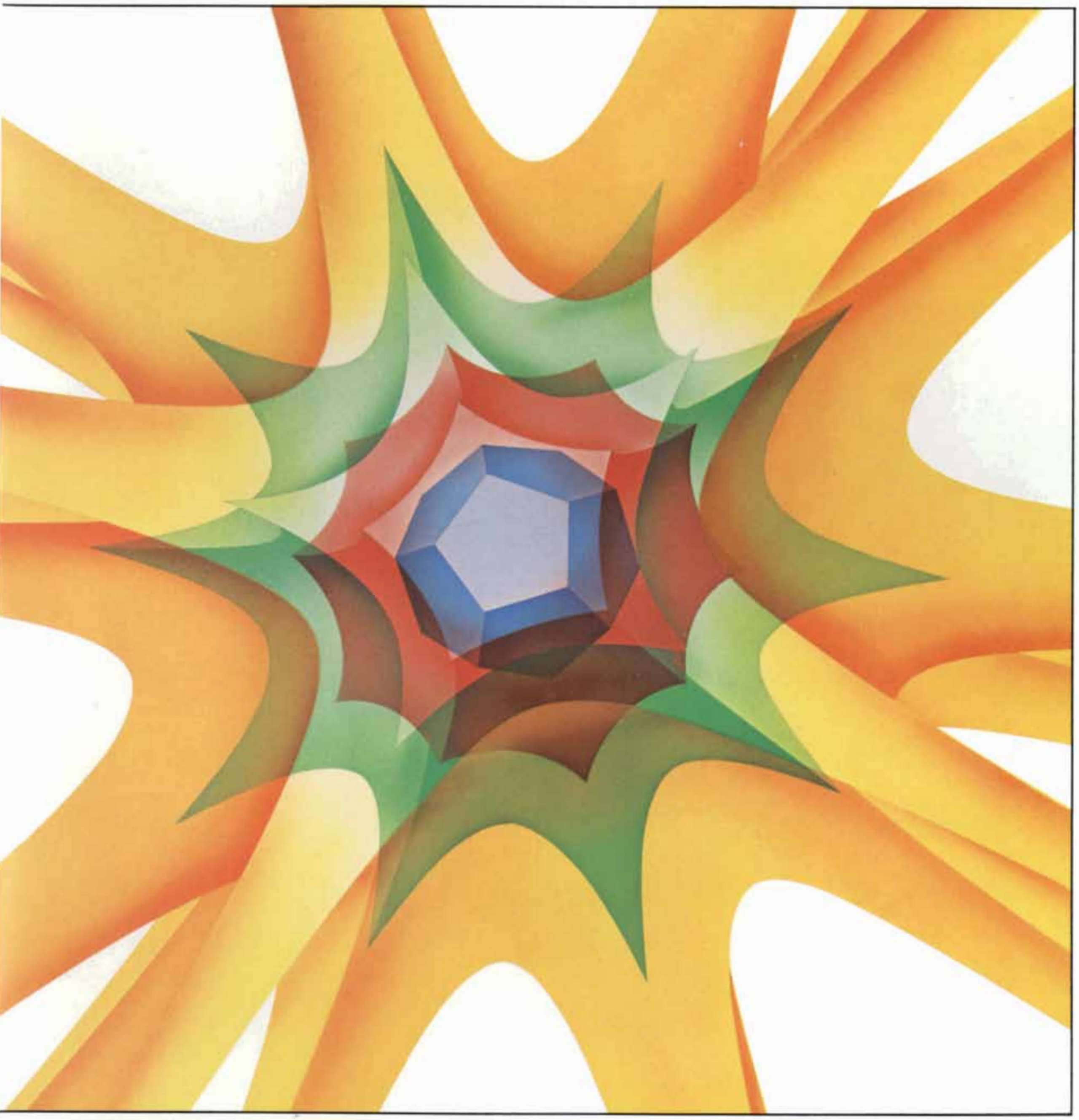
SCIENTIFIC AMERICAN



MATHEMATICAL MANIFOLDS

$2.50

July 1984


TURBO
600 ES TURBO
THERE'S MORE
TO IT THAN MEETS
THE SKY.
Dodge
AN
AMERICAN
REVOLUTION.
charges. †5/50 Protection Plan: 5 years or 50,000 miles, whichever comes first. Limited warranty on engine, powertrain, outer body rust-through. A deductible applies. Excludes
1984 American models.
BUCKLE UP FOR SAFETY

Casio's solar-powered scientific calculators put space-age technology easily within your reach.

Our FX-910 is the logical choice for students and engineers alike. At only $24.95, it gives you algebraic logic, 48 functions and an 8-digit + 2-digit exponent display—in a size that will fit as easily in your pocket as its price will suit your pocketbook.

At the same time, our credit card-size FX-90 ($29.95) has members of the scientific community flipping—over its 49-function flip-open keyboard. Made possible by Casio's innovative sheet key technology, this handy feature makes complicated scientific equations easier to solve because the major function keys are displayed oversize on their own keyboard.

Like our FX-90, our FX-450 ($34.95) has a 10 + 2-digit LCD display and a keyboard with touch-sensitive keys. But the keys are double size and the number of functions increases to 68. Most importantly, it lets you calculate with the speed of light—and eight other commonly used physical constants, including Plank's constant and atomic mass. It also makes computer math calculations and conversions in binary, octal and hex equally easy to use.

Casio has more space-age instruments at down-to-earth prices than there is space for here. However your Casio dealer will gladly let you get your hands on technology that, until now, only the future held.

CASIO
Where miracles never cease®

Casio, Inc. Consumer Products Division: 15 Gardner Road, Fairfield, N.J. 07006 New Jersey (201) 575-7400, Los Angeles (213) 803-3411.

Established 1845 **SCIENTIFIC AMERICAN** July 1984 Volume 251 Number 1

ARTICLES

DEPARTMENTS

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

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 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 $1.25 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036-8733/84. $1.25 + 0.00.

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

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

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

Name ____________________

New Address ____________________

Street ____________________

City ____________________

State and ZIP ____________________

Old Address ____________________

Street ____________________

City ____________________

State and ZIP ____________________

THE COVER

The painting on the cover illustrates a method employed to study the properties of mathematical objects called three-dimensional manifolds (see "The Mathematics of Three-dimensional Manifolds," by William P. Thurston and Jeffrey R. Weeks, page 108). A three-dimensional manifold is generated by a polyhedron whose surface points are mathematically identified with one another in some fixed way. For example, if the faces of a dodecahedron are mathematically "glued together," or identified, the resulting manifold is called the Seifert-Weber space. The study of such a manifold can be greatly simplified if it is smoothly deformed until the curvature is constant throughout the manifold. For the Seifert-Weber space the deformation can be accomplished by inflating the dodecahedron in a "curved" space called hyperbolic space; the greater the inflation, the sharper the angles of the dodecahedron. The painting shows four stages in the inflation. The hyperbolic dodecahedron in red leads to the geometrically simple version of the Seifert-Weber space. Any three-dimensional manifold could serve as a model for the topological structure of the universe.

THE ILLUSTRATIONS

Cover painting by Hank Iken, Walken Graphics

Page	Source
15–27	Ilil Arbel
40–41	Chris Morris, Black Star
42–43	Ilil Arbel
44–45	J. Berndt
46–48	Andrew Tomko
49–50	Roger Hékinian, Centre Océanologique de Bretagne
51	Ian Worpole
52–53	Roger Hékinian, Centre Océanologique de Bretagne
54	Roger C. Searle, Institute of Oceanographic Sciences, Wormley, England
55	Roger Hékinian, Centre Océanologique de Bretagne
57	Arthur E. Frankel, Cetus Corporation
58–64	Bunji Tagawa
73–81	Patricia J. Wynne
85	Lick Observatory (*top*); Goddard Space Flight Center, National Aeronautics and Space Administration (*bottom*)
86–87	George V. Kelvin
88	Goddard Space Flight Center, National Aeronautics and Space Administration (*top*); George V. Kelvin (*bottom*)
89–94	George V. Kelvin
97	Joseph D. Becker
98	Paul Prosise
99–106	Edward Bell
109–120	Ian Worpole
122	Terry S. Reynolds, Michigan Technological University
124	Alan D. Iselin
125–126	Alan D. Iselin (*top*); Terry S. Reynolds, Michigan Technological University (*bottom*)
127	Astor, Lenox and Tilden Foundations, New York Public Library
128–130	Terry S. Reynolds, Michigan Technological University
133–136	Michael Goodman

LETTERS

Sirs:

After reading "The Skill of Typing," by Timothy A. Salthouse [SCIENTIFIC AMERICAN, February], I could not resist the opportunity to toss (gently, of course) a few monkey wrenches into the various theories of typing prowess.

I think the research into single letters, letter phrases, previewing letter clusters and so forth falls somewhat short of the mark. When I type, whether I am copying something or engaging in the actual process of "writing," I do not write letters, clusters or anything of the sort. What I write is ideas. I am not playing semantics here. Bear with me. If there is a clear idea in my head, an idea that must be expressed in words because that is the way my particular skills lead me to express things, my subconscious mind keeps a catalogue of all the words that can serve to express those ideas....

HOWARD R. COHEN

Hollywood, Calif.

Sirs:

Mr. Cohen's point about writing in ideas is a good one and serves to illustrate an important distinction between transcription typing, with which I have been concerned, and composition typing, in which the typist composes the copy while sitting at the keyboard. I suspect that composition typing is much more sporadic than transcription typing but that during the fast bursts typing proceeds in nearly the same way as it does in transcription typing. The major difference is that composition proceeds in parallel with typing, and because one's attention is directed toward the expression of thoughts into words, there may be even less conscious awareness of the specific activities leading up to the execution of keystrokes.

TIMOTHY A. SALTHOUSE

Department of Psychology
University of Missouri
Columbia, Mo.

Sirs:

Edmund J. Sullivan's letter ["Letters," SCIENTIFIC AMERICAN, April] is correct in depicting quantum gravity as being flagrantly speculative. His assertion that this is "symptomatic of a major change" in the relation of theory to experiment, however, betrays a misreading of the history of physics. In many important theoretical advances the "unresolved question that was directly related to experience" could be seen as such only by those who had already accepted the new paradigm.

Consider the Copernican-Newtonian "revolution" that gave rise to modern physics. Before that time the motions of heavenly bodies were hardly an unresolved question, since they could be compounded from circular motions deemed natural to the cosmic realm in which they took place, and compounded to any desired accuracy. It was the insistence on a single physics for terrestrial and celestial motions that made them an "unresolved question." And the crucial terrestrial test of this synthesis, Henry Cavendish's "weighing the earth," was conducted more than a century *after* the theory had been formulated.

In a similar vein James Clerk Maxwell's synthesis of electricity, magnetism and light was in its time both mathematically obscure and not demanded by any unresolved experimental question.

The work of the quantum gravitationists is very much in this spirit. We have today two remarkably rich dynamic syntheses, one in the macroworld and the other in the microworld. Electromagnetism, the one field with a significant role in both domains, has been successfully incorporated into both. Yet the connection between these syntheses remains clouded by both formal and conceptual difficulties.

As an experimental physicist I share Mr. Sullivan's doubts that such questions can be resolved solely by mathematical labors, without experimental input. History shows us, however, that crucial experiments have a way of turning up when they are most needed. The measurement of the Lamb shift in the energy levels of hydrogen was of inestimable value in resolving the problems of quantum field theory, and many similar examples can be cited.

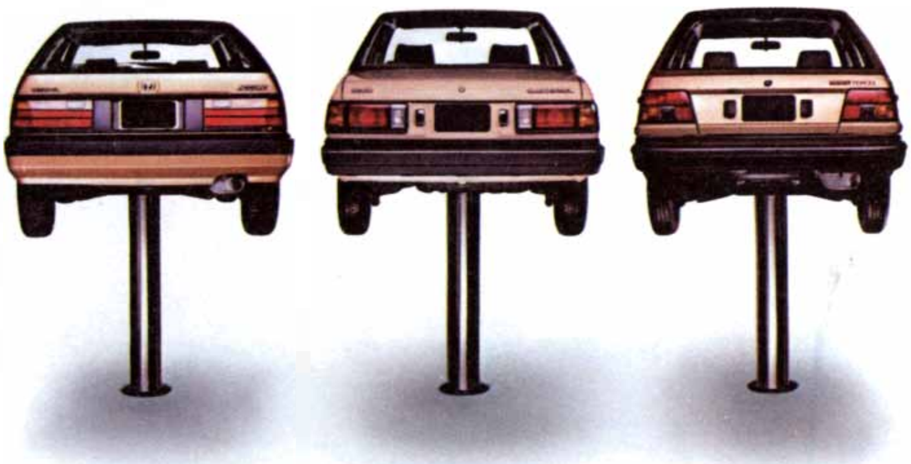
Current searches for baryon nonconservation and magnetic monopoles are motivated by an extrapolation of quantum field theory to a realm in which quantum fluctuations of gravity may have observable effects. If any of these experiments were to succeed, or if help were to come from some unexpected quarter (as it has so often in the past), the situation in quantum gravity could change overnight.

It is curious that gravity, the starting point of modern physics, remains one of its deepest mysteries. Circumstances, however, have changed. Today physics is a mature discipline, and it has a large cadre of well-trained theorists. Surely we can afford the luxury of allocating a small subgroup of these talented individuals to the study of a problem that has been central to our science.

ROBERT H. MARCH

University of Wisconsin
Madison, Wis.

Ford Escort spends less time in the air than these Japanese imports.



Ford Escort has been carefully engineered to help lower your cost of ownership by reducing required scheduled maintenance operations.

Number of Scheduled Maintenance Operations Required*	
Honda Accord	58
Nissan Sentra	54
Toyota Tercel	37
Ford Escort	20

The chart above shows just how well it stacks up against the leading Japanese imports.

Escort's efficient CVH engine will also do its part to help keep your operating costs way down. [37] EPA Est. MPG, 56 Est. HWY.**

So enjoy Ford Escort's economy and front-wheel drive. The smooth ride of its four-wheel independent suspension. The surprising room of its comfortable interior. And spend more time on the road, and less time up on the hoist.

The Best-Built American Cars.

Based on a consumer survey, Ford makes the best-built American cars. The survey measured owner-reported problems during the first three months of ownership of 1983 cars designed and built in the U.S.

Plus a Lifetime Service Guarantee.

Participating Ford Dealers stand behind their work in writing with a free Lifetime Service Guarantee. No other car companies' dealers, foreign or domestic, offer this kind of security. Nobody. See your participating Ford Dealer for details.

