from velocities of the Clouds of Magellan, the nearest galaxies to our own, as estimated by Tadayuki Murai and Mitsuaki Fujimoto of Nagoya University. The sixth point represents independent estimates of the distance and velocities of the Clouds of Magellan made by D. N. C. Lin of the Lick Observatory and Donald Lynden-Bell of the University of Cambridge. The final point is based on the velocity of more remote satellite galaxies as estimated by Jaan Einasto and his colleagues at the Estonian S.S.R. Academy of Sciences. The extent of the vertical lines indicates the range of values for orbits of different geometries. The measurements suggest that rotational velocities in our galaxy lie between 220 and 250 kilometers per second and remain constant out to 80 kiloparsecs, or roughly 10 times the sun's distance from the galactic center. The mass inside 80 kiloparsecs is likewise some 10 times the mass inside the radial distance of the sun, or about $10^{12}$ solar masses.**

account for the disk's gravitational attraction on the stars far out of its central plane. This study offered the first evidence that our galaxy too harbors mass that is not luminous.

More recent evidence comes from the orbital velocities of objects in the plane of the galaxy considerably farther out than the sun. Measurements are difficult, but the velocities have been deduced for a few special cases. For example, Leo Blitz of the University of Maryland at College Park has determined the velocities of clouds of carbon monoxide at a distance of nearly 16 kiloparsecs from the galactic center. These velocities, together with the velocities of hydrogen clouds determined by Blitz and Shrinivas Kulkarni and Carl E. Heiles of the University of California at Berkeley, yield a rotation curve that continues to rise with increasing distance from the galactic center.

In order to deduce the mass at still larger distances the velocities of globular star clusters in the halo of our galaxy, with one sample of clusters at 30 kiloparsecs from the center and another at 60 kiloparsecs, have been measured by F. D. A. Hartwick of the University of Victoria, Wallace L. W. Sargent of Cal Tech, Carlos Frenk of the University of Cambridge and Simon White of the University of California at Berkeley. Their work shows that the mass continues to increase with approximate linearity to the mean distance of the clusters.

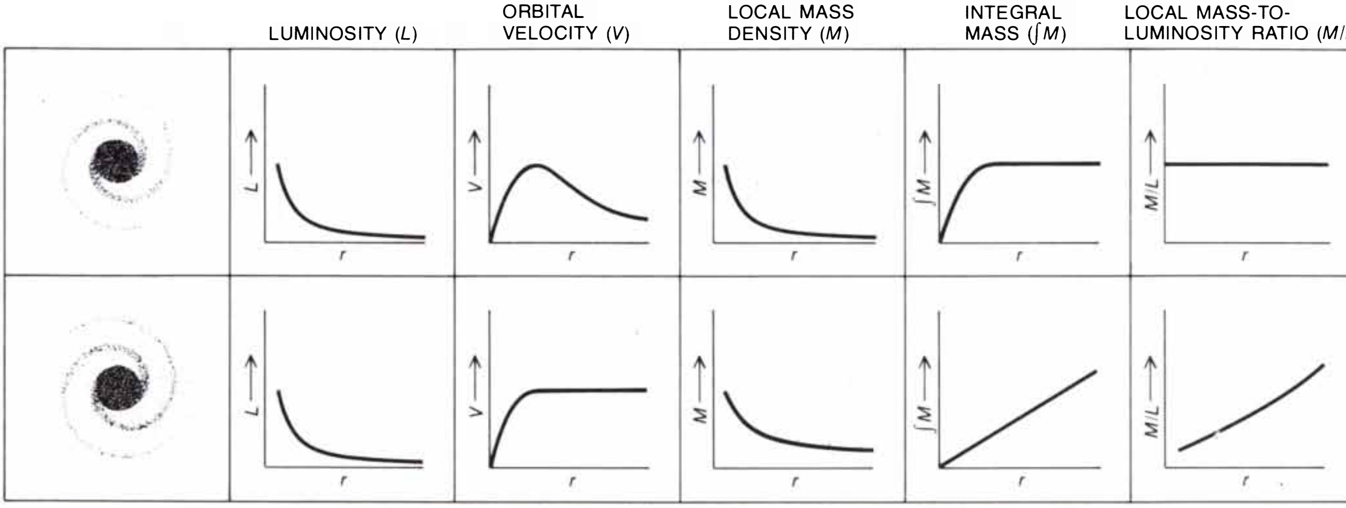
With effort and imagination it is possible to sample the gravitational potential at even more remote distances. Our galaxy is not alone in intergalactic space; it has a retinue of smaller satellite galaxies. The orbits of the two closest satellites, the Large and Small Clouds of Magellan, a little less than 60 kiloparsecs from the center of our galaxy, are highly uncertain. Model orbits have been proposed, however, by Tadayuki Murai and Mitsuaki Fujimoto of Nagoya University, D. N. C. Lin of the Lick Observatory and Donald Lynden-Bell of the University of Cambridge. From the model orbits they deduce values of mass that are consistent with those yielded by the globular clusters.

For still greater distances Jaan Einasto and his colleagues at the Estonian S.S.R. Academy of Sciences have relied on a combination of enormously distant globular clusters and satellite galaxies to deduce the mass to distances beyond 80 kiloparsecs. When the results from such analyses are combined, they indicate a galaxy in which orbital velocities remain in the range of 220 to 250 kilometers per second out to almost 10 times the distance of the sun from the galactic center. Such a mass distribution is mandatory if our galaxy is to resemble all the other spiral galaxies my colleagues and I have studied. It moves the sun from a relatively rural position to a much more urban one.

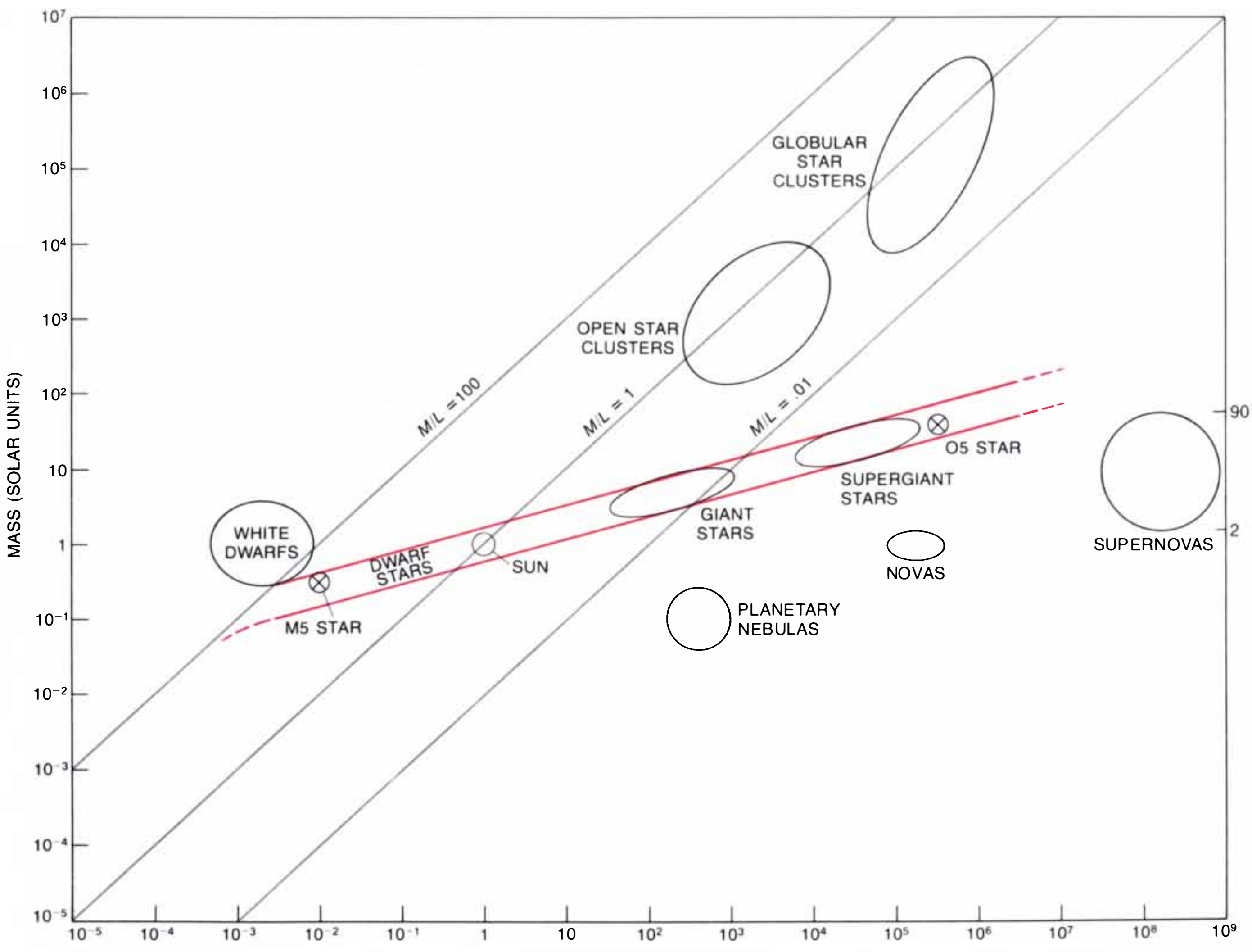
The broad conclusion that can be drawn from all these results is that as the disk of a spiral galaxy is scanned from the center outward the total mass of luminous and dark material falls off slowly and the luminosity (measured in the blue region of the spectrum) falls off rapidly. As a result the ratio of the local mass density to the local (blue) luminosity density, which can be expressed for convenience as the value of the ratio $M/L$, increases steadily with distance from the galactic center. In the nuclear region a lot of luminosity is produced by relatively little mass, whereas at large distances little luminosity is produced by a lot of mass. If there were no invisible material clumped around galaxies, the mass distribution would simply follow the luminosity distribution and the $M/L$ ratio would be approximately constant across the disk from its center to its edge.

If mass and luminosity are measured in units of solar mass and solar luminosity, the $M/L$ ratio of the sun is of course 1/1. In such units (omitting the denominator, which is simply 1) the average $M/L$ ratio near the nucleus of a spiral galaxy has sunlike values of 1 or perhaps 2 or 3. Toward the edge of the visible disk, as luminosity decreases, the $M/L$ value climbs to 10 or 20. Beyond the visible disk, where the luminosity falls essentially to zero and the mass remains high, the average $M/L$ value soars into the hundreds.

In an effort to identify the constituents of the invisible halo we must ask what celestial objects have high values of $M/L$. Stars like the sun are clearly ruled out. The luminous hot young stars that delineate a galaxy's spiral arms are even poorer candidates; their $M/L$ values are about $10^{-4}$. At the other extreme the old red-dwarf stars that populate the nuclear bulge and the outlying regions of the galaxy have both a low mass and a




**HYPOTHETICAL AND ACTUAL GALAXIES** deviate sharply in all their properties except luminosity. The typical actual spiral galaxy at the bottom has a massive nonluminous halo. The hypothetical galaxy at the top has no halo. Its surface brightness decreases rapidly, orbital velocities outside the nucleus decrease in Keplerian fashion, local mass density falls in parallel with luminosity, integral mass reaches a limiting value and the ratio of mass to luminosity stays approximately constant with increasing radial distance. Such were the expected properties of a galaxy. In an actual galaxy the presence of a dark halo changes everything but the galaxy's optical appearance. The orbital velocities remain high, the local mass density falls only slowly, the integral mass increases linearly with radius and the mass-to-luminosity ratio steadily increases as the halo of the galaxy contributes more mass and the luminous disk falls to the threshold of detectability.




**MASS AND LUMINOSITY ARE COMPARED** for typical components of a spiral galaxy such as our own. The mass and luminosity of the sun are taken as unity. In solar units the value of the ratio of mass to luminosity, $M/L$, for normal stars decreases from about 30 for cool, old dwarf stars (type M5) to about $10^{-4}$ for hot, young stars (type O5). Only extremely dense white-dwarf stars have an $M/L$ value in excess of 100. Some other class of objects is needed to populate the halo of a galaxy, where $M/L$ values soar into the hundreds.

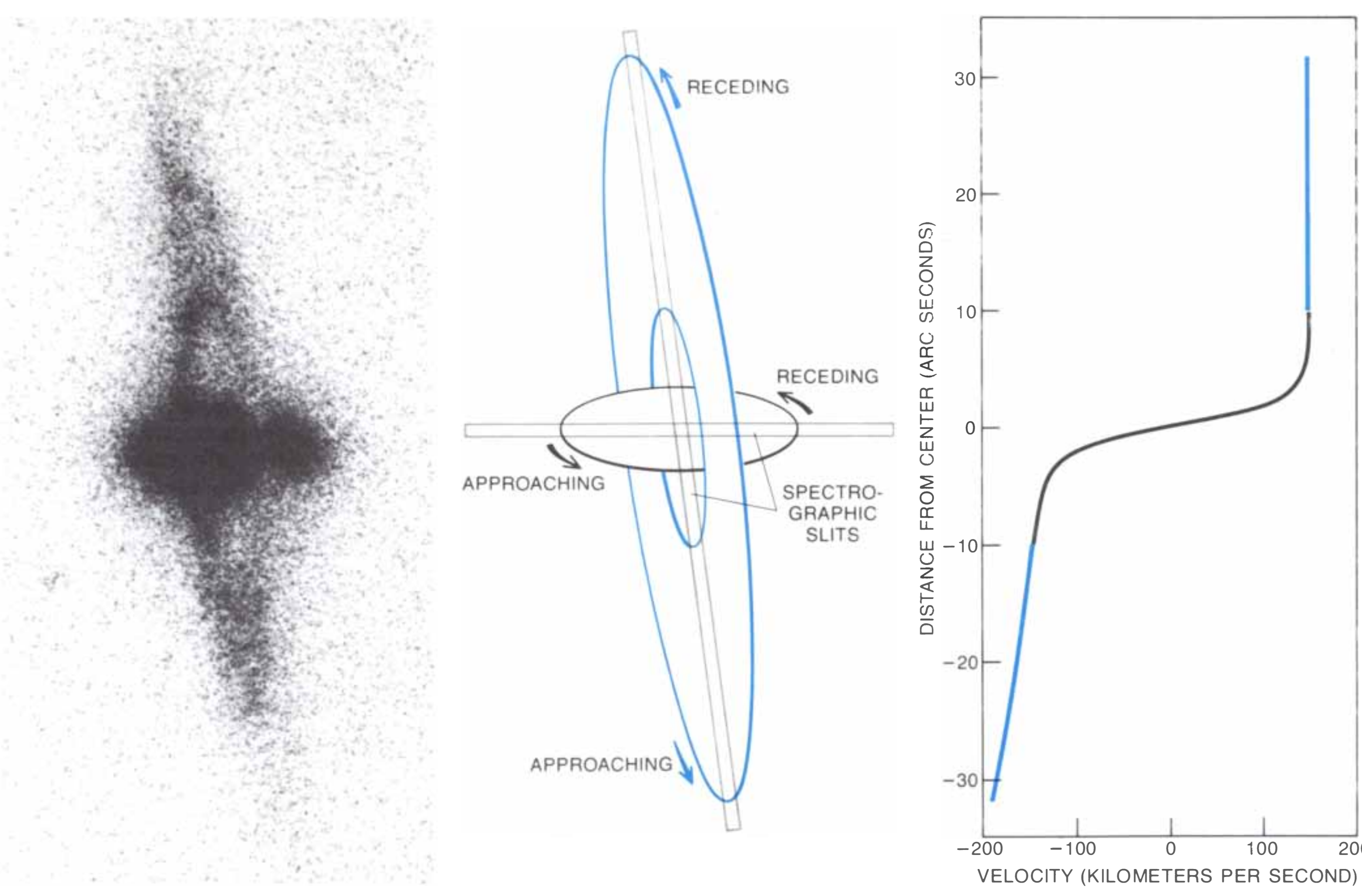
low blue luminosity. Their $M/L$ values, about 20, are still far short of the values needed for the halo. Moreover, a halo consisting of very-low-mass red dwarfs would reveal its presence by radiating strongly in the infrared region of the spectrum. All attempts to detect a halo by its visual, infrared, radio or X-ray radiation have failed.

What candidates are left? Normal stars radiate energy generated by thermonuclear processes, which convert hydrogen and helium into heavier elements. Such nuclear processes are kindled only in bodies whose mass is large enough for the gravitational energy to raise the temperature at the core of the star to several million degrees Kelvin (degrees C. above absolute zero). The minimum mass required is about .085 times the mass of the sun. Jupiter, the largest planet in the solar system, falls short of this value by a factor of nearly 100. A halo of planetlike bodies, perhaps protostars that failed to become stars, is at least conceivable, although rather unlikely. In sum, the only requirement for the halo is the presence of matter in any cold, dark form that meets the $M/L$ constraint, from neutrinos to black holes.

So far I have described the rotational properties of relatively isolated normal spiral galaxies. There is additional observational evidence for high $M/L$ ratios at large distances from the nuclei of other galaxies. Occasionally nature offers an unexpected opportunity to probe its secrets. Recently François Schweizer of the Carnegie Institution, Bradley C. Whitmore of Arizona State University and I have been fascinated by the faint "anonymous" galaxy AO 136–0801, one of a class of spindle galaxies with polar rings. It is called anonymous because it is not listed in any of the standard galactic catalogues; its numerical designation corresponds to its location in the sky.

Our observations of the distribution of light across the spindle show that it is a low-luminosity disk of stars viewed nearly edge on, with little or no gas and dust and no spiral structure. Such galaxies are classed as SO galaxies and represent a significant fraction of all disk galaxies. By our usual methods we have determined the rotational properties of the disk by measuring the Doppler shift of absorption lines from its component stars. A short distance from the center of the object along the major axis of the spindle rotational velocities reach 145 kilometers per second, a value that corresponds closely to velocities measured in type-Sa galaxies of low luminosity. Along the minor axis the orbital velocities show no line-of-sight component, confirming evidence that we are observing a rotating disk of stars.

The unusual feature of AO 136–0801 is a large ring, also seen nearly edge on, that encircles the narrow axis of the spindle by passing almost over the disk's center of rotation [*see illustration below*]. The ring is composed of gas, dust and luminous young stars. The gas reveals itself by its emission-line spectrum, the dust by its absorbing effects where it crosses in front of the spindle




**DISK INSIDE A RING is exhibited by the "anonymous" galaxy AO 136–0801, so called because it is not listed in standard catalogues. (The numbers give its position in right ascension and declination.) The oval central region is a rotating disk of stars seen nearly edge on. The stars and gas in the thin ring are also rotating but in a plane almost perpendicular to the disk. The directions of rotation are indicated by the diagram in the middle, which shows how the slits of the spectrograph were oriented for measuring orbital velocities in the disk and the ring. The two sets of velocity measurements are plotted at the right. At 10 seconds of arc from the center the velocities in the disk (*black*) and in the ring (*color*) are essentially the same. The velocities in the ring, however, can be measured out to nearly three times the optical radius of the disk, and they remain virtually constant. It appears that the mass is continuing to increase linearly out to distances much greater than the disk radius, and that objects in the ring respond to a gravitational potential that is not disklike but spherical.**

Printed circuit boards made of a new material may permit better direct soldering of large leadless ceramic chip-carriers. A Hughes Aircraft Company study proposes using quartz-fabric-reinforced polyimide resin in place of glass-epoxy or glass-polyimide boards. The new material has nearly the same thermal expansion coefficient as ceramic chip carriers. When a leadless carrier is soldered directly to a quartz-fabric polyimide board, there are no shear stresses caused by heating or cooling. Such stresses often cause solder joints to fail on conventional reinforced printed circuit boards.

A semiautomatic production line makes solar cells that are higher performing and much lower in cost than conventional cells. The new facility at Spectrolab, a Hughes subsidiary, creates highly uniform cells and increases productivity. It uses the latest techniques, including photolithographic masking for contact patterns, with subsequent contact plate-up. The first "mechanized" cells have been delivered to Hughes for the Intelsat VI communications satellites.

Advanced electro-optical devices produced for the military by Hughes today have benefited from a 7-to-1 productivity improvement over the last five years. These devices include laser rangefinders, laser designators, and infrared night vision systems. The rise in productivity stems from more automation, computer-aided manufacturing, and employee involvement in solving day-to-day problems.

A wide-field-of-view head-up display can provide pilots with critical sensor and steering information in low-altitude flights at night and under poor visibility conditions. Head-up displays save a pilot from looking down at his instruments by superimposing such data as airspeed, heading, and target information on a glass-like combiner mounted at the pilot's eye level. Hughes pioneered the technology used in its HUD, which incorporates diffraction optics made through a process involving holographic techniques and lasers. In addition to providing the wide field of view, the display is brighter, more transparent, and doesn't obstruct the pilot's forward vision. The display resists glare, reflections, and hot spots caused by the sun. Another important advantage is its ability to display wide-field-of-view scenes from infrared sensors. Test pilots praised the display during evaluation flights in a U.S. Navy F/A-18 Hornet.

Hughes is seeking engineers to develop advanced systems and components for many different weather and communications satellites, plus the Galileo Jupiter Probe. Immediate openings exist in applications software development, data processing, digital subsystems test, microwave/RF circuit design, power supply design, digital communications, signal processing, spacecraft antenna design, system integration test and evaluation, and TELCO interconnection. Send your resume to Ray Bevacqua, Hughes Space & Communications Group, Dept. SE, Bldg. S/41, M.S. A300, P.O. Box 92919, Los Angeles, CA 90009. Equal opportunity employer.

*Creating a new world with electronics*

**HUGHES**

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291

and the stellar component by its knotty, bluish appearance in photographs. The maximum diameter of the ring is several times greater than the long axis of the spindle. As a consequence the motions of the objects in the ring offer a unique opportunity to probe the gravitational field perpendicular to the galactic disk out to distances exceeding the visible radius of the disk.

Our spectrographic observations confirm that the ring is indeed rotating at right angles to the plane of rotation of the disk. It seems improbable that this dynamical configuration could have arisen in the normal evolution of an isolated disk galaxy; the configuration must be the result of some kind of event, such as an encounter with another galaxy or with a disk of gas. By measuring the displacement of emission lines we find that the ring's velocity of rotation is about 170 kilometers per second and that the velocity curve is flat or slightly rising out to a distance of almost three times the radius of the inner disk. If the velocity curves of the disk and the ring are plotted on the same velocity-distance scale, the two are seen to have nearly identical values at the same distance from the center of the galaxy. The high rotational velocity of the ring offers strong evidence for the existence of a massive halo extending at least three times farther than the visible radius of the disk. Moreover, the shape of the halo must be more nearly spherical than disklike. Calculations show that if the halo were as flat as the disk, the velocities above the plane of the disk would be smaller than those in the disk by 20 to 40 percent.

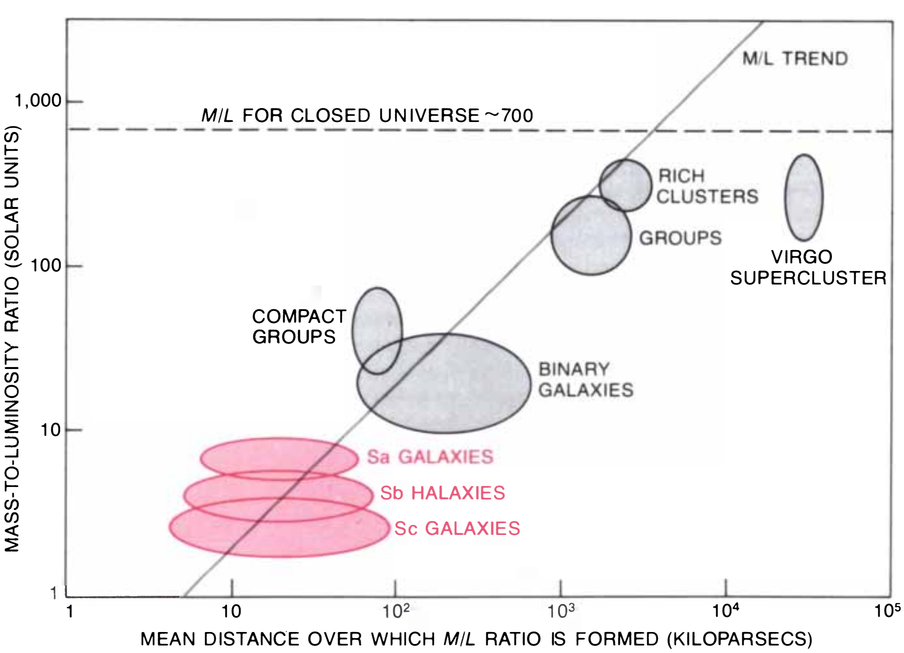
I have been describing determinations of mass made by measuring the velocity of orbiting test objects, objects in the central disk of a galaxy and objects orbiting the pole of an unusual galaxy. Other special instances can help to shed light on the quantity of dark matter in the universe. Galaxies often exist in pairs. In such instances one galaxy can be considered a test object in orbit around the other. The analysis of such a system is complex because both the orientation of the orbit in space and the position of the galaxy in the orbit are unknown. One can, however, resort to the observed properties in a large sample of double galaxies (the difference between the velocities of the two galaxies, their angular separation and their luminosity) to infer from statistical arguments the probable distribution of orbital elements and $M/L$ ratios appropriate to the galaxies.

Independent analyses by Edwin L. Turner of Princeton, Steven D. Peterson, working at Cornell University, Linda Y. Schweizer of the Carnegie Institution and I. D. Karachentsev of the Special Astrophysical Observatory in the U.S.S.R. yield mean $M/L$ values in the range between 25 and 100. These values of $M/L$ are an average over a distance equal to the separation of the galaxies in each pair, a distance generally equal to several galaxy diameters, or on the order of 100 kiloparsecs. This result helps to confirm the view that halos of dark matter, with large values of $M/L$, extend well beyond the optical limits of galaxies.

We can now return to our original question: Does the universe contain enough invisible matter to raise the average density to $5 \times 10^{-30}$ gram per cubic centimeter, the value needed to close the universe and bring its expansion to a halt? As we have seen, such a density would be reached if the density of nonluminous matter exceeded the density of luminous matter by a factor of about 70. Alternatively what would be needed to close the universe can be expressed in terms of the ratio of total mass to luminosity. That value is roughly 700, compared with 1 for the sun.

Is there any evidence that the $M/L$ value of 700 is approached? Averaged over the visible disks of spiral galaxies, the ratio of total mass (luminous and nonluminous) to luminosity is about 5. For SO and elliptical galaxies the $M/L$ value is higher, on the order of 10. For double galaxies and small groups of galaxies the $M/L$ value increases to between 50 and 100. Analyses of galaxy motions in large clusters indicate $M/L$ values of several hundreds. This increase in mean $M/L$ value with increasing distance from the center of the system was first stressed a decade ago by Einasto, Ants Kaasik and Enn Saar of the Estonian S.S.R. Academy of Sciences, and also by Ostriker and Peebles and by Amos Yahil of the State University of New York at Stony Brook. So far there is no evidence for the existence of $M/L$ values above the critical one of 700 needed to close the universe. The highest of the derived values, however, comes tantalizingly close. Some physicists consider it significant that the inferred values seem to be converging on the critical one rather than being orders of magnitude either higher or lower.

Investigations encompassing gigantic distances and vast time scales are made more difficult by this new realization that the distribution of light is an unreliable guide to the distribution of mass in the universe. An unknown fraction of the mass in a spiral galaxy is hidden in a nonluminous constituent, and so is an unknown fraction of the mass in clusters of galaxies. One cannot yet state whether regions of the universe that are devoid of galaxies are simply voids of light or are voids of mass as well. To answer this question astronomers will have to be clever in devising novel observing techniques and physicists will have to determine the properties of exotic forms of matter. Only then will it be possible to establish the nature of the ubiquitous dark matter, to determine the full dimensions and mass of galaxies and to assay the likely fate of the universe.




**RATIOS OF MASS TO LUMINOSITY are plotted for aggregates of matter on progressively larger scales. The plot is based on one devised by Herbert J. Rood of the Institute for Advanced Study. The $M/L$ value of a density of matter sufficient to arrest the expansion of the universe is about 700. For galaxies values are below 10. The value rises with the size of the aggregate.**

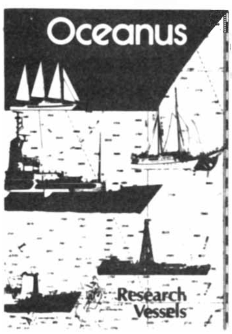
# Oceanus®

## *The International Magazine of Marine Science and Policy*

With offices in one of the world's leading marine research centers and published by the

Woods Hole
Oceanographic Institution,

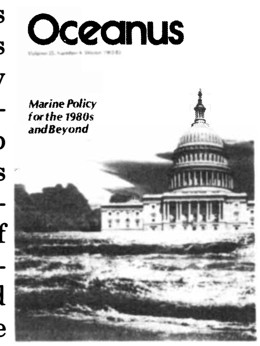
**Oceanus** magazine is in a uniquely favorable position to monitor significant ocean research and related policy issues.




A case in point is the recent discoveries of deep ocean hydrothermal vents at various locations in the Pacific Ocean. These discoveries are causing a revolution in Oceanography.

Many long-standing biological and geological theories are being revised as a result of this exciting research. Exploration of the vents by the submersible Alvin is turning up new life forms (derived chemosynthetically rather than photosynthetically) and possible new sources of metals, such as zinc, iron, and copper. Scientists are getting a first-hand look at how the earth breathes while garnering supporting evidence for the theories of sea-floor spreading and plate tectonics.

Economically, it is a time of great challenge, too. The cost of running submersibles and support ships, for example, is high and rising. And federal funding for marine research all the while is shrinking. Many critical decisions must be made in nearly every area of oceanic concern.




**Oceanus** is meeting this challenge by publishing articles that help our readers grasp the significance of present research and that expose them to the substance of important public policy questions.

Recent thematic issues have dealt with:

- Deep Sea Mining
- Research Vessels
- Sharks
- Coastal Problems
- Pollution
- Oceanography from Space

Upcoming issues will look at:

- Marine Birds
- Oil & Gas Activities
- Industry and the Sea
- Polar Research

and much, much more.

Handsomely illustrated with many sharp photographs and line art, **Oceanus** is edited for all those with a serious interest in the three-quarters of our planet that is covered by water. The general readership includes many students, educators, representatives from government and industry, environmentalists, conservationists, as well as scientists in the disciplines of marine biology, geology and geophysics, chemistry, ocean engineering, and physical oceanography.

In addition to the thematic editorial material, **Oceanus** also features:

PROFILES of selected oceanographers that give a glimpse of the human side of science.

A CONCERNS SECTION that exposes the reader to controversial subjects and opinions.

A BOOK REVIEW SECTION that gives readers a convenient way to choose new and timely reading material.

LETTERS TO THE EDITOR or a forum for debate on views and thoughts expressed in the magazine.

Won't you join the Editors of **Oceanus** — published quarterly — in the exploration of the last great wilderness on earth, our oceans. Subscribe today!

## Oceanus
### SUBSCRIPTION ORDER FORM

Please enter my subscription for:

☐ one year at $20.00
☐ two years at $35.00
☐ payment enclosed ☐ bill me
(prepayment preferred)

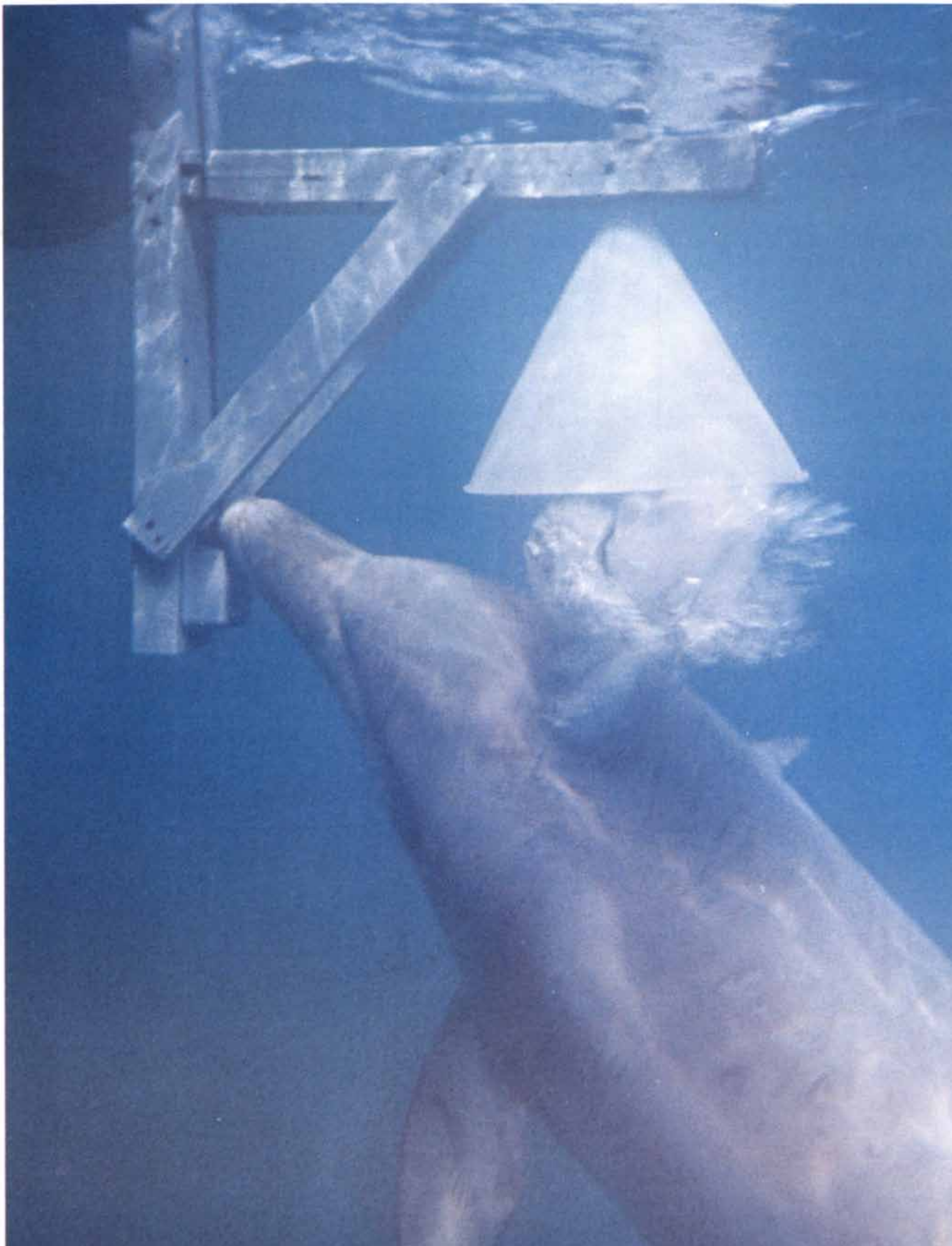
Please send MY Subscription to:

Name (please print)

Street Address

City State Zip

Please make checks payable to Woods Hole Oceanographic Institution. Checks accompanying foreign orders must be made payable in U.S. currency and drawn on a U.S. bank. (Outside U.S., add $3 per year to domestic rates). Mail to: OCEANUS, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543.



**BOTTLENOSE DOLPHIN,** one of the small whales trained by the U.S. Navy for underwater tasks, worked for the authors in experiments on the physiology of breath-holding. This dolphin, known to his trainers as Tuffy, submerged on signal and after various intervals responded to a second signal by nudging a buzzer and exhaling into a funnel, so that his remaining oxygen supply could be measured.

# The Physiological Ecology of Whales and Porpoises

*Like other mammals, these marine species have a high-energy way of life. In pursuing it they evolved particular adaptations of their own, notably the ability to dive deep for long periods*

by John W. Kanwisher and Sam H. Ridgway

It is not generally realized that 70 species of entirely marine mammals—the whales and porpoises—play as crucial a role in oceanic ecosystems as the terrestrial mammals do on land. The broad success of the land mammals at the end of the Mesozoic era (some 65 million years ago) was due in large part to their high-energy way of life, including such features as warm-bloodedness and an expanded central nervous system. When some of these terrestrial mammals later filled a marine ecological niche, they evolved further adaptations of the same high-energy type.

How have the whales and porpoises—the Order Cetacea—overcome the many physiological obstacles presented by living in the ocean? Consider, for one thing, their severely stressed life cycle. Initially expelled from a submerged birth canal into water that can be close to the freezing point, the newborn calf must struggle to the surface unaided before it can even take its first vital breath of air. To be suckled by its mother the calf must hold its breath and return underwater. When eventually it can feed itself, it may have to develop an entirely new behavioral repertory: diving in order to find its prey. Finally, once the cetacean is an adult, it must master a further repertory of complex gymnastics before it can reproduce and thereby give rise to the offspring that initiate the next life cycle.

Consider also the surprising diversity of the cetaceans, both in geographic distribution and in size. Whales and porpoises are plentiful in all the oceans, from the Tropics to the edge of the polar ice, both north and south. Many species even migrate seasonally from cold seas to warm ones and back again. The largest adult blue whale is some 20,000 times heavier than the smallest newborn porpoise, yet both large and small species are found in tropical and polar seas. The toothed whales—the odontocetes—include small porpoises such as the harbor porpoise at one end of their size range and the great sperm whale at the other. How is it that the sperm whale is the cetacean that dives the deepest and stays submerged the longest? How does it avoid the "bends" and other physiological problems human beings encounter when they dive to much shallower depths? The porpoises, being smaller than the smallest of their larger toothed-whale cousins, have a much higher surface-to-volume ratio. How do they manage to stay warm in cold waters? Questions of this kind led us to the first of several investigations: the thermal physiology of porpoises.

Our observations of captive porpoises at Marineland in Florida showed us that the animals were able to maintain a stable internal temperature close to the human one and similarly well regulated. Since water removes heat from an object much faster than air does, we found this observation intriguing. We had thought these thermally stressed mammals might exhibit the more variable body temperatures typical of terrestrial hibernators. This not being the case, we were led to calculate the stringency of the thermal problem faced by a small porpoise living in the cold waters of high latitudes.

For the purpose of heat-flow calculations a cetacean may be regarded as a uniformly warm fleshy core surrounded by a surface layer of insulating blubber. The outward flow of heat from the core, encountering the thermal resistance of the blubber layer, must be great enough to maintain the temperature difference between the surrounding water and the warm inner core. It is known that this difference is equal to the product of the thermal resistance of the blubber times the outward flow of metabolic heat generated in the core. When the insulation is insufficient, more body heat must be generated if the animal is to maintain a constant core temperature. For example, if the core temperature of a terrestrial mammal falls by as little as .5 degree Celsius, the mammal will begin to shiver in order to increase its metabolism. Over longer periods any such imbalance can be corrected by increasing the amount of insulation. The land mammal grows a winter coat and the whale in polar waters develops a thicker layer of blubber.

When the animal is active, such an increased thermal barrier may actually be too much of a good thing. For example, when a whale swims fast, the increase in its metabolic activity will cause its core to overheat. The whale then has recourse to a circulatory stratagem. An increase in blood flow near the body surface, particularly through the flippers and flukes, thermally bypasses the insulating blubber and returns the core temperature to normal. In cetaceans generally a steady body temperature is mainly achieved by such changes in blood flow. The same is true of human beings: when they are too warm, more blood is shunted to the surface of the body, producing the flushed appearance that is the opposite of being "blue from the cold."

There are fundamental differences, however, between the physiology of human beings and the physiology of whales. For example, cetaceans have no sweat glands; evaporative cooling in an aquatic environment is impossible. By the same token, for their surface insulation human beings have only a very poor equivalent of blubber. Human divers have now learned to imitate whales by wearing insulation in the form of foam-rubber suits. Ashore human beings overcome the handicap of their essentially tropical origins by covering themselves with, among other things, winter insulation grown by other terrestrial mammals.

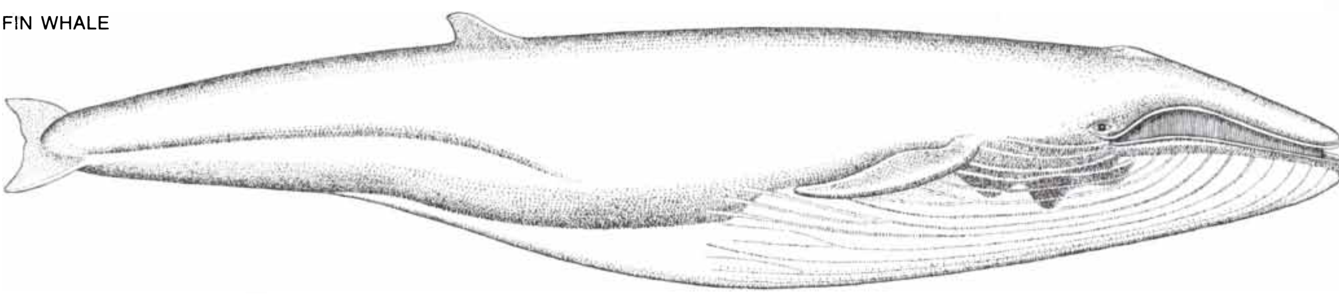
Even before we began our investigations we were well aware that the smallest cetaceans would be the ones that faced the greatest problem in maintaining a normal mammalian temperature. Considering the total surface area that is constantly losing heat to the surrounding water, the smaller the animal, the smaller the core volume that generates the metabolic heat required to maintain the necessary temperature differential across the blubber. In addition to the disadvantageous geometry of surface-to-volume ratios, considerations of linear scaling also work against any small porpoise trying to stay warm in cold water. Because of their smaller dimensions, the smaller animals will tend to have a thinner layer of blubber.

Regardless of their size, cetaceans maintain a normal mammalian internal temperature of about 37 degrees C. In the Arctic, where the water temperature may be −2 degrees C., the temperature difference across the blubber barrier is nearly 40 degrees. In the Tropics, where the water temperature may be 30 degrees C., the temperature gradient across the barrier would be only a fifth of that in Arctic waters. Thus a porpoise adapted to warm water would seem to be headed for serious thermal trouble on entering polar seas. Yet one of the smallest cetaceans, the harbor porpoise (*Phocaena phocaena*), is found mostly in cold northern waters.

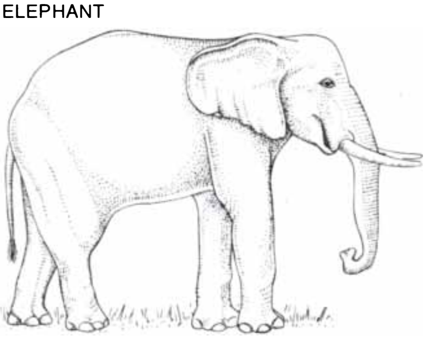
To determine the dimensions of the problem faced by harbor porpoises in cold water we calculated how much metabolic heat they would have to generate in order to keep warm. To determine the total surface area and the thickness of the porpoise's insulating blubber we made direct measurements on a dead harbor porpoise that had washed ashore. When our figures were entered in the formula for heat loss, the only conclusion to be drawn was that with only a normal mammalian metabolic heat supply a porpoise that small could not keep warm in northern waters. Yet there the harbor porpoises were, sporting in the same chill seas that should have killed them.

Such a paradox is upsetting to biologists who have come to expect orderliness when they examine any parameter common to different animal groups. The amount of metabolic heat production in animals of different sizes is one such comparison that seemed well enough understood. For example, the

FIN WHALE



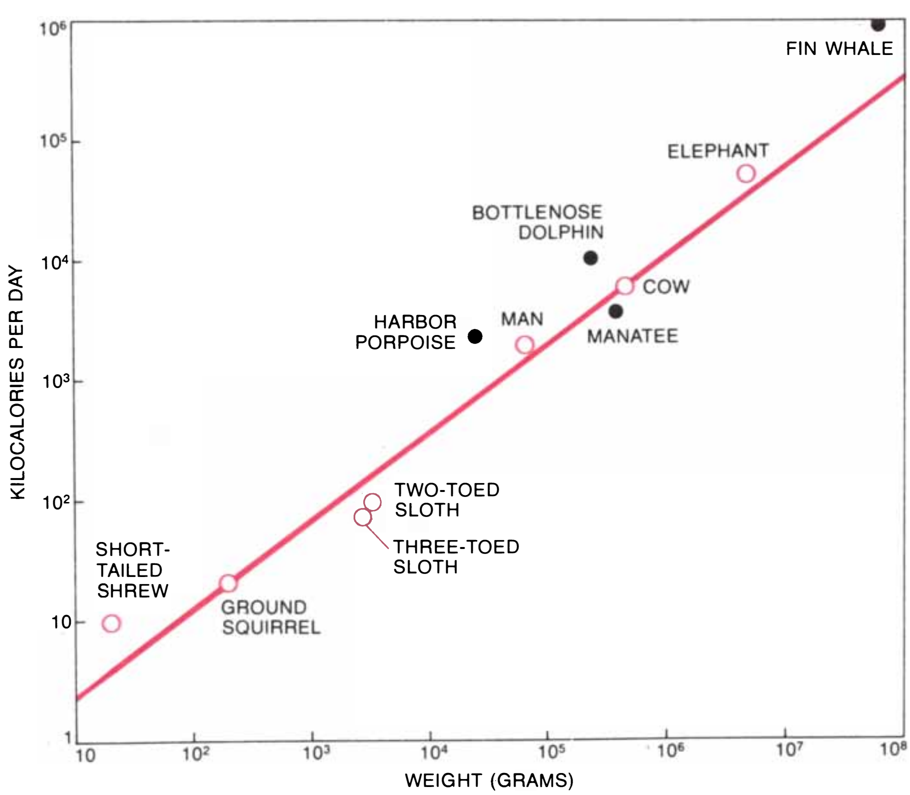
ELEPHANT



BOTTLENOSE DOLPHIN

HARBOR PORPOISE

GROUND SQUIRREL




**"MOUSE TO ELEPHANT" CURVE, shown as a solid line (*color*) in the graph at the right, was devised in the 1930's by Francis G. Benedict of the Carnegie Institution of Washington as a rough means of estimating the basal metabolism of various mammals, expressed in kilocalories per day, on the basis of 75 percent of their weight. Large colored circles show how selected terrestrial mammals closely fit the curve, which is constructed on a logarithmic ordinate and abscissa. (The ordinate has a top value of one million kilocalories and the abscissa extends to 100 million grams.) Direct measurement of the metabolism of harbor porpoises (*Phocaena*) and bottlenose dolphins (*Tursiops*) and an estimate for fin whales (*Balaenoptera*) based on lung volume (*black dots*) all lie well above the terrestrial-mammal curve. Weight for weight, however, the metabolism of fin whales is lower than that of such sluggish terrestrial mammals as sloths.**

metabolic rates of numerous terrestrial mammals have long been arrayed in what is called the "mouse to elephant" curve. It was believed that given the animal's body weight one could closely estimate from this curve the metabolic heat production of any mammal. Our estimate of the metabolic rate needed to keep the porpoise warm in cold water, however, was so high that it was in clear disagreement with the data from terrestrial mammals.

There were two possible ways to explain the paradox. The first was, quite simply, that the metabolism of the cetaceans does not conform to the terrestrial mammals' mouse-to-elephant curve. This is to say the cetaceans may have an intrinsically higher rate of basal metabolism that allows even a small cetacean to survive in cold water. The second possibility is that cetaceans have a "normal" (which is to say a terrestrial) basal rate of heat production but that by exercising all the time when they are in cold water they can generate enough extra heat to keep themselves from getting too cold. To us the only way to choose between these two explanations was to present the question to the porpoises. Which of these thermal strategies had evolution chosen for them?

With the skilled help of Gunnar Sundnes, a physiologist at the Havforskningsinstitutt in Bergen, we set up a heat-loss experiment. First we learned how to keep young harbor porpoises more or less contented in captivity in a large tank. We then transferred them for testing to a smaller box, filled with seawater, that served as a whole-body calorimeter. The young of the species weigh as little as 15 kilograms. Although they are born with blubber that is only one centimeter thick, the difference between their normal body temperature and the temperature of the surrounding seawater can be as much as 25 degrees C.

The porpoise being tested would lie quietly in its box as the metabolic heat it was generating passed through its insulating layer of blubber and warmed the surrounding water. We found we could put a thermistor in the box and with a little amplification produce a curve plotting temperature against time on a chart recorder that was sensitive to within .001 degree C. Taking into account the volume of water in the box, the slope of the temperature curve provided a running record of how much metabolic heat the porpoise was losing at any moment.

Our results proved to be quite clearly in favor of the first of our two possible explanations: these small porpoises had a metabolic rate two to three times higher than the rate of terrestrial mammals of the same weight. This increased heat production is not achieved without cost: the small porpoises must obtain three times as much food as a terrestrial mammal of the same weight and must rise to the surface and breathe three times more often than they would if their metabolic rate were terrestrial.

On the basis of the porpoise data we went on to make similar thermal calculations for larger cetaceans to see how much more easily a large mammal, with its more favorable geometry, could keep warm in cold water. We first considered a theoretical whale weighing 1,000 times more than a harbor porpoise; since the two sea mammals are similar in shape, the whale's linear dimensions must increase about tenfold. The theoretical whale therefore has 1,000 times more tissue to generate metabolic heat but the heat passes out to the surrounding water through a surface area only 100 times larger than the porpoise's.

Consider this diminished heat loss in terms of blubber insulation only. Whereas the small porpoise could balance its heat loss with the aid of a blubber layer two centimeters thick, the 1,000-times-heavier whale, given the same metabolic rate as the porpoise, could make do with a blubber layer no more than two millimeters thick. Actually, however, such a whale would have a blubber layer some 20 centimeters thick and even in polar seas would face a cooling problem. Such a large cetacean appears to be 100 times overinsulated.

If the large cetacean's blubber layer is far thicker than considerations of insulation require, what other functions does it serve? First, the fatty tissue, which is low in specific gravity, provides enough buoyancy to offset the negative buoyancy of the whale's muscle and skeleton. Second, the blubber constitutes a food store large enough to take care of the whale's metabolic needs on the animal's seasonal migrations, often across thousands of miles of food-poor waters.

Because of the general rule that relates the metabolic rate of mammals to their size, we would expect a whale to have a rate of heat production per kilogram of tissue considerably lower than the rate we had established experimentally with harbor porpoises. We could see no way to check this expectation by direct measurement of the metabolic rate of a 40-ton animal, and so we were left with estimating the whale's heat production on the basis of anatomical considerations only.

At Steinshamn, a Norwegian coastal whaling station, we dissected out the lungs of a 48-foot fin whale (*Balaenoptera physalus*). We determined the whale's lung volume by first inflating the organs with compressed air and then emptying them through a gas meter. We assumed that the volume of expelled air was equal to the volume of air the whale inhaled with each breath. From observations at sea we were able to determine how often a fin whale breathes. The normal diving pattern for the species is to surface for five quick breaths and then to stay underwater for five minutes; thus in effect a mature fin whale breathes on the average about once a minute.

From the combination of lung volume and breathing rate we could estimate the amount of air a fin whale breathes: 2,000 liters per minute. Now, mammals in general utilize about 6 percent of the oxygen in the air they breathe (21 percent of the total atmospheric mixture of nitrogen, oxygen and other gases). Because each cubic centimeter of oxygen the animal consumes will generate five calories of heat, we were able to estimate, albeit crudely, the fin whale's rate of metabolism. The figure we got was, like that for the harbor porpoise, significantly above the curve for terrestrial mammals. On a weight-for-weight basis, however, the fin whale's metabolic rate is even lower than that of sloths, which have a rate reflecting their lethargic movements.

When one takes into consideration both the large whales' lower absolute metabolic rate and their ability to accumulate a food reserve in the form of blubber, it becomes evident that all large whales are relatively independent of local food supplies. They can satisfy their energy needs for three to six months simply by living off half of their supply of blubber. This leaves them free to feed in the rich polar seas and to give birth to their young thousands of miles away in warm tropical waters.

In comparison the small harbor porpoise leads a much more restricted life. Its higher metabolic rate makes its need for food much more pressing and limits the time available for migration. Moreover, its higher rate makes it more difficult for it to hold its breath: when the porpoise dives to search for food, it must depend on such oxygen as it can take along. Just how the porpoise makes do with what oxygen it has became the focus of our next investigations.

We already knew that the mammals' high-energy way of life had been made possible by a correspondingly larger consumption of oxygen; this in turn depended on a parallel evolutionary development of respiratory and cardiovascular systems that were more competent than those of amphibians and reptiles. When the terrestrial mammals that were ancestral to the cetaceans abandoned the land (with its unlimited access to oxygen) in favor of an aquatic life, a still further evolutionary development favoring greater oxygen storage and transport was called for. How had this been accomplished?

Consider the sperm whale, the cetacean with one of the most impressive of all underwater capabilities. Whereas the porpoise can stay submerged for only a few minutes, sperm-whale dives can last for as long as an hour. Yet even the porpoises' briefer underwater performance puts an unassisted human diver to shame. The answer, for large and small cetaceans alike, is that they can take more oxygen down with them than a human diver can.

One factor in this performance is the cetaceans' considerably greater volume of blood: up to two or three times more blood per unit of body weight than is the case in human beings. The cetaceans' blood also has a somewhat greater oxygen-carrying capacity than human blood. In addition the whale carries a further store of oxygen in combination with myoglobin in its muscles. The myoglobin (a protein that is actually a subunit of the oxygen-carrying hemoglobin of the red blood cells) is responsible for the muscles' characteristic dark red color. Indeed, in some cetacean species there is enough myoglobin in the muscles to carry more than half again as much oxygen as the animal's red blood cells do.

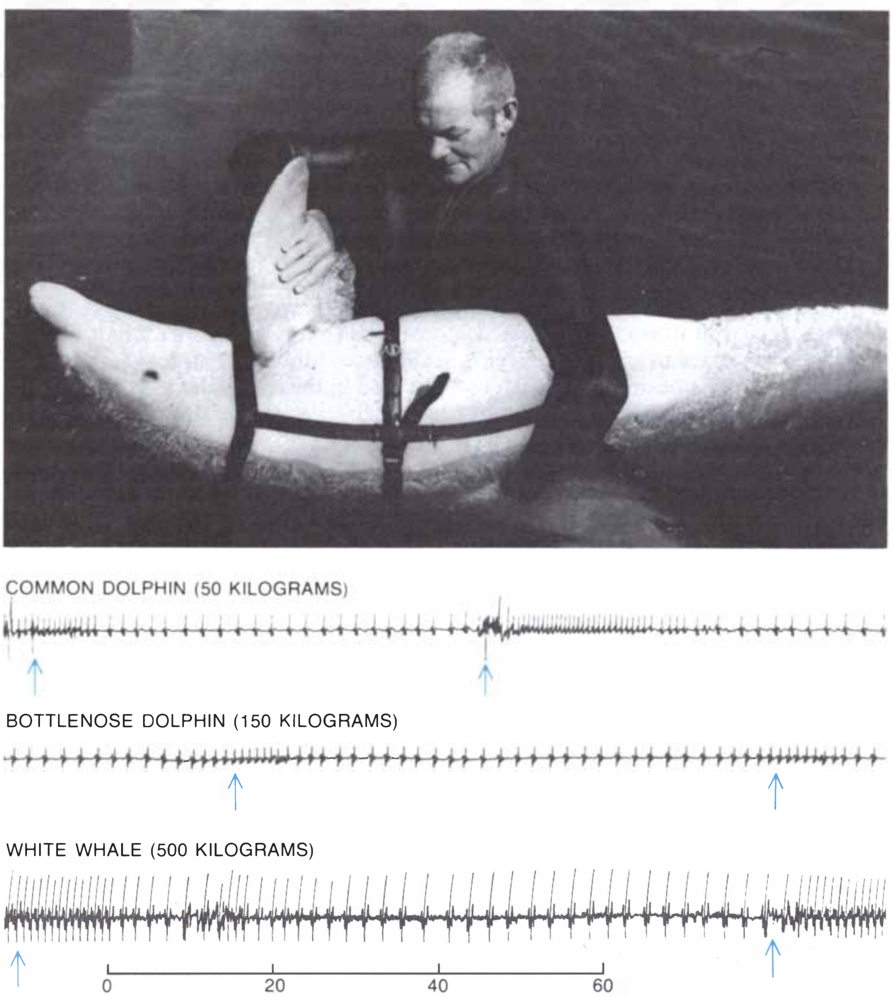
When a whale comes to the surface after an extended dive has depleted both these oxygen stores, its highly efficient vascular system rapidly passes the blood through the lungs for oxygen replenishment with the aid of a particularly powerful heart and a circulatory system that features large networks of capillaries (*retia mirabilia,* or "wonderful nets") and large sinuses in the venous system. Although the exact function of these unusual structures can only be guessed at, it seems likely that they facilitate the rapid reloading of the red blood cells with oxygen. The fast circulation of the blood is reflected in the fact that while the animal is breathing at the surface it has a higher heart rate. This constitutes a fundamental respiratory difference between cetaceans and terrestrial mammals, whose breathing is usually regular and whose blood flow is relatively smooth.

Meanwhile the recharged red blood cells convey part of their oxygen load to any oxygen-depleted myoglobin in the animal's muscle cells. The transfer is easy because myoglobin has a much lower oxygen-loading tension than hemoglobin. Myoglobin thus binds oxygen more tightly than hemoglobin does, inhibiting it from reentering the circulatory system. It is best to think of the myoglobin-bound oxygen as a purely local oxygen resource, available to the surrounding muscle in the course of a dive. Most of it lies within easy diffusing distance of the muscle cells' mitochondria: the intracellular particles where respiration occurs.

Myoglobin is present in the muscles of terrestrial mammals too, but in much smaller amounts. One of its functions in terrestrial mammals is to serve as a local oxygen source for muscle during contraction, when intramuscular pressure tends to throttle capillary blood flow, cutting off the normal supply of oxygen carried by red blood cells. It is evident that the much more abundant supply of oxygen carried by cetacean myoglobin is another example of a diving adaptation that has arisen through the enhancement of a preexisting mammalian capability.

When we began our study of cetacean diving capabilities, a widely accepted explanation of the long dive times characteristic of aquatic animals already existed. It derived from experiments done by P. F. Scholander of the Scripps Institution of Oceanography demonstrating that when a restrained animal (a duck or a seal) was forcibly submerged, its heart rate decreased abruptly and the reduced blood flow was shunted in such a way as to largely shut off the circulation of blood to the animal's muscles and viscera. The reduced cardiac output went almost entirely to the brain and the heart, organs that were considered to have a need for a continuous supply of oxygen. (For example, a few seconds' interruption in the flow of blood to the human brain causes unconsciousness.) All that was available to the oxygen-deprived muscles of Scholander's subjects was an anaerobic energy pathway that resulted in an accumulation of lactic acid as a metabolic by-product.

This logical sequence of events has become the standard textbook explanation of the extended diving capacity not




**TOOTHED WHALES' HEARTBEATS WERE RECORDED, by telemetry for bottlenose dolphins such as the one carrying an ultrasonic transmitter in the photograph at the top and directly in the case of the remaining electrocardiograms. These show the heartbeat of a 50-kilogram common dolphin (*Delphinus delphis*) and that of a 500-kilogram white whale (*Delphinapterus leucas*). Arrows in color indicate the start of inhalation in each instance. The heartbeat of all three animals accelerated as they inhaled, an action that would have accelerated oxygen uptake. Thereafter, whether the animals were diving, were resting on the surface or were out of the water altogether, their heartbeat decelerated until a basal rate was reached. Earlier experiments (conducted by the physiologist P. F. Scholander with seals and ducks under restraint) had shown a radical decrease in heart rate following involuntary submergence, but the three cetaceans did not show such a decrease. The authors' findings suggest that the reaction of animals under restraint was one of alarm as they were involuntarily submerged.**

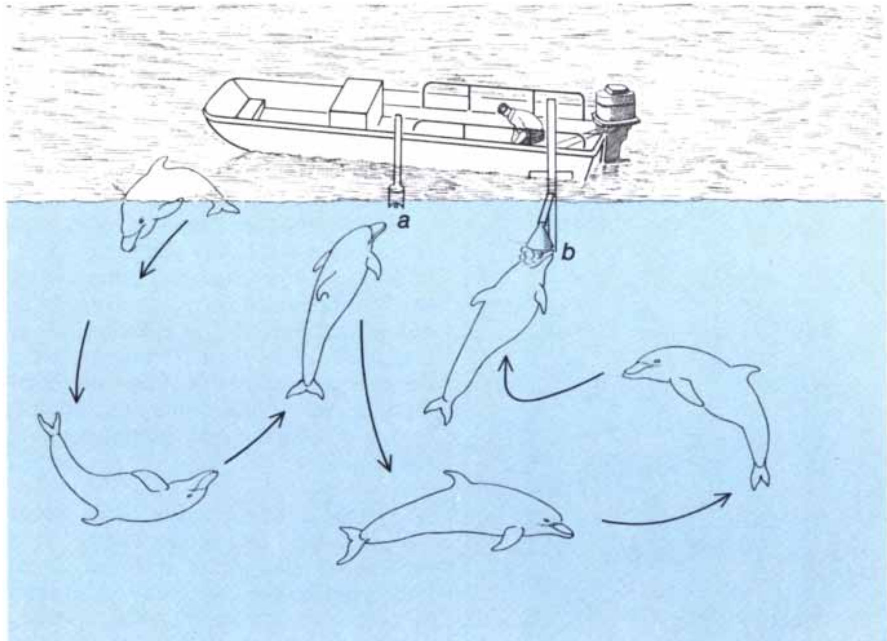
only of ducks and seals but also of other waterfowl and of marine mammals in general. There were, however, virtually no physiological data for cetaceans. Indeed, the one time a restrained porpoise was experimentally deprived of air in this way it went into shock and died. We undertook to collect evidence on this hypothetical diving reflex from free-swimming cetaceans.

As a first step we fitted porpoises with ultrasonic transmitters so that we could monitor the animals' heartbeat as they swam. We were surprised, first of all, to find no evidence of the abruptly diminished heartbeat on diving that was characteristic of forcibly submerged ducks and seals. The porpoises seemed to maintain a relatively high level of general circulation throughout their dive. We therefore set out to find a more direct measure of how quickly the porpoises depleted their store of oxygen while they were submerged.

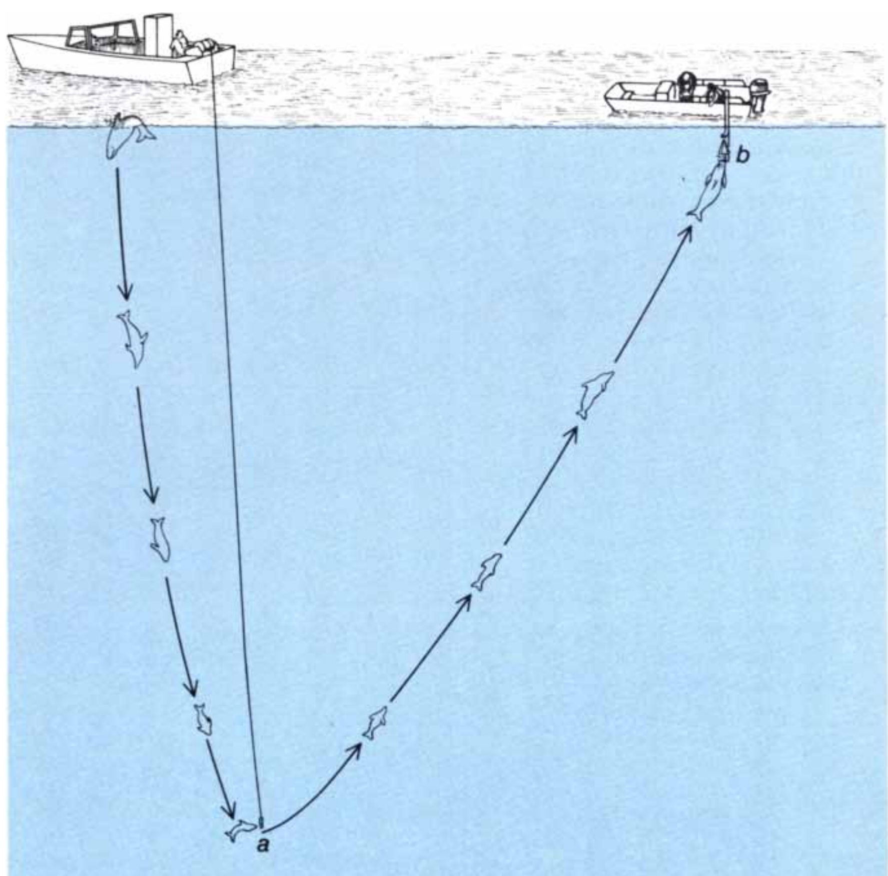
Fortunately for us the U.S. Navy at that time was training bottlenose dolphins (*Tursiops truncatus*) to act as messengers for deep-sea divers. One Navy trainer, William Scronce, undertook to train one of the dolphins, named Tuffy, to exhale on command into an underwater funnel so that we could collect the expired breath for analysis. An underwater buzzer was Tuffy's signal to perform. If he pushed his snout against the buzzer and exhaled into the funnel, he received three fish as a reward. Another buzzer served as a signal for Tuffy to begin his dive. By varying the interval between the sounding of the "dive" buzzer and the "collection" buzzer we could obtain air samples over an entire range of breath-holding times. Tuffy could easily hold his breath for as long as six minutes, but after a much shorter time he grew increasingly impatient as he waited for the "collection" buzzer to sound; looking up through the water at us, he would even gnash his teeth.