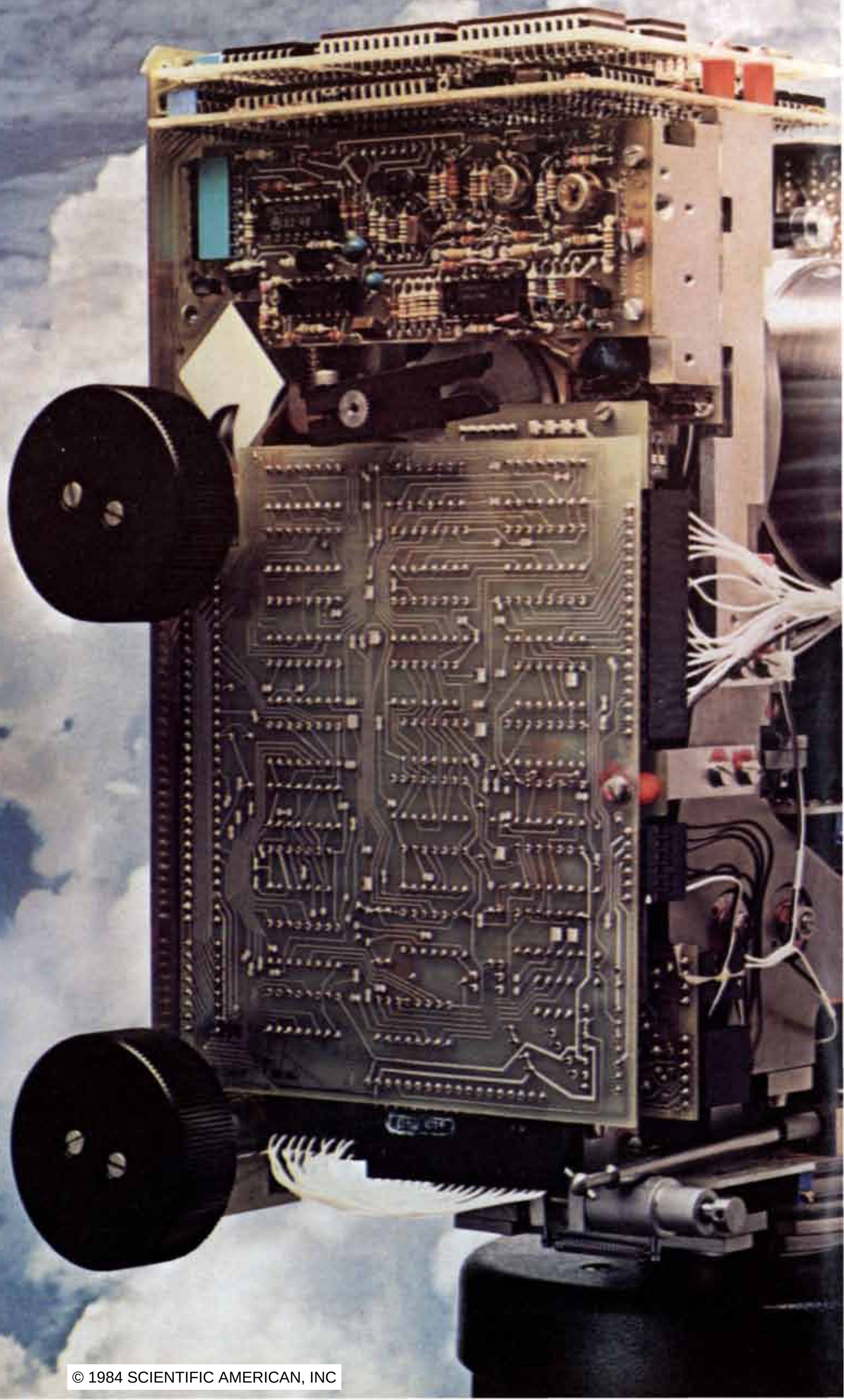
*For five years or 50,000 miles. Scheduled maintenance performed at regular intervals is, of course, essential. But don't forget that other vehicle checks (fluid levels, tire pressure, headlight alignment, etc.) also should be performed periodically by you or a qualified technician. Vehicles in severe use require additional maintenance.

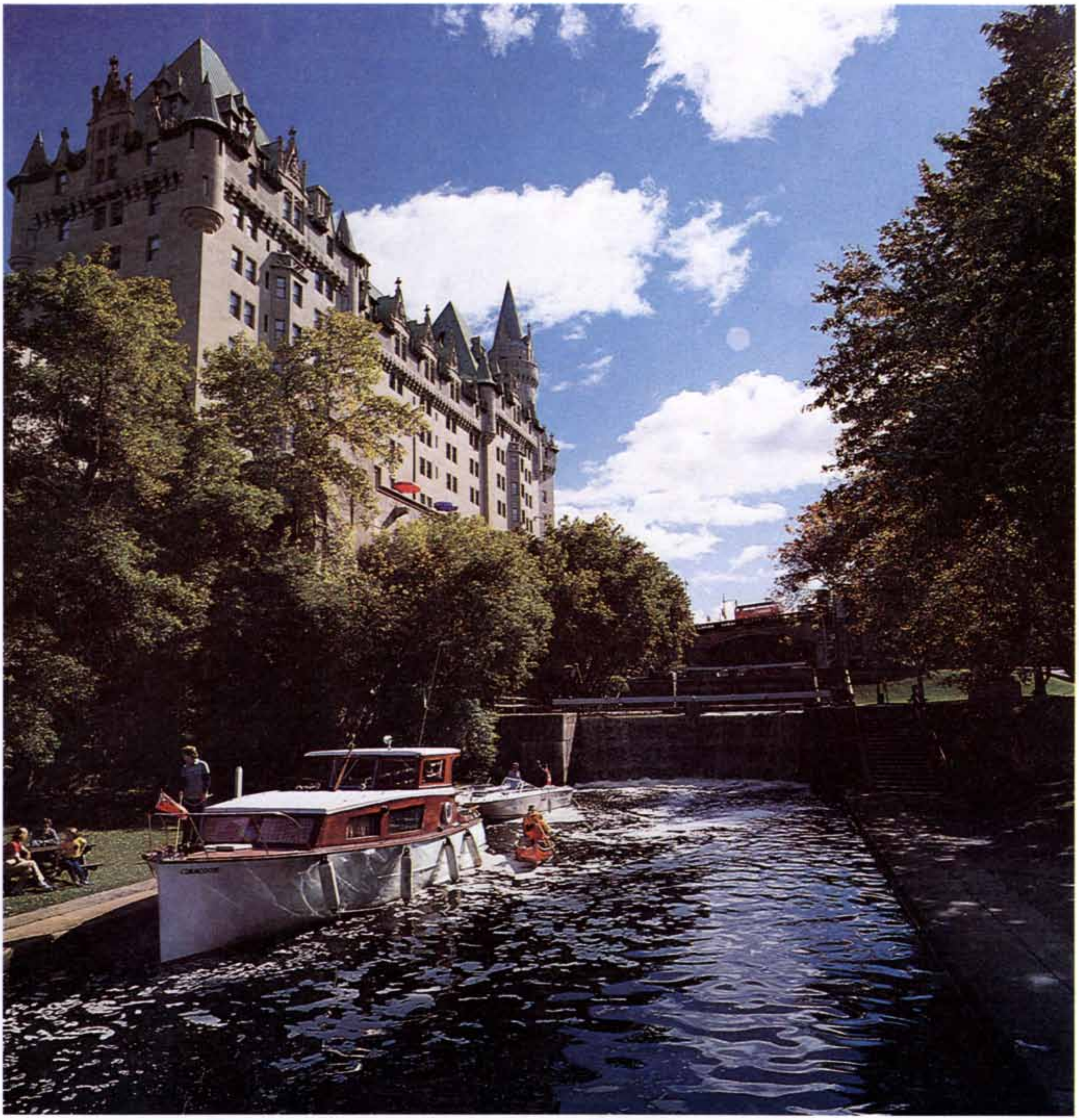
**For comparison. Your mileage may vary depending on speed, trip length, weather. Actual highway mileage lower. Escort mileage applicable to sedans with FS engine and without power steering and A/C. Not available in California.

Have you driven a Ford... lately?

Ford

Get it together —
Buckle up.





Ottawa, Ontario/Canada

Room with a view

It's called a château and looks like a castle, but it's actually a hotel and please be our guest. Canada's capital abounds with far-away feelings from double-decker bus tours and sedate canal cruises to Changing the Guard on Parliament Hill. The splendours of our heritage are yours to discover in Ottawa, Ontario/Canada. Treat yourself royally! Call TOLL FREE 1-800-828-8585 or from New York State 1-800-462-8404 or write: Ontario Travel, Queen's Park, Toronto M7A 2E5, Ontario/Canada.

ONTARIO
yours to discover!

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1934: "One of the major reasons no serious experimenter has yet attempted to build a real moon rocket is the magnitude of the job. Physicists tell us that if we can reach a speed of about 25,000 miles per hour—more exactly 6.664 miles per second—we can then shut off our power. Provided our rocket is outside the earth's atmosphere, it will thereafter coast to its objective on its own momentum. Now, building up a velocity of 6.664 miles per second, particularly in a projectile large enough to carry a crew, will require enormous quantities of energy. If the physicists produce their much-discussed atomic energy, it would help us a great deal. If, as is more likely, we shall have to depend on present materials, the fuels will probably be liquid oxygen and liquid hydrogen or acetylene. These fuels, mingled and fired in a 'blast' chamber, are theoretically capable of giving us the speed we need, provided we can devise a rocket large enough, yet light enough, to carry the cargo."

"A major volcanic eruption in Java or Alaska may fill the upper air with dust so fine that it will float 'round and 'round the world for two or three years before it settles out, and while it is aloft it helps to produce persistently chilly, often rainy weather. This and other effects of great volcanic explosions, such as that of Katmai in Alaska in 1912, which kept the Iowa corn crop from ripening one or two seasons later, were discussed before a meeting of the American Geophysical Union in Washington by William J. Humphreys of the U.S. Weather Bureau. Volcanic dust in the upper atmosphere, Dr. Humphreys said, heralds its presence by strange effects on the weather that follow disturbances in the radiations that reach and proceed from the earth. At such times the sun becomes surrounded by peculiar rings or halos, resulting from the scattering of its rays by tiny dust particles. We might even get the paradox of a frozen earth produced by too much activity by fire-mountains. It would not matter even where the volcanoes were, so long as they blew their dust high enough to set it afloat around the world."

"Of the many strange devices to be found in scientific laboratories perhaps none is stranger than a certain unimposing ring of quartz in the Bell Telephone Laboratories in New York. This crystal ring divides the second into 100,000 parts and therefore is capable of measuring periods of a few minutes with an accuracy of one part in 5,000,000! The standard frequency impulses of this crystal oscillator may be sent by wire anywhere. Thus, for instance, high-precision electric clocks could be checked from any point in the country by transmitting a standard frequency over telephone wires. Likewise, wherever a standard frequency is desired the magic ring stands ready to serve. Many great broadcasting stations now control their transmission frequency by means of it. Indeed, wherever time or frequency measurements are made or compared, this one magic ring may serve as the master for the entire country."

SCIENTIFIC AMERICAN

JULY, 1884: "Throughout the region included in the 'gas belt,' which reaches from the oil regions of Pennsylvania to Moundsville, W.Va., there is just now a good deal of speculation as to the possibilities of the large use of natural gas for fuel. Pittsburg, with its extensive industries, is advantageously situated to realize the full benefit that may be derived should the use of this gas be proved practicable, and it is already in use in some large establishments. There are some drawbacks to its employment, among which are its great unsteadiness of pressure and the ever present doubt as to how permanent may be the flow from any given well. It would seem that the first difficulty might easily be remedied by a proper system of valves and holders, and as the existence of the gas in the earth has been known for even longer than we have known of the petroleum, there is probably as good a reason for counting upon its continued flow as there is for expecting a steady supply of petroleum. The section of the country promising favorably for the boring of gas wells is a comparatively large one, and the successful employment of this natural fuel can hardly fail to have an important bearing on the future of many of our industries, particularly in all branches of the iron manufacture and its related departments."

"The great star Sirius, which was the first to respond to the spectroscopic method of measuring motion, is beginning to give information of a new and most unexpected character. In the spectrum of the star the lines of hydrogen are very strong, so that the astronomer can compare any given line of hydrogen with the corresponding line as given by the glowing gas in one of his laboratory tubes. The comparison so made by Dr. William Huggins showed that Sirius was receding from the earth at the rate of more than 20 miles per second. But of late years the evidence obtained at Greenwich has tended to show that the motion of the star is diminishing. And now it is found that the motion of recession has become so slow that we may expect it presently to change into a motion of approach, as though Sirius were traveling in a mighty orbit with movements alternately carrying him toward and away from our sun. Whatever may be the actual movements of Sirius, it is clear that the new method of measuring motion is capable of giving us such information about these movements as cannot but help us notably in the determination of their true character."

"We scarcely know what services microbes may render us, yet the study of them, which has but recently been begun, has already shown the importance of these organisms in nature. *Mycoderma aceti* converts the alcohol of fermented beverages into vinegar. *Micrococcus ureae* converts the urea of urine into carbonate of ammonia. *Micrococcus nitrificans* converts nitrogenized matters into nitrates. We must also refer to the numerous bacteria that occasion putrefaction in vegetable or animal organisms. These microbes, which float in the air, fall upon dead animals or plants, develop thereon and convert into mineral matters the immediate principles of which the tissues are composed, and thus continually restore to the air and soil the elements necessary for the formation of new organic substances."

"We know of a vigorous foreign policy to which there is no possible objection. It is a policy of peace that misses no opening for an increase of trade between the United States and other countries. It affords scope for the largest statesmanship and for the freest employment of all the arts—save that of war. We have seen it manifested within a year in the building of a railroad between the United States and the heart of Mexico. This one American enterprise, popular in its inception and completion, has done more to promote good will and quicken trade between the two countries than all the legislation of Congress since the Mexican War. The tariff barriers that divide us from Mexico cannot be leveled except with the consent of our Government. Here now is an auspicious occasion for bringing into play a vigorous foreign policy that can hurt nobody, that will cost this country nothing and that will bind Mexico to our interests as tightly as if she were annexed as the result of an expensive war with her. When this is accomplished, it will only remain to apply a similar policy of reciprocal trade to all the states in Central and South America."

THE AUTHORS

ELLEN L. BASSUK ("The Homelessness Problem") is associate professor of psychiatry at the Harvard Medical School. She received a B.A. from Brandeis University in 1964 and went on to earn her M.D. from the Tufts University School of Medicine in 1968. After serving for five years as an intern and a resident in psychiatry at several hospitals in Boston she became a member of the staff of Beth Israel Hospital, where for eight years she was director of the psychiatric emergency service. In June she finished a two-year fellowship at the Mary Ingraham Bunting Institute of Radcliffe College. She is a member of the Mental Health Policy Working Group of the Division of Health Policy, Research and Education at the John F. Kennedy School of Government at Harvard University.

ROGER HÉKINIAN ("Undersea Volcanoes") is a marine geologist working in the department of geology, geophysics and geochemistry at the Centre Océanologique de Bretagne (CNEXO) in Brest. He attended the University of Pisa as an undergraduate before coming to the U.S. to continue his education. He studied oceanography at the Lamont-Doherty Geological Observatory of Columbia University; Columbia awarded him his M.S. in 1966. He earned his doctorate from the State University of New York at Binghamton. Hékinian has participated in projects staffed by French and U.S. workers that explored the midocean ridges of the Atlantic and the Pacific in manned submersibles.

R. JOHN COLLIER and DONALD A. KAPLAN ("Immunotoxins") are respectively a microbiologist and a molecular geneticist who did much of the work described in their article when both were at the University of California at Los Angeles. Collier received a B.A. from Rice University and his Ph.D. in biology from Harvard University. He writes: "I have been in the department of microbiology at U.C.L.A. since 1966. My career has centered on the study of the structures and modes of action of the bacterial toxins, in particular diphtheria toxin, *Pseudomonas* toxin and cholera toxin. After I had done some of the early work on immunotoxins Donald Kaplan, whom I had known for many years at U.C.L.A., joined my group and spearheaded a project to clone and sequence the gene for diphtheria toxin. Toward the end of the project Kaplan moved to the Cetus Corporation. My group has continued to pursue the detailed structure-activity relations of toxic proteins." Kaplan received a B.A. and his Ph.D. in biology from U.C.L.A. After serving as research molecular geneticist in the department of microbiology at U.C.L.A. he became senior scientist and project manager of immunotoxins at Cetus in 1981. He left Cetus earlier this year to take up a job as senior associate scientist at the Dow Chemical Company.