Studying the data Tuffy provided us, we were surprised to see how quickly the oxygen concentration in the porpoise's lungs dropped off. For example, after only three minutes the gas in Tuffy's lungs had fallen from its normal atmospheric level of 21 percent oxygen to less than 2 percent. For the remaining three minutes of these "easy" dives, therefore, he was virtually anaerobic. After such a breath-hold he would not perform again until he had spent four or five minutes breathing rapidly on the surface, obviously repaying the oxygen debt he had just incurred. After dives of two minutes or less, however, he would readily dive again following only a brief breathing interval at the surface.

We could now understand why we had failed to observe an abruptly diminished heart rate when our ultrasonically equipped porpoises dived. Like Tuffy, they had made only a partial effort, if




**NAVY-TRAINED DOLPHIN TUFFY** was taught as is shown here to submerge when one buzzer (*a*) sounded and to stay underwater until a second buzzer (*b*) sounded. Tuffy then swam to the second buzzer, touched it with his snout and exhaled into a funnel. After he exhaled he received a reward of three fish. The study showed that Tuffy could remain underwater comfortably for six minutes but had exhausted almost all his oxygen supply within three minutes.




**DEEP-DIVING ABILITY OF THE DOLPHIN** was tested in a second experiment. First a buzzer and a switch (*a*) were lowered to a depth of 150 meters. Turning the buzzer on was Tuffy's signal to dive and switch it off. The dolphin then returned to the dinghy, to touch a second buzzer (*b*) and claim his reward. Tuffy had no difficulty with 150-meter dives but needed time at the surface to recover after turning off the "dive" buzzer at a depth of 300 meters.

any at all, to conserve their initial store of oxygen. That store was quickly consumed; thereafter an oxygen debt developed in all their tissues, including the brain. We found that all the aquatic air breathers we observed (diving birds, seals and manatees in addition to our cetaceans) tried to avoid such an oxygen debt by limiting their dives to short periods in order not to exhaust their supply of oxygen. For example, when Tuffy and his dolphin companions swam without instructions to the contrary, dives lasting for more than 2.5 minutes were the exception rather than the rule. Submerging for shorter intervals, however, the dolphins might continue diving for hours.

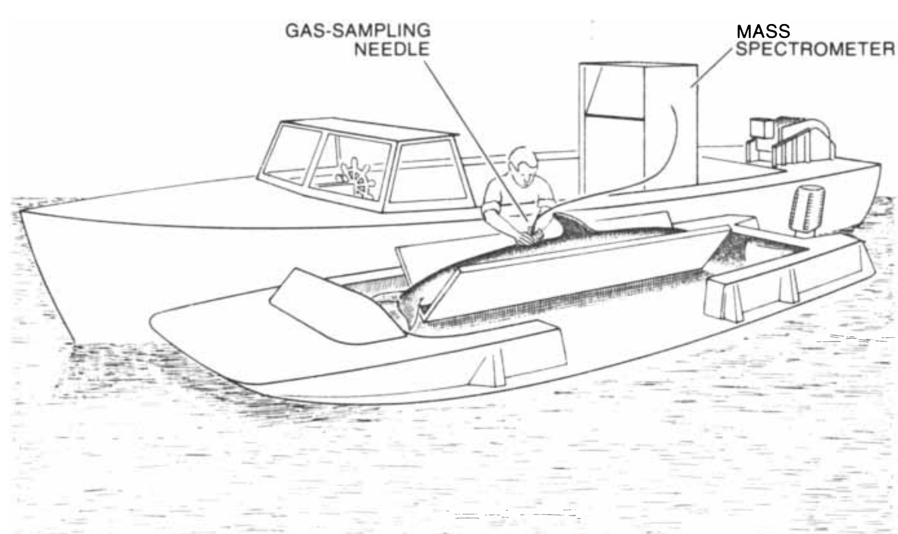
When a human diver has spent enough time underwater for a substantial amount of nitrogen to go into solution in his blood and then he comes up too rapidly, the nitrogen forms tiny bubbles that can cause painful symptoms ("the bends") and even death. How did our porpoises, let alone the large whales that may dive deeper than a kilometer for more than an hour, avoid the bends? With this question and related ones in mind we arranged some deep-diving experiments for our trained bottlenose dolphin.

One buzzer, controlled from the surface, we lowered on a line to one or another of a variety of chosen depths. Tuffy was trained to swim down to the buzzer when we sounded it and turn it off by pushing a switch with his snout. When the animal turned off this "dive" buzzer, Scronce immediately sounded the "collection" buzzer at the surface. Tuffy then quickly swam up from his dive to the "collection" buzzer and exhaled into the collection funnel.

The scientific literature on porpoises suggests they do not dive very deep. When we began our experiments, we lowered the "dive" buzzer to a depth of 150 meters. Tuffy easily managed trips to this depth, came up to give us his gas sample and sometimes demanded all three of his reward fish before even bothering to inhale; he was clearly far from his depth limit. When we lowered the "dive" buzzer to 300 meters, we found that Tuffy needed a much longer recovery time. He was obviously approaching his limit. (Navy trainers in Hawaii have since trained a related species to descend to nearly 600 meters.)

Tuffy seemed to be untroubled by the increasing water pressure at greater depths (as much as 30 atmospheres, or 440 pounds per square inch, at a depth of 300 meters). We expected that at such a high external pressure the animal's lungs would be almost completely collapsed. By rigging the "dive" buzzer with a flash camera we were able to record Tuffy's appearance as he turned off the buzzer 300 meters below the surface. We found that his rib cage was indeed much deformed by the hydrostatic pressure pushing on his lungs.

We next tried to give another trained dolphin the bends by directing it to dive repeatedly. As soon as its surface breathing indicated its recovery from the previous dive we would send it down




**THIRD EXPERIMENT was aimed at discovering whether a bottlenose dolphin is susceptible to "the bends" (the painful symptoms that can result from the absorption of atmospheric nitrogen by the blood in the course of a long dive). On signal the dolphin dived repeatedly to a depth of 100 meters, spending more than 20 minutes out of an hour at an average depth of 50 meters. The animal then swam onto a floating couch next to a boat containing a mass spectrometer; the experimenters pushed a gas-sampling needle through its blubber layer and into its muscle. The dolphin's muscle proved to be saturated with nitrogen to an extent dangerous to a human diver. The authors attribute the dolphin's freedom from the bends to a diving adaptation.**

again, until in the course of an hour it had remained below 50 meters for more than 20 minutes. The dolphin still showed no apparent sign of distress. Then, with the help of Robert S. ("Red") Howard of the Scripps Institution, we measured the actual amount of nitrogen accumulated in the animal's muscles as the result of these many consecutive submergences. It was no easy task. First the dolphin was trained after such a sequence of dives to swim onto a floating couch alongside our boat. That enabled us to insert a small gas-sampling needle through its blubber layer into the underlying muscle. A tube led from the needle to a mass spectrometer (aboard a second boat) that could directly measure the concentration of nitrogen dissolved in the muscle. Doing all this while at sea in small boats makes a substantial demand on the experimenter's patience.

Again we got a surprise. The concentration of dissolved nitrogen in the dolphin's muscle tissue was indicative of a degree of supersaturation that in a human diver would have been dangerous. We could only conclude cetaceans can tolerate nitrogen concentrations that would give a human diver the bends. How they do so we cannot yet say, but the capability is, of course, what one should probably expect of a well-adapted order of deep-diving mammals.

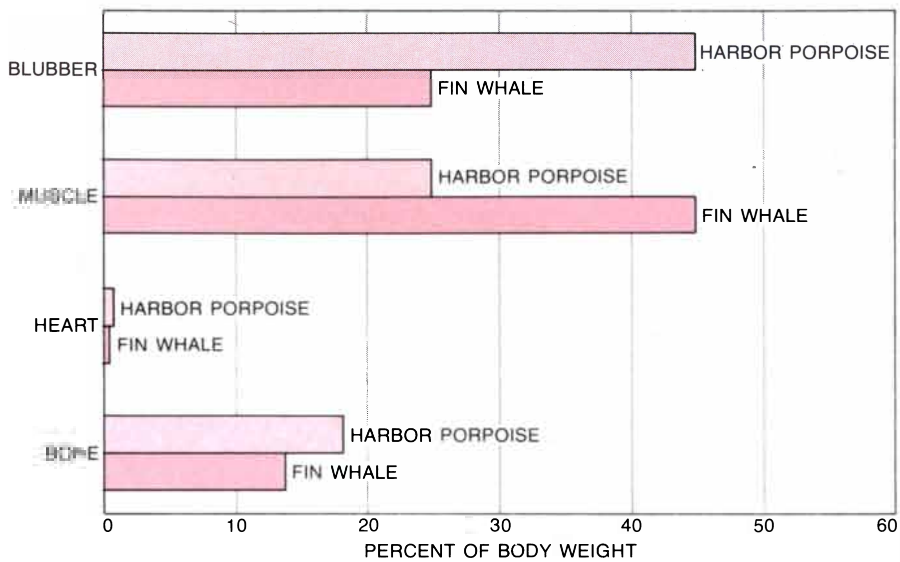
What has the evolution of this superior diving capacity meant to the ecology of the world's oceans? Some aquatic mammals, for example the sea otter and the manatee, have evolved to more or less nibble at the edges of the oceanic food chain. Others, such as the carnivorous suborder of pinnipeds (seals, sea lions and walruses), are more at home in the open ocean but still need to go ashore for mating and reproduction. Only cetaceans have evolved to deal with a fully oceanic way of life. What they gained thereby is a vast food supply in the form of other marine animal life. The whales and porpoises have established themselves as harvesters at the very top of the marine food chain.

The success of these physiologically adept marine predators may be judged by their numbers. Many species of cetaceans are commonly thought to be in imminent danger of extinction; this conclusion is based largely on a misperception of their current population levels. For example, the total number of blue whales is greater by at least one order of magnitude than the total number of such endangered terrestrial species as the orangutan or the giant panda. In fact, most cetacean species exist in substantial numbers, and some populations are even reaching the level where they serve as entertainment. Whale watching is a growing winter sport on the California coast, and in the Atlantic excursion boats putting out from the Boston area

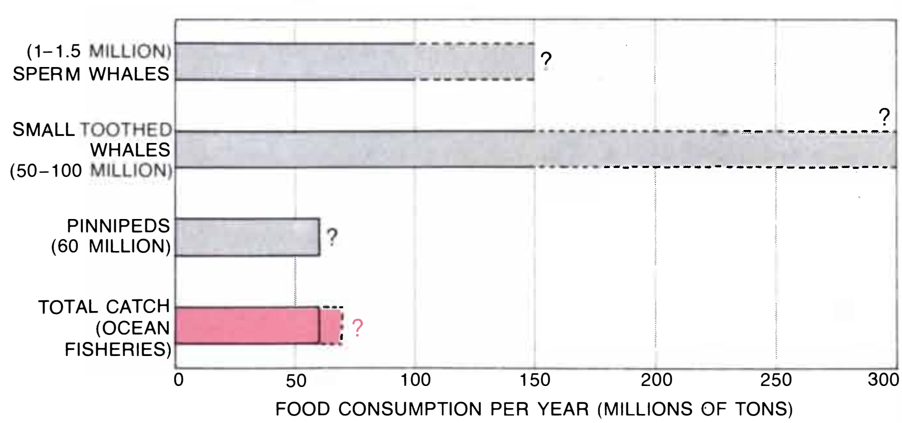
assure tourists of frequent whale sightings after a trip of only an hour or two.

Sperm whales were probably twice as numerous in the generations before modern whaling began than they are today, but even so their modern number is far from low. The British cetologist Malcolm R. Clarke reckons the present population of sperm whales at about a million (others say 1.5 million) and estimates that they annually consume some 100 million tons of deep-sea squid, an oceanic resource that is otherwise scarcely touched. It is instructive to compare the predation of this single cetacean species with the total of the world's annual fishery yields: 60 to 70 million tons.

An annual-consumption estimate for the next-ranking aquatic mammals, the pinnipeds, has been made by Michael E. Q. Pilson of the University of Rhode Island. Pilson's estimate has the advantage of better-known population numbers, since pinnipeds must come ashore for breeding and can be more accurately counted than the 70 cetacean species. Censuses have been taken at most pinniped breeding areas, and their annual food needs are known from the feeding of captive species. Pilson estimates that the seals, sea lions and walruses consume 60 million tons of fish per year, or approximately the amount harvested by man. Extrapolating from the number of porpoises killed annually by the opera-




**DIFFERING PROPORTIONS OF TOTAL BODY WEIGHT devoted to various components are shown in this graph. A small odontocete, the harbor porpoise, is compared with a large mysticete, the fin whale of cold polar waters. Only about a fourth of the porpoise's weight consists of muscle, whereas nearly half of its weight consists of blubber. The geometry of surface-to-volume ratios provides the fin whale with a thick blubber sheath accounting for only a fourth of its total weight, whereas nearly half of the total consists of its muscle tissues.**




**ECOLOGICAL ROLE OF MARINE MAMMALS appears to be at least as significant as that of the land mammals. The graph compares the marine mammals' estimated "harvest" of other marine organisms with the total recorded international catch of fish. A million sperm whales consume some 100 million tons of squid each year (more if sperm whales are actually more numerous). The 50 to 100 million porpoises and other small odontocetes consume between 100 and 300 million tons of fish. Estimates of consumption by whales other than the sperm are less reliable, but seals, sea lions and walruses consume 60 million tons of fish annually. The total catch by marine fishermen worldwide is somewhere between 60 and 70 million tons (*color*).**

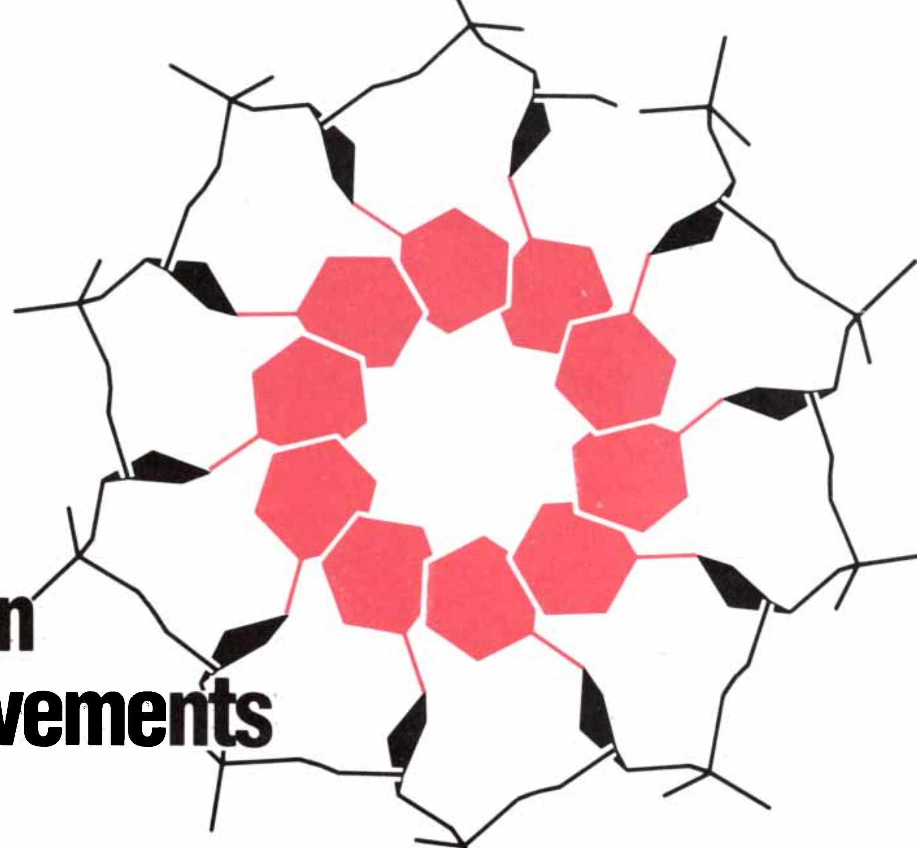
tions of the tuna fishery in the eastern Pacific, the worldwide stock of porpoises and dolphins may approach 100 million. Even if the population is only half this size, its annual food consumption must total some 150 million tons.

Such large numbers may come as a surprise, and we hasten to add that they must be considered tentative. Indeed, unless oceanic animals are harvested by man either accidentally or deliberately, the size of their populations is not likely to be well known. Two examples will suffice. Until recently Fraser's dolphin was known only from a single skeleton collected in 1954. Now that people are aware of the species, Fraser's dolphins have been sighted in schools numbering in the hundreds in Philippine waters, and the animals appear to be widely distributed in the tropical zones of the Pacific and Indian oceans. The North Pacific population of Dall's porpoise was estimated at only a few tens of thousands. Since it became known that Japanese tuna fishermen kill as many as 10,000 Dall's porpoises every year, new surveys have raised the population estimate to a million or more.

Thus it is evident that the cetaceans' role in marine food chains is far greater than man's. This does not mean that man's overfishing has not heavily depleted local fish stocks, as it has the stocks of some whale species, particularly in the Antarctic. It does indicate, however, that human harvesting, being largely confined to a few highly productive areas, has probably not seriously affected the gross nutrient dynamics of most ocean areas.

One feature of gross nutrient dynamics is upwelling: the lifting of nutrients from deep waters such as that accomplished by the rising currents off southern Africa in the Atlantic and off South America in the Pacific. Cetaceans play an analogous role on a surprisingly large scale. All other life in the sea tends to move nutrients downward. Tied to the surface by their air-breathing, marine mammals tend to move nutrients upward. Even the whales' fecal output does not move downward: because it is liquid, it tends to disperse rather than sink when it is released.

To those of us who are interested in the whales' well-being for emotional as well as biological reasons there is new cause for hope. Hunting these animals in distant seas is an energy-intensive activity. As the commercially desirable baleen whales have become harder to find, the harvesters' costs, largely in fuel consumed, have risen to a point where economics will increasingly discourage the activity. Ironically the Organization of Petroleum Exporting Countries (OPEC) may in the end do more than conservationists can to relieve hunting pressure on these remarkable marine mammals.



# A celebration of the achievements of a science

# BIOCHEMISTRY

SECOND EDITION

## LUBERT STRYER

*Stanford University*

*949 pages*
*983 illustrations*
*$32.95*

Lubert Stryer's *Biochemistry* has become the standard against which all other biochemistry books should be measured. It has shaped and continues to shape the subject it teaches.

First published eight years ago, *Biochemistry* immediately set new standards in the teaching and presentation of biochemistry. Emphasizing the process of discovery, Stryer aimed "to provide an introduction to the principles of biochemistry that gives the reader a command of its concepts and language." Reviewers commented that it was "doubtful whether there will be another text which will appeal to the student as strongly as this one does" (*Nature*).

The Second Edition, "bigger and even more splendid," is "a most handsome and thought-provoking new edition of a fine textbook" (*TIBS*). "It continues the tradition of excellence established in his first edition" (*American Scientist*), brilliantly achieving the author's aim as stated in the Preface: "to weave new knowledge into the fabric of the text...to enhance the book's teaching effectiveness by centering the exposition of new material on common themes wherever feasible and by citing recurrent motifs....and to convey a sense of the intellectual power and beauty of the discipline of biochemistry."

With a clarity and conciseness achieved through the unified narrative of a single author, "....the brevity of the text often belies its depth: the author has a unique ability to expound difficult concepts in a few words." (*Nature*)

Stryer's *Biochemistry*, Second Edition—giving up-to-date coverage of new developments while communicating the excitement and progress of biochemical frontiers.

If you are not using Stryer, compare *Biochemistry* with your present text. After the comparison, even *more* the choice.

**W. H. Freeman and Company**
41 Madison Avenue
New York, N.Y. 10010

# Georg Cantor and the Origins of Transfinite Set Theory

*How large is an infinite set? Cantor demonstrated that there is a hierarchy of infinities, each one "larger" than the preceding one. His set theory is one of the cornerstones of mathematics*

by Joseph W. Dauben

The nature of the infinite has always been a controversial topic. In antiquity it appeared in the famous paradoxes of Zeno of Elea, who argued with discomfiting clarity that motion is impossible because it requires an object to pass through an infinite number of points in a finite time. The success of Newtonian physics is in large part a result of Newton's introduction of the calculus of infinitesimal rates of change, yet no mathematically rigorous formulation of this awkward but effective idea was offered for more than 200 years. In modern times problems associated with the infinite have appeared in the abstract theory of sets, a theory that provides a foundation for virtually all contemporary mathematics. Moreover, throughout history the idea of the infinite has had theological overtones that have played a role in the acceptance or rejection of the concept and the mathematical and philosophical doctrines associated with it. All these currents of thought converge in the life and work of the German mathematician Georg Cantor.

The substance of Cantor's lifework is well known: in developing what he called the arithmetic of transfinite numbers he gave mathematical content to the idea of the actual infinite. In so doing he laid the groundwork for abstract set theory and made significant contributions to the foundations of the calculus and to the analysis of the continuum of real numbers. Cantor's most remarkable achievement was to show, in a mathematically rigorous way, that the concept of infinity is not an undifferentiated concept. Not all infinite sets are the same size; consequently infinite sets can be compared with one another. For example, the set of points on a line and the set of all fractions are both infinite sets. Cantor was able to prove, in a well-defined sense, that the first set is greater in size than the second. So shocking and counterintuitive were Cantor's ideas to his contemporaries that the eminent French mathematician Henri Poincaré condemned the theory of transfinite numbers as "a disease" from which mathematics would someday be cured. Leopold Kronecker, one of Cantor's teachers and a formidable member of the German mathematics establishment, even attacked Cantor personally, calling him a "scientific charlatan," a "renegade" and a "corrupter of youth."

It is also well known that Cantor suffered throughout his life from a series of "nervous breakdowns," which became increasingly frequent and debilitating as he got older. The breakdowns were probably the symptoms of organic mental disease; a recent study by the British historian of mathematics Ivor Grattan-Guinness, who relies on the evaluation of Cantor's medical records by psychologists at the Halle Nervenklinik (the mental hospital in the East German city of Halle), suggests that Cantor was a victim of manic-depressive psychosis. Nevertheless, it was all too easy for Cantor's early biographers to present Cantor, who was trying to defend his complex theory but was suffering increasingly long periods of mental breakdown, as the hapless victim of the persecutions of his contemporaries.

Such accounts distort the truth by trivializing the genuine intellectual concerns that motivated some of the most thoughtful contemporary opposition to Cantor's theory. They also fail to credit the power and scope of the defense Cantor offered for his ideas. At first he resisted the transfinite numbers because he believed the idea of the actual infinite could not be consistently formulated and so had no place in rigorous mathematics. Nevertheless, by his own account he soon overcame his "prejudice" regarding the transfinite numbers because he found they were indispensable for the further development of his mathematics. Because of his own early doubts he was able to anticipate opposition from diverse quarters, which he attempted to meet with philosophical and theological arguments as well as mathematical ones. Moreover, when he was called on to respond to his critics, he was able to muster his ideas with considerable force. His mental illness, far from playing an entirely negative role, may well have contributed in its manic phases to the energy and single-mindedness with which he promoted his theory.

Georg Ferdinand Ludwig Philipp Cantor was born on March 3, 1845, in St. Petersburg (now Leningrad). His mother, Maria Anna Böhm, came from a family that included several gifted musicians; most notable among them was her uncle Joseph Böhm, the director of a conservatory in Vienna and the founder of a school for violinists that trained many virtuosos of the time. Georg's father, Georg Woldemar Cantor, was a successful tradesman and a devout Lutheran who communicated deep religious convictions to his son. According to Eric Temple Bell's widely read book *Men of Mathematics,* first published in 1937, the younger Georg's insecurities in later life stemmed from a ruinous Freudian conflict with his father, but surviving letters and other evidence concerning their relationship indicate quite the contrary. Georg's father appears to have been a sensitive man who was attentive to his children and took a special but not coercive interest in the welfare and education of his eldest son.

When the young Cantor was still a child the family moved from Russia to Germany, and it was there he began to study mathematics. In 1868 he received his doctorate from the University of Berlin for a dissertation on the theory of numbers; two years later he accepted a position as privatdocent, or instructor, at the University of Halle, a respected institution but not as prestigious for mathematics as the universities at Göttingen or Berlin. One of his colleagues at Halle, Heinrich Eduard Heine, was then working on the theory of trigonometric series, and Heine encouraged Cantor to take up the difficult problem of the uniqueness of such series. In 1872, when Cantor was 27, he published a paper that included a very general solution to the problem, together with the seeds of what was later to become the theory of transfinite sets.

The problem suggested by Heine arose from the work of the French mathematician Jean Baptiste Joseph Fourier. In 1822 Fourier had shown that the graph of any "reasonably smooth" curve (that is, a curve that exhibits at most a finite number of points of discontinuity) can be represented throughout an interval as the sum of an infinite trigonometric series. In other words, by superposing an infinite number of sine and cosine waves on one another every point on the "reasonably smooth" curve except the points of discontinuity can be approximated to any desired degree of accuracy [*see illustration on next page*]. Such a series is said to converge to the curve or function except at a finite number of points, or "almost everywhere." Fourier's result was of signal mathematical importance because it suggested that certain complicated functions could be represented, or replaced, by a sum of sines or cosines, which is much easier to manipulate mathematically. In order to justify such replacement, however, some guarantee was required that only one such trigonometric series converges to the function. Cantor began to investigate the conditions under which a trigonometric series that converges to a function is unique.

In 1870 Cantor achieved his first result: If a function is continuous throughout an interval, its representation by a trigonometric series is unique. His next step was to relax the requirement that the function be continuous throughout the interval. Suppose, for example, the function to be approximated by the series is defined as follows: The graph of

**PORTRAIT OF CANTOR and his wife was made in about 1880, when Cantor was at the height of his career. He had already begun to upset the German mathematics establishment by showing that the infinite set of real numbers, represented by the continuum of points on a line, is larger than the infinite set of all fractions. He also showed that infinite quantities called transfinite numbers could be defined that describe such differences. A few years after this photograph was made Cantor suffered a major attack of manic-depressive psychosis, which ultimately put an end to his creative work in mathematics. The original photograph is in the private collection of Egbert Schneider.**

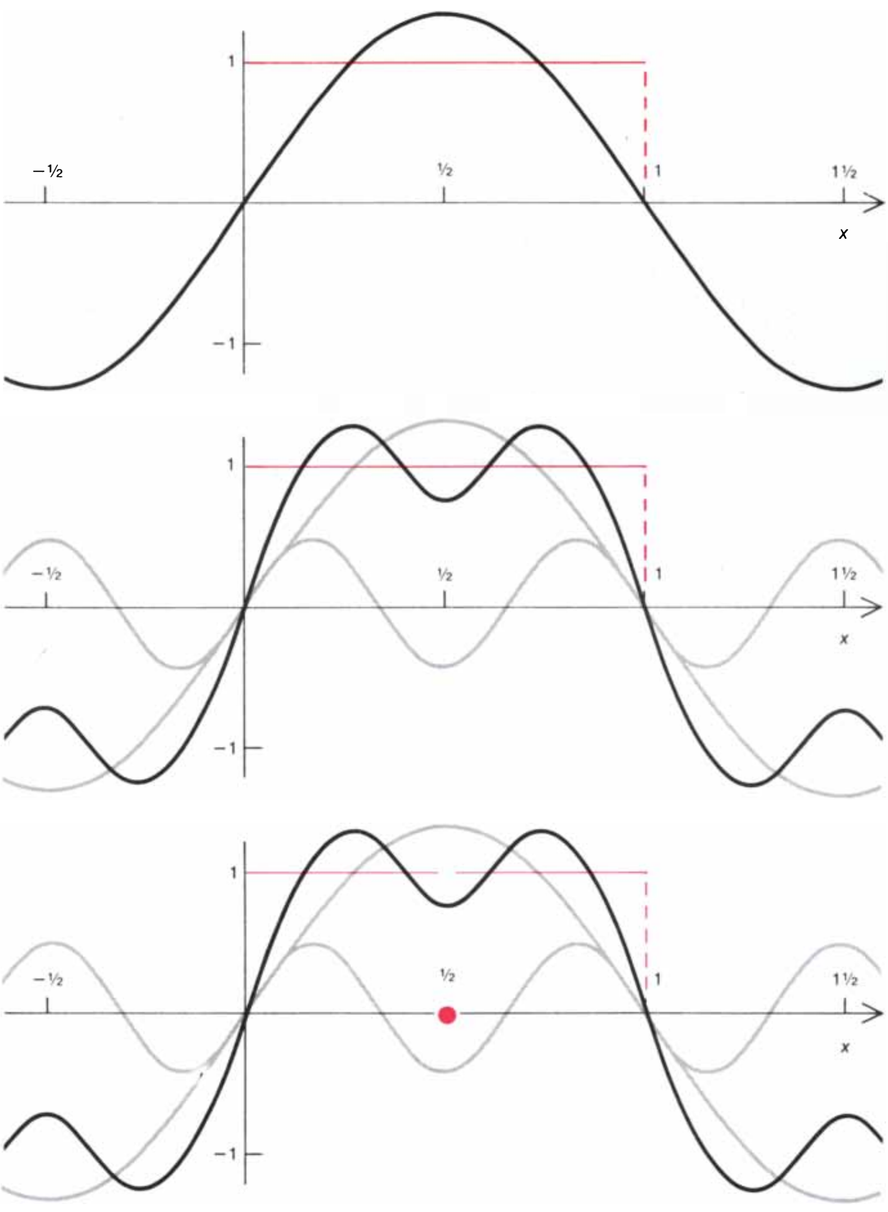
the function is a straight line parallel to the *x*, or horizontal, axis of the graph except at the point on the *x* axis corresponding to 1/2. For the point 1/2 on the *x* axis the function assumes the value 0 instead of 1. Cantor was able to show that if convergence is given up at the point where *x* is equal to 1/2, there is still a unique trigonometric series that converges to the function everywhere else. No other trigonometric series of a similar form exists that also approximates the function. It was then a simple matter for Cantor to extend his result to include all functions having any finite number of points of discontinuity, which Cantor called exceptional points.

In 1872 Cantor's search for a more general statement of his uniqueness theorem led to the publication of a remarkable discovery. He found that as long as the exceptional points are distributed in a carefully specified way on the *x* axis there can be infinitely many of them. The most important step in the proof of the new result was the precise description of the infinite set of exceptional points, and for this purpose Cantor realized he needed to provide a satisfactory way of analyzing the continuum of points along the *x* axis. Thus Cantor's study of trigonometric series gave rise to an important transition in his thought: he began to focus his attention on the relations among points in the continuum instead of on theorems about trigonometric series.

Cantor regarded it as axiomatic that for any point on a continuous line there corresponds a number, which is called a real number in order to distinguish it from the "imaginary" numbers that are multiples of $\sqrt{-1}$. Conversely, for every real number there corresponds exactly one point on a continuous line. Hence the problem of describing the continuum of points on a line is equivalent to the problem of defining the real numbers and exploring their properties. One of the major accomplishments of Cantor's paper of 1872 was a rigorous account of these properties.

The main obstacles to a theory of the real numbers are numbers such as $\pi$ and $\sqrt{2}$, which are not rational numbers. (A rational number is a number that can be expressed as the quotient of two whole numbers. It was known in antiquity that $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$ and many other roots are irrational.) Because the legitimacy of the rational numbers was not in question, Cantor followed an approach suggested by Karl Weierstrass, one of his former teachers at the University of Berlin. Cantor proposed that any irrational number could be represented by an infinite sequence of rational numbers. The number $\sqrt{2}$, for example, can be represented by the infinite sequence of rational numbers 1, 1.4, 1.41 and so on. In this way all irrational numbers can be understood as geometric points on a real-number line just as the rational numbers can.

In spite of the advantages of Cantor's approach, it troubled some mathematicians because it presumes the existence of sets or sequences of numbers having infinitely many elements. Philosophers and mathematicians had rejected the concept of completed infinities since the time of Aristotle, primarily because of the logical paradoxes they inevitably seemed to generate. Galileo, for example, pointed out that if completed infinite sets are admissible in mathematics, there must be as many even integers as there are even and odd integers together. Every even integer can be uniquely matched with the integer that is half its magnitude and so there is a one-to-one correspondence between each member of one set and each member of the other.