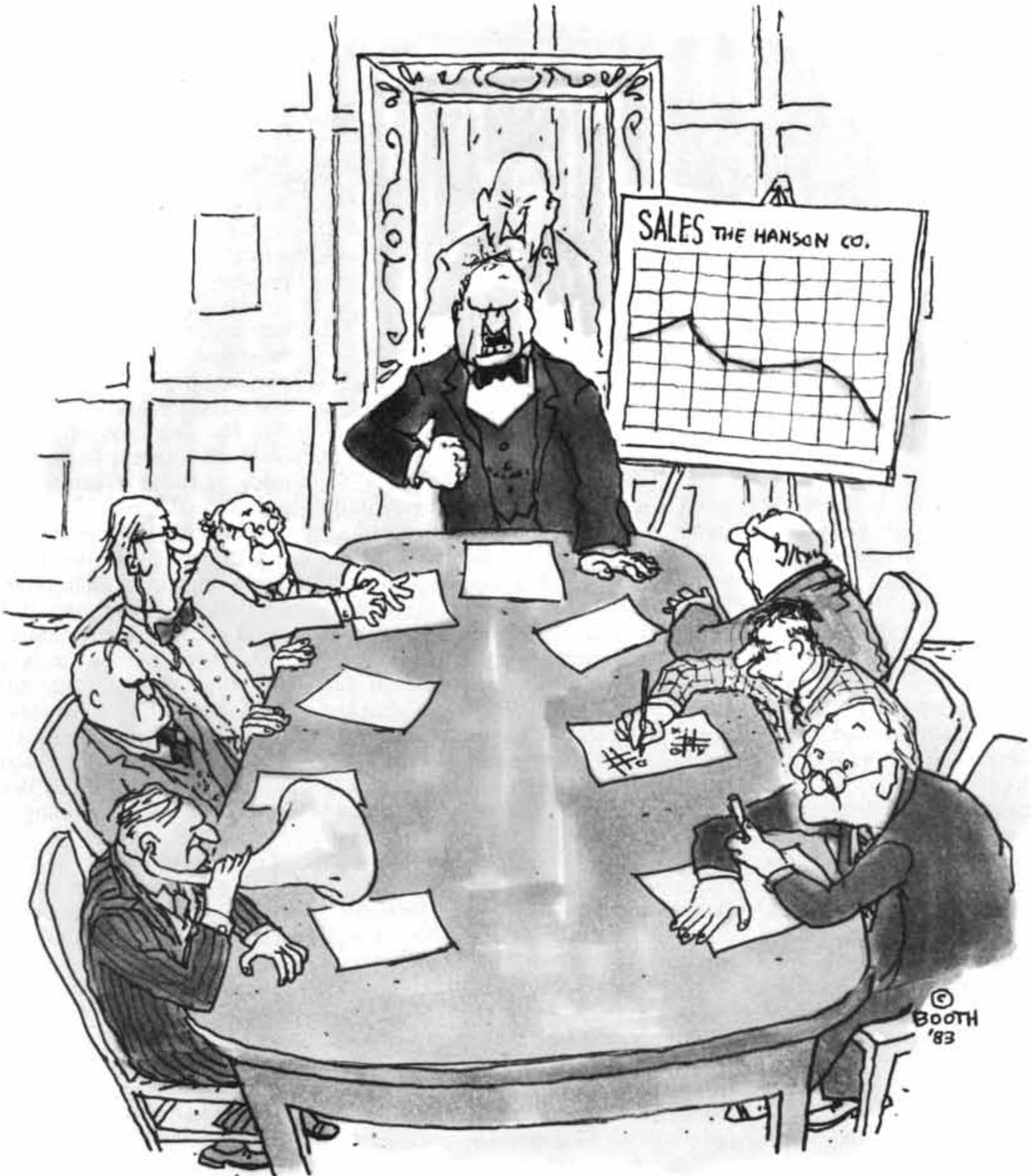
PAUL W. WEBB ("Form and Function in Fish Swimming") is professor of natural resources at the University of Michigan in Ann Arbor. A native of the United Kingdom, he received his education at the University of Bristol, which granted him a B.S. in 1967 and his Ph.D. in zoology in 1971. After earning his doctorate he moved to Canada with a fellowship from the Canadian National Research Council that enabled him to work at the Pacific Biological Station of the Fisheries Research Board at Nanaimo, B.C. In 1972 he joined the faculty of the School of Natural Resources at Michigan, where he has remained with the exception of a year spent as senior research associate at the Southwest Fisheries Center of the U.S. Marine Fisheries Service. Webb is director of research for Keron Productions, a company that makes educational computer software.

MINAS KAFATOS and ANDREW G. MICHALITSIANOS ("Symbiotic Stars") are astrophysicists who share an interest in the subject of their article. Kafatos is associate professor of physics at George Mason University in Virginia. His B.A. (1967) is from Cornell University and his Ph.D. in physics (1972) from the Massachusetts Institute of Technology. After getting his doctorate he worked at the Goddard Space Flight Center of the National Aeronautics and Space Administration for several years before moving to George Mason. Michalitsianos is head of the observatories section of the Laboratory for Astronomy and Solar Physics at the Goddard Space Flight Center. He was graduated from the University of Arizona at Tucson with a bachelor's degree and went on to earn his Ph.D. in astrophysics from the University of Cambridge in 1973. After a stint as research fellow in astrophysics at the California Institute of Technology he moved to the Goddard Center. His work on stars with peculiar emission lines and on late-type binary stars is mainly done with the instruments for ultraviolet spectroscopy mounted on board the International Ultraviolet Explorer satellite.

JOSEPH D. BECKER ("Multilingual Word Processing") writes: "I am currently manager of international advanced development at the Xerox Office Systems Division in Palo Alto, Calif. I got a B.S. from the Massachusetts Institute of Technology in 1966 and a Ph.D. in computer science from Stanford University in 1970. I worked at Bolt, Beranek & Newman, Inc., on semantic networks and intelligent robots until 1975 and then went traveling around the world, stopping in Taiwan to study Chinese for several months. In 1976 I joined the Xerox Palo Alto Research Center to help create the Fuji Xerox '8012-J Star' Japanese-language professional work station. I spent two years living in Tokyo testing the cultural acceptability of J Star there. After returning to the U.S. I helped to bring Arabic, Chinese, Korean, Russian and other languages to the Star system. After having devoted nearly 10 frustrating years to the pursuit of machine translation and artificial intelligence I was pleasantly surprised to find challenging practical problems in multilingual word processing."

WILLIAM P. THURSTON and JEFFREY R. WEEKS ("The Mathematics of Three-dimensional Manifolds") are mathematicians who collaborated in work on the subject of their article at Princeton University under Thurston's guidance. Thurston is professor of mathematics at Princeton. He was graduated from New College at Sarasota, Fla., in 1967 as a member of that college's first graduating class. After obtaining his Ph.D. from the University of California at Berkeley he spent a year at the Institute for Advanced Study at Princeton and a year at the Massachusetts Institute of Technology before taking up his current job. For several years he worked on the theory of foliations; a foliation is an onionlike mathematical pattern on a manifold. He also worked on the homeomorphisms of surfaces before beginning the study of three-dimensional manifolds. Weeks is instructor of mathematics at Stockton State College in New Jersey. He writes: "My degrees, an A.B. from Dartmouth (1978) and an M.A. from Princeton (1980), are in mathematics. I began work on a Ph.D. under Bill Thurston but was soon distracted by the pleasures of teaching. I have spent the past few years developing ways to teach nonspecialists about mathematical manifolds. I hope to finish my Ph.D. next year and then return to teaching."

TERRY S. REYNOLDS ("Medieval Roots of the Industrial Revolution") is associate professor of science, technology and society and director of the program in that subject at Michigan Technological University. His Ph.D. in the history of science was awarded by the University of Kansas in 1973. He taught at the University of Wisconsin at Madison for 10 years before taking up his current job.



"What do they mean our training program is outdated? We all went through it!"

Back in Calvin Coolidge's day, American business moved along at a nice comfortable pace. When things started picking up though, the employee training program seemed to fall behind.

But now, thanks to a company called Wicat, the power of computers can provide intense, individualized training for everyone. From clerks to line workers to managers.

CUT TRAINING TIME AND COST IN HALF.

Some of the world's largest corporations put a Wicat learning station in front of each trainee. And train them better, in less than half the time. For less than half the cost.

We've also eliminated weeks of programming. With our WISE authoring system, a non-programmer can computerize a course in days.

THE TOTAL TRAINING SOLUTION.

We can assure you a successful implementation by providing a complete system, including the hardware, authoring system, program management software, videodisc option and customer support. We can even develop custom courseware for any application.

Call us at (801) 224-6400, Ext. 209 for details. Or write to Wicat Systems, P.O. Box 539, 1875 South State Street, Orem, Utah 84057.

We'll show you how up-to-date a training system can be. When it's designed by a company that really means business.

WICATsystems™

Lessons created using WISE are compatible with the IBM PC.

COMPUTER RECREATIONS

A program that plays checkers can often stay one jump ahead

by A. K. Dewdney

"Is it possible to write a checkers-playing program that will never lose to a human player—ever?" The boy was no more than 10 years old and he looked up from the terminal with that expression of innocent wisdom only the very young can have. He was playing checkers against a program written by some students of mine. We had set the program's look-ahead factor to a moderate level, and the lad was doing rather well.

The boy's question, asked at a departmental open house several years ago, has different answers depending on which game of checkers one is talking about. The question also touches on themes lying at the heart of artificial intelligence, the study of intelligent behavior in computer programs. I shall return to these themes below.

At the time the boy asked the question he was playing 6 × 6 checkers, one of an infinite number of checkers games. Here I shall examine 4 × 4, 6 × 6 and 8 × 8 checkers (the standard game) and compare the prospects for programs designed to play them. In general one can define *n*th-order checkers on a 2*n* × 2*n* checkerboard. In each case the men, or checkers pieces, occupy all the dark squares except those in the middle two rows of the board. The rules of *n*th-order checkers are discussed in more detail in the box on page 22.

For 4 × 4 checkers the answer to the boy's question appears to be "Yes." For 6 × 6 checkers the answer is "Very likely" and for 8 × 8 checkers "Probably." These answers are based on a combination of personal experience, reading and pure conjecture.

Consider the 4 × 4 game. Each side has two men and a game rarely goes beyond 10 moves, at least when experts are playing. Novices may go 20 moves or more before recognizing a drawn position. Speaking of drawn games, a recent 4 × 4 checkers tournament in Damenspielsburg produced six draws in a row, the seventh game going on a technicality to grand master Samuel Jensen Truscott: his opponent accidentally touched an unplayable piece twice.

In any event, 4 × 4 checkers can be analyzed by hand. It appears to be a draw, and a reasonably alert player never has to lose. The same is true of a reasonably good program.

In a game of 6 × 6 checkers each side has six men and a typical game runs for about 20 moves. When our student program was set to maximum look-ahead, it took from five to 10 minutes to choose a move. Up until the end game it played seemingly invincible checkers at this setting. Usually it was so far ahead in material or position that it would go on to win or at least draw. Occasionally, however, one of us would survive with enough pieces to beat it in the end game, where its play was surprisingly weak.

The game of 8 × 8 checkers has 12 men on a side, and games between experts can last for 30 moves or more. In tournament play in standard checkers it is now the usual practice to draw opening moves by lot. The practice was adopted around the turn of the century in England because of a tendency of master players to stick with favorite openings and familiar lines of play, which led to an increasing proportion of drawn games. Whether draws are common because checkers is inherently a drawn game or because of conservatism on the part of the experts is hard to say. Nevertheless, 8 × 8 checkers remains a fascinating and difficult game. Mathematically it is somewhat less complex than chess, but there are players expert in both games who do not rate chess a "superior" game.

The two best-known checkers-playing programs were developed by Arthur L. Samuel of the International Business Machines Corporation in the 1960's and by Eric C. Jensen and Tom R. Truscott of Duke University in 1977. In 1962 Samuel's checkers-playing program defeated Robert Nealey, a former Connecticut state champion. Although the program plays a very strong game, it is apparently not rated an expert. The Jensen-Truscott program is stronger still and may rate among the top players in the world.

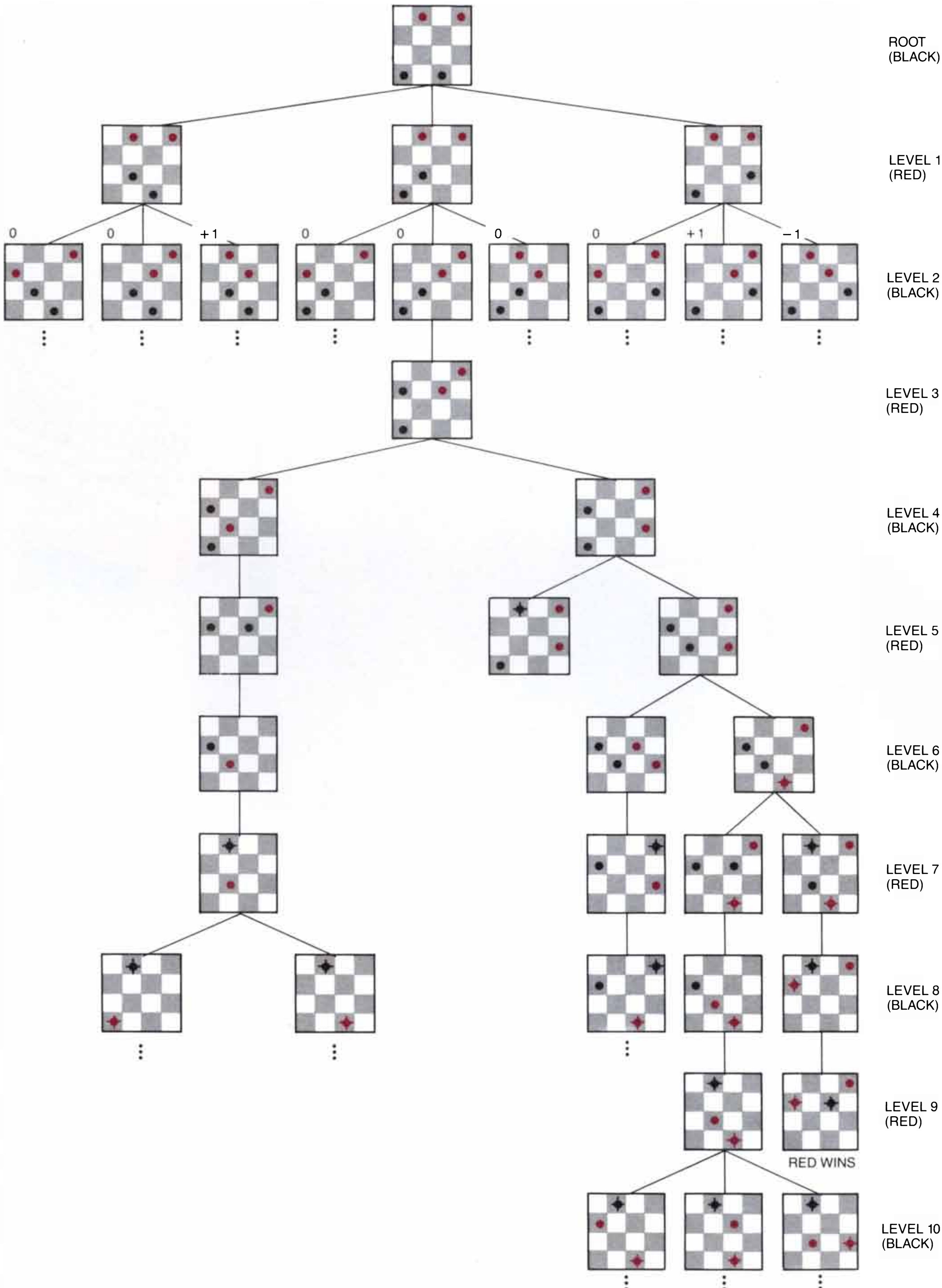
In 1979 the Jensen-Truscott program played Elbert Lowder, who is generally acknowledged to be one of the best checkers players in the U.S. In five games the program drew twice, won once and lost twice. The losses seemed to be due to deficiencies in the end game: when Lowder realized that two of the program's men were kept on the back row as guards, he exploited this weakness and won the last two games. It must be stated that Lowder's attitude while playing the Jensen-Truscott program was experimental and somewhat sportive. His one loss apparently resulted from an attempt to lead the program into unorthodox positions—something one should probably not try with a checkers program of this caliber.

Meanwhile Marion F. Tinsley, the world checkers champion, looks on, amused, from his home in Tallahassee, Fla. I do not know whether Tinsley has ever played one of the better checkers programs. On the other hand, the two leading programs have played each other. In two games against the Samuel program the Jensen-Truscott program proved to be "dramatically superior," in Truscott's words, winning both games.

Just about every successful game-playing program has three major parts: a move generator, a board evaluator and a minimax procedure. Another element of central importance is the game tree: it has a role in all three program segments. The move generator develops the tree; the board evaluator is consulted at the end of each branch of the tree, and the minimax procedure is applied to the tree as a whole.

A game tree has the initial board as its root, and every branch represents a board position that can be reached by making a single move. Oddly, the tree is usually drawn upside down, with the root at the top. It can be arranged in levels, where the boards at the *n*th level represent all the possible moves of one player on the *n*th turn. Although game trees are generally not very tall (or deep), they can be excessively bushy; as a rule the number of boards at each level is more than double the number at the previous level. Even if each board had only two successors, the number of branches would soon grow to be unmanageably large. A tree that forked in this way 64 times would have more branches than all the trees on the earth. (In nature trees do not fork much more than a dozen times from trunk to leaves.) The exponential explosion in the number of game-tree branches makes it impossible for a computer to examine more than a small fraction of the continuations of a game such as 8 × 8 checkers.

The checkers tree is created by the move generator, a program that is given



The game tree of 4 × 4 checkers, complete to the second move and with selected continuations

The new IBM DisplayWrite

Series is here. Spread the word.

In 1980, IBM introduced the Displaywriter System. Today, it's become the best-selling stand-alone text processor in the world. One reason for this success is the Displaywriter's function-rich software.

If you're looking for software like that, but working on an IBM personal computer, you don't have to look any further. Because the IBM DisplayWrite Series is here.

And it will put many of the features and capabilities of a dedicated word processor to work for you when you're writing.

It runs in the family.

You'll find two word processing programs in this series: There's DisplayWrite 1, for IBM personal computers—including PC*jr.* And DisplayWrite 2, with added functions for your PC, PC/XT or *Portable* PC.

You'll also find DisplayWrite Legal, a dictionary of about 16,000 words that a lawyer might need to check.

And you'll find DisplayComm, which lets your IBM PC send and receive text to and from other IBM PCs. If you're writing at the office, this program could also let you send text to an IBM Displaywriter down the hall. (From there, it could be sent on to an IBM host computer for distribution.)

Some words on high function.

The DisplayWrite word processing programs give you the time-saving features you'd expect from IBM. Justified margins, centered lines and pagination, for example. You'll even have prompts and messages to help guide you along.

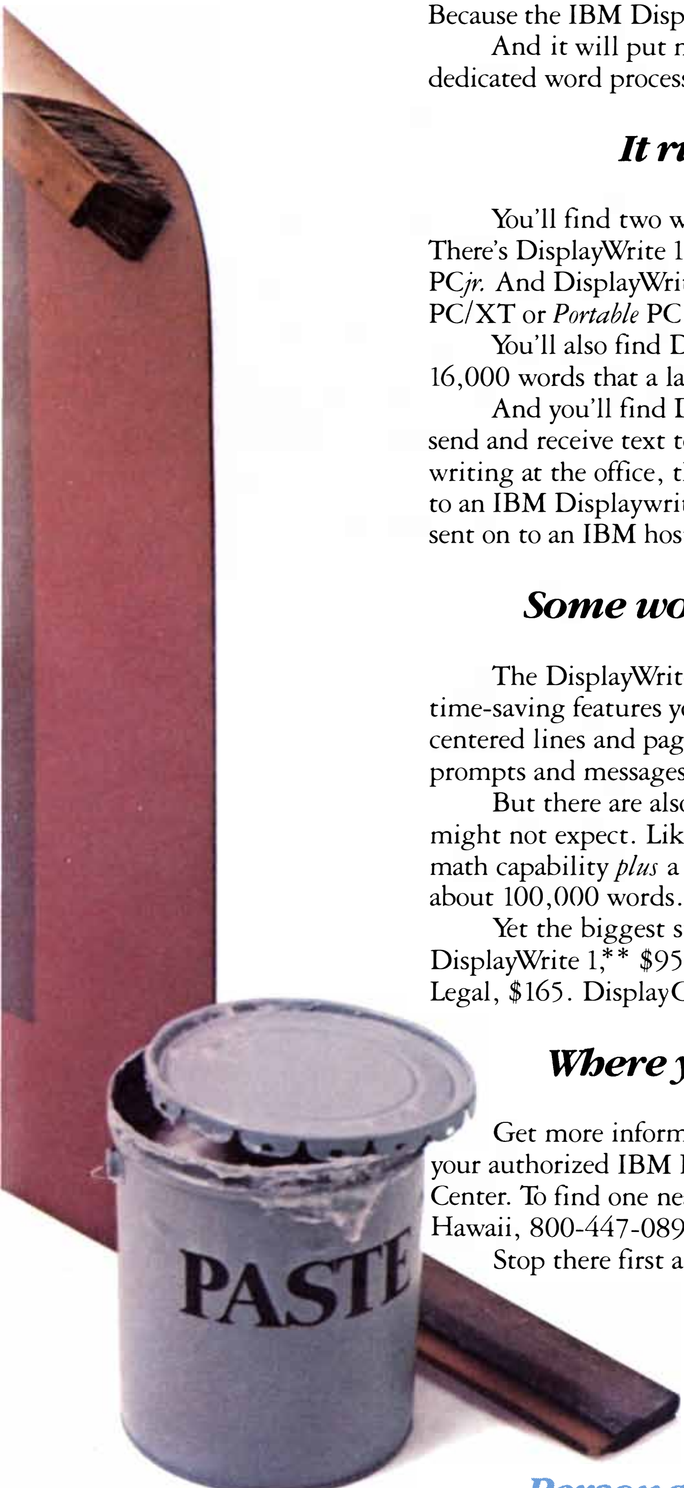
But there are also some features in DisplayWrite 2 you might not expect. Like easy column formatting, four-function math capability *plus* a spelling checker based on a dictionary of about 100,000 words.

Yet the biggest surprise of all may be the price.* DisplayWrite 1,** $95. DisplayWrite 2, $299. DisplayWrite Legal, $165. DisplayComm, $375.

Where you can find all four.

Get more information about the IBM DisplayWrite Series at your authorized IBM Personal Computer dealer or IBM Product Center. To find one near you, call 800-447-4700. In Alaska or Hawaii, 800-447-0890.

Stop there first and get the last word.

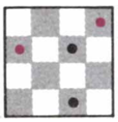


IBM®

Personal Computer Software

a particular board as input, along with an indication of whose turn it is to move. The output of the move generator is a list of all the boards that can be reached from the given position by a legal move. Consider the following 4 × 4 board, assuming it is Black's turn:

Black has two possible moves, and so the move generator lists the following two boards:



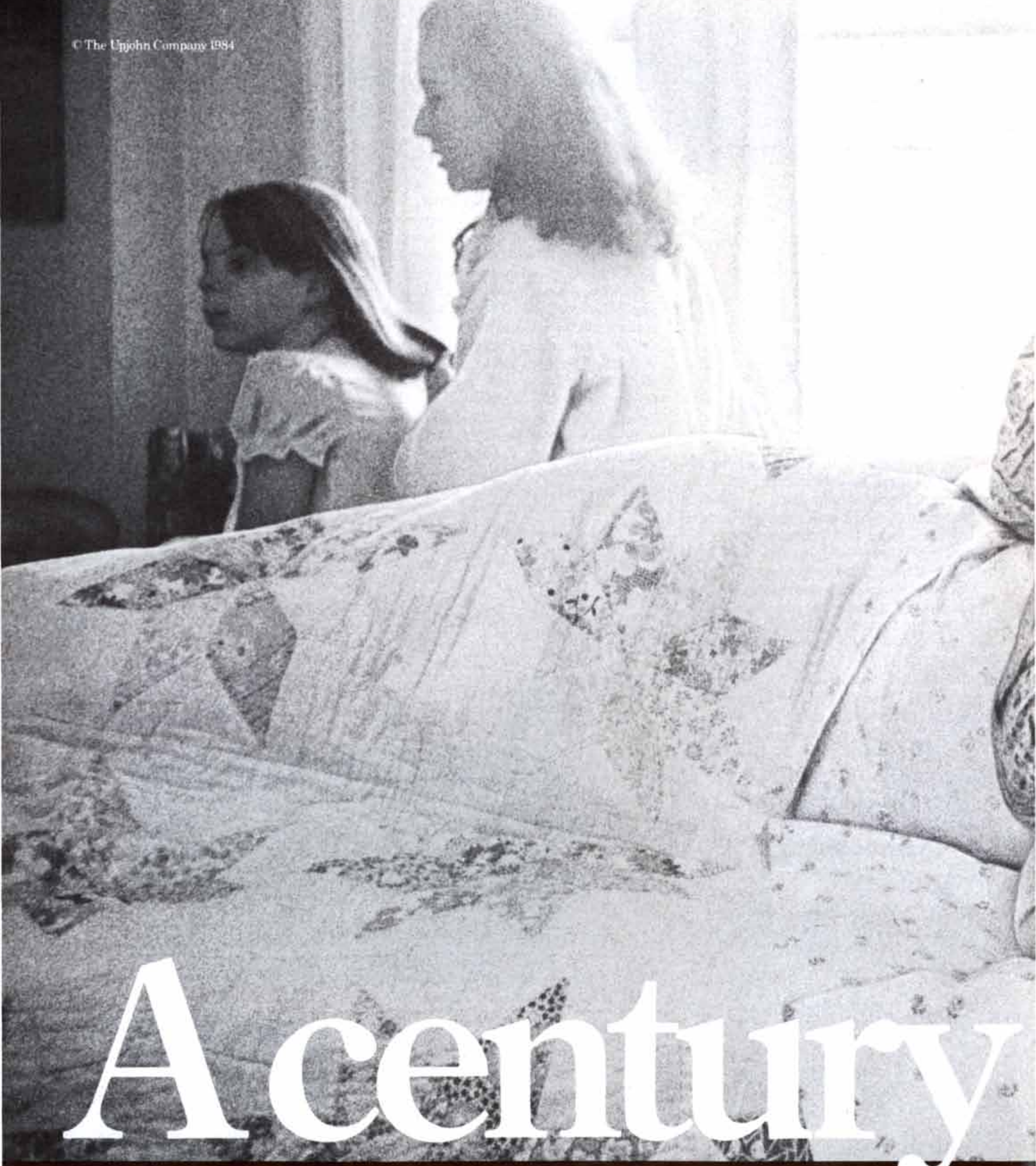

The move generator could then be applied to these boards in turn.

Readers who would like to explore the action of a move generator are invited to develop the 4 × 4 checkers tree from this level downward. At the next level, for example, there are only two boards because in each case the rules force Red to capture a piece. At the level below that there are still two boards, then three (with a Red win), then two, then five, then 10, then 24. Those who want to go still further are advised to follow two principles. First, scan each level for duplicate boards and develop the tree at only one of them. Second, scan upward from each board to see if the position has occurred before. If it has, terminate that branch and label it "Draw."

How does a move-generating program work? The first issue to be addressed is how the board is to be represented. The simplest scheme is a matrix, or two-dimensional array, whose entries identify the occupying pieces. For example, an empty square might be represented by a 0, a square with a man by a 1 and a square with a king by a 2; the sign of the entry would denote the side, say plus for Black and minus for Red. The first 4 × 4 board shown above would then be encoded in the matrix:

$$\begin{matrix} 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \end{matrix}$$

The legal moves for each piece depend on its color and position and on whether the diagonally adjacent squares are occupied. In this case the move generator scans the array row by row until it comes to the positive (Black) entry at row 3 and column 2, a position designated (3,2). Since Black is to move and the piece is a man (magnitude 1) rather than a king, the move generator examines squares (2,1) and (2,3). Square (2,1) is already occupied by a piece (and no

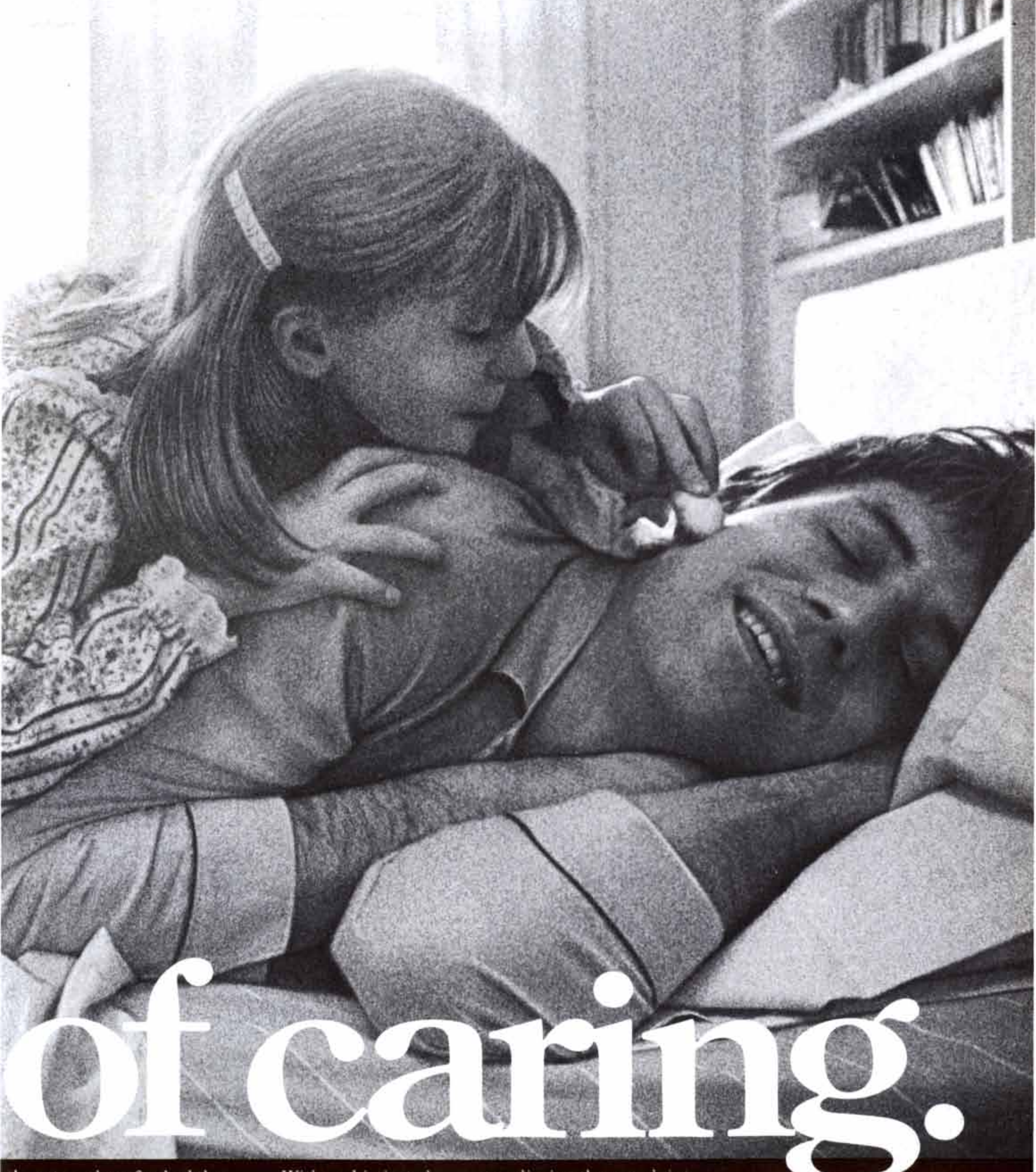


To help people with insomnia wake up to a bright new day, Upjohn searches deep within the mysteries of sleep.