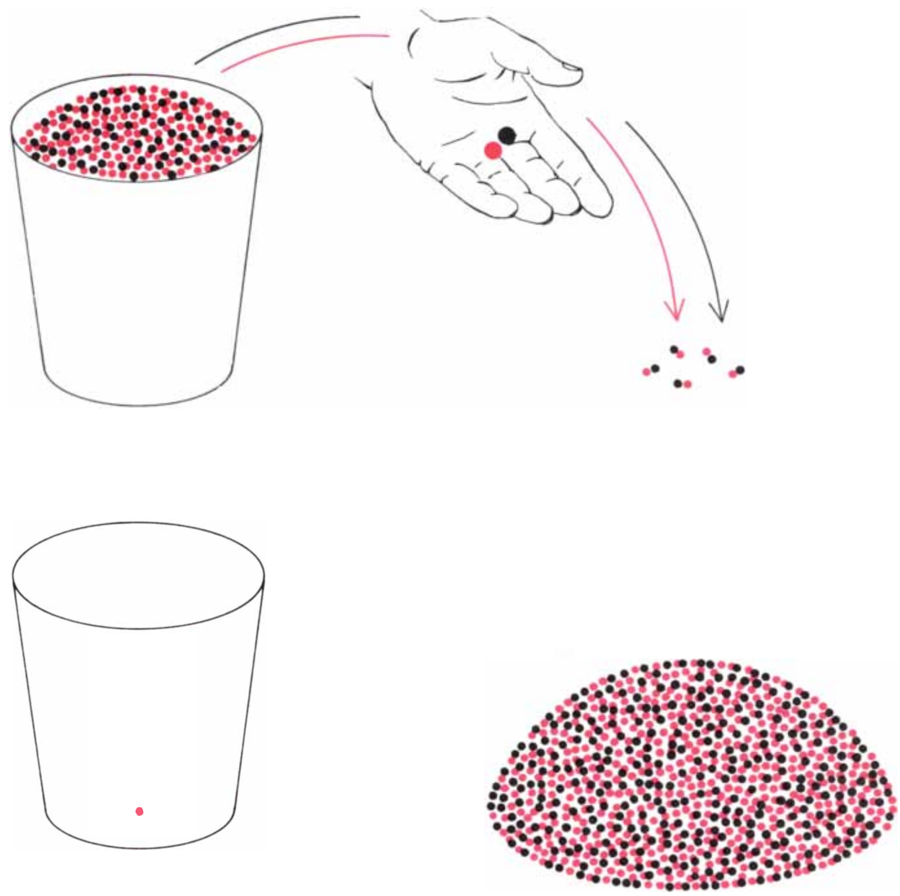



**SMOOTH, CONTINUOUS GRAPH whose height depends on the value of a corresponding point on the *x* axis can be approximated to any desired accuracy by a trigonometric series, that is, by a sum of sines and cosines. For example, a straight, horizontal line one unit above the *x* axis (*color*) can be approximated by superposing sine waves on one another (*gray curves*); the first two stages of the approximation are depicted (*black curves at top and middle*). The trigonometric series that approximates the graph is unique. Even if the graph is not continuous, however, it can often be approximated by a unique trigonometric series. For example, if the height of the graph is one unit everywhere except at the point where *x* is equal to 1/2, the trigonometric series that converges to the continuous line also converges to the broken line except at the point 1/2 (*bottom*). Cantor showed that a graph can be approximated by a unique trigonometric series even if the number of points where the graph is not continuous is an infinite number, provided the points of discontinuity are distributed along the *x* axis in a specific way. He was thereby led to consider the structural properties of abstract infinite sets and the infinitely many ways in which elements of infinite sets can be ordered with respect to one another.**

Theologians such as St. Thomas Aquinas, representing another voice of traditional opposition to the completed infinite, regarded the idea as a direct challenge to the unique and absolutely infinite nature of God.

In order to meet such objections mathematicians had customarily drawn a sharp distinction between the infinite regarded as a completed quantity and the infinite regarded as potential, represented by an indefinite sum or series of numbers that tends toward some limit. Only the potential infinite were they willing to countenance. In 1831 Carl Friedrich Gauss expressed his opposition to the use of completed infinities in terms Cantor once described as authoritative. In a letter to Heinrich Schumacher, Gauss wrote: "But concerning your proof I protest above all the use of an infinite quantity as a *completed* one, which in mathematics is never allowed. The infinite is only a *façon de parler,* in which one properly speaks of limits."

By speaking of limits it was possible to avoid the paradoxes associated with actual infinities. For example, by adding additional digits to the decimal expansion of $\pi$ one can approximate the true value of $\pi$ with increasing precision. Gauss insisted, however, that one must never assume all the terms of the decimal expansion could be given, so that the numerical value of $\pi$ would be determined exactly. To do so would be to grasp an infinite number of terms in its entirety, or in other words an actually infinite set of numbers, an operation Gauss refused to allow.



**TWO SETS can be compared in size by matching the elements in the first set one for one with the elements in the second. For example, in order to determine whether or not there are more colored marbles than black marbles in a bucket the marbles can be removed from the bucket in pairs of one colored and one black marble. When no more pairings can be made, the marbles remaining in the bucket, if any, are the basis of the comparison. Cantor applied the same elementary principle in order to describe how the sizes of infinite sets can be compared.**

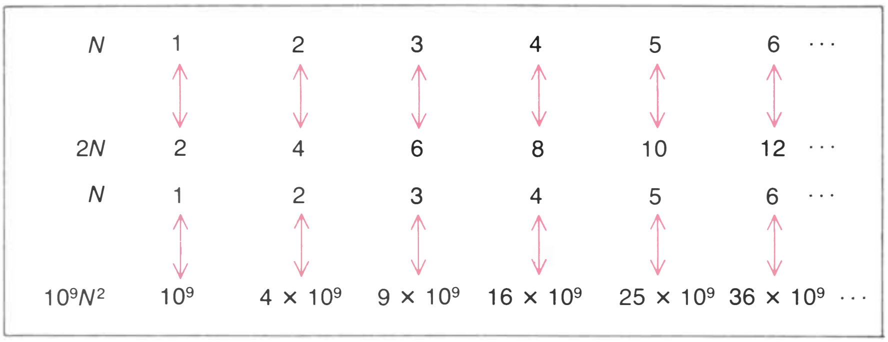
Cantor was not alone in studying the properties of the continuum in rigorous detail. In 1872, the same year Cantor's paper appeared, the German mathematician Richard Dedekind also published an analysis of the continuum that was based on infinite sets. In his paper Dedekind articulated an idea that Cantor later made more precise: "The line is infinitely richer in point-individuals than is the domain ... of rational numbers in number-individuals." The property can be put into perspective if one considers the distribution of the points on a line segment that correspond to rational numbers, or rational points. No matter how small the segment, there are infinitely many rational points on it. Thus the gist of Dedekind's remark was that in spite of the density of the rational points on a line segment there is still room to pack an infinite number of irrational points into the line. Irrational points such as $\sqrt{2}$ fall between the rational points, and so the set of rational points, although dense, is still riddled with holes and fails to be continuous.

Dedekind's statement is consistent with a correct understanding of the continuum, but it conceals a serious weakness. If anyone had asked Dedekind how much richer the infinite set of points in the continuum was than the infinite set of rational numbers, he could not have replied. Cantor's major contribution to the question was published in 1874 by August Leopold Crelle's *Journal für die reine und angewandte Mathematik,* also referred to as *Crelle's Journal* and perhaps the most respected mathematical periodical of its day.

In effect Cantor borrowed the paradox cited by Galileo and turned it into a means of comparing the size of infinite sets. He defined two sets as equivalent if a one-to-one correspondence can be established between the members of each set. Such a correspondence provides a natural way of comparing size. For example, imagine a bucket filled with black and colored marbles, and suppose one wants to compare the number of black marbles with the number of colored ones. The simplest way is to remove the marbles from the bucket in pairs of one black and one colored marble. If every marble can be paired with a marble of a different color, the two sets are equivalent. Failing that, the marbles remaining in the bucket are the basis of the comparison.

By applying the same principle of correspondence Cantor showed that the property Galileo regarded as a paradox was actually a natural property of infinite sets. The set of even numbers is equivalent to the set of all positive whole numbers, even and odd together, because the pairings between members of each set can go on forever without omitting any members from either set. Cantor was also able to exhibit an ingenious and sophisticated method whereby the set of all rational numbers could be paired with the whole numbers [*see bottom illustration on next page*]. Any set of numbers whose members can be matched one for one, or in effect counted by the set of positive whole numbers, Cantor called a denumerable set.

Given the density of the rational numbers on a line and the relative sparseness of the whole numbers, it may seem grossly counterintuitive that the two sets turn out to be the same size. Cantor's next move, however, was even more striking. He showed that no such one-to-one correspondence can exist between the set of whole numbers and the set of all points on a line, that is, the real numbers; in short, the real numbers form

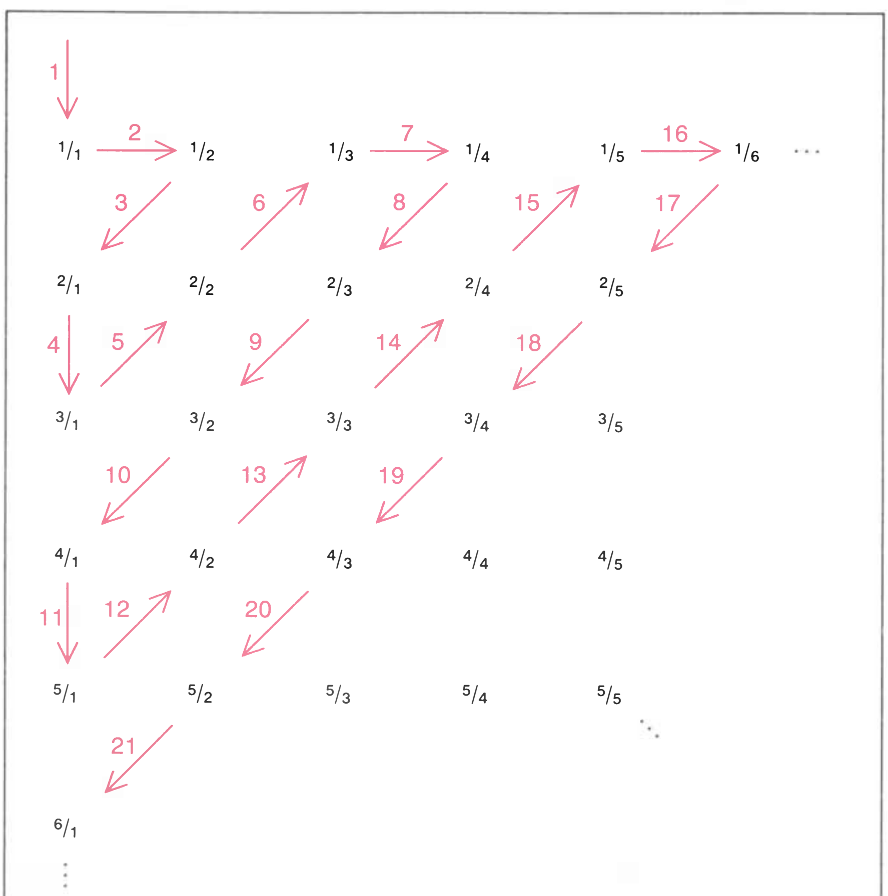



**WHOLE NUMBERS can be paired one for one with the even numbers without exhausting either set of numbers. Hence although it may appear there must be more whole numbers than even numbers, the two infinite sets actually have the same number of elements. Many other infinite sets, such as the perfect squares multiplied by a billion, can also be compared one for one or, in effect, counted by the whole numbers. Such sets are called countable or denumerable sets.**

a nondenumerable set. Cantor gave a somewhat complicated proof for this assertion in his paper of 1874; I shall instead present the main idea of a simplified and much more powerful version he gave in 1891.

Cantor began his proof by supposing there does exist a correspondence between the set of real numbers and the set of whole numbers. The argument then proceeds by showing that the supposition leads to a contradiction; it therefore follows that the original supposition must be false and that no such one-to-one correspondence is possible. The argument can be simplified by examining only the real numbers between 0 and 1; if this set of real numbers is greater than the set of whole numbers, the set of all the real numbers is greater than the set of whole numbers as well.

Suppose, therefore, that the real numbers between 0 and 1 can be paired one for one with the whole numbers. Setting up such a correspondence is equivalent to making a list of the real numbers, each one represented as an infinite decimal. It is then possible to define a new real number that has not been included in the list. Consider the first digit of the first decimal expansion in the list of real numbers. If the digit is a 1, write a 9 in the first decimal place of the defined number. If the first digit in the list is not a 1, write a 1 in the first place for the defined number. Continue constructing the new number to be defined by changing the second digit of the second decimal expansion in the list, the third digit in the third decimal expansion and so on. The newly constructed number must differ in at least one decimal place from every real number already in the list, but it will still represent some real number between 0 and 1. A number has therefore been constructed that is not in the list of real numbers, and so the supposition that all the real numbers can be listed leads to a contradiction.




**INFINITE SET OF RATIONAL NUMBERS (that is, numbers that can be expressed as the quotient of two whole numbers) might seem much larger than the set of whole numbers. For example, between any two consecutive whole numbers such as 0 and 1 there are infinitely many rational numbers. Nevertheless, Cantor showed in 1874 how the rational numbers can be paired one for one with the whole numbers. Every possible rational number can be listed in a square array as shown in the illustration; each rational number can then be associated with a whole number by tracing the path shown in color. Thus the set of rational numbers is denumerable.**

In August, 1874, Cantor married Vally Guttmann; the couple spent the summer in the Harz Mountains, where they also met with Dedekind. The period was an exceedingly fruitful one for Cantor's work. Earlier that year he had posed his next important problem to Dedekind in a letter: "Might it be possible to correspond a surface (a square, perhaps, including its boundaries) with a straight line (perhaps an interval with the inclusion of its endpoints) so that to each point of the surface one point of the line corresponds, and conversely?" Although Cantor thought the answer must be negative, he could offer no reason for his belief, and neither could Dedekind.

By 1877, however, Cantor was able to send Dedekind the startling news that, contrary to prevailing mathematical opinion, a one-to-one correspondence between lines and planes is not impossible. The proof is a matter of representing each point on a square by a pair of decimal coordinates. The decimal representations are then shuffled together in a strictly specified way to generate a single decimal expansion; this decimal is associated with a point on a line segment. The entire process can also be reversed [see illustration on page 129]. Cantor himself was so unprepared for the result that it prompted him to exclaim, "I see it but I don't believe it!"

Cantor immediately prepared a manuscript that described his new discovery, and he sent it, as he had his paper of 1874, to *Crelle's Journal*. Although the result was of major significance, the paper was also the first occasion for open hostilities between Cantor and his former teacher Kronecker. As an editor of the journal, Kronecker was in a position to block the publication of any article, and by 1877 he was so appalled at the direction of Cantor's work that he did just that. Although Cantor had submitted his manuscript on July 12, no steps were taken to prepare it for publication and it did not appear in the volume for 1877. Cantor, suspecting Kronecker's intervention, wrote a bitter letter to Dedekind complaining about the treatment of his paper and mentioning the possibility of withdrawing it from the journal. Dedekind, recalling his own experience in such matters, persuaded Cantor to wait. As it turned out, Dedekind was right. The article appeared in the volume for 1878, but Cantor was so offended by the incident that he refused to publish in *Crelle's Journal* again.

The controversy that had arisen between Cantor and Kronecker was clouded with personal animosity, but at its roots were sharply divergent views on the philosophy of mathematics. Such views are still reflected in the debate between constructivist and formalist mathematicians. Kronecker, a forerunner of the constructivists, is well known for his quip that captures the essence of his position: "God made the integers; all else is the work of man." In this spirit he advocated constructing all mathematics from the integers and from finite arithmetic combinations of them. In the early 1870's he began to reject any of the limiting processes in the traditional calculus, and he opposed all attempts to define mathematical objects in terms of limits. Thus even the irrational numbers, which had been accepted by mathematicians for centuries, were to be read out of mathematics unless some way could be found to construct them, as the rational numbers are constructed, out of the integers.

Cantor, who had written two major papers under Kronecker in his student days at the University of Berlin, was well aware of Kronecker's extreme position and was sensitive to its advantages. It guaranteed the maximum certainty and correctness to a mathematical proof, and it could be applied as a corrective to unfettered speculation in mathematics. Cantor nonetheless argued that to accept Kronecker's position would be to rule out many of the most promising developments in mathematics; moreover, it could burden innovative mathematical research with narrow and ultimately barren methodological scruples.

The definition of the irrational numbers given in Cantor's 1872 paper was tantamount to accepting the existence of completed infinite sets. Cantor embraced a formalist position on the existence of the irrationals and he argued that the only grounds on which their legitimacy was to be judged in mathematics were their formal and internal consistency. "In introducing new numbers," he once wrote, "mathematics is only obliged to give definitions of them, by which... they can definitely be distinguished from one another. As soon as a number satisfies all these conditions it can and must be regarded as existent and real in mathematics."

This point of view on the irrational numbers was crucial for Cantor's justification of his introduction of transfinite numbers. In his 1872 paper he had defined sets of exceptional points by introducing the concept of a limit point. The irrational number $\sqrt{2}$, for example, is a limit point of the sequence 1, 1.4, 1.41 and so on. More generally, a point is a limit point of an infinite set if there are an infinite number of members in the set that lie arbitrarily close to the point.

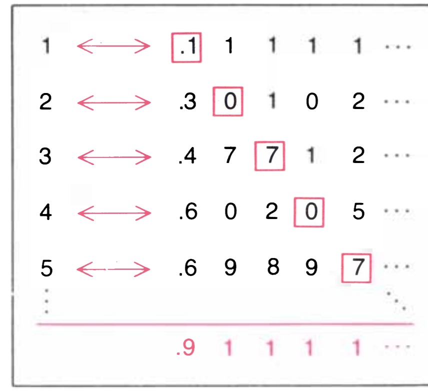
Given any infinite set $P$, Cantor defined the derived set $P^1$ as the set of all the limit points of $P$. Similarly, if $P^1$ were also an infinite set, its derived set $P^2$ was defined as the set of all the limit points of $P^1$. Cantor showed that the relation of being a subset determines a natural ordering for the sets: it turns out that every member of $P^2$ is also a member of $P^1$ and so $P^2$ is a subset of $P^1$; similarly, $P^3$ is a subset of $P^2$ and so on.

It can turn out that for some finite whole number $n$ the derived set $P^n$ is a finite set; if this condition holds, the infinite set of points $P$ that gives rise to the set $P^n$ is precisely the set of exceptional points needed to prove the general version of Cantor's theorem about the unique representation of functions by trigonometric series. On the other hand, it can also happen that no derived set in the sequence $P^1, P^2, P^3, \ldots$ is a finite set. Cantor argued that in this case it still makes sense to consider the set of points that are common to all the derived sets $P^1, P^2, P^3, \ldots, P^n, \ldots$. He designated the set of points common to all these derived sets $P^\infty$, and in 1880 he began to refer to the superscript $\infty$ as a transfinite symbol. Moreover, if $P^\infty$ were an infinite set of points, one could then form its derived set $P^{\infty+1}$, which could in turn lead to an entire sequence of derived sets, $P^{\infty+2}$ and so on.

Cantor might have added that the superscripts $\infty$, $\infty + 1$, $\infty + 2$ and so on actually constitute a new kind of number, but at first he did not do so. In 1872 he had been careful to discuss the irrational numbers only in terms of sequences of rational numbers; similarly, he initially regarded the symbols $\infty$, $\infty + 1$, $\infty + 2$ and so on only as tags for identifying sets. In 1883, however, he abandoned his reticence and presented them as transfinite numbers, an autonomous and systematic extension of the natural numbers.

The immediate reason for their introduction, Cantor maintained, was that they were necessary for further progress in set theory and for the study of the real numbers. Nevertheless, in order to respond to critics such as Kronecker, Cantor argued from his formalist philosophical position: Once the self-consistency of the transfinite numbers was recognized they could not be refused a place alongside the other accepted but once disputed members of the mathematical family, such as the irrational numbers. By formulating a theory of the infinite that was able to avoid the well-known mathematical paradoxes, Cantor believed he had removed the only rational objection mathematicians could raise for refusing to consider the completed infinite in their work.

The transfinite numbers ultimately introduced by Cantor are widely known by the notation he adopted for them in later years, namely the first letter of the Hebrew alphabet ℵ (aleph). The alephs designate the cardinality, or number of elements, in an infinite set, and so the



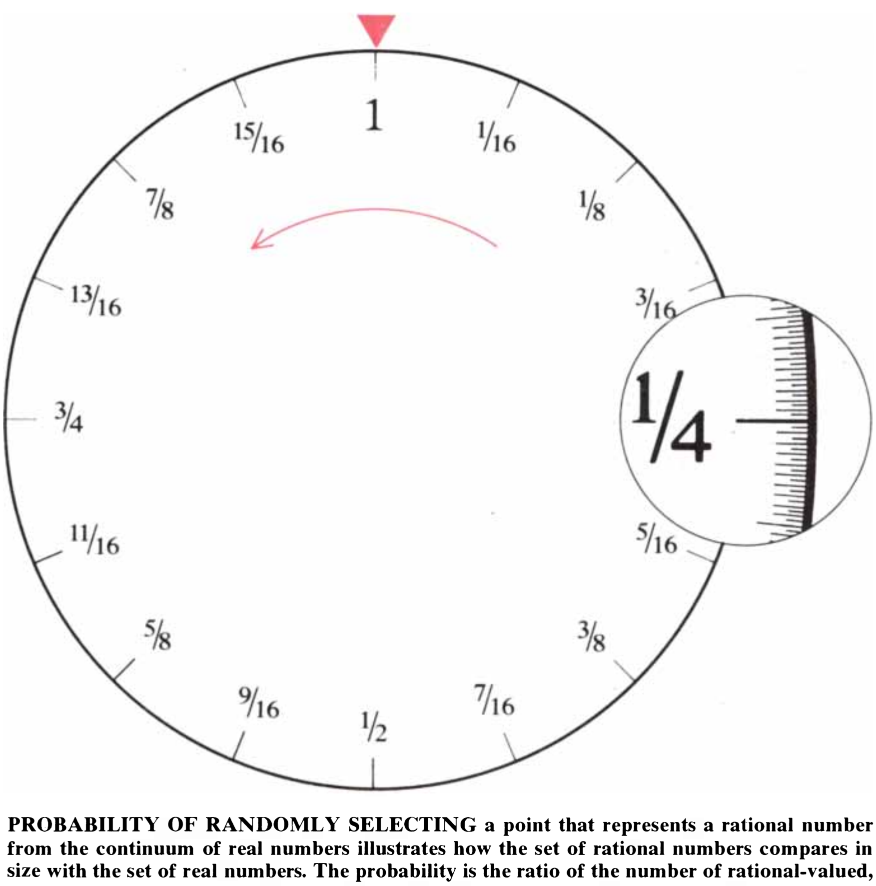

**SET OF REAL NUMBERS, represented by the continuum of points on a line, is not denumerable. If it were, the real numbers between, say, 0 and 1 could be paired one for one with the whole numbers. Each real number in the list can be represented by an infinite decimal expansion. (Infinite decimals such as .5000... are to be represented by an equivalent infinite decimal, such as .4999....) No matter how such infinite decimals are listed, however, a new decimal can be constructed that defines a real number not included in the list: write a 9 in the first decimal place of the constructed number if the first digit in the decimal expansion of the first real number in the list is a 1; otherwise write a 1 in the first decimal place of the constructed number. Continue by changing the second decimal digit of the second real number, the third decimal digit of the third real number and so on. The constructed decimal expansion represents a real number between 0 and 1, but it must differ in at least one decimal place from every real number in the list. Hence the supposition that the real numbers can be paired with the whole numbers leads to a contradiction and so must be abandoned. The main idea employed in the proof is called the diagonalization method.**

equivalences among infinite sets that Cantor demonstrated in the 1870's are frequently stated in terms of the transfinite cardinal numbers, the alephs. It is therefore of considerable historical interest that the first transfinite numbers to be introduced were not cardinal numbers but ordinal numbers.

An ordinal number is defined by its order or position in a list. The ordinal number associated with a finite set corresponds to the cardinal number of the set. For example, any set having five elements (that is, any set whose cardinal number is five) can in a certain sense be thought of as the immediate successor to any set having four elements. In other words, the ordinal number of the set is also five; it is the fifth set in a list of sets. The ordinal number of an infinite set, however, must be distinguished from its cardinal number. Cantor showed it is possible to construct an infinite number of infinite sets having different ordinal numbers but the same cardinal number. Indeed, Cantor later found it is possible to turn this property of infinite sets into a criterion for distinguishing them from finite sets. A set is finite only if its cardinal number and its ordinal number are the same.

Cantor pointed out that the ordinal number of a sequence of finite sets of increasing size 1, 2, 3 and so on is based on the repeated addition of units. There is no greatest ordinal number associated with the sequence of finite sets, but just as it is possible to define $\pi$ as the limit of a sequence of rational numbers without supposing $\pi$ is itself a rational number, so, Cantor believed, it is proper to define a new, transfinite ordinal number $\omega$ as the first number following the entire sequence of ordinary ordinal numbers 1, 2, 3 and so on. Once $\omega$ is defined it is possible through the repeated addition of units to generate additional transfinite ordinal numbers, $\omega + 1$, $\omega + 2$, $\omega + 3$ and so on. Because there is no largest element for this sequence either, one could imagine another ordinal number $\omega + \omega$, or $2\omega$, as the first ordinal following the sequence $\omega + 1$, $\omega + 2$, $\omega + 3$ and so on. By alternately repeating these two principles of generation Cantor was able to define a hierarchy of progressively greater transfinite ordinal numbers [*see illustration on page 130*].

How does one distinguish between, say, the ordinal number $\omega$ and the ordinal number $\omega + 1$? The difference is determined by the order of the elements within the sets to which $\omega$ and $\omega + 1$ correspond. For example, the set of natural numbers in its familiar sequence (1, 2, 3, ...) has the ordinal number $\omega$, which represents the entire sequence of natural numbers in its usual order. The set of all the natural numbers in the rearranged sequence (2, 3, 4, ..., 1), however, or the set of all natural numbers in the sequence (10, 30, 40, ..., 20) has the ordinal number $\omega + 1$. In other words, the distinction depends on the consecutive order of the elements in the sequence and on the placement of the infinitely long gap, represented by the ellipsis; if one number is displaced to the end of the sequence, the ordinal number of the new sequence is $\omega + 1$. The sequence (2, 4, 6, ..., 1, 3, 5, ...) has two infinite gaps, and its ordinal number is $\omega + \omega$, or $2\omega$. Note that all the sets have the same number of elements, that is, their elements can always be placed in a one-to-one correspondence with each other and with the positive whole numbers. Hence their cardinal numbers are the same even though their ordinal numbers are quite different.




**PROBABILITY OF RANDOMLY SELECTING** a point that represents a rational number from the continuum of real numbers illustrates how the set of rational numbers compares in size with the set of real numbers. The probability is the ratio of the number of rational-valued, or rational, points to the total number of points along some interval. Here the interval between 0 and 1 is represented by the circumference of a wheel of fortune. (On the wheel 0 and 1 are identified.) It is assumed that by spinning the wheel and allowing it to stop any point on the wheel can be selected at random. The points representing the rational numbers are infinitely dense, in the sense that along any arc between two rational points on the circumference, no matter how short, there must be an infinite number of rational points. A few such points are marked. Nevertheless, the set of all the points along the circumference is infinitely larger than the set of rational points; the probability that the wheel stops at a rational point is zero. More precisely, the probability is smaller than any positive quantity that can be specified in advance.

Once the transfinite ordinal numbers were defined Cantor proceeded to describe their arithmetic properties. An important distinction between the transfinite and the ordinary numbers had to be made with respect to the commutative property of addition and multiplication. For two ordinary numbers $A$ and $B$ the commutative property expresses the fact that $A + B$ is equal to $B + A$ and $A \times B$ is equal to $B \times A$. When addition or multiplication are defined to include the transfinite numbers, however, the commutative property cannot be assured. For example, $\omega + 2$, which represents the sequence (1, 2, 3, ..., 1, 2), is not equal to $2 + \omega$, which represents the sequence (1, 2, 1, 2, 3, ...).

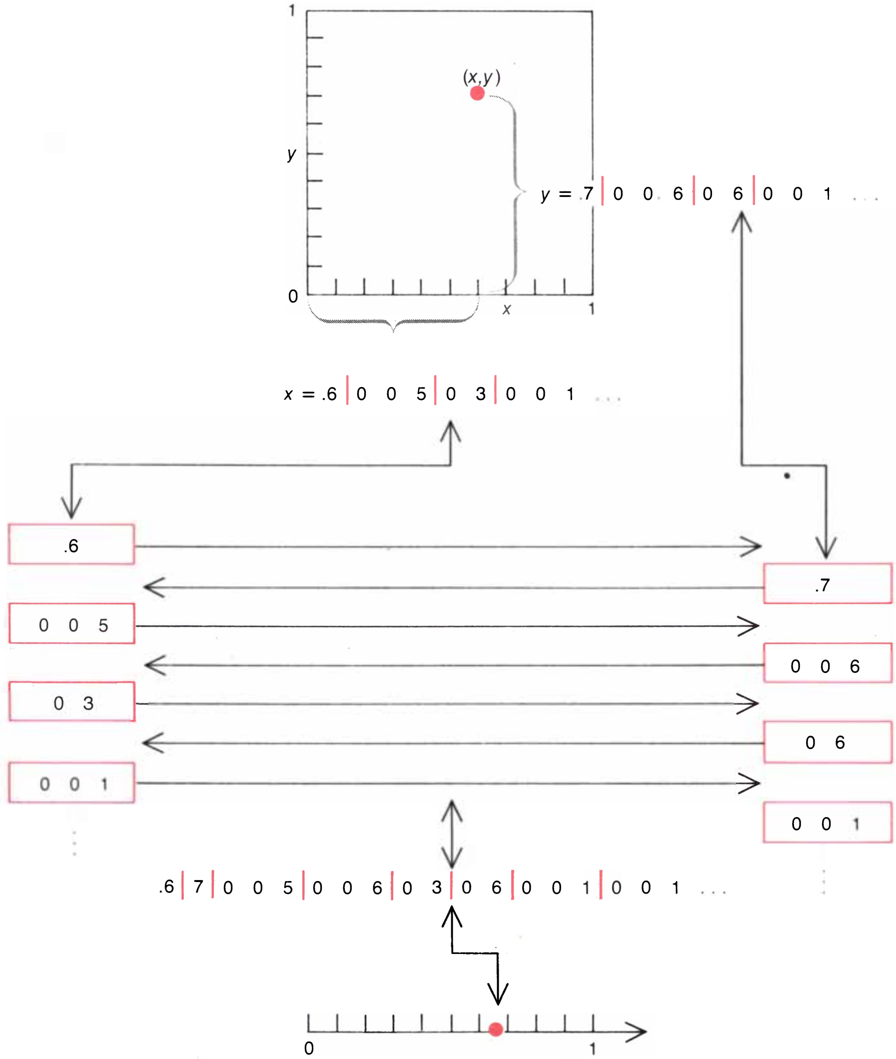
Although the difference between ordinal number and cardinal number is indistinct for finite sets, it helped to explain how the application of the concept of number to an infinite set could lead to confusion and paradox. Because the concepts of ordinal and cardinal number are fundamentally distinct for infinite sets, any argument that discusses the number associated with an infinite set without making the distinction is subject to ambiguity. Thus it is illegitimate to extend the seemingly unambiguous properties of finite sets to infinite sets, as Galileo and others had done.

In spite of the results achieved by Cantor in the 1880's there remained a serious lacuna. The question of the cardinal number (or power, in Cantor's early terminology) of the continuum of real numbers was still unanswered. Remember that in his paper of 1883 Cantor had defined the sequence of transfinite ordinal numbers according to two principles of generation. In order to introduce natural divisions in the sequence he had added a third principle. Consider the set of all finite whole numbers, which Cantor called the first number class. Its power or cardinal number is greater than the power associated with any element in the set. Similarly, he noted, one could also consider the set of all the transfinite ordinal numbers corresponding to denumerably infinite sets, or in other words sets whose power is the same as the power of the set of all whole numbers. Cantor called this set of transfinite ordinal numbers the second number class. It turns out that the power of the second number class is greater than the power associated with any of the transfinite ordinal numbers that make up the set. In short, the second number class is a nondenumerable set. Although Cantor was never able to prove it, he was convinced that the power of the second number class is equivalent to the power of the continuum of real numbers.

The conjecture has come to be known as Cantor's continuum hypothesis, and it has never been proved. In 1963 Paul J. Cohen of Stanford University, building on the work of Kurt Gödel of the Institute for Advanced Study, showed that although the continuum hypothesis is consistent with the axioms of a standard version of set theory, it is also independent of them. In fact, the continuum hypothesis plays the role for set theory that Euclid's parallel postulate plays for geometry. Different versions of set theory can be constructed by assuming the truth or falsity of the continuum hypothesis, just as Euclidean or non-Euclidean geometry can be constructed by assuming the truth or falsity of the parallel postulate [see "Non-Cantorian Set Theory," by Paul J. Cohen and Reuben Hersh; SCIENTIFIC AMERICAN, December, 1967].

Cantor's unsuccessful efforts to prove the continuum hypothesis caused him considerable stress and anxiety. Early in 1884 he thought he had found a proof, but a few days later he reversed himself completely and thought he could disprove the hypothesis. Finally he realized he had made no progress at all. Throughout the period he had to endure mounting opposition and threats from Kronecker, who said he was preparing an article that would show "the results of modern function theory and set theory are of no real significance."