For many, the worst nightmare of a troubled sleep begins the morning after. The ensuing fatigue, irritability, and other complaints can affect productivity and strain relationships at home and at work.

Persistent insomnia is a complex, 24-hour problem of great concern to the people at Upjohn.

For temporary changes in sleep-wake rhythm, medication may be a useful tool. Upjohn researchers are now concentrating on ways to help patients get to sleep faster and allow



them to wake refreshed the next day.

The key to the future lies in finding out more about how we sleep. So Upjohn is at the leading edge of this effort, conducting its own research and working with sleep laboratories around the country.

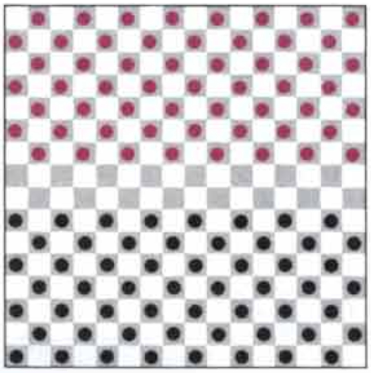
With sophisticated instruments, clarification of physiology during sleep can lead to the most informative diagnosis of a sleep disorder and point the way to other treatments, such as behavior modification, which go beyond medication and bring us even closer to relieving the complaint altogether.

For nearly 100 years, Upjohn has been helping people live a healthier, more productive life. And challenges such as this promise to make our second century of caring even more exciting than the first.

For booklet "What You Should Know About Insomnia," write INSOMNIA, Dept. IS, P.O. Box 2497, Kalamazoo, Michigan 49003.

Upjohn

The Rules of *n*th-Order Checkers



The squares of a $2n \times 2n$ board are given alternating light and dark colors and each player puts $n^2 - n$ men on the dark squares on opposite sides of the board. The middle two rows are empty of men. Each player has a "single corner" (one dark square occupying the corner) at the left; the other corner is called the double corner. Black moves first.

A man may move diagonally forward to an adjacent unoccupied square. If that square is occupied by an enemy piece but the square beyond it is not occupied, the man may jump the piece and remove it from the board. All such opportunities for a capture must be taken; if two or more captures are possible, at least one must be made. If, after making a capture, a man has an opportunity to make another, it must do so.

When a man advances to the last row on the opposite side of the board, it is made into a king. A king can move both forward and backward (still diagonally) and can capture in both directions.

The game ends in victory for one player when the other player cannot make a legal move because all his pieces have been captured or immobilized. If both players agree that a win is impossible, a draw is declared.

There are further rules about fiddling with the pieces, smoking and swearing, but a computer is unlikely to transgress them and so they are omitted here.

The board above shows the starting position of eighth-order checkers, played on a 16×16 board with 56 pieces on each side.

jump is possible), but (2,3) is free. Consequently the program writes a new array with the +1 deleted from square (3,2) and entered at (2,3):

$$\begin{array}{cccc} 0 & 0 & 0 & -1 \\ -1 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & 0 \end{array}$$

The move generator then continues its scan of the array until the next positive entry is found.

In order to keep track of the boards and the corresponding arrays, they can be numbered consecutively as they are generated. Any square on any board can then be specified by giving the board number and the row and column. As the move generator constructs the checkers tree, it must keep a record of the connections between boards. When the program is choosing a move, it searches the game tree down several levels, then up, then down again and so on. To traverse the tree in this way the program must be able to identify both the "parent" of a board and its "children." To this end the program maintains an array called a pointer table. The pointer for each board is a list of numbers, the first number pointing to the board's parent and the subsequent numbers pointing to its children.

The next major component of the checkers program, the board evaluator, is given a board as input and computes a number reflecting the value of the board for Black. A high number means Black is likely to win if the position is reached in actual play. A good board evaluator is difficult to design because there are many factors to consider in judging a position. In most games, including checkers, it is not always clear what weight should be given to the various factors. Here is a short list of relatively primitive factors.

MEN: add 1 for each Black man on the board; subtract 1 for each Red man.

KINGS: add 1 for each Black king on the board; subtract 1 for each Red king.

CENTER: add 1 for each Black man occupying one of the four center squares; subtract 1 for each Red man there.

ADVANCE: add 1 for each Black man in the second row, 2 for each Black man in the third row, 3 for each Black man (king) in the fourth row. Subtract the corresponding numbers for Red.

MOBILITY: add 1 for each move available to Black and subtract 1 for each move available to Red.

What is the relative importance of these factors? Should KINGS get double the weight of MEN? Does ADVANCE already include the information available in KINGS and MEN?

Samuel answered such questions by allowing his program to determine how much weight to give each factor based on its playing experience. Relying on 39 factors in all (many suggested by checkers experts and books), Samuel's board evaluator initially gave them all the same weight and simply added the numbers to obtain the value of a board. Each lost game was analyzed by the program, and the factor that led to the losing move was isolated. The weight of that factor was then reduced. The process was speeded by automating the analysis and having the program play itself a number of times.

Board-evaluation procedures are never perfect. If they were, the other components of a game-playing program would hardly be necessary: the program would only have to examine the boards one move ahead, evaluate each board and then choose the best move. There is a trade-off of sorts between the accuracy of the board evaluator and the number of levels the program must look ahead. If the evaluator could do little more than recognize a won or lost position, the checkers tree would have to be examined all the way to the bottom!

The third major component is the minimax procedure, essentially a tree-searching program that uses the board values at the lowest level searched to assign values to boards higher up in the tree. At some point in the search all the boards at a certain level must be rated by the board evaluator. A board *B* on the level above that one is given a value according to the following simple rule: If *B* has Red to move, select the minimum value found among *B*'s children as the value of *B*; if *B* has Black to move, select the maximum value.

The reasoning supporting this rule should be clear. If it is Red's turn to move, Black may as well assume that Red will choose a move minimizing the value of the resulting board for Black. By the same token, if it is Black's turn to move, the board of maximum value will be chosen. The minimax procedure therefore starts at the deepest level of the tree currently being examined, evaluates all the boards at that level and then carries the values back up the tree. In going from one level to the next-higher one it alternately minimizes and maximizes, depending on the turn. Ultimately it arrives at the current board and presents Black with a value for each possible move. Naturally Black chooses the move with the highest value.

A checkers program consists of more than these three components. It must interact with a human player, accept moves from a keyboard and print the moves it decides on. In addition it may display a picture of the current board with symbols representing the two sides and their pieces. In many cases the program also allows the human player to select a "level of expertise" at which the program is to play. The level

For a pro like Andy Bean, golf is more than a game... sometimes it's hard work. Especially when he's caught in the back of a bunker, or buried deep in the rough. Those are the times he'd rather jump in his Jeep® Grand Wagoneer and take the easy way out.

With a quick switch into the sure-footed traction of 4-wheel drive, his Grand Wagoneer is capable of taking him just about anywhere–like his "secret" fishing spots, off-limits to ordinary vehicles.

The confidence you experience behind the wheel of a Grand Wagoneer is comforting, but there are many other comforts that make this vehicle the most luxurious 4-wheel drive wagon around. Standard features like supple leather and plush upholstering, power steering, power windows and door locks, and a premium stereo sound system, to name just a few.

The Grand Wagoneer does so many things so well. But you may appreciate it most when conditions are the worst. That's when it smooths out the rough and drives on like a pro!

Jeep® Grand Wagoneer. The Ultimate Wagon.

AVAILABLE FOR PURCHASE OR LEASE AT PARTICIPA...

SAFETY BELTS SAVE LIVES.


ROYAL PAISLEY
ROYAL PAISLEY
FROM CHIVAS BROTHERS
BLENDED
SCOTCH WHISKY
100% SCOTCH WHISKIES
BLENDED & BOTTLED IN SCOTLAND
CHIVAS BROTHERS
The taste is guaranteed right on the bottle.

is generally a number that reflects the program's look-ahead: the number of levels in the checkers tree it is to explore from the current position.

When the look-ahead is set to a fairly deep level, the program invariably takes longer to move; each additional level of the tree to be explored may double the time needed. To reduce the amount of searching the program must do, the minimax procedure may incorporate a clever idea called alpha-beta pruning, discussed in the box below. Alpha-beta pruning can be done at virtually every level of the tree and can eliminate a tremendous amount of unnecessary searching and evaluation.

The checkers-playing algorithm is not difficult to program. Indeed, Truscott feels there is no need to have a large mainframe computer to compete effectively. "The microprocessor is well suited for developing checker programs," he writes, "and industrious hobbyists will probably make significant contributions in this challenging, rewarding and still completely open area."

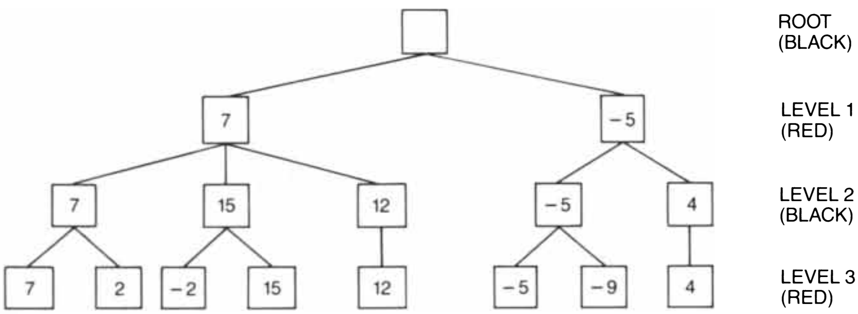
For readers with some programming experience the foregoing account would be complete enough for them to design and write a checkers program. Other readers may still be uncertain how to proceed. The construction of a move generator and that of a board evaluator are reasonably straightforward projects, but the minimax procedure may call for discussion in greater detail. How does a program move up and down the game tree, remembering where it has already been and deciding where to go next? Some advice on traversing the tree is given in the box on the opposite page.

Once a game-playing program has been designed, written and debugged, its limitations generally become apparent when it is run with the look-ahead set to a large value. Moves can take too long or the amount of space needed to store the game tree can exceed the available memory. There are many tricks for reducing demands on time and space. For example, I have represented each checkerboard by a two-dimensional array in which half of the elements (those corresponding to the light squares) are invariably zero. The board can be reduced to a one-dimensional array with the empty spaces eliminated. When that is done, however, the rule for generating moves becomes more complex.

The question of whether it is possible to write a checkers-playing program that never loses to a human player boils down to a question about the nature of checkers itself. In *n*th-order checkers is there a strategy by which Black can always win? Can Red always force a win? Is the game a draw? An analysis of *n*th-order checkers is possible by hand only when *n* is 2. I have analyzed 4 × 4 checkers by laying out the game tree down to an average of 10 moves. At this level there are a few wins for Black and an equally small number of losses. The remaining positions all seem to be drawn ones in which kings chase one another eternally around the cramped quarters of the 4 × 4 board. I have not confirmed draws in all cases, but if my evaluations are correct, the minimax procedure gives the opening board a value of zero—a draw.

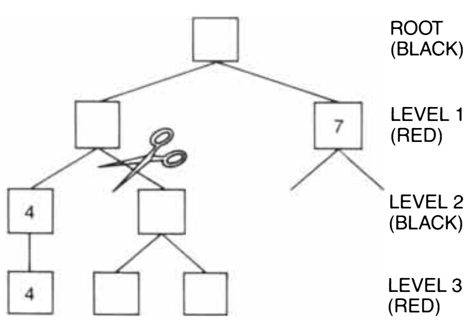
Is 6 × 6 checkers a draw? When the program written by my students played against itself at large look-ahead set-

The Minimax Procedure and Tree Pruning



The minimax procedure is applied to a game tree after all the legal moves have been generated to some preset depth and the board evaluator has estimated the value of the position reached at the end of each branch. Here it is Black's turn to play, the tree has been searched to a depth of three levels and the board evaluator has assigned to each terminal board an integer reflecting its relative value to Black. The task of the minimax procedure is to select a move for Black based on these values.

Note that the sequence of moves most favorable for Black leads through boxes labeled 7, 15 and 15; Red, on the other hand, would favor moves yielding the sequence −5, −5 and −9. The minimax procedure takes into account these conflicting preferences. The program first examines all the level 3 moves that are "children" of a given "parent" move at level 2. Since it is Black's turn to move at level 2, the parent is assigned the value of the highest-rated child. The same procedure is used to assign values to the boards at level 1, but because it is Red's turn, the lowest values are chosen. If both players evaluate the boards in the same way, play would follow a path through moves rated 7, 7 and 7.



Search trees deeper than a few levels can be enormous, and only by pruning them can a thorough search be made. Here it is Black's turn to move, and a partial search has already given a value of 7 to one of Black's choices at level 1. In investigating the other choice the program has found that one of Red's possible responses at level 2 has a value of 4. It follows that there is no reason to explore the rest of the tree; since Red will choose the lowest-value move, Black's alternative at level 1 cannot have a value greater than 4, and so the move assigned a value of 7 is preferable. The elimination of paths that Black can ignore is called alpha pruning; the corresponding process for Red is beta pruning.

tings, the games seldom reached a conclusion. On this basis alone one might be tempted to conjecture that 6 × 6 checkers is a draw. Naturally, if both 4 × 4 and 6 × 6 checkers turn out to be draws, the shadow of suspicion will fall heavily on 8 × 8 checkers. Perhaps someone reading this has that rare combination of skill and daring (or is it foolhardiness?) to write a 6 × 6 checkers-tree analyzer. On a mainframe computer the project may just be possible.

On the subject of nth-order checkers there are some interesting results from theorists of computing. It turns out the following problem is computationally intractable: "Given an arbitrary legal placement of Black and Red kings and men on a $2n \times 2n$ board, decide whether Black can win." In 1978 the problem was shown to be "*P*-space hard" by Aviezri Fraenkel of the Weizmann Institute of Science in Israel and by Michael R. Garey and David Johnson of AT&T Bell Laboratories. For a problem to be *P*-space hard is even worse than for it to be *NP*-complete, and most theorists think problems classified *NP*-complete are unlikely to have any practical general solution. The finding suggests that if a program could play perfect nth-order checkers, the average time to generate a move would increase faster than any polynomial function of n.

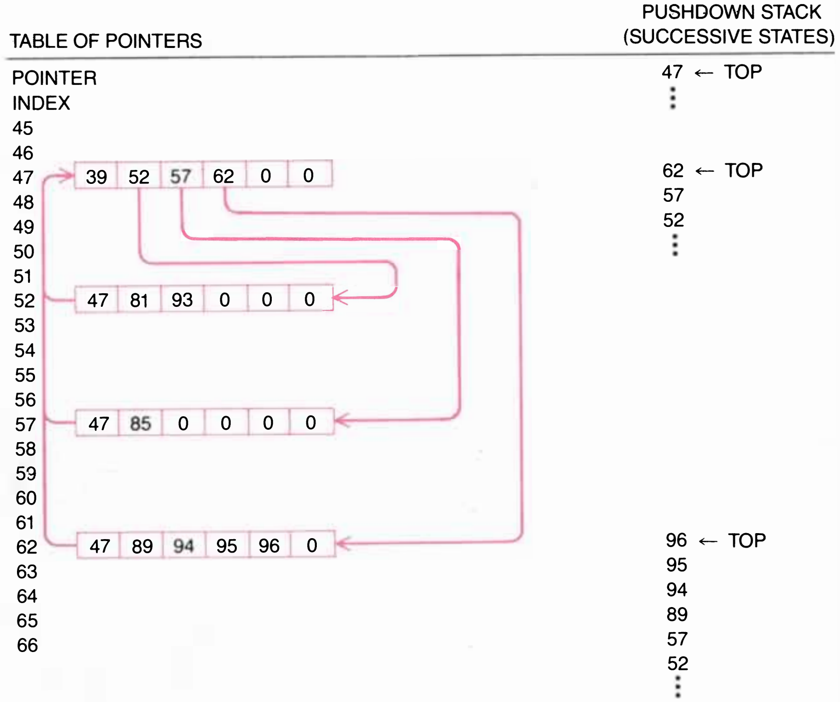
A program that would never lose a game of checkers would also depend heavily on computational power. Even if checkers were inherently a drawn game, a successful program would undoubtedly have a fairly large look-ahead factor, other things being equal. Truscott writes of the program he developed with Jensen: "Averaging five seconds per move (on an IBM 370/165), the program is just able to defeat its authors. Averaging 20 seconds per move, the program is perhaps the 100th-strongest checker player in the United States. Averaging 80 seconds per move, the program is perhaps the 10th-strongest player in the world."

Sometimes it is disillusioning to read a description of a game-playing program in which its internal operations are laid bare. It is easy enough when playing against a program opponent and knowing nothing of how it works to impute to it marvelous intellectual powers it simply does not have. For many people the fantasy is enjoyable. One can only hope that disappointment over its loss is replaced by delight in the structure and operation of game-playing programs.

The boy had asked whether it is possible to write a checkers-playing program that would never lose to a human player. The question touches on two themes central to artificial intelligence: What *is* intelligence, and how intelligently can programs be made to behave? When such questions are brought to focus on restricted forms of human "intelligent" activity such as game playing, the question seems to dissolve into a mass of technical details. Is the answer there?

The search for a theory of intelligence continues. Some A.I. workers have suggested that constructing a machine that thinks may be somewhat like constructing a machine that flies. A kind of "intellectual airfoil" theory may be possible. To fly it is not necessary to build a mechanical bird; an airplane performs quite adequately. Will we ever have an AIrplane?

Traversing the Game Tree



When a checkers-playing program is selecting a move, it needs to explore the game tree to some fixed depth, visiting each branch exactly once. A simple method of traversing the tree employs a table of pointers, which record the relations among branches, and a data structure called a pushdown stack, which monitors the program's current position in the tree and maintains a list of the branches yet to be visited.

As each new board is generated it is assigned a number, which serves as an index to the array of pointers. The pointer associated with a board is itself a list of board numbers. The first number in the list is that of the board's parent; all the other numbers point to its children. If the program has reached board 47, the contents of the pointer indicate that it should continue the search with boards 52, 57 and 62. If the program is at board 62 and it must return to the parent, the first entry in the pointer directs it back to board 47.

It is the function of the pushdown stack to keep track of the program's progress through the table of pointers. The stack itself can be created by setting up an array called STACK and a variable called TOP, which holds the address of the top item on the stack. Whenever an item is added to the stack (pushed) or deleted from it (popped), TOP is incremented or decremented.

The item at the top of the stack is the number of the board to be examined next. If 47 is the top item, the tree-traversal algorithm pops that value off the stack and then examines the pointer list for board 47. The first number in the list (referring to the parent of board 47) is ignored, but the remaining nonzero entries are pushed onto the stack and the value of TOP is adjusted accordingly. Hence board 62 is selected for investigation next.

Each time the program reaches a board at the maximum allowed depth, the board evaluator is called on to give it a rating. The minimax procedure can then consult the pointer list to find the parent of the board and give it a provisional minimum or maximum value. On later visits to sibling boards the procedure will have an opportunity to revise the value.

Which?

Which South American companies import electronic components from the U.S.?

How?

How can I find all the orthodontists in Toledo for my direct mail campaign?

What?

What are the mechanical properties of shape memory materials?

Has?

Has the cost-effectiveness of teleconferencing changed in the last year?

Who?

Who are some of the experts in medical sonar scanning techniques?

all these questions by 9 a.m.?

mation, you can store it on disk or print it out. In-Search is compatible with most word processing programs.

How to get the answers. To see a demonstration or to purchase In-Search, call (408) 986-1200 for your nearest dealer. Or send this coupon for a demonstration disk designed for IBM, IBM-compatible, or Texas Instruments personal computers.

In·Search™

Use your VISA or MasterCard to charge the purchase of your $5.00 diskette (California residents please add applicable sales tax.)

Menlo Corporation™
4633 Old Ironsides, Suite 400
Santa Clara, CA 95050

☐ Enclosed is $5.00 for my In-Search demonstration diskette.
☐ Please send me a free booklet with more details.

NAME
COMPANY
ADDRESS
CITY STATE ZIP
MY OCCUPATION IS

69

BOOKS

Green manures, sport physics, Southwest Indians, radioactivity in the environment

by Philip Morrison

BIOLOGICAL NITROGEN FIXATION TECHNOLOGY FOR TROPICAL AGRICULTURE, edited by Peter H. Graham and Susan C. Harris. Centro Internacional de Agricultura Tropical, Cali, Colombia. Available in the U.S. from UNIPUB, 345 Park Avenue South, New York, N.Y. 10010 ($35). GENETIC ENGINEERING OF PLANTS, a conference summary by Leslie Roberts. Board on Agriculture, National Research Council. National Academy Press, Washington, D.C. 20418 ($9.50). The dwarfed but heavy-headed stalks of golden grain flourish now from Mexican, Egyptian and Punjabi wheat fields to the paddies of South China. Along with them march the high towers of the new ammonia plants, from Chihuahua to Baroda to Changsha and beyond. Given the inputs of modern plant varieties, fertilizer and enough water, skilled farmers have over the past 25 years carried out their Green Revolution. Cruel Prince Famine is now in exile, perhaps forever, from the cities and the wide fields; he is reduced to harsh guerrilla raids during African, and perhaps Brazilian, droughts. But the tyrant King Poverty reigns still over most of his old domains, wherever men and women must till acreage too narrow or rack-rented.

A few years ago a new campaign began. Its aim is to reach out with modern agricultural development even to those too poor to afford the expensive tube wells and nitrogen fertilizers that make the new high-yielding strains worthwhile. The new volume from Cali, one of the international chain of agricultural research centers that are the arsenals of the Green Revolution, is the report of a workshop held there a few years back. It offers brief, varied accounts, often in a compact jargon, by almost 200 participants from some three dozen countries. We can dub the new campaign a Blue-Green Reform, since its single purpose is to search out new ways to do more widely what the blue-green algae have always done unseen for wetland rice.

Practical demonstration of the work is compelling; one much-cited example comes from the Rice Research Institute, facing the high volcanic cone of Mount Mayon in southern Luzon. That laboratory was the original link in the research chain. The agronomists carefully measured the nitrogen balance for a certain trial paddy plot during 23 successive crops, a good yield of 3,000 pounds of rice per acre maintained over the years without any added nitrogen. Allowing for inputs of free nitrogen in rain and irrigation water, it was clear that about 50 pounds per acre of newly fixed nitrogen had been supplied to each crop of rice, more or less the amount you might have wanted to add in chemical fertilizer. It has been recognized for 50 years that this is the work of free-living nitrogen-fixing bacteria thriving in the flooded land. Among the grains only rice offers that possibility today. Of course, added nitrogen often grants still higher yields.

For the rest of the cereals nitrogen must be somehow supplied in good amount. Farmers have long understood this need. Crop rotation or even interplanting the cereal and pasture grasses with the legumes that enjoy a tight symbiosis with nitrogen-fixing bacteria in their roots is very old. Everyone knows animal manures also enrich the crop; an alewife in every hill of New England corn was the famous tip to the Pilgrims from the first farmers of Massachusetts. In the intensive agriculture of China even today a major part of the nitrogen taken out in crops that supply a billion people with essential amino acids comes from manures, mostly the wastes of the foraging pig. Night soil—human manure—plays a much-remarked but rather subordinate role.

There are plenty of other paths to biological nitrogen. For 1,000 years Chinese rice growers have made use of a green manure: *Azolla,* a tiny aquatic fern that covers a freshwater surface with a mosslike mat. At the base of each minute leaf a cavity holds symbiotic blue-green algae, nitrogen fixers. *Azolla* itself yields a crop in three or four weeks; at the price of labor and perhaps some extra area of paddy the nitrogenous fraction of chemical fertilizer is replaced. For dry-land crops, of course, the same host role is played by the legumes, whose nitrogen-fixing bacterial symbionts form tiny nodules among the root hairs of crops such as soybeans or alfalfa, and indeed most other plants of the family. In the Tropics there are even tree legumes, fine soil-enriching woody perennials, much used as a part of traditional cropping systems by Mexican farmers.

In the U.S. there has been built up for a couple of generations an important microbial industry: mass culture of the nodular blue-green strains best suited as symbionts for the crop legumes. These helpful cyanobacteria are added to carriers such as peat or even granules of coal and sown along with the seed. Soybeans and black beans, cowpeas and peanuts are all reported to respond to deliberate inoculation of the right bacterial symbiont by an increased yield, particularly under poor conditions in dry, hot soils, in lands far from the original home of the crop legume. We know now that many other crops, among them sugarcane, sorghum, amaranths, even potatoes, support specific nitrogen-fixing bacteria among their roots, even though nodular colonies are not present. Inoculation with the right symbionts can be good husbandry for those crops too; the technique is young, and there is a lot to learn.

Fixing nitrogen is not free, even for plants with cyanobacterial allies. The enzymes are there, but it all costs energy. It takes a couple of tons of average coal to yield a ton of ammonia in modern engineering practice; it costs somewhat less in carbohydrate yield from the photosynthesizing crop to win fixed nitrogen biologically, by one estimate maybe a ton and a half of starch per ton of fixed nitrogen. The ruling fact is that a ton of rice, 85 percent of which is good food and not water, requires only 50 indispensable pounds of nitrogen.

The favorable tradeoff is well worthwhile. It was the sharp rise in oil prices that first called new attention to biological nitrogen sources. Clearly if soil inoculation or the right intercrop legume or some green manure old or new gives an increase in yield with little outlay of cash, the hardworking farmer with little capital and little contact with the outside world can gain from tending to the blue-greens. That is the goal of this complicated endeavor.

In a way the point made by *Biological Nitrogen Fixation Technology for Tropical Agriculture* is simply that what appears to be one specific crop is in fact two crops. One is microbial, the other a green plant. Both respond to a knowing husbandry. *Genetic Engineering for Plants,* the second work reviewed here, takes another approach: solve the problem by making the two crops into one. The 90-page brochure is a brief connected summary of a two-day convocation held in Washington about a year ago, all-American save for one Australian investigator. The aim of the meeting was

"If they ever make computers as easy to use as the telephone, I'll buy one."

OK, what size

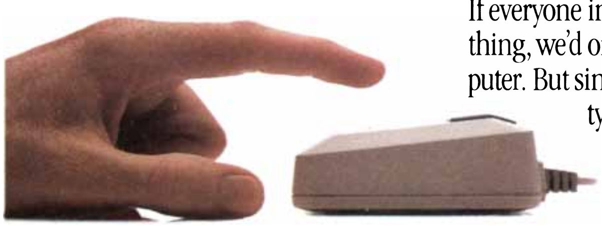
Macintosh $2495

Lisa 2 $3495

Until now, only 7 out of 100 office workers could work a computer.

The reason is simple: most business computers aren't.

They make you wade through



If you have a finger, you can run the mouse that runs the computer.

manuals thicker than the Manhattan Yellow Pages. Spend weekends in computer classes. Memorizing things like "Q-Slash-Z-Control-A-Return."

But now, all that's history. Thanks to the most sophisticated family of computers in history. And their "mice," those friendly-looking devices pictured above that make using a computer easier than phoning your branch office in Biloxi.

Introducing Apple 32 SuperMicros.

If everyone in your office did the same thing, we'd only need to make one computer. But since managers manage, typists type and senior vice presidents do who knows what, one computer simply won't fit all.

It takes four.

Macintosh,™ Lisa™ 2, Lisa 2/5, Lisa 2/10: the Apple® 32™ SuperMicros system. A compatible family of 32-bit microcomputers that can make everyone in your office more productive. In minutes, not days.

Because they all utilize revolutionary mouse technology. Moving the mouse on your desk moves a pointer on the screen. Point to an object, click the button on the mouse, and the computer does what you want.

There are no cursor keys. No complex commands. Which means computers are now as easy to operate as any other office tool. Like the telephone. Making it possible for you to overcome the single biggest problem known to business computers:

Getting people to use them.

And once they do, you'll see computers can be as useful for communicating as the telephone.

Using MacWrite's™ different type styles and sizes, for instance, put out memos people will actually read. And with MacPaint,™ illustrate your point with diagrams and schematic drawings.

Lisa's incredible power lets you shuffle programs on the screen the same

would you like?

Lisa 2/5 $4495

Lisa 2/10 $5495

way you shuffle papers on your desk.

As quick as a mouse, you can go from data base management to data communications. "Cut" numbers from your mainframe, turn them into a graph, and "paste" the graph into a presentation. All while printing out an in-basket full of sales reports.

A few of our family features.

If Apple 32 SuperMicros seem like uncommon computers, it's because of all they have in common.

Like brains. The same lightning-fast, 32-bit MC68000 microprocessor. It's twice as powerful as run-of-the-mill 8088's found in ordinary computers.

And they all use new 3½″ micro-disks which store more than the old 5¼″ floppies. You can even run Macintosh disks on any Lisa, giving the entire family access to the hundreds of programs being written by leading developers.

Built-in serial ports let you add printers, modems and other peripherals. Without adding $150 interface cards.

And they can all communicate with your company's mainframe in fluent TTY, VT52 and VT100. And, with extra hardware, IBM® 3278 protocols.

All of which means you should be communicating with an authorized Apple dealer. Since there are over 1,800 of them, phoning one should be easy.

Almost as easy as using a computer.

to examine policy as much as science; in attendance were college administrators, laboratory directors and Federal officials as well as biologists from governmental, industrial and academic laboratories. Of particular interest is the up-to-date account of the status of genetic engineering in agriculture.