Soon thereafter, in May, 1884, Cantor had the first of his serious nervous breakdowns. Although frustration over his lack of progress on the continuum hypothesis or stress from Kronecker's




**POINTS IN THE PLANE can be matched one for one with the points on the line. Each point in the plane is represented by a pair of infinite decimals, and the infinite decimals are broken up into groups. Every digit except 0 in the decimal expansions begins a new group. The groups are then combined by alternating them into a single infinite decimal, which represents a point on the line. The entire procedure can also be reversed. A similar proof shows that the number of points in any finite-dimensional space is equivalent to the number of points on the line.**

attack may have helped to trigger the breakdown, it now seems clear that such events had little to do with its underlying cause. The illness took over with startling speed and lasted for somewhat more than a month. At the time only the manic phase of manic-depressive psychosis was recognized as a symptom; when Cantor "recovered" at the end of June, 1884, and entered the depressive phase, he complained that he lacked the energy and interest to return to rigorous mathematical thinking. He was content to take care of trifling administrative matters at the university and felt capable of little more.

Although Cantor eventually returned to mathematics, he also became increasingly absorbed in other interests. He undertook a study of English history and literature and became engrossed in a scholarly diversion that was taken with remarkable seriousness by many people at the time: the supposition that Francis Bacon was the author of Shakespeare's plays. Cantor also tried his hand without success at teaching philosophy instead of mathematics, and he began to correspond with several theologians who had taken an interest in the philosophical implications of his theories about the infinite. The correspondence was of special significance to Cantor because he was convinced that the transfinite numbers had come to him as a message from God. He was anxious that his views be carefully studied by theologians so that his mathematical concept of the infinite could be reconciled with church doctrine.

More important, Cantor was instrumental in creating a professional society for the promotion of mathematics in Germany, the Deutsche Mathematiker-Vereinigung. He believed his own career had been greatly damaged by the premature and prejudiced rejection of his work by the mathematics establishment, and he hoped the independent organization would give younger mathematicians encouragement and a fair hearing for new and perhaps radical ideas.

There was one last element of transfinite set theory with which Cantor had to come to terms, namely the nature and status of the transfinite cardinal numbers. The evolution of his thinking on the subject is curious, because the transfinite cardinal numbers were the last part of his theory to be given either rigorous definition or a special symbol. Indeed, it is difficult in the clarity of hindsight to reconstruct the obscurity within which Cantor must have been groping, and I have discussed his work up to now as if he had already recognized that the power of an infinite set could be understood as a cardinal number. In fact, although Cantor already understood it is the power of a set that establishes its equivalence (or lack thereof) with any other set, he initially avoided any suggestion that the power of an infinite set could be interpreted as a number.

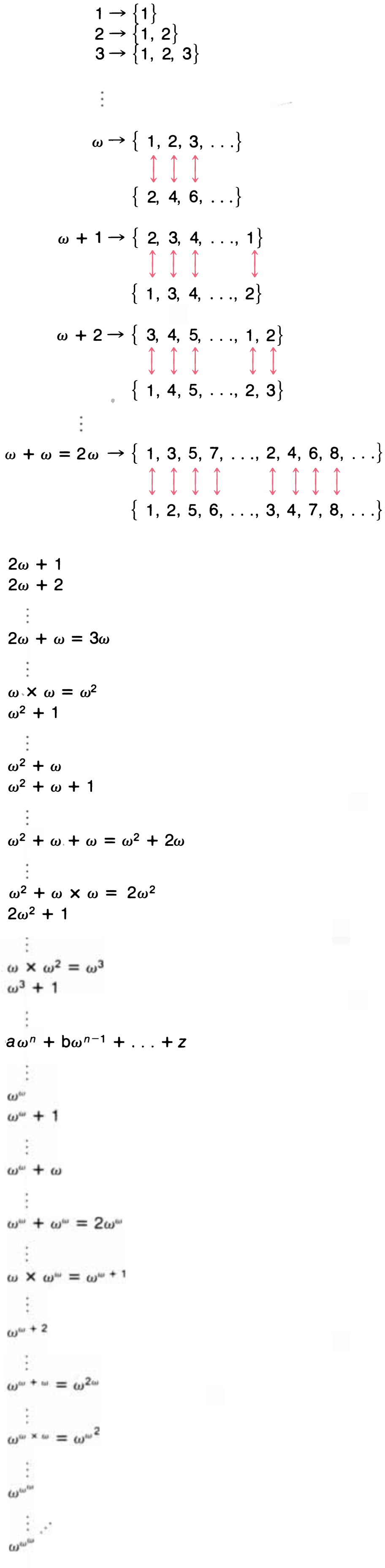
He began to equate the two concepts in September, 1883; still, however, no symbol was provided for distinguishing one transfinite cardinal number from another. Since he had already adopted the symbol $\omega$ to designate the least transfinite ordinal number, it is clear that the ordinal numbers were much more important than the cardinals for the early conceptual development of Cantorian set theory. When Cantor finally introduced a symbol for the first transfinite cardinal number, it was borrowed from the symbols already in service for the transfinite ordinals; the first transfinite cardinal was written $\overset{*}{\omega}$.

Cantor did not settle on the alephs as symbols until 1893. By then the Italian mathematician Giulio Vivanti was preparing a general account of set theory, and Cantor realized it was time to adopt a standard notation. He chose the alephs for the transfinite cardinals because he thought the familiar Greek and Roman alphabets were already too widely employed in mathematics for other purposes. His new numbers deserved something distinct and unique. Thus he chose the letter $\aleph$, which was readily available in the type fonts of German printers. The choice was particularly clever, as Cantor was pleased to admit, because the Hebrew aleph was also a symbol for the number 1. Since the transfinite cardinal numbers were themselves infinite unities, the aleph could be taken to represent a new beginning for mathematics. Cantor designated the cardinal number of the first number class $\aleph_0$ (aleph-null), the number he had previously called $\overset{*}{\omega}$; the cardinal number of the second number class was designated $\aleph_1$ (aleph-one).

Cantor made his last major contributions to set theory in a pair of articles published in 1895 and 1897. He had proved in a paper delivered before the first meeting of the Deutsche Mathematiker-Vereinigung in 1891 that the cardinal number of any set is always smaller than the cardinal number of the set of all its subsets. (A version of the proof is given in the illustration on the opposite page.) A few years later he derived a corollary to this result: the cardinal number of the continuum is equal to a cardinal number Cantor designated $2^{\aleph_0}$. He hoped this result would soon lead to a solution of the continuum hypothesis, because the hypothesis could now be stated in clear algebraic form: $2^{\aleph_0} = \aleph_1$.

The arguments in Cantor's proof about the cardinal number of the set of subsets, however, led to far different conclusions. The most important of them was reached by Bertrand Russell in 1903; Russell showed that a paradox can be derived in set theory by considering all sets that do not include themselves as members. Russell's paradox suggested there was something fundamentally wrong with Cantor's definition of a set, and the consequences of this realization have become an important problem in 20th-century mathematical logic. Nevertheless, such developments have not overturned any of Cantor's major results in transfinite arithmetic.

Unfortunately by 1903 Cantor was suffering attacks of manic depression with increasing frequency, and no evidence has been found to suggest he ever became aware of Russell's result. In fact, illness prompted Cantor to apply for a leave of absence from the University of Halle during the fall term of 1899, and the request was granted. In




**TRANSFINITE ORDINAL NUMBERS are identified by their order or position in a list. The list is generated according to two principles. First, each new transfinite ordinal is derived from the immediately preceding one by adding one unit, just as if one were "counting" beyond the transfinite ordinal $\omega$ that is associated with the set of whole numbers arranged in their natural order. Second, whenever a succession of transfinite ordinals exists for which there is no largest number, a new transfinite ordinal is defined as the next number larger than all the others. Such new numbers are listed immediately after a vertical mark of ellipsis; for example, $2\omega$ is the next transfinite ordinal number larger than all the numbers $\omega$, $\omega + 1$, $\omega + 2$ and so on. It is the placement of the infinite gaps (*horizontal marks of ellipsis*) within the sets to which the ordinal numbers correspond that distinguishes the transfinite ordinals from one another. Thus in the diagram two examples of sets are given that correspond to the ordinal numbers $\omega$, $\omega + 1$, $\omega + 2$ and $2\omega$. Every infinite set represented by an ordinal number in the list, however, has the same cardinal number, namely $\aleph_0$, or in other words every set includes the same number of elements.**

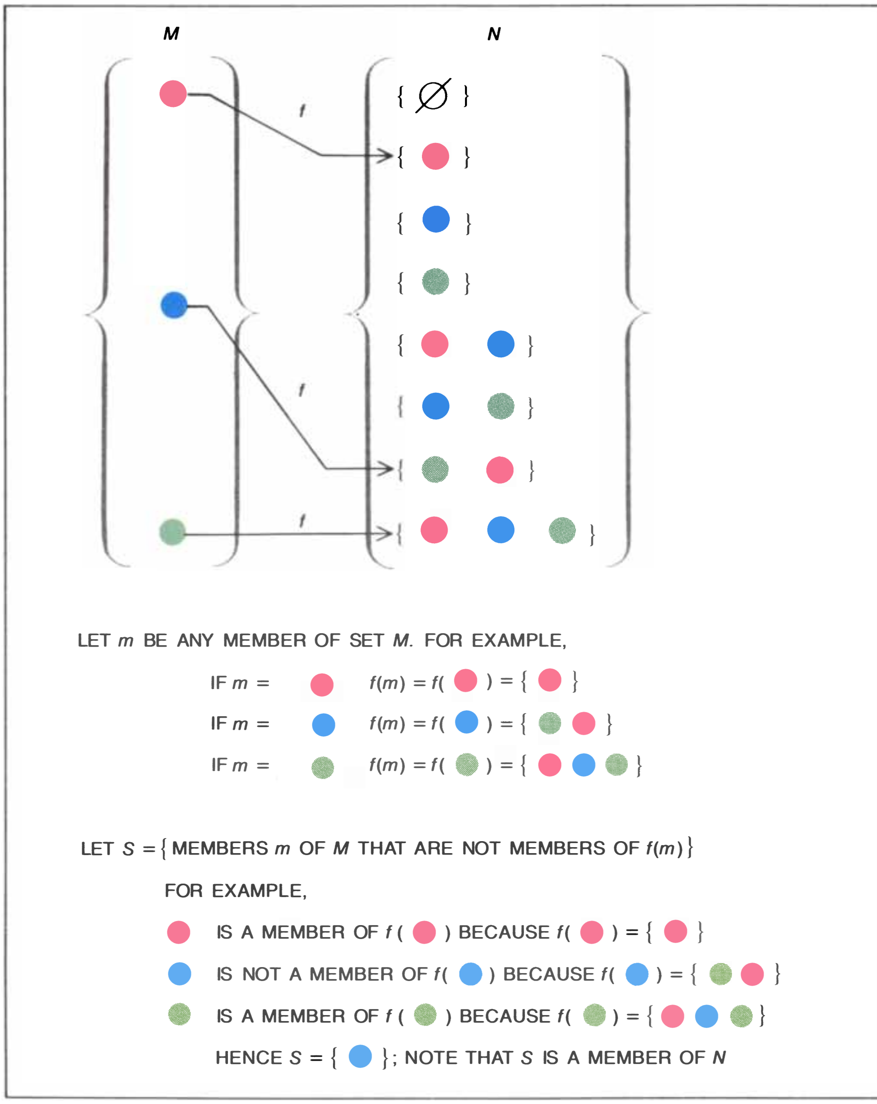
November of the same year he notified the German Ministry of Culture that he wanted to give up his professorship completely. As long as his salary was not reduced he would be content with a modest position in a library. Among his qualifications he emphasized his publications on the Shakespeare question, and he concluded with the extraordinary request that the Ministry issue its reply within the next two days. If he were offered no alternative to teaching, he wrote, then as a person born in Russia he would seek to enter the service of the Russian diplomatic corps.

Nothing seems to have come of Cantor's request, and he did not enter the service of Emperor Nicholas II. Nevertheless, the entire episode fits the earlier pattern of his behavior in 1884, when he had seriously considered giving up mathematics for philosophy after his first major breakdown. Similarly, he was hospitalized for manic depression late in 1899, again in the winter terms of 1902 and 1903 and thereafter for increasingly long and frequent periods. He died of heart failure on January 6, 1918, in the Halle Nervenklinik.

There is an important connection between Cantor's mental illness and his mathematics. Certain documents suggest that on occasion the illness afforded him periodic respite from daily affairs, during which he was still able to pursue his mathematical ideas in the solitude of the hospital or quietly at home. The illness may also have supported his belief that the transfinite numbers had been communicated to him from God. Following a long period of hospitalization in 1908, he wrote to a friend in Göttingen, the British mathematician Grace Chisholm Young. As he described it, his manic depression took on a strikingly generative quality: "A peculiar fate, which thank goodness has in no way broken me, but in fact has made me stronger inwardly, happier and more expectantly joyful than I have been in a couple of years, has kept me far from home—I can also say far from the world.... In my lengthy isolation neither mathematics nor in particular the theory of transfinite numbers has slept or lain fallow in me."

Elsewhere Cantor described his conviction about the truth of his theory in quasi-religious terms: "My theory stands as firm as a rock; every arrow directed against it will return quickly to its archer. How do I know this? Because I have studied it from all sides for many years, because I have examined all objections that have ever been made against the infinite numbers and above all because I have followed its roots, so to speak, to the first infallible cause of all created things."

Later generations might dismiss the philosophy, look askance at the abundant references to St. Thomas or to the church fathers, overlook the metaphysical pronouncements and miss entirely the deeply religious roots of Cantor's later faith in the absolute truth of his theory. But all these commitments contributed to his resolve not to abandon the transfinite numbers. Opposition seems to have strengthened his determination. His forbearance, as much as anything else he might have contributed, ensured that set theory would survive the early years of doubt and denunciation to flourish eventually as a vigorous, revolutionary force in the scientific thought of the 20th century.




**INFINITE SEQUENCE OF SETS, each set larger than the one before it in the sequence, can be constructed by considering the set of all the subsets of any given set. An ingenious variation of Cantor's method of diagonalization can be employed to show that, by assuming a one-to-one correspondence *f* between any set *M* and the set of all its subsets *N*, it is possible to construct a subset *S* that is not included in the one-to-one correspondence, whatever *f* may be. In order to understand the construction consider the finite set *M* that includes a red, a blue and a green disk. The set has eight subsets (including the empty set Φ, which contains no elements). Let *S* be defined as the set of all the elements *m* of *M* that are not members of the subset *f*(*m*) to which *m* corresponds. For the example in the illustration *S* includes only the blue disk. Because *S* is a subset of *M* and because *f* is assumed to be a one-to-one correspondence, there must be some element *a* of *M* to which *S* corresponds, or in other words for which *f*(*a*) is identical with *S*. Either *a* is a member of *S* or it is not. If *a* is a member of *S*, it must also be a member of *f*(*a*), since *f*(*a*) is equal to *S*; on the other hand, if *a* is a member of *S*, it cannot be a member of *f*(*a*) because of the way *S* is defined. Hence *a* cannot be a member of *S*. Again, however, by the definition of *S*, if *a* is not a member of *S*, *a* must be a member of *f*(*a*), and since *f*(*a*) is equal to *S*, *a* must also be a member of *S*. No matter what the status of *a* is, therefore, the supposition that the set *M* can be paired one for one with the set of all its subsets leads to a contradiction, and so the supposition must be abandoned. In the same way it is proved that even if a set is infinite, the set of all its subsets is larger than the original set. An infinite sequence of progressively larger infinite sets can be constructed by forming the set *N* of all the subsets of any infinite set *M*, then forming the set *P* of all the subsets of *N* and so on. There is no largest set in the sequence.**

# The Slide Fastener

*Otherwise known as the zipper, it took hold, so to speak, some 60 years ago. It is the creation of a host of inventors going back to Elias Howe of the sewing machine, and the end is not in sight*

by Lewis Weiner

The slide fastener, perhaps more widely known as the zipper, appears on objects ranging from suitcases to trousers and dresses, and it would be unusual for anyone in a Western society to get through the day without encountering one. Yet this common and usually reliable mechanism goes remarkably uncelebrated. Out of curiosity I looked in seven leading encyclopedias, five published in the U.S., one in Britain and one in Germany. Some of them have in their latest edition an entry of a few lines on the subject; none has what could be called a full-scale article. The origin and development of the slide fastener are matters few people know about, even in the companies that make it. Having been associated with the industry for many years, partly as an inventor of machinery for making zippers and partly as a consulting engineer for manufacturers, I am moved to put forward an account of this simple and ubiquitous device.

In 1943 Frank B. Jewett, chairman of the Bell Telephone Laboratories and president of the National Academy of Sciences, gave an address at New York University titled "The Promise of Technology." Remarking that "the real creative ideas" originate with individuals and that no one can foretell what the ideas will be or where they may arise, he said: "As an example of how impossible it is, even with simple things, to forecast the future, I have often thought of how infinitesimally small would have been the chance of any man or group of men, except the one who actually had the idea, planning to invent the common zipper."

The one who actually had the idea was Elias Howe, who is much better known for his contribution to the invention of the sewing machine. In 1851 he received a U.S. patent for an automatic, continuous clothing closure. Describing it in the patent, he wrote: "My invention consists of a series of clasps united by a connecting cord running or sliding upon ribs." Howe's device anticipated many of the features of the modern slide fastener, but for reasons that are not clear he never put it on the market.

Hence the man who usually gets credit for inventing the zipper is Whitcomb L. Judson, who in 1893 received two U.S. patents for an automatic closing device actuated by a slide mechanism. Whether Judson had a limited vision of the applications that might be found for his fastener or simply wanted to state a specific application for the purpose of obtaining the patent is not known, but his title for the first patent was "Clasp Locker or Unlocker for Shoes." The fasteners differed from the modern zipper mainly in that their movement was perpendicular to the opening that was to be closed, like the buckles on the galoshes made before zippers were put on them. Judson's scheme consisted of a series of separate clasps, each of which consisted in turn of two interlocking parts, one attached to each side of the opening that was to be closed. They could be fastened by hand, but Judson also provided for a sliding device to close and open them in sequence.

In 1894 Judson and Lewis A. Walker organized the Universal Fastener Company to exploit the Judson patents. Walker was a lawyer who had a flair for corporate organization. The company began promoting one form of fastener under the name Universal in 1896. Sales were poor, partly because the fasteners had a tendency to pop open under stress and partly because they were rather sharp and sometimes tore fabrics that came in contact with them.

In 1904 the company was reorganized as the Automatic Hook and Eye Company and offered an improved fastener under the name C-Curity. (An advertising slogan was "A Pull and It's Done.") The company had the advantage of a machine Judson had patented in 1902 to make the fasteners, which previously had been made by hand. It had the further advantage in 1906 of employing Gideon Sundback, an engineer who over a period of years greatly improved the fastener.

Simultaneously other inventors were also improving it. In 1911 Catharina Kuhn-Moos and her partner Henri Foster received a Swiss patent showing a device closely resembling a modern metallic zipper in that it had no hooks. It appears these inventors were too far ahead of their time; their fastener was not a commercial success. In 1913 P. A. Aronson, the superintendent of Automatic Hook and Eye, was awarded a patent that shows the principle of the modern separating zipper: a fastener that opens at both ends.

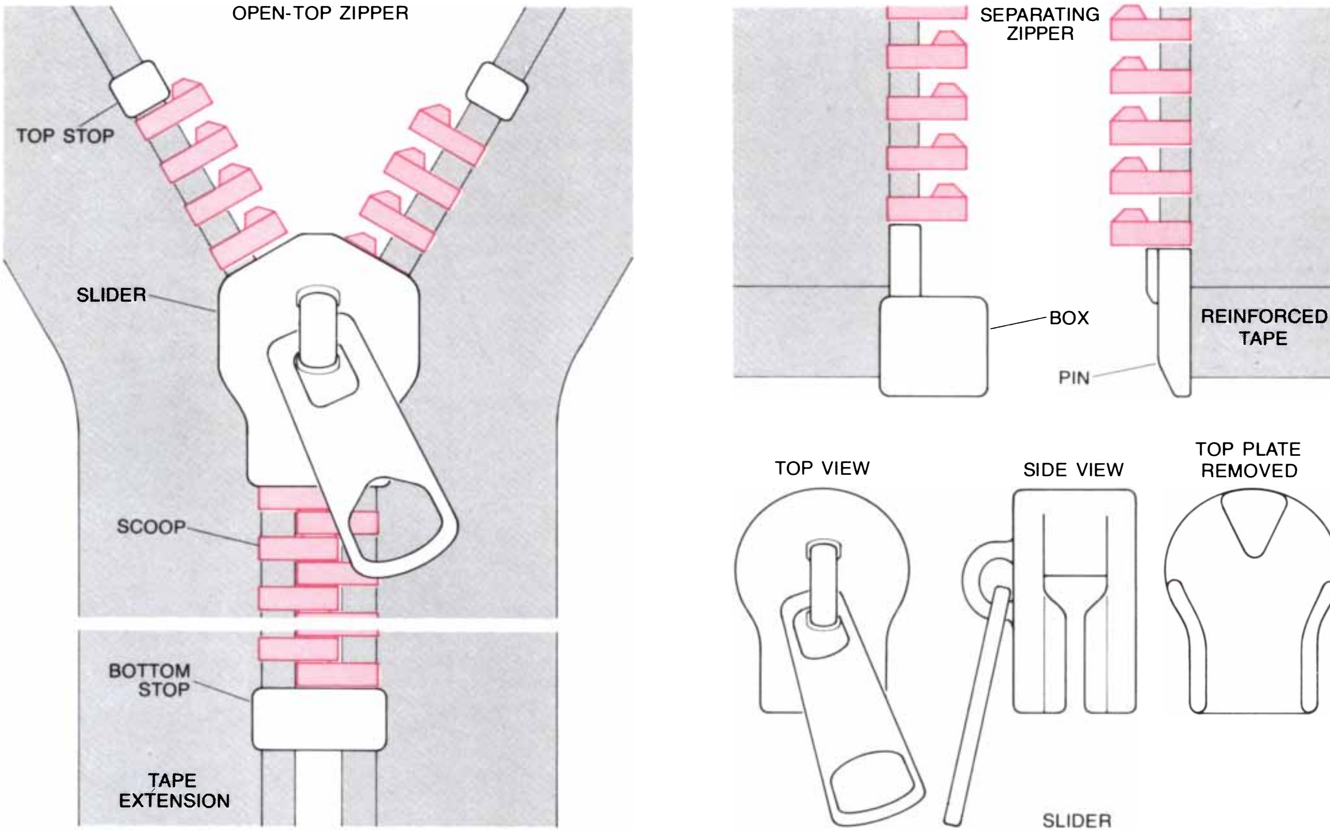
Sales continued to be poor, however, and the company verged on bankruptcy. It survived by taking on work other than fasteners. Then in 1917 Sundback received a patent on a metallic fastener that had all the features of the modern metallic zipper. In an interview he said: "The great need was to eliminate the hooks." The invention was so important to the company that it changed its name to the Hookless Fastener Company. Walker raised more capital and reincorporated the business with himself as president, a position he held until his death in 1938.

Gradually business improved. The first item to be manufactured in quantity with a slide fastener was a money belt developed in 1917 by a tailor in New York; it was popular among sailors in World War I. Most of the 24,000 fasteners the company sold that year were for such belts. In 1918 the fasteners were incorporated in some 10,000 flying suits made for the Navy.

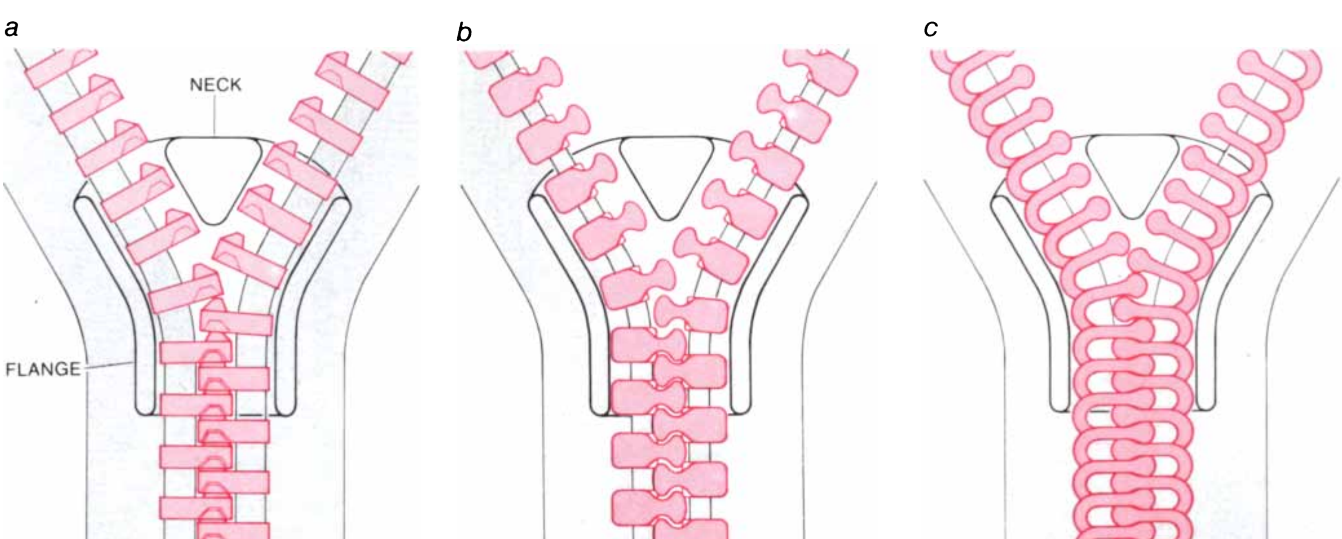
Soon after the war the introduction of slide fasteners on gloves and tobacco pouches helped to make the device familiar to civilian consumers. Probably the biggest impetus, however, came from the introduction of slide fasteners on galoshes in 1923 by the B. F. Goodrich Company, which also originated the name zipper and registered it as a trademark for the Goodrich line of rubber footwear.

Expansion was now rapid. Hookless Fastener, which was the sole maker of slide fasteners in the U.S. from 1917 to 1926, saw its sales rise from 24,000 in 1917 to more than 60 million in 1934. By that time the company had changed the name of its fasteners to Talon, and in 1937 it also changed its corporate name to Talon, Inc. The U.S. production of zippers peaked at about 2.3 billion some 10 years ago; it has declined since then as a result of competition from abroad and stood at less than 1.8 billion in 1981.

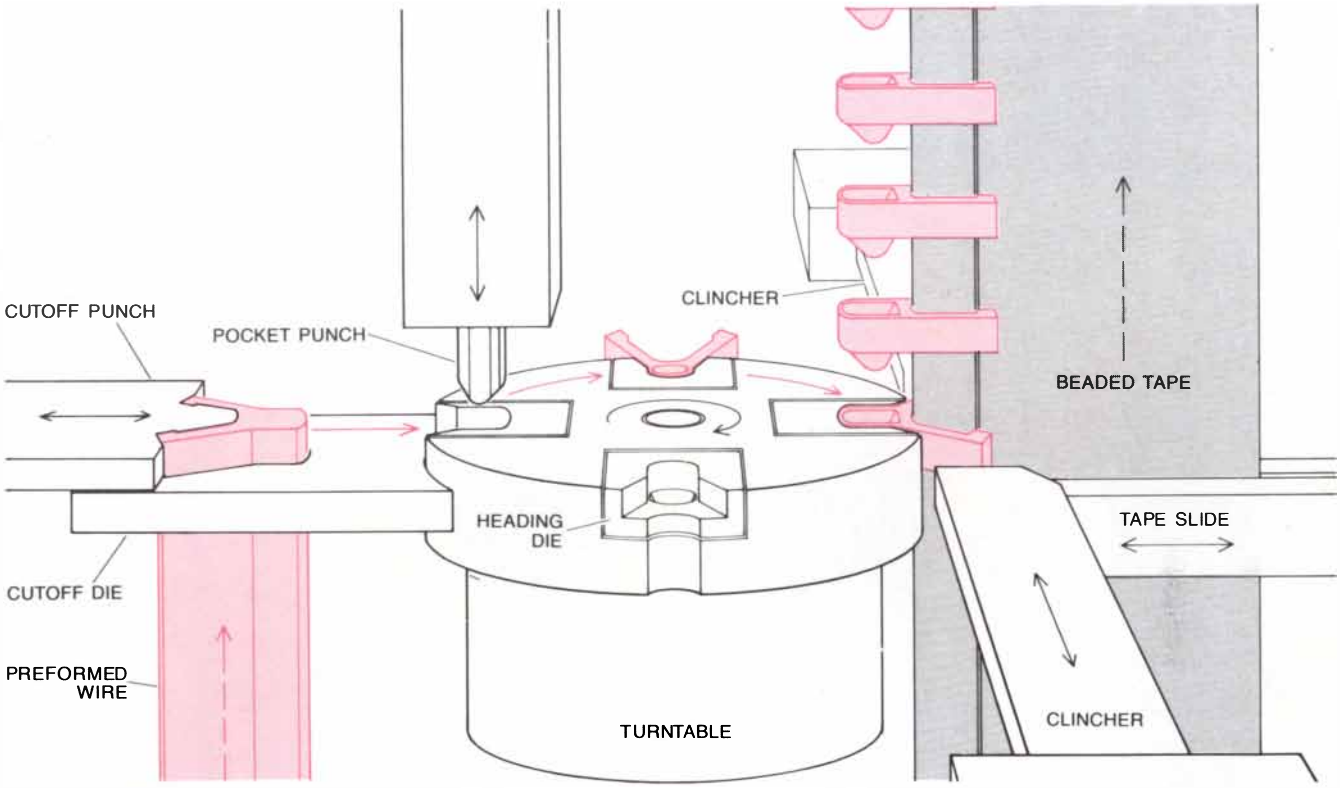
The basic components of a metallic zipper are the teeth, the tape, the "hardware" and the slider. The teeth are the individual units, arrayed on a piece of tape. The tape fitted with teeth is called a stringer. Two stringers joined side by side make up a chain. The hardware consists of parts that prevent the slider



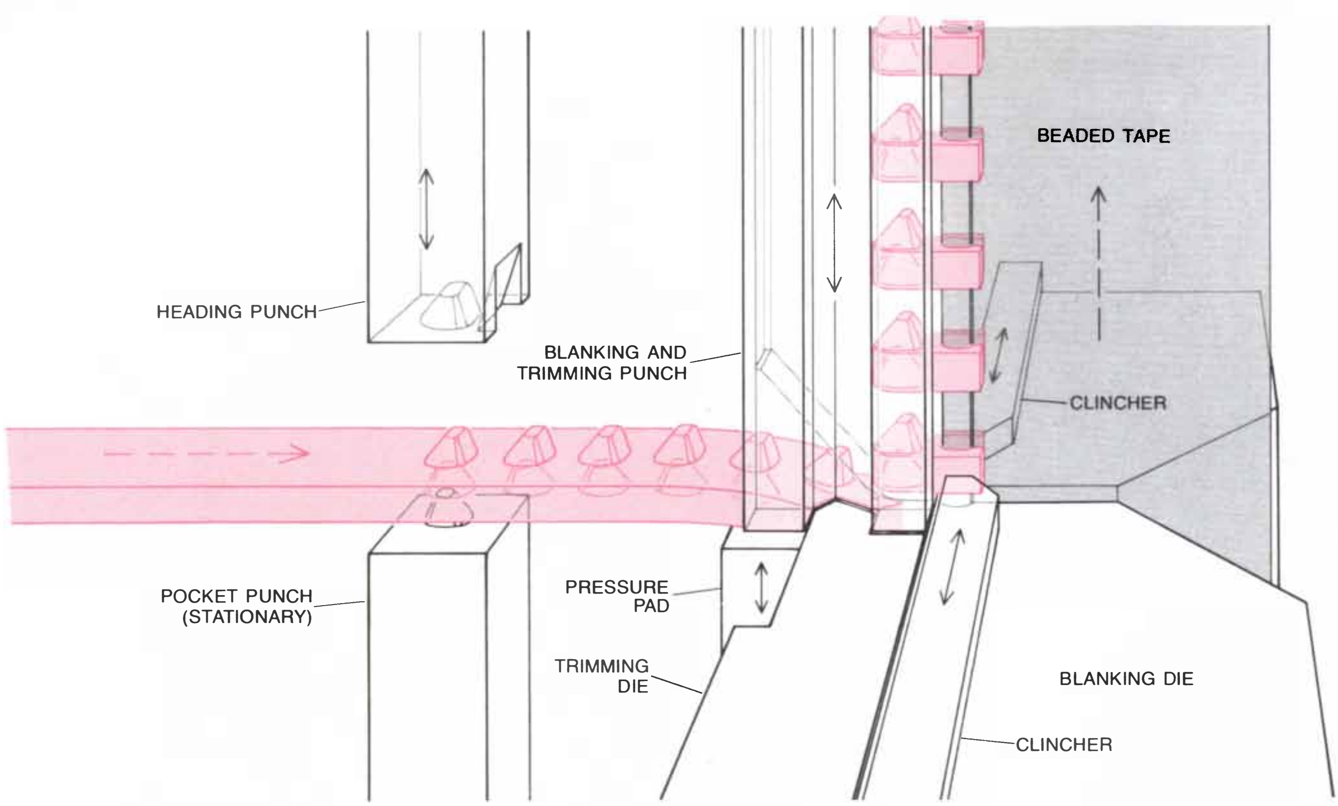

**COMPONENTS OF A ZIPPER are portrayed for an open type of the kind put on trousers and for a separating type such as is needed on a jacket. The teeth are sometimes called scoops because a metal tooth in profile resembles a scoop. The fabric tape serves not only to hold the teeth in position but also as the means of sewing the zipper to a garment, a step further served by the tape extension. A tape fitted with teeth or spirals is termed a stringer. Two stringers joined are a chain. A chain provided with a slider and stops is a zipper.**




**OPERATION OF ZIPPER is shown with the front of the slider removed. In a zipper with metal teeth (*a*) each tooth has a dome or ball on the top of one end and a pocket on the bottom. As the slider is raised its neck positions the teeth at the appropriate angle and distance for the ball of one tooth to fit into the pocket of the opposite tooth; the flanges push the teeth together. In opening a zipper the neck pushes the teeth apart. Plastic spirals or teeth (*b, c*) are made in many different shapes, but the opening and closing principle is the same.**




**METAL STRINGER** is made in several ways, one of which is depicted here. The process begins with wire that has been preformed into a *Y* profile (*left*). As the wire is moved upward in the machine a cutoff punch slices off what will become individual teeth. A sliced piece is pushed into a heading die that is rotating clockwise. There a pocket punch forms the ball and pocket. After two 90-degree turns of the die the piece reaches the tape, which is pushed between the legs of the *Y*. The legs are clamped on the beaded tape by the clincher and then the tape advances. The bead, which does not appear on all tapes, provides more material for the legs of the tooth to grip.




**ALTERNATIVE PROCESS** for making metal zippers operates on a strip of flat wire. The strip is fed into the machine by rollers. A heading punch works against a stationary pocket punch to form the ball and pocket. Then a blanking punch cuts off notches to form the basic *Y* and at the same time cuts off the front tooth. Even before the front scoop is fully off the strip the clinchers bend the legs of the scoop around the cord of the tape. Then the stringer advances into the next position. The machine can make 50 scoops per second.

from slipping out. The chain is closed and opened by the slider.

If you look closely at a zipper tooth made of metal, you will see a tiny dome on the top of the outer end and a corresponding pocket on the bottom. Together they give the unit something of the configuration of a scoop, which is what a zipper tooth is often called. The dome or ball of a scoop fits into the pocket of the corresponding scoop on the opposite side of the chain; this geometry along the length of the chain is what holds the two stringers together when the zipper is closed.

If you now look closely at a slider, you will see that its left and right sides form a flange. If you are able to peer inside the slider at the top, you will see a solid shape in the form of a *V;* it is called the neck. As you close a zipper the inner *V* and the flanges acting together move the scoops of opposing stringers far enough apart so that they are separated and held at an angle allowing the end of each scoop to slip in between two scoops on the opposite side. The flanges of the slider as it moves upward then push the two stringers together, scoop by scoop. When you open a zipper, the flanges hold the scoops at the angle that allows opposing scoops to slip apart, and the descending *V* separates them.

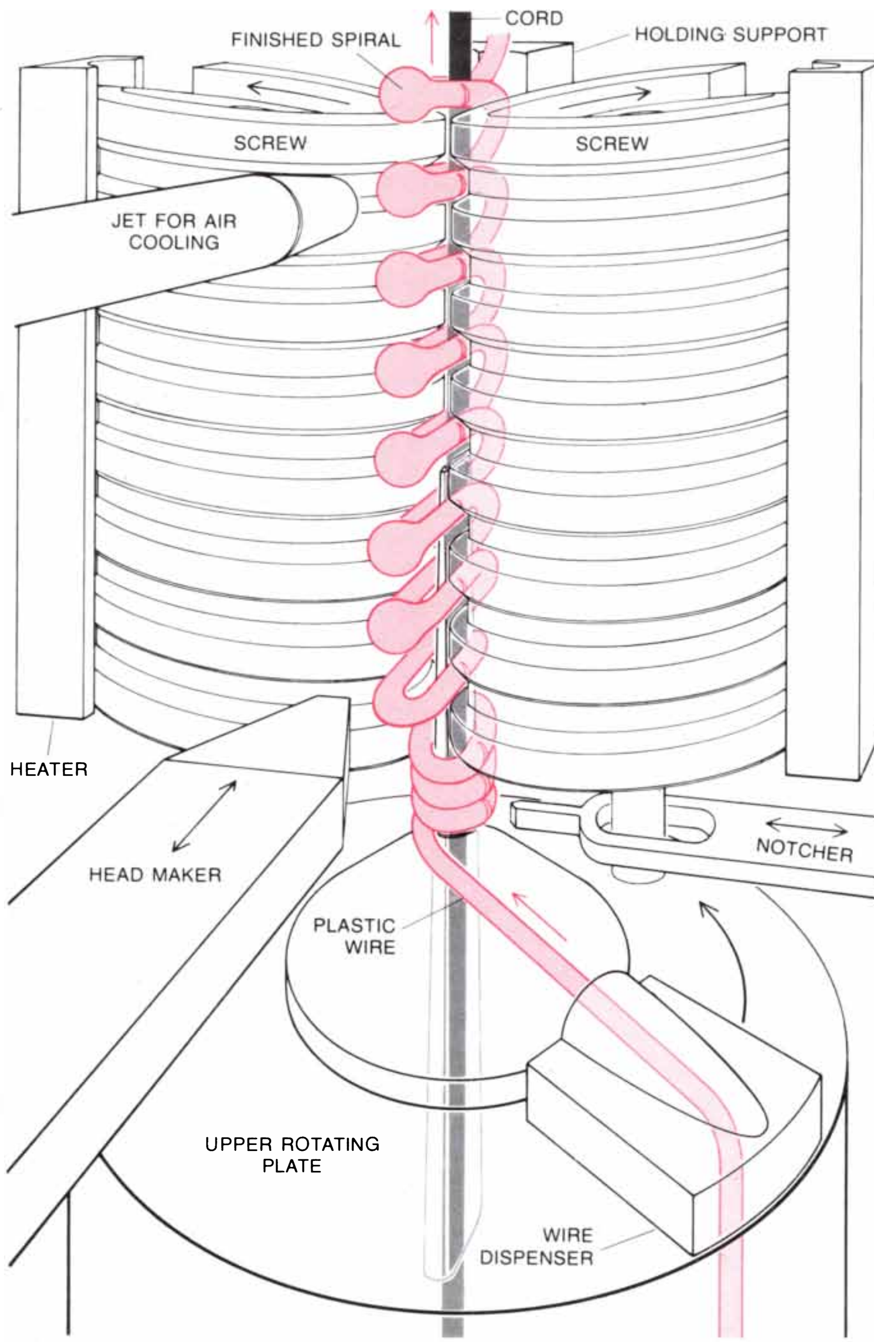
At one time zippers were made only of metal. After World War II the improvement of polymers and of methods for forming them gave rise to a variety of plastic zippers: spirals, coils, ladders and other configurations. In some of them the opposing spirals are held together by geometries that differ from the ball-and-pocket one, but the principle whereby the slider opens and closes the two sides of the chain is much the same.

Slide fasteners are made in various ways because certain types serve better than others in a specific application. For example, the zipper on a dress must be pliable and not likely to irritate the skin, and a plastic zipper in which the stringers consist of fine spirals seems to be the best choice. On the other hand, a zipper on heavy luggage must be strong in order to withstand substantial loads. A large zipper with metallic scoops or plastic coils or teeth is the best.

The size of a chain is identified by a measurement across the joined scoops or spirals. The numbers for metallic chains range from 2 (.135 inch) to 7 (.35 inch). For plastic chains the numbers range from 3 to 9, which correspond roughly to millimeters.

A zipper chain is made in one continuous length and is later cut up into individual zippers, to each of which are added a slider and appropriate stops to limit the travel of the slider according to the type of zipper. A separating zipper, as on a warm-up jacket, has a stop at the top of each stringer and a box and pin at the bottom. An open-top zipper, as on a pair of trousers, has a single stop at the bottom of the chain and a stop at the top of each stringer.

Let us consider the manufacture of zippers in two categories, according to whether the fastening element is made of metal or of plastic. I shall describe three methods of making a metallic chain. In each case a fully automatic machine does the job, and one operator can tend several machines. Essentially what happens in the first two methods is that individual scoops are made in rapid




**COILER USING ONE ROUND WIRE makes a modified spiral for a plastic stringer from polyester or nylon wire. Two joined spirals are sewn simultaneously on two tapes. Here one sees the upper part of the process, beginning where the plastic wire emerges from the dispenser and is wound around a tapered mandrel by the action of a rotating plate. At the same time a cord is pulled up through the center of the spiral; it helps to secure the spiral to the tape. As the spiral advances up the mandrel it is notched from the side; the notches also help to secure the spiral to the tape. Two heated screws, rotating in opposite directions, seize the plastic and stretch it so that the front part becomes horizontal and properly spaced. The front of the spiral is compressed to make the head. The spiral is heated for a period of time at a set temperature and is then quickly cooled so that it will maintain the desired shape. This is the process called thermosetting. Left-hand and right-hand units work side by side to make matching stringers.**

succession and clamped onto an unrolling ribbon of fabric tape. The edge of the tape is beaded so that the legs of the scoop grip firmly.

The first process begins with a round wire made of brass, aluminum or a nickel-silver alloy. The wire goes through several passes in a rolling mill that form it into a *Y* profile, so that as it is later sliced like a loaf of bread a series of individual *Y*'s is produced. Each of them becomes a scoop; the branches of the *Y* are the legs that clamp to the beaded tape.

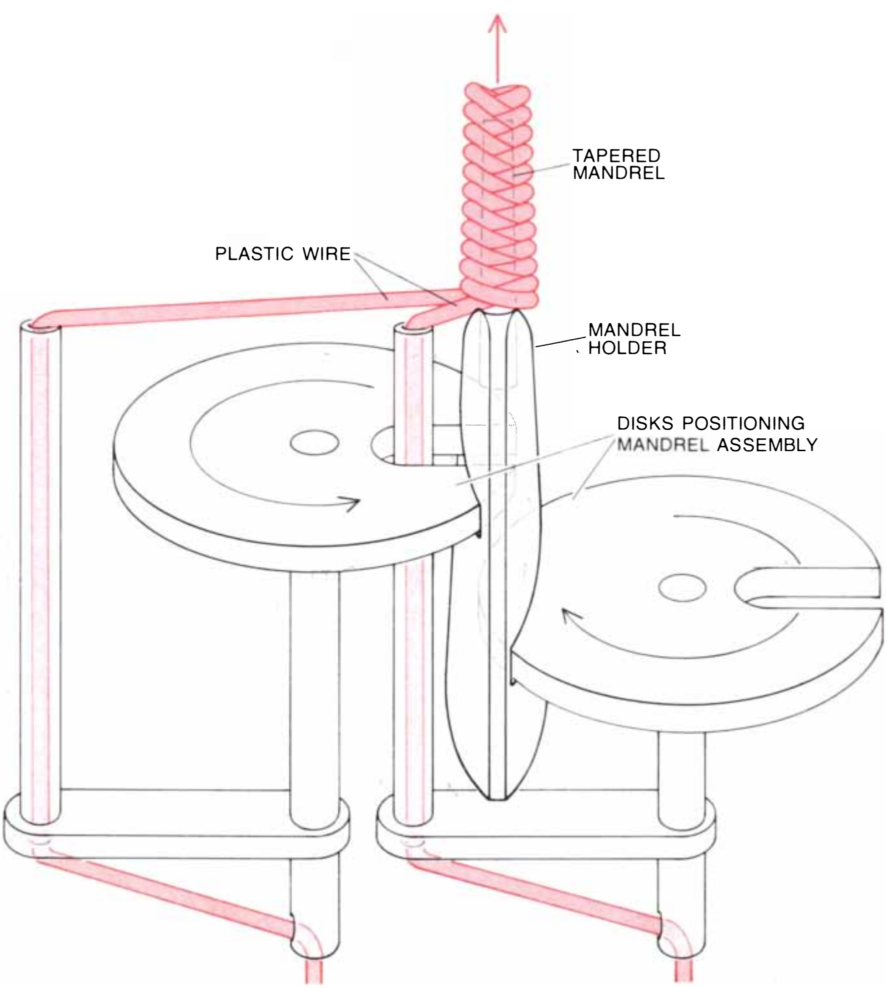
As the profiled wire is fed into the zipper machine a cutting punch slices a blank *Y* and pushes it into a heading die, one of four dies on a turntable. A pocket punch descends on the *Y* and forms the ball and pocket. After two 90-degree rotations of the turntable, during each of which a blank is added to another die, the blank reaches the beaded tape. The beaded edge of the tape is pushed into the scoop, and clinchers close the legs of the scoop around the bead. The tape is then raised by approximately twice the thickness of the scoop plus about 10 percent, a distance calculated to make room for the next scoop and also for the companion scoop from the other side that will slip between them when the zipper is closed.

This method of making a metal zipper chain is some 60 years old. Indeed, the machine is a direct descendant of the first automatic chain-making machine, which was invented by Sundback in 1923. Before that time the general method of making zippers was to stamp out scoops individually. The scoops were tumbled to eliminate sharp edges, plated and inserted manually into a fixture. When the fixture was filled, a corded tape was threaded through the legs of the scoops. The legs were then closed on the tape by a power press. Later the manual procedure was mechanized, but the process was still slow and unreliable. Moreover, it was wasteful: the stamping of the scoops left some 40 percent scrap. Sundback's automatic chain-making machine left no scrap, and it was a tremendous success. Round-wire processes, however, are relatively slow and therefore not much used today.

In the second process, which is traced back to a patent received in 1940 by Frederick Ulrich of the Conmar Products Corporation, the machine works not on a round wire but on a flattened strip of wire. Two rollers feed the wire into the machine, where a heading punch works against a stationary pocket punch to form the head and pocket on what will become the individual tooth of the stringer. Next a blanking punch cuts off notches on both sides of the wire and at the same time cuts the front scoop from the flattened wire. Before the blank is fully cut its legs are bent around the bead of the tape. A well-designed machine of this kind operates at a rate of about 50 scoops per second.

In the third process, which originated with a patent received in 1932 by Gustav Johnson, the teeth are not stamped but are cast directly on the tape. The tape moves through an open two-part mold, which has a series of cavities corresponding to the teeth. When the mold is closed, molten zinc is injected into it under pressure. The mold is water-cooled, and when the molten metal solidifies, the mold opens. The tape with the teeth moves into the next position and the dross is cut off.

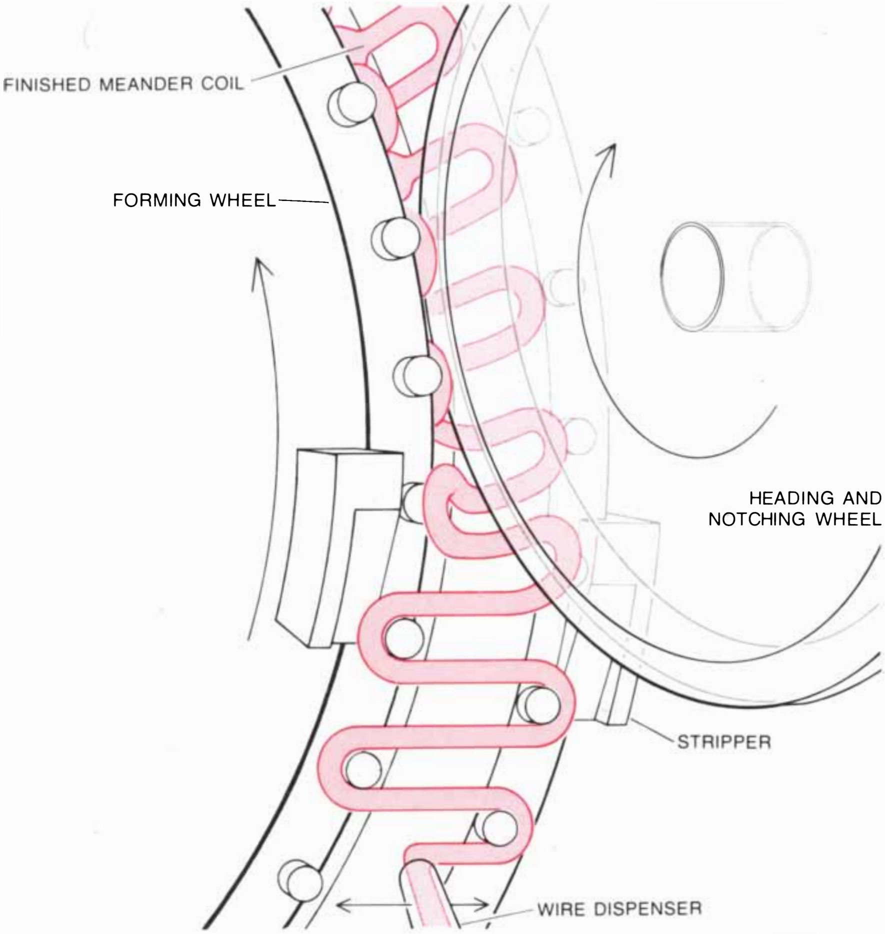
Once a metallic stringer has been made by any of these three processes it is put through certain finishing operations. Two stringers are joined into a continuous chain by a closing fixture that resembles the slider on a standard zipper. The closed chain passes through rollers that press it to the proper thickness, and then it goes through a series of wire brushes that remove sharp edges. The next station is a tank of starch. Emerging from that, the chain is wrung out by squeezing rollers and passed over hot drying cans. The purpose of this operation is to make the tape flat. Finally the scoops receive an application of hot wax, which makes the slider operate smoothly when the zipper is new. (When the zipper has been broken in, the slider will operate smoothly without wax.) The continuous finished chain is rolled onto a spool and is ready to be assembled into a zipper.



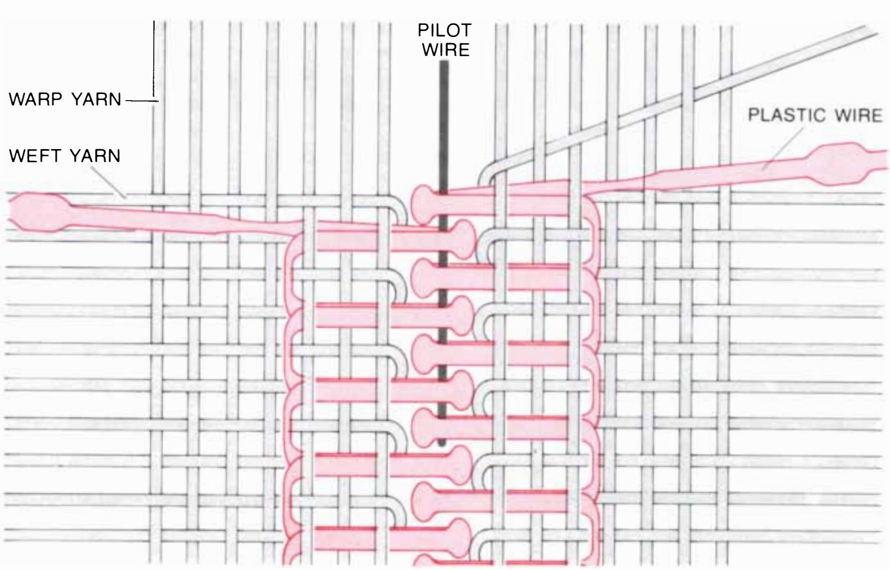

**COILER FOR TWO WIRES winds them on one mandrel. This double-wire coiler operates on plastic wire that is profiled rather than round. The wires are wound in opposite directions on a tapered mandrel. After the plastic coil is removed from the mandrel and thermoset on another wheel it is woven, each coil separately, into the stringer. The method is sparing of material.**

The plastic zipper of today made its first appearance in Germany after World War II. The German zipper manufacturers had lost factories and equipment during the war, and what had survived was antiquated. Faced with the prospect of having to rebuild their industry, the Germans decided to develop plastic zippers.

The pioneering company was the West German firm Opti-Werk GMBH. As always, however, the development of such zippers was the work of many inventors. Among them were two Ameri-




**ROUND PLASTIC WIRE is formed into a ladder (or meander) coil by a pair of wheels. The wire dispenser moves intermittently and crosswise to the forming wheel, and the forming wheel also moves intermittently so that the wire is laid between the pins. Stationary strippers lift the wire from the pins. The second wheel presses the ladder formation into a *U* on the forming wheel, the plastic is thermoset and the heads are formed. Finally the coils are sewn on the tapes.**




**LOOM WEAVING of notched plastic wire into a zipper chain is depicted. Two shuttles hold yarn and two hold notched round plastic wire. The pilot wire is attached to the frame of the loom, and as the weaving progresses it slips out of the chain. Chain made in this way is of the best quality and is particularly suitable for garments. The operation is slow, because only one of the four shuttles works while the other three are waiting, and so the manufacturing cost is quite high. Loom weaving is not done in the U.S., but many zippers of this type are imported.**

cans. In 1942 Alden W. Hanson received a patent on a method of sewing a plastic coil on a tape; in 1951 Nicholas A. Wahl received one on a method of winding two profiled plastic wires on a single mandrel. Thereafter J. R. Ruhrman and his associates were awarded a German patent on a plastic ladder chain. A. Gerlach and his associates, and independently the firm William Prym-Werke, were awarded patents on a notched plastic wire that could be woven into a tape. Another important patent was granted in 1968 to the Australian inventor E. E. Cuckson and his associates for a system that would continuously cast scoops in plastic zippers. A coiler that folds two round plastic wires into a thermosetting wheel was invented in the U.S. in 1969.

A plastic stringer has spirals made of polyester or nylon formed as a modified spiral, a meander or other shapes. In most zippers today the plastic spirals are sewn or woven on the tape, except in one version where the teeth are cast directly on the tape. Both polyester and nylon have a melting point above 204 degrees Celsius (400 degrees Fahrenheit), so that a garment with a plastic zipper can be safely ironed.

For the purpose of this discussion I shall describe two of the machines that make plastic zipper chain. One is the machine that winds round plastic wire on a mandrel. Its essential parts are the stationary tapered mandrel and a stationary cord section, both mounted on an offset shaft. (The cord runs vertically through the spiral of the stringer; it helps to secure the spiral to the tape.) A rotating spool pays out the plastic wire, and the cord is pulled from a cone. Other important parts are two heated screws, a notcher and a head maker.

The plastic wire comes off the rotating spool and passes through two rotating plates and into a dispenser, which winds it on the mandrel. At the same time the cord is pulled through the offset shaft and a slot in the mandrel, so that it ends up passing through the center of the spiral. The cord is not always used. With each revolution of the dispenser the spiral advances up the mandrel. As it advances it is notched from the side; the notches are another means of helping to secure the spirals when they are sewn on the tape.

The two heated screws, rotating in opposite directions toward the notched spiral, seize it and stretch it so that its front end is horizontal and properly spaced. A stationary bar behind the screws holds the spiral in the desired position. The screws are heated so that they set the thermosetting plastic of the spiral. After the stretching the front of the spiral is compressed to make the heads. At the end of its travel the spiral is cooled by a jet of air. Often the screws are replaced by gears to stretch


# Authorized Trek Dealers

This is a partial listing of Trek dealers. If there is no dealer in your area, call 800-558-TREK. (In Wisconsin call 414-478-2195.)

**Alabama**
Auburn/Freewheeler 887-8117
Birmingham/Cahaba Cycles 969-0476

**Arizona**
Phoenix/Happiness House Schwinn Cyclery 995-8474
Phoenix/Landis Cyclery 264-5681
Tempe/Landis Cyclery 839-9383
Tucson/Bicycles West 887-7770

**California**
Anaheim/Anaheim Schwinn Cyclery 535-3231
Berkeley/Missing Link 843-7471
Bonita/Bonita Schwinn Cyclery 475-7205
Camarillo/Open Air Bicycles 484-4433
Claremont/Bud's Bike Shop 626-3285
Fullerton/Fullerton Bike Shop 879-8310
Goleta/Open Air Bicycles 967-9669
Huntington Beach/Two Wheel Transit Authority 536-6940
Isla Vista/Open Air Bicycles 968-5571
Laguna Beach/Laguna Beach Cyclery 494-1522
Los Altos/Bicycle Outfitter 948-8092
Los Angeles/I. Martin Imports Bicycles 653-6900
Northridge/Cycle World 349-6900
Palo Alto/Palo Alto Bikes 328-7411
Port Hueneme/Open Air Bicycles 985-5045
Redwood City/Chain Reaction Bicycles 366-7130
Sacramento/City Sport Works 332-6453
San Francisco/Presidio Bicycle Shop 751-3200
Santa Barbara/Open Air Bicycles 963-3717
Santa Rosa/The Cycle Path Ltd. 526-7121
Ventura/Open Air Bicycles 653-1100

**Colorado**
Aspen/Hub of Aspen 925-7970
Boulder/High Wheeler 442-5588
Colorado Springs/Criterium 475-0149
Denver/Denver Spoke 777-1720
Denver/Life Cycle 572-8405
Denver/Turin Bicycles, Ltd. 837-1857
Durango/Four Faces 247-0500
Estes Park/Cosmic Wheels Cyclery 586-2975
Fort Collins/The Bike Broker 484-2453

**Connecticut**
Bloomfield/Bloomfield Bike & Repair 242-9884
Old Greenwich/Buzz's Bicycle Shop 637-1665

**Delaware**
Newark/Two Wheeled Cycle 368-2685

**Florida**
Orlando/Orange Cycle Works 422-5552
South Miami/Bicycle Store of South Miami 665-0125

**Georgia**
Athens/Dixon's Bicycling Center, Inc. 549-2453
Atlanta/Vinings Schwinn 952-7731
Dunwoody/Dunwoody Schwinn 455-3171

**Idaho**
Boise/George's Lightweight Cycles 343-3782

**Illinois**
Arlington Heights/Bikes Plus 398-1650
Chicago/Village Cycle 751-2412
Evanston/Turin Bicycle Co-op 864-7660
Niles/Spoke 'N Pedal 692-4240
Oak Park/Oak Park Cyclery 386-0282

**Indiana**
Bloomington/Bicycle Garage 339-3457
Fort Wayne/Denny's Bike Shop 432-7636
West Lafayette/Hodson's Bay Company 743-3506

**Iowa**
Davenport/World of Bikes 323-8227
Des Moines/Bike World 255-7047
Iowa City/International World of Bikes 351-8337

**Kansas**
Manhattan/Pathfinder 539-5639
Overland Park/The Bike Rack 642-6115
Wichita/Bicycle Peddler 262-8029

**Louisiana**
New Orleans/Bikesmith, Inc. 891-2453
New Orleans/Gus Betat & Son 821-6386
Shreveport/Sunset Bicycle 635-5079

**Maryland**
Gaithersburg/Adam's Bicycles of Gaithersburg 840-0484

**Massachusetts**
Fall River/Paul's Schwinn 674-6343
Fitchburg/Gamache's Cyclery 343-3140
Worcester/O'Neil's Bicycle Shop 798-0084

**Michigan**
Ann Arbor/Great Lakes Cycling Center 663-1604
Kalamazoo/Alfred E. Bike 349-9423

**Minnesota**
Bloomington/Penn Cycle 888-1427
Hopkins/Now Sports 827-2838
Lake Elmo/North Country, Inc. 739-3500
Mankato/The Bike Shop 625-2453
Minneapolis/Freewheel Bike Co-op 339-2219
Richfield/Penn Cycle 866-7540
St. Paul/Now Sports 644-2354

**Missouri**
Bridgeton/Touring Cyclist 739-5183
Florissant/Touring Cyclist 921-1717
Maplewood/Maplewood Bicycle Sales & Service 781-9566
Richmond Heights/Touring Cyclist 781-7973

**Montana**
Billings/The Spoke Shop 656-8342

**Nebraska**
Lincoln/Cycle Tech, Inc. 466-2921
Omaha/Bicycle Specialties 556-2453

**New Hampshire**
Durham/Durham Bike 868-5634

**New Jersey**
Brielle/The Brielle Cyclery 528-9121
Montclair/Montclair Bikery 744-7252
Morristown/Marty's Reliable Cycle Center 538-7773
Ridgewood/Ridgewood Cycle Shop 444-2553

**New Mexico**
Albuquerque/Albuquerque Bike Co-op 265-5170
Los Alamos/Strings 'N Spokes 662-3456

**New York**
Brooklyn/Sheepshead Bicycle Shop 646-9430
Mt. Kisco/Bicycle World 666-4044
New York/Bicycle Habitat 691-2783
Rochester/Towner's 271-4453

**North Carolina**
Asheville/Liberty Corner Bicycles 253-7092

**Ohio**
Centerville/International Pro Bike Shop 433-6687
Columbus/Cross Country Cycle 294-4343
Findlay/The Bicycle Shop 423-2729
Yellow Springs/Yellow Springs Bike Shop 767-1956

**Oklahoma**
Oklahoma City/Pro Bike, Inc. 721-6707

**Oregon**
Eugene/Collins Cycle Shop 342-4878
Portland/The Bike Gallery 281-9800

**Pennsylvania**
Horsham/Pedal Pusher Ski & Sport 672-0202
State College/The Ski Station 237-2655

**Texas**
Austin/World Cycle 477-2142
Dallas/University Cycle 368-2035
Fort Worth/Bikes Etc. 927-2421
Lewisville/Lewisville Bicycle Center 221-2744

**Utah**
Salt Lake City/Joe Fisher's Bike Shop 466-3971
Salt Lake City/Sid's Ski & Cycle 485-3204

**Virginia**
Norfolk/Colley Avenue Bike Shop 622-0006

**Washington**
Seattle/Angle Lake Cyclery 878-7457
Seattle/Aurora Cycle 783-1000
Seattle/Northwest Cycle 329-2453
Spokane/Wheel Sport 326-3977

**Wisconsin**
Appleton/Janus 731-2253
Green Bay/Wheels & Boards 468-4745
Hales Corners/Wheel & Sprocket 425-7930
Madison/Yellow Jersey 257-4737
Milwaukee/Rainbow Jersey Bicycle Shop 961-1198
Monona/Village Pedaler 221-0311
Sun Prairie/Bicycle World 837-3440

**Wyoming**
Lander/Freewheel Sports 332-6616

**Canada**
Toronto/Bloor Cycle 536-9718

TREK®

# and find more.

so you can get any information you want, in any order you specify, and get it a lot faster from an Apple than from a file cabinet.

You can selectively retrieve your information—like the name and address of your top salesperson on the East Coast. Or you can generate a

*With Apple's Quick File software, you can easily create electronic files that match the paper files you're used to working with. But that's the end of the similarity because...*

variety of reports—like a listing of all your accounts ranked alphabetically, or by profitability, size, age, or any other way you specify.

## Give yourself instant recall.

It's easy to understand the value of data base management to the IRS or the Pentagon. But what can it do for you?

Yes, it can shrink your paper files. Apple's ProFile™ storage system can hold up to 3000 pages of double-spaced text.

Yes, it can put the information you need at your fingertips—sales records, personnel files, expense reports, account lists, appointment calendars, price lists and inventories.

Those benefits alone have paid for thousands of Apple systems, and provided smaller businesses and departments with efficiencies that were once available only to the largest organizations. But for you, the manager or proprietor, it means much more.

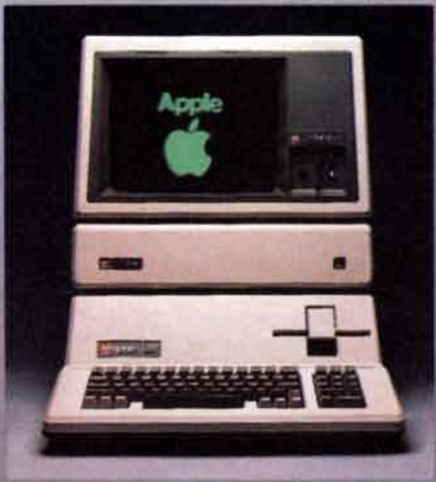
It means increased personal productivity, because you spend less time searching for information and

*Not only can you retrieve information selectively in an instant—whether a single name or all the addresses in a certain zip code—but you can also generate a variety of reports instantly, like this alphabetical inventory or...*

more time deciding how to use it.

It means a totally private information library that you can use as a management tool.

And it means the flexibility to consider information easily in many different ways, allowing you to develop an instinctive feel for the facts and make better decisions.