Can the 17 genes that control nitrogen fixation in bacteria be transferred en masse to corn and wheat? Along with them must go the collection of regulatory signals that allow expression of the enzymatic recipes of the genes. The entire gene cluster has already been transferred to yeast, but the genes were not expressed. The enzyme system requires an environment rich in energy and poor in oxygen; it is not certain where in a cell of corn the system should be positioned, the more because it makes demands for iron and molybdenum beyond the plant-cell levels. The commonest nitrogen-fixing bacterium has a second set of genes that determine nodule formation through the induced growth of plant root cells, including the construction of thin tubular infection threads that conduct bacteria from one plant cell to another within the nodule. The system has been transferred to the bacterium that induces crown gall, whose tumor-inducing plasmid is the vector favored for the transfer of genetic material to plant cells. The newly engineered bacterium induced root nodules and infection threads properly (an electron micrograph here shows them fairly clearly), but once again the nitrogen-fixing genes were not expressed, even though they were transferred.

We are not as informed about the genetics of development in higher organisms as we need to be. No doubt the more hopeful task is modifying the bacteria to enjoy their new host, rather than trying to modify the much more complex host. We may need to work at both ends. There are surprises already. These new plant breeders work with single-cell cultures in Petri dishes, millions of cells to a dish, screening as many individuals as a plant breeder of the classical kind can find in 150 acres of corn. The technique begins with a piece of scar tissue, the soft, undifferentiated callus that first forms when a shoot, say, is damaged. That clump of cells will grow on the right medium into an entire clone of plants.

The method is already in commercial use for orchids. If the callus is gently dispersed in a liquid medium, each of its single cells, or clumps of a few cells, each develops into an embryo that forms an entire plant. The scheme has succeeded for corn, although it has not worked well for grasses or legumes in general. But there is more afoot. The plants that arise from this kind of treatment are usually not the identical clones expected, individual organisms all very much the same because they have the same genome. In all the useful species tested, the process of going from a differentiated cell through a callus culture to a separated redifferentiation induces a "genetic earthquake." The regenerated plants show a wide diversity of new traits, which are stably inherited. The simplest interpretation of this unexpected genetic jumbling is some failure of the neat chromosomal dance; the diploid state of the starting organism is not maintained. Chromosomes break, duplicate, rejoin. The real molecular mechanism is not yet known. The breeders are already exploiting this mysterious windfall of diversity to select varieties of wheat through cell culture.

The genetic engineering of bacteria, mostly for medical or biochemical applications, is well under way. Work with plants, whose cells contain as much DNA as those of human beings, is still groping among millions of genes, only a few percent of which are active at any single stage of the plant's development. How those genes are regulated in the course of development remains outside what we now know of molecular biology. Perhaps some of the promising techniques will succeed without much deeper understanding; perhaps the pull of the market or the push of other successes in microbial molecular biology will lead enough able workers into plant studies to solve the difficult problems soon. Certainly there is enthusiasm, and the smell of money is as strong in these laboratories as that of solvent vapors. One would do well to remember, however, that the intricate natural systems have a very long history; exploiting the entire system functionally may yet be the shortest path to the set goal of Blue-Green Reform, bringing adequate nitrogen fixation under the hot sun even to the smallholder and the tenant.

SPORT SCIENCE: PHYSICAL LAWS AND OPTIMUM PERFORMANCE, by Peter J. Brancazio. Simon and Schuster ($18.95). A basketball flies at low speed toward the hoop, feeling little effect from air drag. Fit for the simple analysis of projectile motion dear to introductory physics, its flight parabola is a cinch to calculate. Our guide here is a Brooklyn College professor with a physics textbook to his credit, who has long been a serious if wistful aspirant to excellence on the court, where he plays a few times a week all year round. He well knows why and how to compute the entire family of parabolas that link the hand to the hoop, the low-angle line-drive shot no less than the high rainbow one.

Which is best? Everyone admires the cool shooter with a soft touch, the player who uses the economical minimum of force to drop the ball in. The problem is subtler: that minimum-speed trajectory is also the one with the widest tolerance for error in the angle of launch, seven times greater than the error allowed to the flattest trajectory. After 20 years of mediocre line-drive shots our aging physicist-forward improved his performance dramatically once he understood the virtues of the right arch. Practice, with a friend or a coach to the side to help judge the height of the arch, will lead to success, based on "the way it looks and the way it feels." The arched shot calls for an exquisite touch, but it is plainly wise to cluster your long trial efforts around the right optimum.

This engaging book is actually not a how-to-do-it, although because it seeks a rationale for every technique it treats, it can help any serious athlete who has not gone beyond its modest scope. It is instead a volume to satisfy curiosity: how sports work, a book simple and timely enough for television viewers of the summer Olympics and the brightly lighted night baseball diamonds. It deals only briefly with muscle mechanisms and the kinematic evolution of running records. It has nothing much to say about the operations research of sports or the statistical choice of competitive tactics. The focus is closely on matter in motion: bodies, balls and bats, and the resisting medium they traverse. The reader is taken almost at once into the introductory physics of forces and torques, of moment of inertia and center of gravity, of energy and power. The treatment is concrete, so that the stereotypes of the elementary physics texts, all those ladders and falling stones, are replaced with point and freshness by examples that play a natural role on the athletic field.

Only the last chapters go beyond freshman topics; they offer a clear introduction to drag and lift forces in air and in water, treated offscreen, so to speak, by the author's home computer. His results are freely cited, but the book never demands from the reader much mathematical proficiency. Seventeen boxes present sequestered equations and numerical results, from unit conversions to terminal speeds in air. The agreeable way in which "customary units"—miles per hour, say—are employed, like the attractive pages on which commonplace Aristotelian intuitions are gently transformed into working Newtonian dynamics, are the mark of the able and experienced teacher.

The quantitative study of collisions first caught up the savants in good King Charles's golden days, although the modern beginner in physics is likely to be set the same questions. Colliding protons work at the frontier of physics, but little laboratory blocks sliding on their low-friction track are unmitigated pedagogy. Let a bat collide with a ball and the outcome becomes important to many more observers. A couple of dozen pages consider this urgent matter. All the parameters take on meaning. The ball's coefficient of restitution is itself prescribed by high baseball law: "The rebound... shall be 54.6 percent of the initial velocity, with a tolerance of ± 3.2 percent." Soccer balls are livelier; tennis balls are deader. A good description is given of the bat-ball encounter, including some discussion of the effective mass to be ascribed to the player's hand grip.

Is it better to use a fast but light bat or a slower but heavier one? An early study of bat and ball collisions set an optimum mass much below the weights of bats actually in service. The problem is overconstrained; the best choice for ball speed is a long, thin, light bat. When the batter is at the plate, this choice is out, because such a bat has so small a hitting surface that few batters will risk it. The long, light fungo bat with which the coach hits practice fly balls to his distant outfielders is one realization of the best bat mass, but the coach has only to hit a ball he himself has tossed lightly into the air. The mighty Babe had wrists strong enough to swing a bat with weight half again the big-league norm, to magnificent effect.

Nowadays the aluminum bat has taken over amateur play in schools and colleges. It is indestructible, whereas wood bats often break. Direct experiment shows that the aluminum bat adds 10 percent to the length of a line drive. Moreover, the hollow metal bat ends the tight coupling between weight, moment of inertia, length and striking surface, freeing the batter to roam a parameter space undreamed of in the old days of ash. The statistics in the college leagues bear clear testimony to the new bat; batting averages are way up, and earned runs too. There is rough justice here. The golfer, the hockey player, the cricket batsman and the racquet wielders are all armed with flat striking surfaces. Only in baseball must the batter confront a fast-flying sphere with a circular cylinder. Making a hit is one of the most difficult routine plays in any sport; even the finest batters succeed only about one time in three. So far professional baseball will admit only the traditional wood bat. In the big leagues the war between pitcher and hitter is waged under arms control.

Air is the medium of most sports. The 16-pound iron shot can ignore air drag, which can shorten its flight only by a few inches out of 70 feet. Of the dozen balls whose terminal velocities are tabulated here, only those of golf, baseball and football enter often on long flight paths where aerodynamic effects shape the play. Professor Brancazio has computed drag effects rather simply, using plausible drag coefficients in a trial-and-error fashion to fit experience. Lift is even more complicated, particularly when it arises from spin. The low launch angle of a long drive in golf is a consequence of the high backspin-induced lift. A golf ball just off the driver head may well be spinning at 8,000 revolutions per minute. What are hook and slice but axes of spin slightly askew?

The dimples on the golf ball induce a turbulent boundary layer. In the prevailing regime of flight they decrease drag by reducing the size of the wake. At the same time the lift is increased. The dimpled ball flies better when it is well struck, but it also tends to hook and slice more easily. A golf ball with dimples on an equatorial band only, leaving the poles smooth, is a duffer's compromise. It is said that it curves much less often and falls short by only 10 percent.

It is baseball that "lives and dies" by the curved flight of a spinning ball in air. Wind-tunnel studies have shown there is a significant lateral deflecting force on a baseball even when it is not spinning. The origin of this force is the unsymmetrical wake induced by the raised seams on the surface of the ball. A very slow rotation can lead to a varying wake and a rather erratic flight: the magic of the knuckleball and the black magic of the spitball. There remains reasonable doubt whether or not the fast ball hops or the curve ball breaks. What is at issue is not the fact of curvature, of course, but the sharpness of those departures.

Both perspective effects and increased deflection with time cause the curvature to increase as the ball approaches the batter. Ballplayers say a good overhand curve looks as though the ball is "falling off a table." No smooth increase in curvature fills the bill. Perhaps the witnesses are overwrought. It is possible, however, that the stability regime of the wake is affected by the changing orientation of the seams. We admire the aerodynamic control of men such as Hoyt Wilhelm, an "outstanding knuckleballer" who pitched with success in the major leagues until he was 49; a noisy wake can bring unpredictable perturbations even to a ball thrown at modest speed by an old arm.

This informally written volume and its references also open a substantial recent literature of theory and experiment on the physics of the crawl stroke and cricket, of karate strikes and the angular momentum of divers in the air, of the "sweet spot" and the wonderful soaring flight of that massive Frisbee the discus, and much more. This is a fine place to begin the analysis; surely the end of these rich applications of physics lies well over the fence and far away.

HANDBOOK OF NORTH AMERICAN INDIANS, William C. Sturtevant, General Editor. Twenty volumes. Volumes 9 (1979) and 10 (1983), Southwest, Alfonso Ortiz, Volume Editor. Smithsonian Institution, U.S. Government Printing Office (Volume 9, $23; Volume 10, $25). In Teddy Roosevelt's day the

Smithsonian issued two thick volumes under a title quite similar to this one. For a long time they served as a standard reference for the history and culture of the peoples treated, a respected handbook. Congress put up the money for a new version of the encyclopedic work in 1971, and the scholars have been planfully busy ever since. The modern set—it can be called a handbook only by analogy—will consist of 20 fat books, of which five have appeared, the first in 1978, with Volume 10 the latest.

The heart of the work is a subset of 11 volumes covering the 10 distinct cultural regions north of the urban civilizations of central Mexico. (Only the Southwest region rates two volumes, those under review here.) There will be four volumes of a more general nature, one volume to treat arts and crafts arranged by type rather than by place of origin, one volume on languages (promising a sample of 14 representative grammars given in some detail) and two volumes of historical biographies of thousands of individual Indian and Eskimo men and women. Finally, Volume 20, to be issued sometime in the 1990's, will be a comprehensive index. American ethnography today, like art, is long.

The two volumes on the Southwest are related, although they are distinct enough for each to be useful without the other. They share the same editor; a few articles that open Volume 9 are relevant to both books, just as a few overall summaries close Volume 10. About 100 articles lie between, by about as many expert authors. Brief accounts are given one by one of the history and culture of two dozen Rio Grande Pueblo villages, from Taos out to Acoma, all lively settlements still, not to forget a piece on the abandoned sites. The larger and more distant Zuñi villages draw half a dozen essays, and the Hopi, the only people who remained on their ancient arid lands when around the time of Dante the rest withdrew toward the river valleys, are discussed in nine more essays.

Volume 10 is given over to those cultures of the Southwest that are not Pueblo. It begins with a few small Arizona groups, such as the Havasupai, who dwell near the floor of the Grand Canyon. Ten pieces discuss the Pima and the Papago, a few more the Yaqui and the Tarahumara and smaller groups in northern Mexico. The volume ends with the Navajo and their Apache kin. The Navajo are much the most numerous of all the peoples covered in the 20 volumes; certainly there are today 100,000 or more people who speak the Navajo language.

A work so encyclopedic is plainly for the reference shelves. But the diversity of treatments, like the diverse authorship, gives it a readability that will repay many general readers. The volume is nowhere forbiddingly technical, even though a few pieces enter the statistical domain of modern mathematical taxonomy or resort to linguistic and kinship diagrams of some intricacy. On the contrary, the prose is lightened by a generous provision of maps, drawings and many photographs old and new. The concreteness of the specialized articles is more agreeable than a long string of comprehensive accounts would be.

Most of the articles, of course, are tersely matter-of-fact, carefully and tightly organized for clarity. Some, however, are personal, even eloquent. Two pieces in particular, one by a Pima author and one by a Navajo, address the reader directly. The Pima piece reports the recorded proposals of two old men, using their own words as examples of the Pima style of open speech in the public good. The Navajo writer candidly describes the shortcomings of the Navajo nation in which he himself lives, ending with the strong admonition, "Whites: Hear us well! There will be no more Kit Carson . . . no more interference . . . no more brutality." For him the end of colonial dependence entails the right to make one's own mistakes and learn from them.

It would be a pity to close with only this external description. There is no room here for a fair sampling of the 1,500 crammed and varied pages. One may nonetheless offer a taste. There is a remarkable pair of photographs: on one page we see a long pole made from a giant cactus rib, serving in 1968 on Papago lands for the Noceo family to harvest the sweet ripe red fruits from the top of the tall saguaro cactus. A few pages farther along there is a photograph that shows a row of stately saguaros. Each of them bears on its top a cross, and the row faces the San José Mission, its old facade redecorated in a striking black-and-white pattern by a modern Papago artist evoking traditional Papago pottery designs. In the two pictures the complex history of the Southwest and the stability of its cultures are compact.

The Hopi name four cardinal directions in their sun-baked land: the sunrise and sunset points at the two solstice times. They are an original people. For three centuries from about A.D. 1300 the Hopi strip-mined coal on their mesas, taking out a couple of pounds a day per head, which they used for cooking, for heating houses and kivas, and for firing pottery. Over the years they mined more than 100,000 tons, which largely or entirely supplanted firewood. They gave up the winning of coal just after the time of contact with the Franciscans. We do not know why; it may be that accessible coal was running out at the very time newly introduced donkeys and iron axes made the gathering of firewood easier.

Finally, it is good to read that the health of the Navajo people rapidly improved in the decades after World War II; "a fairly hale and hearty people were sometimes surprised to hear themselves described as being among the world's least healthy" by well-wishers a little out of date. In 1945 a Navajo infant had a 50 percent chance of surviving to school age; by 1967 that chance was better than 98 percent. Tuberculosis has been sharply reduced. There is a felt need for traditional medicine men, and it is being filled through training in a context of both ancient and modern psychiatry. But the long-hoped-for "flood of Navajo physicians" has not yet risen to follow the first pioneers; an American Indian medical school near Window Rock is still only an aspiration.

The next volume is promised for the fall; it will treat the remarkable people who dwell along polar shores from the Bering Sea to Greenland.

ENVIRONMENTAL RADIOACTIVITY, edited by Thomas M. Koval. Proceedings of the Nineteenth Annual Meeting, 6–7 April, 1983, of the National Council on Radiation Protection and Measurements, 7910 Woodmont Avenue, Bethesda, Md. 20814 ($17). Relevant and often surprising, a dozen papers, a review lecture and a number of terse accounts of important publications in the field make up this readable report from the world of the experts on radiation doses and what they mean.