## More ways to establish order.

There are more ready-to-use data management programs available for Apples than for any other personal computer you can buy. Although many are specialized packages for specific types of businesses, Apple

*With a few keystrokes, you can look at the same information in chronological order. Or, for that matter, in any other order you want—by location, billing, merchandise sold, or whatever you specify.*

offers several data base packages adaptable to any office.

Quick File, for example, is designed for individual use by managers or professionals. It's extremely flexible, easy to use, and well-suited to a wide variety of applications.

## Put this in your memory bank.

For your information, Apple has more ways to obtain, store and retrieve information than any other personal computer. We also have more dealers to show you how it works and keep it working—over 1300 of them.

So call the number below to find the authorized Apple dealer nearest you.

And learn how to manipulate the facts to your advantage.

apple

The most personal computer.

Call (800) 538-9696 for the location of the authorized Apple dealer nearest you, or for information regarding corporate purchases through our National Account Program. In California (800) 662-9238. Or write Apple Computer Inc., Advertising and Promotion Dept., 20525 Mariani Ave., Cupertino, CA 95014. 

# Our Friends Indeed:

## Marlboro Music Salutes Its Corporate And Foundation Supporters And Invites You To Join Them!


RICHARD STOLTZMAN, RONALD LEONARD, ANDRE-MICHEL SCHUB, FELIX GALIMIR

This unique musical community, presided over by Rudolf Serkin, has been described as "perhaps the highest concentration of artistic activity, of whatever kind, that takes place anywhere, anytime, in the United States and possibly the world." The alumni of this internationally acclaimed center for advanced musical studies have taken their places in each succeeding generation of musical leaders where their inspiration and vitality have helped shape our musical lives.


CECILE LICAD, DAVID SOYER, RUDOLF SERKIN

Marlboro reviews with gratitude and pride the list of corporations and foundations which have supported its programs during its first thirty years, and acknowledges the National Endowment for the Arts for its annual support and for the Challenge Grant which played such an important role in the recent Endowment Campaign.

For information about Music from Marlboro concerts or radio broadcasts in your area, Marlboro Recording Society releases or the summer Music Festival in Vermont, please write Marlboro Music, 201 West 54th Street, Suite 4C, New York, New York 10019.


PABLO CASALS

Benjamin and Elizabeth Abrams Foundation, Inc.
The Alford Foundation, Inc.
American Broadcasting Company
Arwood Foundation, Inc.
Ashe Foundation, Inc.
Baer Securities Corporation
The Bell System*
Burlington Industries, Inc.
CBS Foundation
Cargill Incorporated
Cricket Foundation
Charles E. Culpeper Foundation
Dolby Laboratories
The Dover Fund, Inc.
Exxon Corporation
Ford Foundation
Foundation for Advanced Education in the Sciences, Inc.
The Fromm Music Foundation
William Randolph Hearst Foundation
Heineman Foundation for Research, Educational, Charitable and Scientific Purposes
Gunter Henle Verlag
Hochschild Fund
IBM, Essex Junction, Vermont
The Island Foundation
The JDR 3rd Fund, Inc.
L.A.W. Fund, Inc.
The Joe and Emily Lowe Foundation, Inc.
Maya Corporation
The Andrew W. Mellon Foundation
The Charles E. Merrill Trust
Henry and Lucy Moses Fund, Inc.
National Broadcasting Company, Inc.
Walter W. Naumburg Foundation, Inc.
The New York Community Trust
The May and Morris Newburger Foundation
Edward John Noble Foundation
William Penn Foundation
Philip Morris Incorporated
The Presser Foundation
The Rockefeller Foundation
The Martha Baird Rockefeller Foundation for Music, Inc.
Rolled Alloys
The Scherman Foundation, Inc.
Scientific American, Inc.
The Franz W. Sichel Foundation
Spanel Foundation, Inc.
Steinway & Sons
The Florence Louchheim Stol Foundation
The Jean Tennyson Foundation, Inc.
Time, Inc.
von Siemens Fund
The Sulzberger Foundation, Inc.
United Technologies Corporation
Bruno Walter Memorial Foundation
Anonymous Donors

**Special funds given in connection with the Bell Telephone Hour program "Casals at Marlboro."*

and form the spiral and then the thermosetting is done by induction heat.

In a chain of the spiral type the two stringers must have oppositely wound spirals. Therefore two machines are necessary to wind a left coil and a right one. Normally they operate side by side so that the two spirals can be joined into a double spiral. The joined double spiral is ready to be sewn onto the tape.

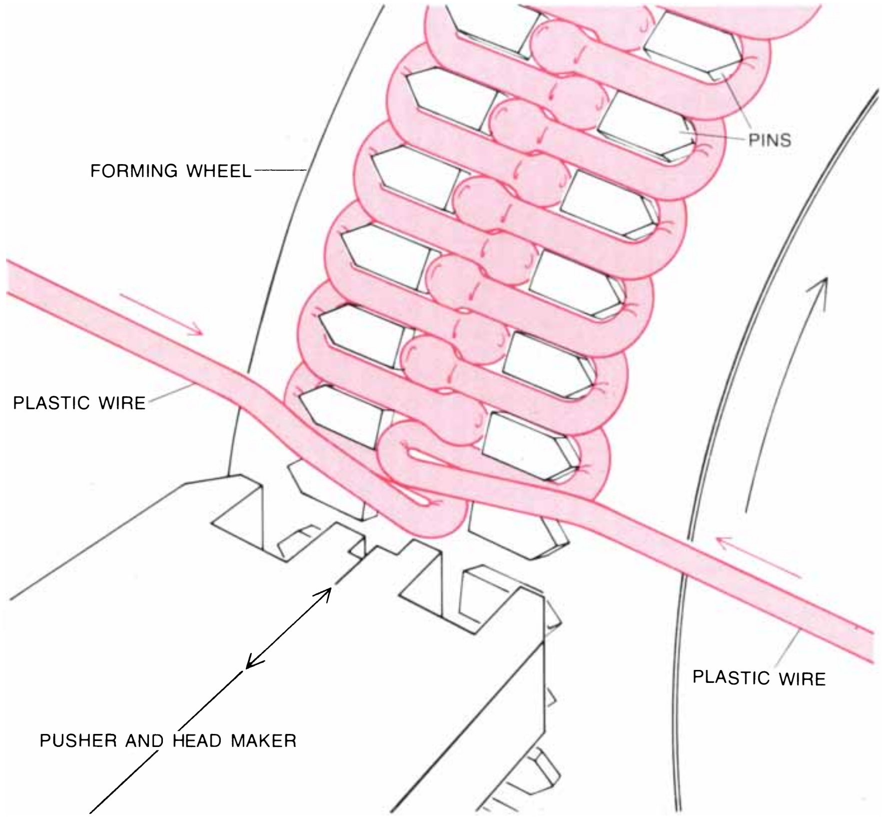
The continuously cast plastic scoops are made on a rotating casting wheel that has along its perimeter uniformly spaced cavities corresponding to the shape of the teeth. A pair of cords are fed to each line of teeth. The cords hold the series of teeth together and facilitate the removal of the string of units from the wheel. Semimolten plastic material is delivered to each passing scoop mold on the rotating wheel. An apron holds the material in place until it solidifies, and then a stationary knife separates the surplus plastic from the units. Next the units go through a folding machine, where they are folded into the proper shape and thermoset. The next step is to sew them on the zipper tape.

A plastic chain goes through finishing operations after it is made. It is put through a starch tank, squeezing rollers and a drying apparatus to make the tape flat. Finally the continuous chain is wound on a paper spool.

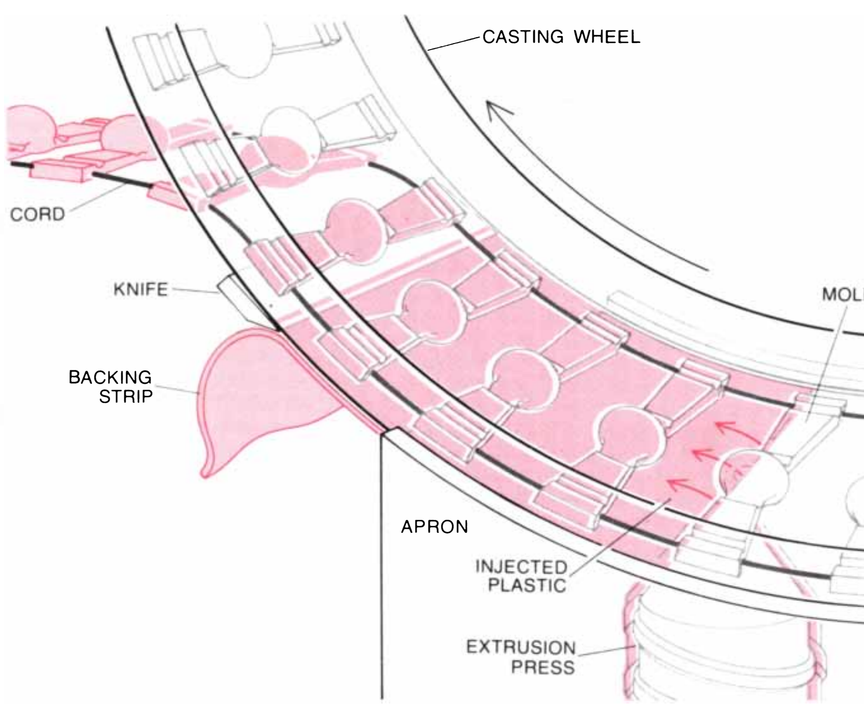
In order to close or open a zipper a slider is needed. The slider body and the pull tab are either stamped from a strip of mild steel or die-cast out of zinc. In general the stamped-out sliders are put on metal chains and the die-cast sliders are put on plastic chains. A slider on a plastic chain must have ribs inside its body to force the spiral into the right position when the zipper is being closed. It is difficult to make the ribs in a stamping process, which is why the die-cast slider is commonest on plastic chains.

To make the slider rustproof the component parts are plated with brass or zinc. An exception appears in what are called automatic sliders. Such a slider has a spring that holds the slider in place until the pull tab is lifted; the spring is likely to be made of stainless steel. Often the slider is enameled to match the color of the chain. The body, the pull tab and (in automatic sliders) the spring are assembled in a special machine.

The last step in making a zipper is to assemble the chain, the slider (two sliders for the type of zipper that can be opened and closed from each end) and the stops. The assembling machinery is different for metallic and plastic zippers, but the methods are essentially the same. A typical assembly operation begins with the continuous zipper chain on a spool. A pair of circular pullers unrolls the chain, pulling it first through a gapper that removes teeth at certain in-




**COILER FOLDING TWO ROUND WIRES makes the right and left spirals at the same time. The two round plastic wires are fed from each side into the forming wheel, in a confined space where they form loops. The pusher and head maker force the wire between the pins. After thermosetting the product is taken off the wheel as two closed spirals, which are sewn on the tapes.**




**CASTING OF PLASTIC TEETH is done on a rotating wheel that has uniformly spaced cavities corresponding to the shape of the teeth. The two cords fed to each line of teeth connect them and make it easier to remove individual units from the wheel. Semimolten plastic is delivered from the extrusion press to the cavities, where it is held in place by the cooled apron until it solidifies. The knife removes the surplus plastic. In a separate operation (not shown) the teeth are folded into the shape of a *U;* they are then ready to be sewn on a flat tape.**

tervals to provide a stretch of empty tape and simultaneously attaches the bottom stop; when the gap is cut across at the midpoint, it determines the length of the zipper and provides the bit of extra tape needed to properly sew the zipper into a garment. The gap can also be made on the chain machine by omitting the teeth. The chain then goes through a machine that slips the slider through the gap in the chain, and the top stop is put on. The last step is pinking, that is, cutting across the chain in the middle of a gap. Sometimes the operations are combined. The result is an individual slide fastener, ready to be attached to a garment or bag or whatever product it was designed for.

The manufacturing process is slightly different for a separating fastener. A cotton tape, laminated with nylon, is fused across the gap as a reinforcement and top stops are attached. The laminated tape is severed along an extension of the axis of the zipper. The slider is slipped on one stringer and the box is also mounted there. A pin is mounted on the other stringer. When the zipper is opened all the way, the pin pulls out of the box through the slider and the fastener is fully separated, with the slider held on its stringer by the box.

One might suppose little in the way of new developments could be expected in a device as simple and standardized as the zipper. Actually, however, several possibilities are in sight, particularly for plastic zippers. One is for zippers both much larger and much smaller than the ones now on the market. The larger ones would be for heavy-duty action in fastening, say, curtainlike dividers that allow small meeting places to be made in a large room. The smaller zippers would be for products that reflect the current trend toward miniaturization, such as for carrying cases to fit the Walkman type of portable radio.

Another prospect is for an inexpensive zipper that would serve on disposable paper garments and other objects designed to be used once and then discarded. In a more exotic application one can envision a zipper that is strong, airtight and yet reasonably flexible to be put in garments for astronauts. If such a zipper is developed, it would probably turn up soon in consumer products that need fasteners with those attributes.

A final example is an application that is still looking for a zipper. In surgery there is a need for an airtight and chemically inert zipper that would serve in place of stitches in an incision that might have to be reopened, say, for access to an artificial organ such as a heart pacemaker. Several surgeons have expressed a desire for such a device; to design and make it will be a challenge for medical investigators and the zipper industry.

# THE AMATEUR SCIENTIST

## *What causes the color in plastic objects stressed between two polarizing filters?*

by Jearl Walker

The failure of materials as a result of stress is both common and costly. One way to guard against such a failure is to test a sample object by applying enough stress to destroy it; when an identical object is put in service, the stresses are kept well below the destructive level. Another way is to make a model of the object out of plastic or some similar material and then to subject the model to photoelastic analysis, which yields a pictorial representation of the stresses in it.

Recently Frank R. Seufert of Cleveland showed me some examples of his photoelastic analysis of models of various objects. When someone is concerned about the stresses on an object, Seufert makes a model of it in Lexan or Tuffak plastic about an eighth of an inch thick. If the object is large, the model is a scaled-down version. Seufert mounts the model in a wood frame and undertakes to mimic the stress on the real object by stressing the model with rubber bands or machine screws. The screws work better because they can be tightened slowly and controlled more easily.

Seufert puts the model in front of a polarizing filter and illuminates it with light that passes from a 200-watt lamp through a ground-glass diffuser and then through the filter. Between the model and a 35-millimeter camera (a single-lens reflex model) he positions a long lens hood made out of cardboard. Attached to the camera is a Vivitar 2X teleconverter with its lenses removed. It serves as an extension tube to move the lens farther from the film; a different type of tube would serve as well. Seufert's setup includes a 135-mm. telephoto lens rated at *f*/2.5. A second polarizing filter and a light blue filter (designated 80A) are mounted on the lens. The camera is held steady with the top piece from a tripod; the piece is attached to a block of plywood.

To make a photograph Seufert orients the first polarizing filter with its axis of polarization 45 degrees from the vertical. Then he puts his plastic model in front of it and turns on the lamp. As he sights through the camera he rotates the polarizing filter in front of the lens until a meaningful pattern is superposed on the image of the plastic model. He records the scene with a cable release that also triggers an electronic flash positioned near the lamp. The extra light is needed for a good photograph.

The film is Fuji color negative rated at ASA 100; Seufert has glossy prints made from the negatives. The colored lines in the accompanying photographs reveal the stress patterns in the plastic models. Regions under stress stand out. They are the places where the object is most likely to fail.

The first polarizing filter in Seufert's setup polarizes the light. The passage of this kind of light through stressed plastic encodes information about the stresses. The second polarizing filter (the one mounted on the camera) makes the coded information visible to an observer. The top illustration on page 149 depicts the stresses on a tiny, thin element that is part of a model being investigated by photoelasticity. Each edge of the element is under stress because of forces pulling perpendicular to the edge. Each edge is also under shear as material on opposite sides of the element attempt to slide in opposite directions.

Such is the nature of the stresses on a randomly chosen element. The picture is simplified if a differently oriented element is chosen at this region in the material. The new element is special in that it is square and has two important axes called the principal stress axes. The advantage of considering such a specially oriented element is that its edges are not affected by shear. The only stress on an edge is the perpendicular one. The orientation of the principal stress axes is what is revealed in a photograph of a stressed model illuminated with polarized light.

To understand the interaction of polarized light with the principal stresses in the plastic it is necessary to understand the nature of polarized light. According to classical physics, light is a wave composed of oscillating electric and magnetic fields. It is a peculiar kind of wave in that nothing material participates in the wave motion. Water waves accord more with intuition because a wave involves oscillations of the water surface. Something material participates. In the wave concept of light, however, the oscillation is of immaterial electric and magnetic fields.

The electric components specify the polarization of light. An electric field is a vector (of a specified size and direction) assigned to a point to be examined, as is shown in the middle illustration on page 149. The concept is useful when one considers how a charged particle might behave when it is placed at that point. It is also useful in forming a mental picture of light.

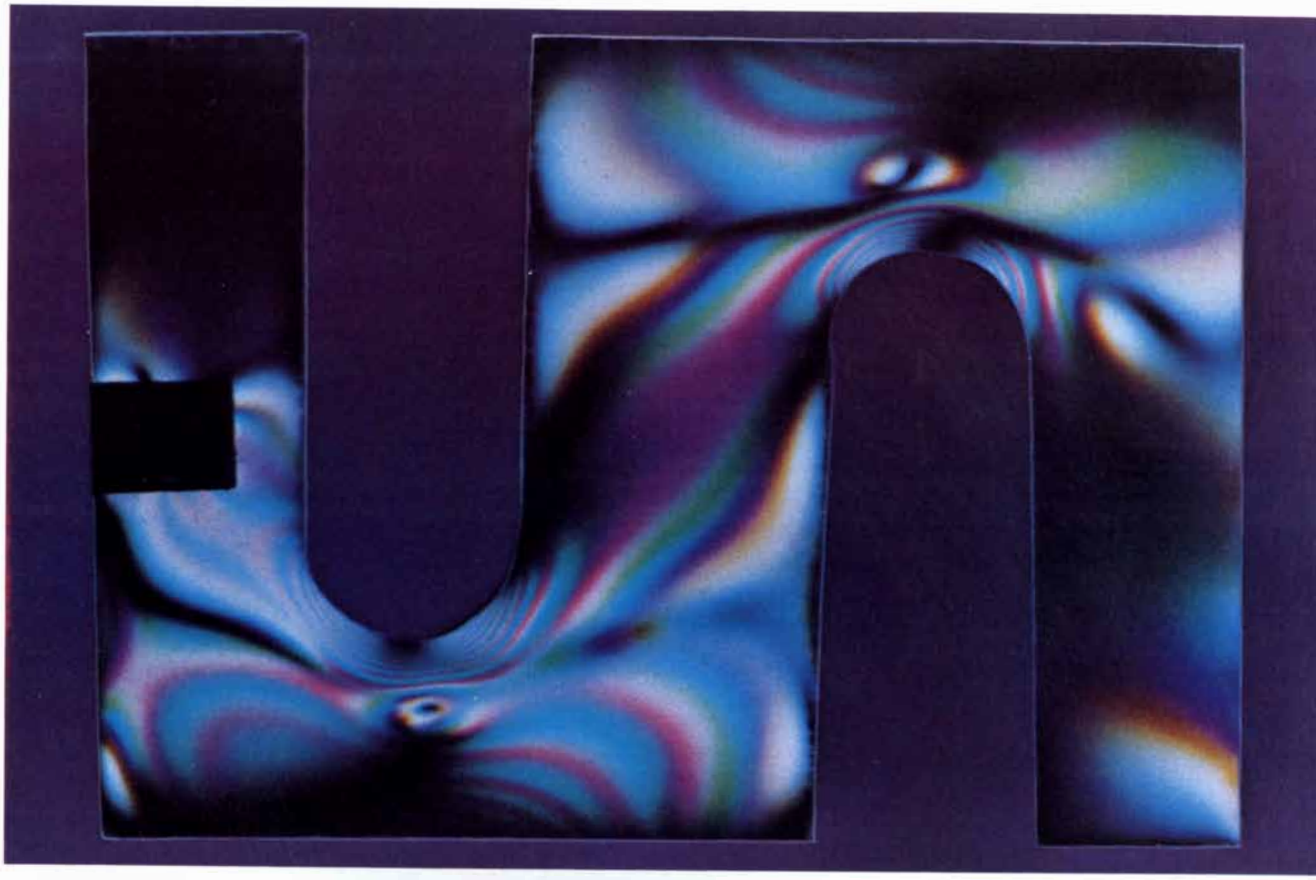
The illustration can be regarded as a snapshot of a light wave. Superposed on the ray, which indicates the direction of travel, are some of the electric vectors assigned to points along the ray. One point is singled out for scrutiny. At that point in the first snapshot the electric vector is long and points upward. A positive particle there would "feel" a strong force upward.

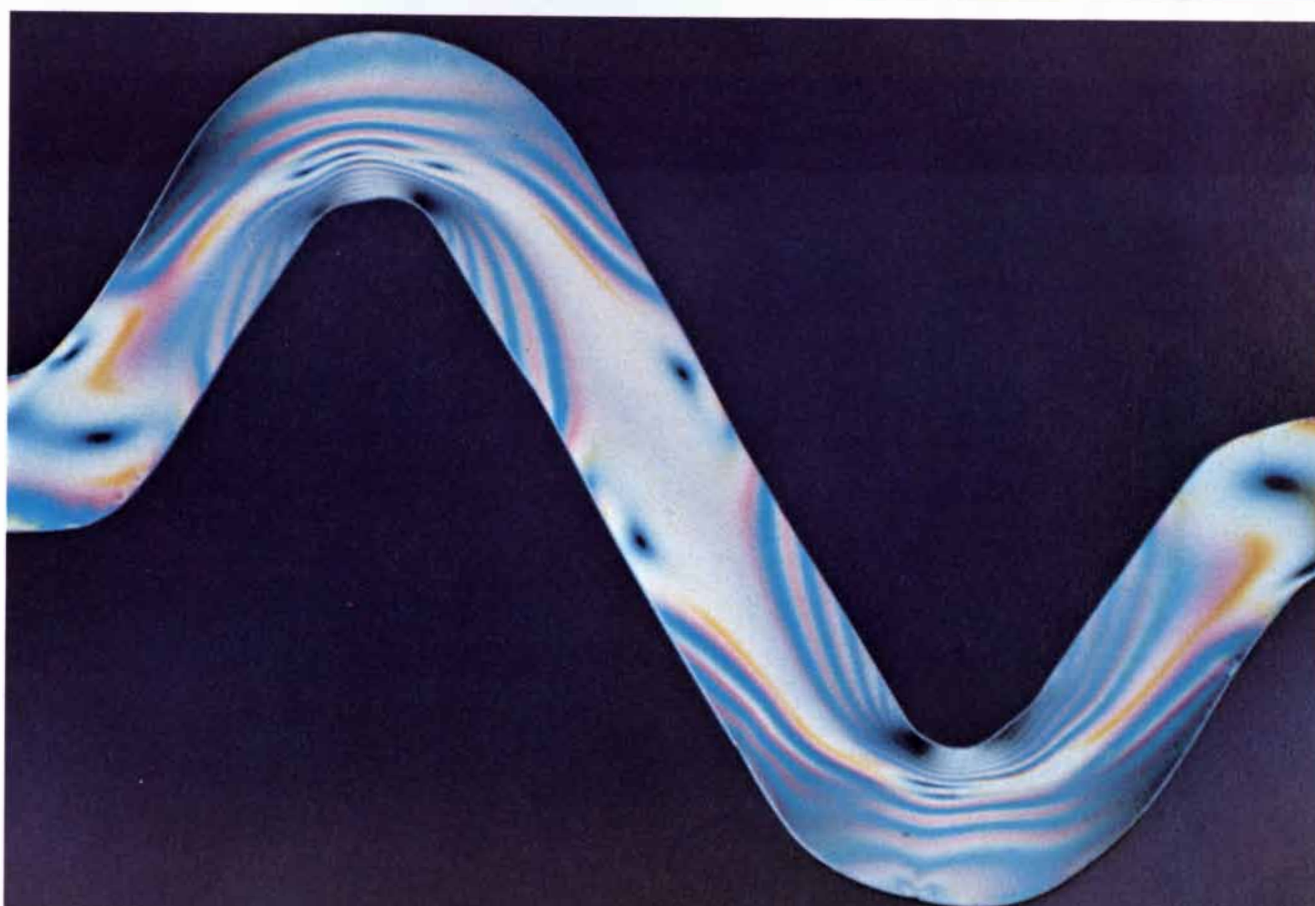
The light continues to move toward the right after the first snapshot is made. The electric field at the chosen point changes quickly, literally at the speed of light. A new snapshot reveals the change. The electric vector at the designated point is now downward. Since it is not one of the largest vectors in the illustration, the electric field is not as large as it could be and not as large as it will be immediately after this snapshot. A positive particle at this point would now "feel" a moderate downward force.

As light sweeps past the point, the electric vectors oscillate in direction and strength. Do not, however, be misled by the pictorial representation. The vectors are not material. A ray has no vectors sticking out of it like thorns on a rose stem. The electric vectors are bits of imagination stuck on a line called a ray.

Fictitious though the electric vectors are, they are almost essential to an understanding of the polarization of light. In the two snapshots of light the electric vectors are all flat in the plane of the page. Light from most common sources has vectors that are not as limited in direction. The vectors are necessarily perpendicular to a ray of the light, but they can point in any direction in a plane perpendicular to the ray. Such light is said to be unpolarized.

If light passes through a polarizing filter, the oscillation of the vectors is strictly limited to a single axis. Such light is said to be linearly polarized. (Some call it plane-polarized light.) If unpolarized light from an electric light bulb strikes a filter, the emerging light has its electric vectors confined to a single axis in a plane perpendicular to the ray. The direction of polarization of the light is the





*Two of Frank R. Seufert's plastic models under compression* orientation of that axis. If the axis is vertical, the light is said to be vertically polarized.

Polarizing filters work by a process of elimination. The filter incorporates molecular chains that can be visualized as long, parallel absorbers. As light reaches them the electric vectors that oscillate parallel to them are eliminated; the perpendicular vectors pass through.

For example, if the filter's long molecules are stretched out horizontally (parallel to the *x* axis), only the horizontal components of the electric vectors in the light are eliminated. The bottom illustration on the opposite page presents a shorthand notation in which two double vectors represent unpolarized light; the filter eliminates the horizontal one and passes the vertical one. The result is vertically polarized light.

Normally the orientation of the molecules in a filter is not specified. Instead the filter is assigned a polarization axis perpendicular to the length of the molecules. This fictitious axis is parallel to the polarization of the emerging light.

Suppose a ray of vertically polarized light encounters a second polarizing filter. Will the light pass through? That depends on the polarization axis of the filter. If the axis is vertical, it is parallel to the polarization of the light approaching it. All the light is transmitted. If the axis is horizontal, none of the light is transmitted.

When light moving through air at a speed of $3 \times 10^8$ meters per second enters any transparent material, its effective speed decreases. The reason is that the light interacts with the molecules in its path. At each encounter with a molecule the light is absorbed; after a brief delay it is reemitted. Between molecules its speed is $3 \times 10^8$ meters per second, the value it has in a vacuum. Since it is intermittently delayed, however, it takes longer to travel through the material than it does to travel an equal distance in a vacuum. The light is said to be traveling slower in the material.

This effect was measured indirectly long before anyone knew about molecules. In order to tabulate the effect each transparent material was assigned the number called the index of refraction. A glass with an index of 1.6 transmits light slower than one with an index of 1.5. (In either case the actual time of transmission is so incredibly short that the difference is immaterial in everyday life.)

In 1816 David Brewster discovered how the index of refraction can aid in the analysis of stress in a transparent material. When he applied a stress to a sheet of glass illuminated with linearly polarized light, he found a change in the index of refraction of the glass. Moreover, the index then depended on how the light was polarized.

To follow Brewster's experiments visualize a vertical sheet of glass uniformly compressed by forces applied at the top and bottom in such a way that the principal stress axes are vertical and horizontal. When the light illuminating the glass is polarized vertically, it encounters a smaller index of refraction and moves with higher velocity than it does when it is polarized horizontally. If the glass is under tension (stretched by forces at the top and bottom), the results are just the opposite. A material of this type, with a speed of transmission that depends on the polarization of the light, is said to be birefringent, or doubly refracting.

A simple example will demonstrate how birefringence can aid in stress analysis. A vertical sheet of plastic is uniformly compressed to make the principal stress axes vertical and horizontal. The way polarized light interacts with the stressed plastic depends on whether the polarization of the light is aligned initially with one of those axes. You can create such an alignment by illuminating the plastic through a polarizing filter with its axis vertical. On the opposite side place another polarizing filter with its axis horizontal. The first filter is often called the polarizer, the second one the analyzer; the two are said to be crossed.

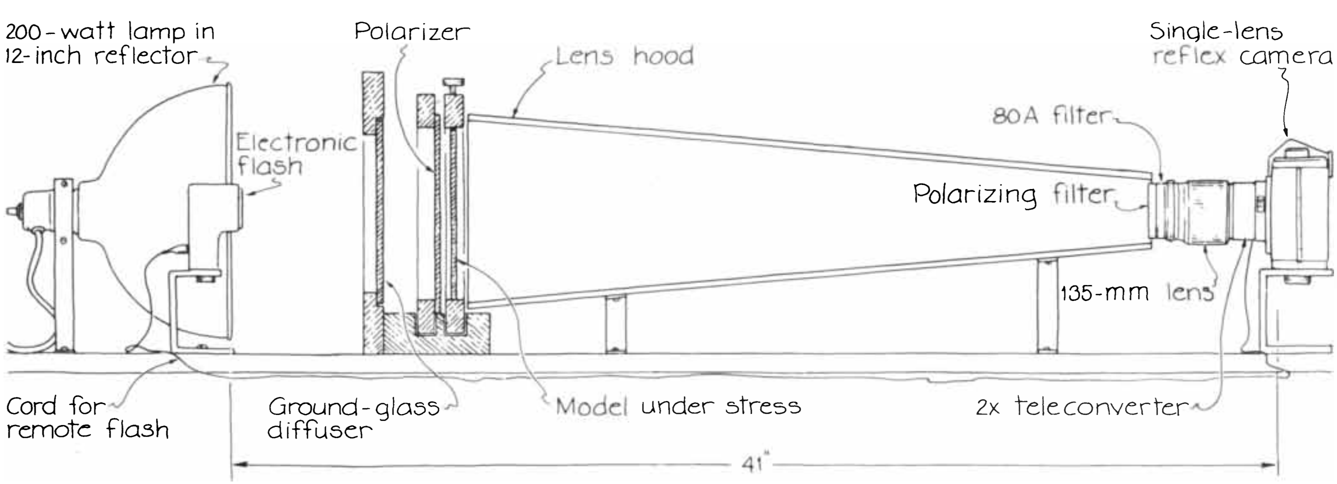
The light passes through the sheet of plastic at whatever speed the stress allows. The light emerging from the plastic is still vertically polarized and is therefore stopped by the analyzer. An observer looking through the analyzer sees darkness. In the experimental setup transmission is eliminated whenever the plastic is illuminated with light polarized parallel to a principal stress axis.

Next rotate each filter by 45 degrees in the same direction. The plastic sheet now receives light polarized at 45 degrees from the vertical. The course of light through the sheet is a bit harder to follow because the polarization must be considered in two components parallel to the principal stress axes. The two components travel through the plastic at different speeds because the indexes of refraction along the two axes differ.

When the components emerge from the plastic, they in effect recombine. The light is then likely to have a new direction of polarization. Whether or not the light will pass through the analyzer depends on how the polarization has been changed.

To determine the change one must consider how the wavelength of the light is altered by the plastic. For simplicity assume that the light is of a single wavelength. When light passes from air into a transparent material, its wavelength shortens. The larger the index of refraction, the shorter the wavelength of the light in the material.

Since stressed plastic has a different index for each principal stress axis, the extent to which the wavelength shortens depends on the polarization of the light. If the light is vertically polarized, its wavelength is divided by the index of refraction associated with the vertical principal stress axis. If the light is horizontally polarized, the wavelength is divided by the index associated with the horizontal axis. When the polarization lies between the two axes, it must be considered in components. The vertical




*Seufert's experimental setup*

component is shortened in wavelength by one amount and the horizontal component by a different amount.

As the two components proceed through the plastic with their different wavelengths they oscillate a different number of times. For example, the component polarized parallel to the vertical stress axis might have 1,000 oscillations (wavelengths); the other component, being of shorter wavelength, might have an additional oscillation, for a total of 1,001. The two components begin their trip exactly in step and emerge again in step even though one of them has an additional oscillation.

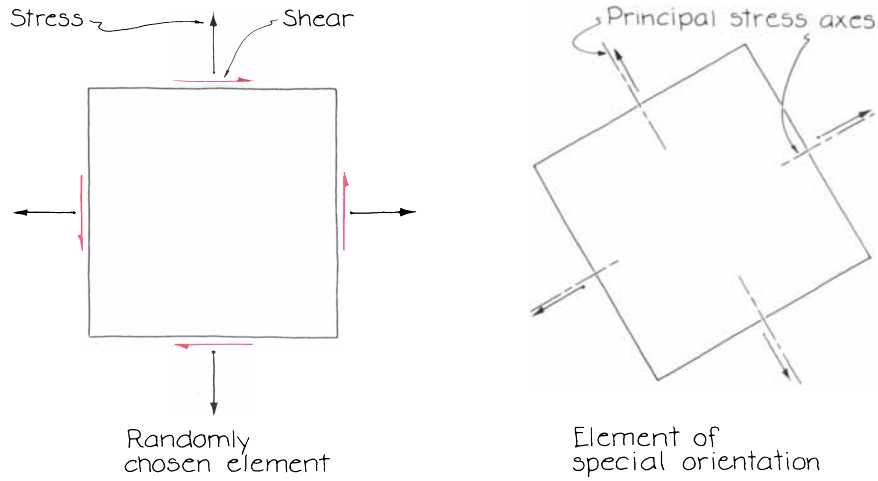
When one mathematically recombines the two emerging components, the electric vectors of the light are found to be oscillating exactly as they were before they entered the plastic. The polarization is tilted from the vertical by 45 degrees. Since the polarizing filters are crossed, the light reaching the analyzer is blocked.

If instead one component has an additional half oscillation, recombination yields a polarization rotated by 90 degrees. This new polarization is parallel to the polarization axis of the analyzer and so the light passes through.

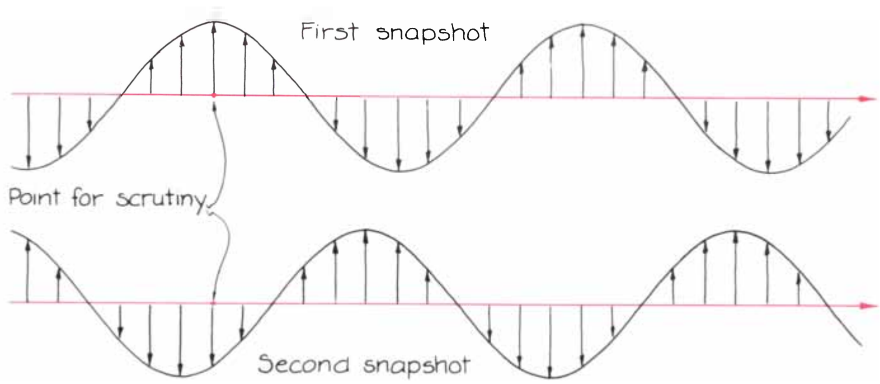
Any intermediate result is also possible. Recombination is then harder to visualize. The components do not yield linearly polarized light. Instead the polarization rotates continuously about the ray of light. The electric vector rotates from upward to downward and then back to upward. Light with a rotating polarization is said to be elliptically polarized. (In the special case where the maximum size of the electric vector remains the same throughout the rotation, the light is said to be circularly polarized.) Elliptically polarized light is partly transmitted by the analyzer. The component of polarization that is parallel to the polarization axis of the analyzer is passed whereas the other component is blocked.

The point of this discussion is that what the observer sees through the analyzer depends on the angle between the polarization of the light and a principal stress axis in the plastic. Alignment of the polarization with an axis yields darkness at the analyzer. If in misalignment the plastic releases light of unchanged polarization, the analyzer again blocks it. In other instances at least some light gets through the analyzer.

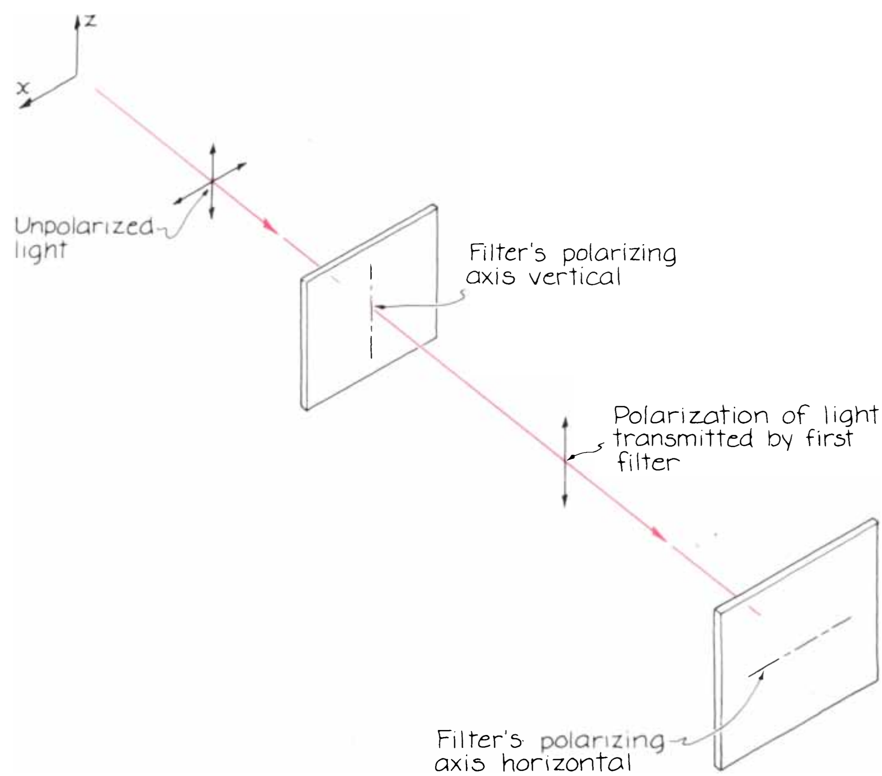
The example is simple in that the stress is uniform and the principal stress axes are vertical and horizontal at any point in the plastic sheet. If the plastic is subjected to various stresses, the axes will be oriented differently for each point in it. The purpose of photoelasticity is to discover how they are oriented; in this way one might be able to locate a section in the model that might fail under the stress. When a complex and




*Stress and shear on an element*

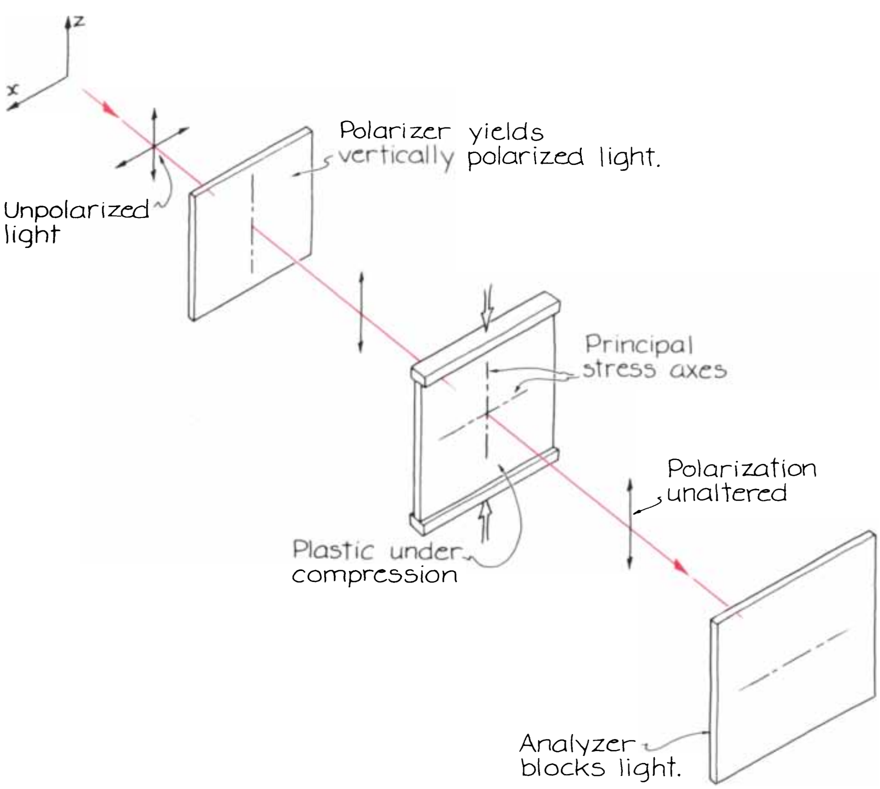



*"Snapshots" of electric vectors along a ray of light*

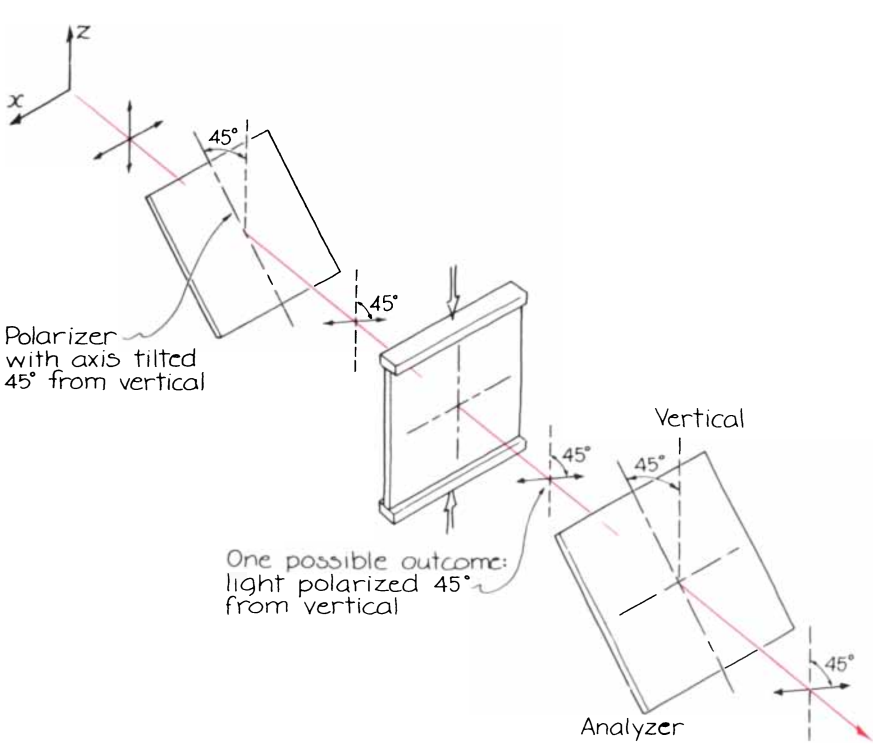



*How crossed polarizing filters block light* unknown arrangement of stresses is applied to the plastic, the model as seen through the analyzer has a superposed pattern of the dark and bright lines called fringes.

A dark fringe marks points within the plastic from which the emerging light is polarized in exactly the way that keeps it from going through the analyzer. The reason is either that the light passing through one of these dark points is polarized parallel to one of the principal axes or that it has two components (one parallel to each principal stress axis there) that recombine to yield the same polarization the light had when it entered the plastic. In either case the analyzer blocks the light.




*An alignment of polarization with a principal stress axis*




*One of the possibilities when the polarization is tilted*

Usually the first condition creates most of the dark fringes. A fringe caused in this way is called an isoclinic. The pattern of isoclinics reveals the orientation of the principal stress axes within the plastic.

To map the principal axes you photograph the plastic through the analyzer at a particular orientation of the two polarizing filters. Suppose the axis of the first filter is vertical and the axis of the analyzer is horizontal. The plastic is thus illuminated with vertically polarized light. The isoclinics photographed through the analyzer mark the points on the plastic that have one of their principal stress axes vertical. (Since the stress axes are at right angles to each other, the other axis must then be horizontal.) The pattern in the photograph is traced on paper, and the principal axes are drawn superposed on several points on the isoclinics.

Next the filters are rotated 10 degrees, say, and another photograph is made. The isoclinics mark the points with a principal stress axis tilted from the vertical by 10 degrees. These isoclinics are added to the trace and again the principal axes are sketched in at a few places. After a few more photographs have been made the trace shows how the principal stress axes are oriented at many places in the plastic.

Now you add to the map lines that connect points having equal stress values. For example, a line would be drawn from a principal axis at one point to another principal axis at a nearby point. It is likely to be a curved line. Although a certain amount of guesswork is inescapable here, the result is a rough map of the lines of principal stress.

The dark fringes resulting from recombination are called isochromatics. They are usually masked by the isoclinics, but they can be employed to assign values to the principal stresses revealed by the isoclinics. To make use of them you must first eliminate the isoclinics.

You can do it by putting two additional filters, called quarter-wave plates, in the path of the light. One plate goes between the first polarizer and the plastic and the other one goes between the plastic and the analyzer. The function of a quarter-wave plate is to transform linearly polarized light into circularly polarized light.

The plate works something like the stressed plastic being analyzed. The plate is birefringent, that is, it has two orthogonal axes ("fast" and "slow") that pass light at different speeds. Suppose the first polarizer yields light that is polarized vertically. The first quarter-wave plate is set with its fast axis tilted from the vertical by 45 degrees. The light reaching the plate thus has components along both the fast and the slow axes. As the two components pass through the plate they oscillate a different number of times. The polarization of the light emerging from the plate depends on that difference. Since the plates are designed to make one component complete a quarter of an oscillation more than the other component (hence the name quarter-wave plate), circularly polarized light emerges.

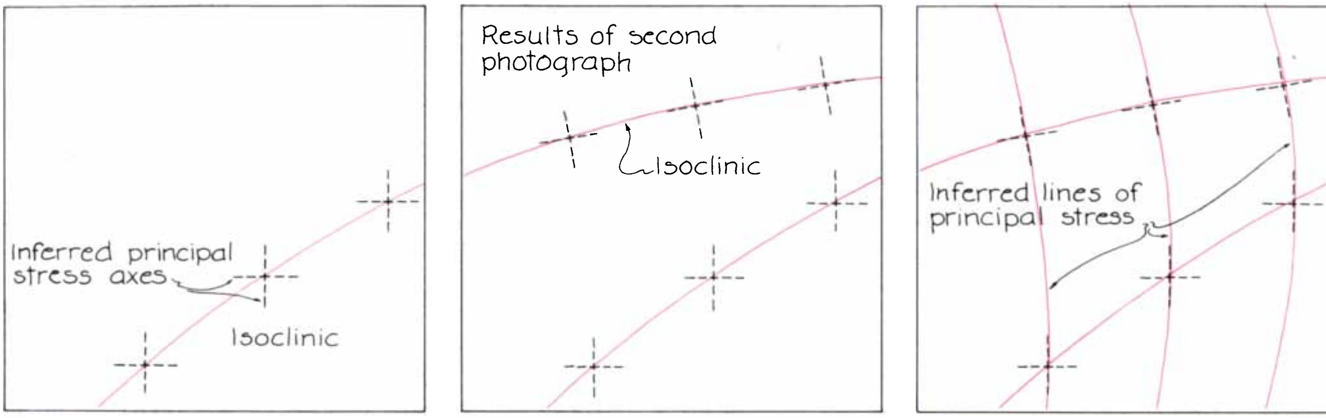
The second quarter-wave plate is arranged with its fast axis perpendicular to that of the first one. The function of the second plate is to subtract the quarter-wavelength difference imparted to the two components of light by the first plate. The net result may seem hardly worth the trouble, but a useful end is gained. When the circularly polarized light passes through the stressed plastic, the polarization cannot be strictly parallel to a principal stress axis. Therefore the isoclinics resulting from such an alignment are eliminated. What one receives through the analyzer is a pattern of the isochromatics.

The advantage of this pattern is that the fringes are related to the size of the stresses in the plastic. At any given point in the plastic the difference between the principal stresses determines the polarization of the light. Thus the stress difference determines whether the point ends up as part of a dark fringe or as part of a bright one. In principle the stress values for any point in the plastic can be computed by examining the isochromatics.

Since the isochromatic pattern depends on the shortening of the wavelengths of the two components of light transmitted through any point in the plastic, the pattern depends on the wavelength of the light illuminating the plastic. With white light each color creates its own pattern. At one point in the plastic one of the colors might leave with the same polarization it had when it entered. That color is eliminated by the analyzer. The other colors passing through the same point are changed somewhat in polarization and so are at least partly transmitted by the analyzer. The observer sees that point on the plastic as being neither white nor dark but a color. The composite of the points generating the same color forms a pattern of colored isochromatics.

The actual color the observer will see at a particular point is difficult to predict. It depends on how the various colors are altered in polarization. It also depends on the perception of the observer or the color response of the film in a camera. Color prediction becomes even more complicated when the stress differences at each point in the plastic increase. When the differences are large enough, the colors begin to wash out into white.

The photographs made by Seufert are from experiments where linearly polarized light is sent through stressed plastic. Both the isoclinics (which are only bright and dark) and the isochromatics

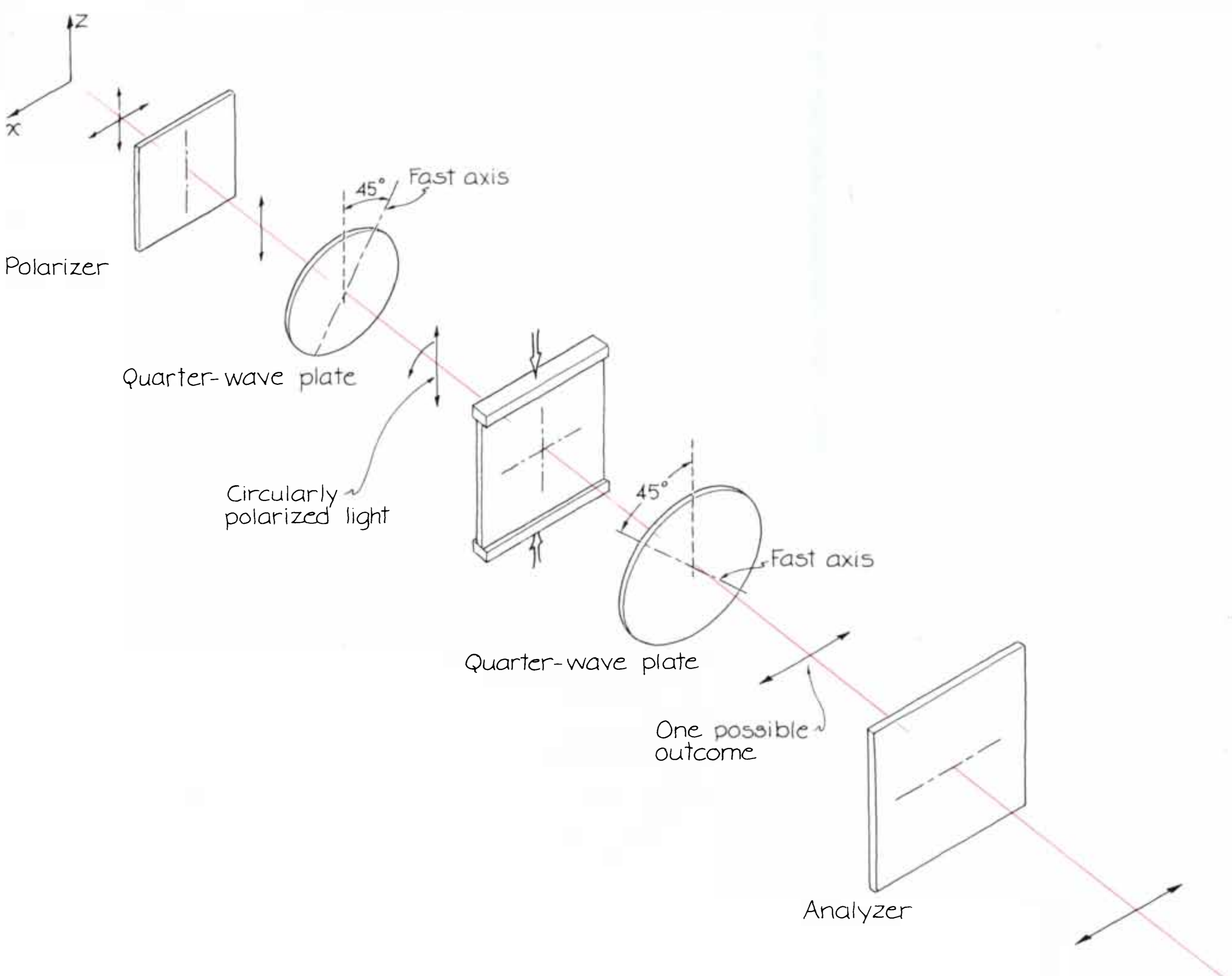



*The analysis of isoclinics*

(which are color-dependent) are included in the photographs.

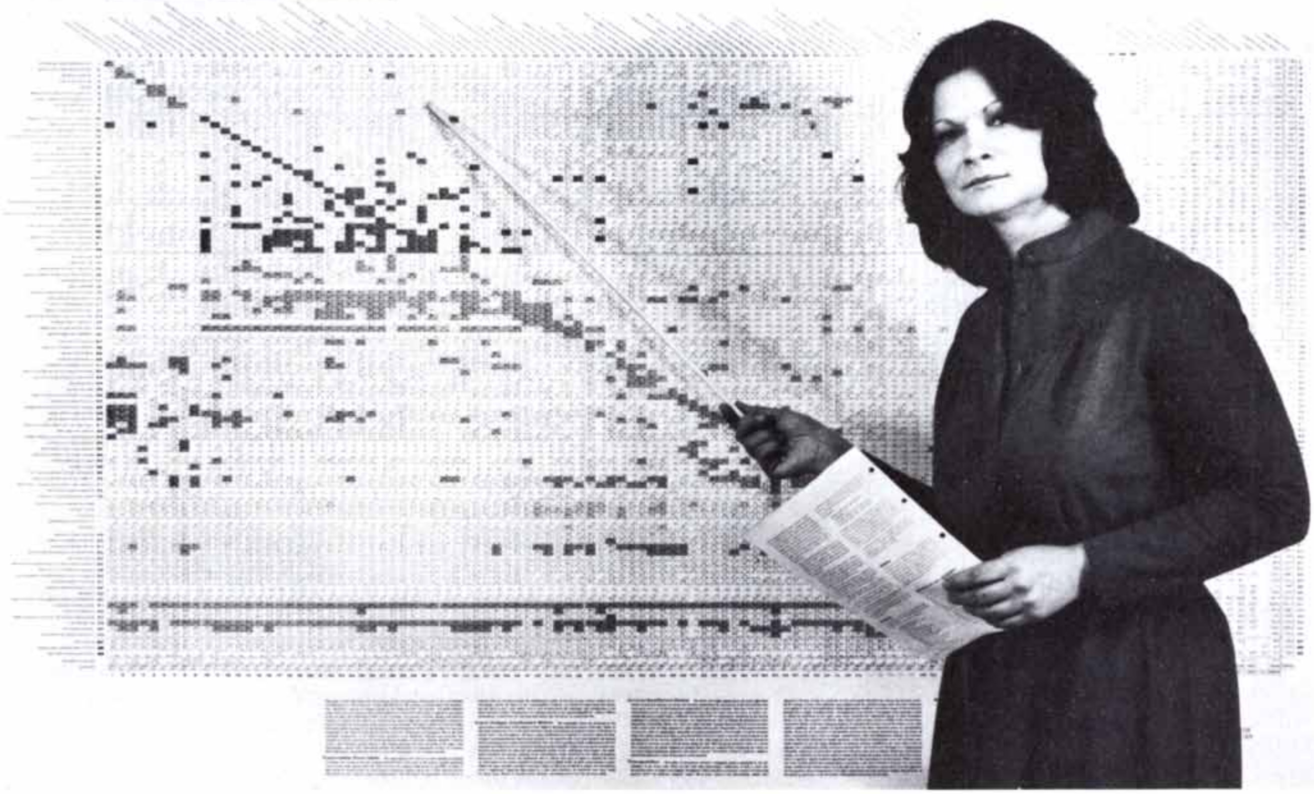
The two photographs on page 147 show plastic models with concentrations of stress in their concave regions. The isochromatics are closely spaced there, indicating that the stress differences at points in those regions vary considerably. Since the models are under compression from forces on both the left and the right, the concave regions are under compression. In the top photograph the convex region is being stretched. The more uniform distribution of color in the straight sections indicates that the distribution of stress is more uniform there than it is in the curved sections.

The polarizing filters and quarter-wave plates for photographs of this type can be obtained from the Edmund Scientific Co. (101 East Gloucester Pike, Barrington, N.J. 08007) or Jerryco, Inc. (601 Linden Place, Evanston, Ill. 60202). Readers wanting to write to Seufert may do so at 2050 West Boulevard, Cleveland, Ohio 44102.




*How to eliminate isoclinics in order to see isochromatics*

## THE INPUT/OUTPUT STRUCTURE OF THE UNITED STATES ECONOMY



# WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

**SCIENTIFIC AMERICAN**

415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $____ for ____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $____ for ____ chart(s) at $40 each.

Name __________

Company __________

Street Address __________

City, State, Zip Code __________

*Charts shipped international require one fold. 6/83

Make check payable to SCIENTIFIC AMERICAN. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.

# BIBLIOGRAPHY

*Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.*


## METAMAGICAL THEMAS

ESCAPE FROM PARADOX. Anatol Rapoport in *Scientific American,* Vol. 217, No. 1, pages 50–56; July, 1967.

NEWCOMB'S PROBLEM AND THE PRISONER'S DILEMMA. Steven J. Brams in *The Journal of Conflict Resolution,* Vol. 19, No. 4, pages 596–612; December, 1975.

THE PRISONERS OF INSECURITY. Bruce Russett. W. H. Freeman and Co., 1983.

## PHYSICAL DISABILITY AND PUBLIC POLICY

DESIGN FOR INDEPENDENT LIVING: THE ENVIRONMENT AND PHYSICALLY DISABLED PEOPLE. Raymond Lifchez and Barbara Winslow. University of California Press, 1979.

ENVIRONMENTAL ACCESSIBILITY AND INDEPENDENT LIVING OUTCOMES: DIRECTIONS FOR DISABILITY POLICY AND RESEARCH. Gerben DeJong. University Center for International Rehabilitation, Michigan State University, 1981.

DISABLED PEOPLE AS SECOND CLASS CITIZENS. Edited by Myron Eisenberg. Cynthia Griggins and Richard Duval. Springer Publishing Company, 1982.

## A VECTOR FOR INTRODUCING NEW GENES INTO PLANTS

PLASMIDS SPECIFYING PLANT HYPERPLASIAS. Eugene W. Nester and Tsune Kosuge in *Annual Review of Microbiology,* Vol. 35, pages 531–565; 1981.

MOLECULAR BIOLOGY OF PLANT TUMORS. Edited by G. Kahl and J. S. Schell. Academic Press, 1982.

T-DNA OF THE AGROBACTERIUM TI AND RI PLASMIDS. Michael Webster Bevan and Mary-Dell Chilton in *Annual Review of Genetics,* Vol. 16, pages 357–384; 1982.

PROSPECTS IN PLANT GENETIC ENGINEERING. Kenneth A. Barton and Winston J. Brill in *Science,* Vol. 219, No. 4585; pages 671–676; February 11, 1983.

## GIANT VOLCANIC CALDERAS

RESURGENT CAULDRONS. Robert L. Smith and Roy A. Bailey in *The Geological Society of America: Memoir 116,* pages 613–662; 1968.

THEORETICAL MODELING OF THE GENERATION, MOVEMENT, AND EMPLACEMENT OF PYROCLASTIC FLOWS BY COLUMN COLLAPSE. R. S. J. Sparks, L. Wilson and G. Hulme in *Journal of Geophysical Research,* Vol. 83, No. B4, pages 1727–1739; April 10, 1978.

YELLOWSTONE PARK AS A WINDOW ON THE EARTH'S INTERIOR. Robert B. Smith and Robert L. Christiansen in *Scientific American,* Vol. 242, No. 2, pages 104–117; February, 1980.

## PEKING MAN

ON A LOWER MOLAR HOMINID TOOTH FROM THE CHOUKOUTIEN DEPOSIT. Davidson Black in *Palaeontologia Sinica,* Series D, Vol. 7, Fascicule 1, pages 1–28; 1927.

ON AN ADOLESCENT SKULL OF *SINANTHROPUS PEKINENSIS* IN COMPARISON WITH AN ADULT SKULL OF THE SAME SPECIES AND WITH OTHER HOMINID SKULLS, RECENT AND FOSSIL. Davidson Black in *Palaeontologia Sinica,* Series D, Vol. 7, Fascicule 2, pages 1–144; 1927.

NOTICE OF THE DISCOVERY OF QUARTZ AND OTHER STONE ARTIFACTS IN THE LOWER PLEISTOCENE HOMINID-BEARING SEDIMENTS OF THE CHOUKOUTIEN CAVE DEPOSIT. Pei W. C. in *Bulletin of the Geological Society of China,* Vol. 11, No. 2, pages 109–139; December, 1931.

THE *SINANTHROPUS* POPULATION OF CHOU KOU TIEN. Franz Weidenreich in *Bulletin of the Geological Society of China,* Vol. 14, No. 4, pages 427–461; 1935.

THE CAVE HOME OF PEKING MAN. Chia Lan-po. Foreign Language Press, 1975.

## DARK MATTER IN SPIRAL GALAXIES

DYNAMIC EVIDENCE ON MASSIVE CORONAS OF GALAXIES. Jaan Einasto, Ants Kaasik and Enn Saar in *Nature,* Vol. 250, No. 5464, pages 309–310; July 26, 1974.

MASSES AND MASS-TO-LIGHT RATIOS OF GALAXIES. S. M. Faber and J. S. Gallagher in *Annual Review of Astronomy and Astrophysics,* Vol. 17, pages 135–187; 1979.

THE LARGE-SCALE STRUCTURE OF THE UNIVERSE. J. P. E. Peebles. Princeton University Press, 1980.

## THE PHYSIOLOGICAL ECOLOGY OF WHALES AND PORPOISES

PHYSIOLOGY OF A SMALL CETACEAN. J. Kanwisher and G. Sundnes in *Hvålradets Skrifter,* No. 48, pages 45–53; 1965.

THERMAL REGULATION IN CETACEANS. John Kanwisher and Gunnar Sundnes in *Whales, Dolphins and Porpoises,* edited by Kenneth S. Norris. University of California Press, 1966.

RESPIRATION AND DEEP DIVING IN THE BOTTLENOSE PORPOISE. Sam H. Ridgway, B. L. Scronce and John Kanwisher in *Science,* Vol. 166, No. 3913, pages 1651–1654; December 26, 1969.

DOLPHIN LUNG COLLAPSE AND INTRAMUSCULAR CIRCULATION DURING FREE DIVING: EVIDENCE FROM NITROGEN WASHOUT. Sam H. Ridgway and Red Howard in *Science,* Vol. 206, No. 4423, pages 1182–1183; December 7, 1979.

FREE AND FORCED DIVING IN BIRDS. John W. Kanwisher, Geir Gabrielsen and Nancy Kanwisher in *Science,* Vol. 211, No. 4483, pages 717–719; February 13, 1981.

## GEORG CANTOR AND THE ORIGINS OF TRANSFINITE SET THEORY

WHAT IS CANTOR'S CONTINUUM PROBLEM? Kurt Gödel in *Philosophy of Mathematics: Selected Readings,* edited by Paul Benacerraf and Hilary Putnam. Prentice-Hall, Inc., 1964.

PROBLEME DES UNENDLICHEN: WERK UND LEBEN GEORG CANTORS. Herbert Meschkowski. Vieweg, Braunschweig, 1967.

TOWARDS A BIOGRAPHY OF GEORG CANTOR. I. Grattan-Guinness in *Annals of Science,* Vol. 27, No. 4, pages 345–391; December, 1971.

GEORG CANTOR: HIS MATHEMATICS AND PHILOSOPHY OF THE INFINITE. Joseph Warren Dauben. Harvard University Press, 1979.

## THE SLIDE FASTENER

THE INVENTION AND INTRODUCTION OF THE ZIPPER. P. J. Federico in *Journal of the Patent Office Society,* Vol. 28, No. 12, pages 855–876; December, 1946.

THE SLIDE FASTENER INDUSTRY IN 1950: A REPORT FOR SLIDE FASTENER ASSOCIATION, INC. Prepared by Business Information Division, Marketing Services Company. Dun and Bradstreet, Inc., 1951.

TALON, INC.: A ROMANCE OF ACHIEVEMENT. James Gray. Meadville Printcraft Corp., 1963.

## THE AMATEUR SCIENTIST

PHOTOELASTICITY: PRINCIPLES & METHODS. H. T. Jessop and F. C. Harris. Dover Publications, Inc., 1960.

POLARIZED LIGHT. William A. Shurcliff and Stanley S. Ballard. D. Van Nostrand Company, Inc., 1964.

PHOTO-ELASTIC ANALYSIS. A. W. Hendry. Pergamon Press, 1966.


FOR YEARS, THE NEWNESS
OF A CAR WAS DETERMINED
BY WHAT YOU SAW.
Get it together—buckle up.