It is almost 10 years since the Council on Radiation Protection published its official summary of natural background radiation as measured in the U.S. A few pages bring the subject up to date as far as the strictly natural component is concerned. The cosmic-ray estimate "has remained firm," but the external radiation we absorb from terrestrial sources has been reduced a little to take account of our penchant for living indoors. The effective dose from inhalation of natural radioactive gases has been increased sixfold; measurements and interpretation were always uncertain, since the dose is entirely absorbed within the delicate bronchial epithelium. All this exposure, most of it as old as the human species, seems to contribute perhaps one in 40 of the cancer deaths we now count.

There is a new survey of radon in the home. The radioactive gas is still not very well studied, but the data here, collected from 100 frame houses in the western suburbs of Chicago, not far from the Argonne National Laboratory where the physicists work, are unexpected. Forced-air heating spreads the radon uniformly through the house from its source in soil gases coming from near and around the foundations. In a house with hot-water radiators the basement is more than 10 times as radioactive as the first floor, at least outside the heating season. The variations are not understood; they may reflect the level of the water table or some other soil variable.

What happens to the householder's lungs from the radioactive gas is also not clear; it may depend critically on the concentration of airborne dust. The alpha particles emitted by radioactive elements are part of a complex microenvironment we do not yet control. Houses of modern construction clearly leak less air than usually freshened the drafty houses of old. To that degree the radon-caused dose is a modern artifact. The effects are still small, although it is to be remarked that the highest 10 percent of the ordinary houses measured in Chicago were as radon-contaminated as the houses and schools of Grand Junction, Colo., where the town contractors routinely used uranium mill tailings as backfill.

The old headlines march past one by one; there are papers on radioactivity from Three Mile Island ("of minor significance"), on the contaminated atolls of the Marshalls, on the fallout downwind from the Nevada weapons tests, on the Windscale reactor fire of 1957 in England, on an estimate of the radiological risks of a future nuclear fuel cycle and on a variety of more technical questions of methods and criteria. A physicist from the New York State Department of Health looks carefully at the radiological risk of high-level active waste-disposal schemes. Salt-bed disposal is fine, if the formation remains unbreached. Wet hard-rock sites look barely acceptable; their long-term safety depends on the slow travel of water subsurface. (There are no safely dry sites.)

This expert opts for the burial of stable canisters in certain self-healing wet sediments—like chocolate mousse—on the ocean floor. You simply heave the canisters overside in the open sea, so that they fall slowly through deep dark waters to oozy graves 30 meters into the floor of the ocean, scattered far from shore and from one another. Radiological safety is assured at low cost if it all works as planned, if the mushy sediments do seal above the heavy canisters, if they all penetrate to full depth, if accidents in port or in the course of shallow-water transport do not occur, and so on.

These look like small risks, even if they are real ones. In this volume, however, there is a cheerful emphasis on a peaceful world. Only the senior physicist who delivers the ceremonial lecture on the human environment in general addresses the looming problem: "Without any doubt I would place nuclear war at the top of my hierarchy of threats to the environment." Few could disagree with Dr. Merril Eisenbud. A map of Eniwetok Atoll appears early in the volume, although it does not record the ocean-filled crater now in the ring where before the first thermonuclear test in 1952 was an islet named Elugelab.

Supra tames tortuous hairpins, undulating curves and tantalizing straightaways. Unleash one. And take control.

Its incredible engine is an electronically fuel-injected 2.8 liter Twin Cam six. By raising the compression ratio, enlarging the intake manifold and streamlining the exhaust system, you'll find 10 more horses than last year.

Revel in the raw power of the fastest Supra ever.

The road is putty in your hands when you've got standard features that are far from standard. MacPherson struts, independent rear suspension, variable assist power rack-and-pinion steering, 225/60HR14 wide radial tires on 14" x 7" aluminum alloy wheels, and a front air dam to channel the wind give Supra precision to conquer the curves, tame the turns and chew up the straightaways.

Unrelenting evolution has created the ideal driving environment inside, too. Supra's famous 8-way adjustable driver's Sport Seat has every conceivable adjustment, from lumbar support to hip and thigh bolsters. The fully instrumented cockpit has an AM/FM/MPX cassette stereo with 5 speakers, 7-band graphic equalizer and power antenna also.

Look around.
Isn't there a road you'd like to own?

OH WHAT A FEELING!
TOYOTA

BUCKLE UP—IT'S A GOOD FEELING!

WHY OWN JUST A CAR WHEN YOU CAN OWN THE ROAD. THE 1984 TOYOTA SUPRA.

Established 1845 SCIENTIFIC AMERICAN July 1984 Volume 251 Number 1

The Homelessness Problem

Many of the homeless people wandering the streets of American cities and crowding into emergency shelters are mentally ill. They require adequate housing and appropriate psychiatric care

by Ellen L. Bassuk

More Americans were homeless last winter than at any time since the Great Depression. Estimates of the size of the vagrant population vary widely. The National Coalition for the Homeless puts the figure at 2.5 million for 1983, an increase of 500,000 over the preceding year. The Federal Department of Housing and Urban Development (HUD) estimates that only 250,000 to 350,000 are homeless nationwide. Whatever the number is, everyone agrees it is growing.

Particularly in the past five years government officials and private groups in cities around the country have responded by opening emergency shelters to try to meet the immediate needs of the homeless. Beds in these shelters fill as soon as they become available, and still only a fraction of those in need are provided for. Some of the rest seek temporary refuge elsewhere, for example in hospitals, but most probably fend for themselves on the streets, huddling in doorways or over subway ventilation grates. When the weather turns cold, some die.

At night in New York City 18 public shelters house some of the thousands of men and women who roam the streets during the day; 16 of these shelters did not exist before 1980. Private groups in New York have also stepped up their efforts. In 1982, 10 churches offered a total of 113 beds to homeless people; by the end of 1983, 172 churches and synagogues were providing a total of 650 beds in 60 shelters. In Boston two large shelters recently doubled their capacity. Nevertheless, on a snowy night in January, Boston's largest shelter, the Pine Street Inn, reported a record number of "guests": the 350 beds were filled, as always, and 267 people crowded onto the Inn's bare cement floors.

Who are these people? Unfortunately there are no reliable national data on the homeless, even though they have always been numerous in American cities. Anecdotal evidence suggests that in the decades before 1970 most of the homeless were unattached, middle-aged, alcoholic men—the denizens of Skid Row. Since about 1970 the population appears to have been getting progressively younger. Moreover, the sparse literature on the subject and my own experience as a psychiatrist working with homeless people in Boston leads me to believe a more important change has taken place: an increasing number—I would say a large majority—of the homeless suffer from mental illness, ranging from schizophrenia to severe personality disorders.

At a time when the accepted solution to the homelessness problem is to establish more shelters, this finding has disturbing implications. Shelters are invaluable: they save lives. The trouble is that many shelters do little more, and the mentally ill need more than just a meal and protection from the elements. Those whose disorders are treatable or at least manageable require appropriate psychiatric care, which they do not get at shelters. The chronically disabled people who will never be able to care for themselves deserve better than to spend their lives begging on the streets and sleeping on army cots in gymnasiums. Shelters have been saddled with the impossible task of replacing not only the almshouses of the past but also the large state mental institutions. At this task they must inevitably fail, and thus American society has failed in its moral responsibility to care for its weakest members.

The statement that a majority of the homeless are mentally ill does not in itself explain why their number is growing or why a particular individual joins their ranks. Without reliable data it is difficult to answer the first question, but several factors may have contributed to the recent swelling of the homeless population. The most obvious one is the re-

HOMELESS PEOPLE sleep in an armory in New York City. As the number of homeless

cession. Unemployment reached a peak of 10.7 percent in November, 1982, its highest level since the 1930's. Some of those who lost their jobs and incomes undoubtedly lost their homes as well.

The effects of unemployment are intensified by another problem: the dearth of low-cost housing. According to an analysis of the Federal Government's Annual Housing Survey by the Low Income Housing Information Service, the number of renter households with incomes below $3,000 per year dropped by about 46 percent, from 5.8 to 2.7 million, between 1970 and 1980; at the same time, however, the number of rental units available to these households at 30 percent of their income fell by 70 percent, from an estimated 5.1 to about 1.2 million (excluding dwellings for which no cash rent was paid). As the "housing gap" widened, the median rent paid by households in the lowest income bracket rose from $72 a month in 1970 to $179 a month in 1980. That works out to 72 percent of an annual income of $3,000 and leaves $71 a month to cover all other household needs. A family devoting such a large fraction of its income to rent is in a precarious position: it may easily be dislodged by a drop in its income or by a further rise in its expenses. Unemployment and the lack of low-cost housing help to account for the increasing number of homeless families (as opposed to individuals), which once were rare.

Recent cuts in government benefit payments may also have thrown some people onto the streets, although the evidence is inferential. One of the Federal Government's most controversial measures in this area has been its effort to reform the Social Security Disability Insurance program, which in 1983 provided monthly benefits to a total of 3.8 million disabled workers and their dependents. To receive payments a worker must be physically or mentally unable to perform any kind of "substantial gainful work" for which he is qualified, regardless of whether such work is available where he lives. Following a report by the General Accounting Office that as many as 20 percent of the beneficiaries might be ineligible under the law, the Reagan Administration launched a "crackdown on ineligibility" in March, 1981. Between 150,000 and 200,000 people lost their benefits before the Administration halted its review of the beneficiary rolls in April, 1984, amid charges that truly disabled people, including some who were too mentally disabled to respond to termination notices, had been stricken from the rolls. Again, a lack of data makes it impossible to draw definite conclusions, but it seems not unreasonable to infer that the loss of disability benefits reduced some people to not being able to pay for their housing.

Far more important, however, in its impact on the homeless population has been the long-term change in the national policy for dealing with the mentally ill. A little more than 20 years ago state and county mental institutions began releasing large numbers of patients, many of whom suffered from severe illnesses. The "deinstitutionalization" movement followed the widespread introduction in the 1950's of psychoactive drugs, which seemed to offer the possibility of rehabilitating psychotic people within a community setting, under better living conditions and with greater respect for their civil rights. It was also thought the "community mental health" approach would be cheaper than operating large state hospitals. The movement was launched in 1963 when Congress passed a law promising Federal funding for the construction of community mental health centers.

Deinstitutionalization was a well-intentioned and perhaps even enlightened reform, but it has not proceeded according to the original plan. The first step has been accomplished: the patient population at state and county mental hospitals

has swelled over the past few years governments and private groups have responded by providing emergency shelters. Although the food and refuge they offer save lives, most of these shelters offer no psychiatric care and thus do not meet the needs of many of the homeless.

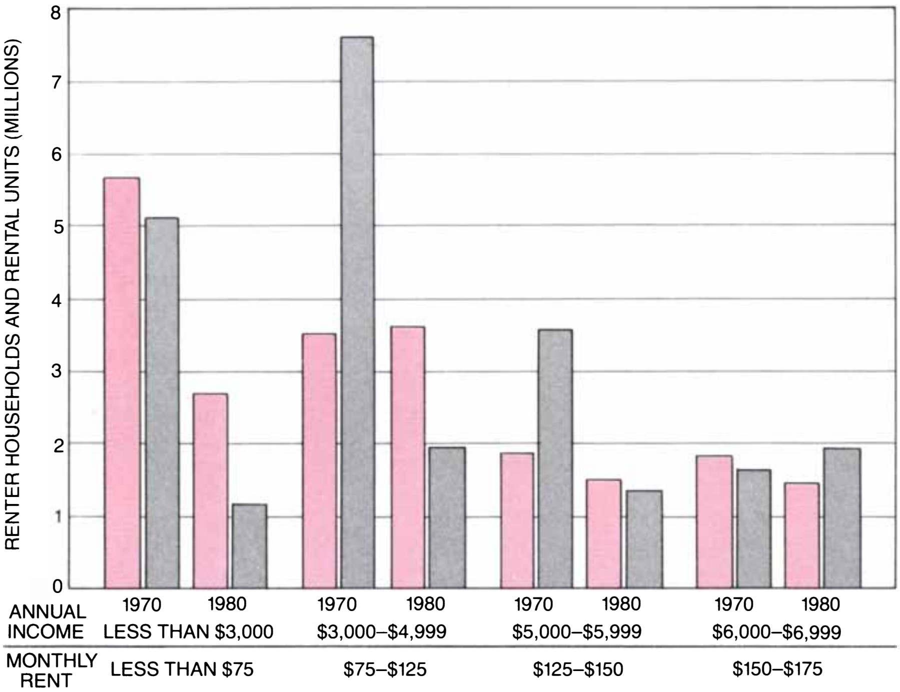
is now less than one-fourth of its 1955 peak level of 559,000. By and large, however, the various levels of government have not taken the second step: they have not provided enough places, such as halfway houses or group homes, for discharged patients to go. Other factors contributing to the problems of the system include the fact that fewer than half of the community mental health centers needed to cover the entire U.S. population have been built; moreover, existing centers often do not coordinate their activities with those of the institutions that are discharging their patients.

The inadequacy of the care available to deinstitutionalized patients is suggested by the large increases since the early 1960's in the rate of admissions to state mental hospitals and by the fact that a growing majority of admitted patients have been hospitalized before. The drop in the resident population of the institutions is accounted for by shorter average stays. Younger ill people who might have been institutionalized 15 years ago now receive only brief and episodic care; one major reason is that the courts have decided only those among the mentally ill who are dangerous to themselves or to others may be committed involuntarily. In the absence of alternatives to the institutions, respect for the civil rights of the disturbed sometimes conflicts with the goal of providing them with humane treatment and asylum. Chronically disturbed people are sent out into the community, often to empty lives in single-room-occupancy hotels and Skid Row rooming houses. With the growing unavailability of even these housing options many of the people end up on the streets.

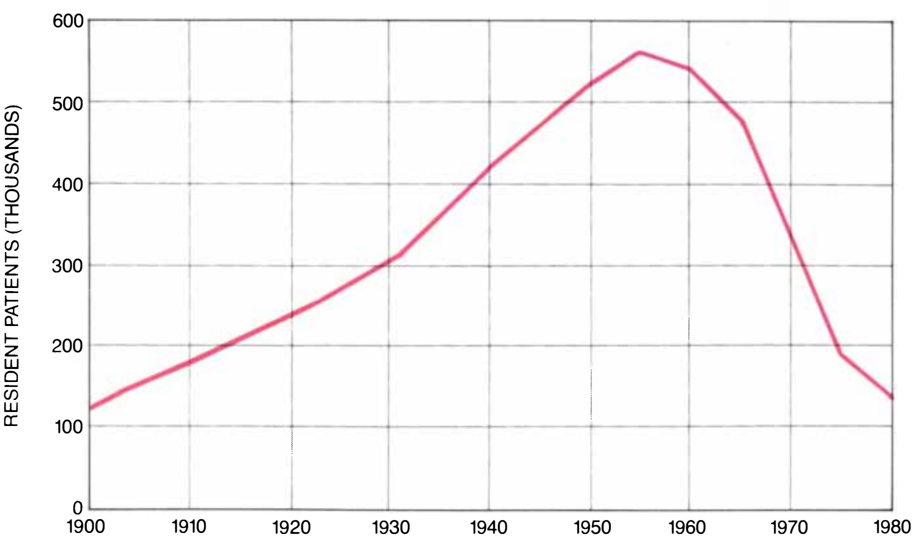
Thus it should not be surprising to find that a significant fraction of shelter residents are mentally ill. In fact, a clinical study I designed and implemented last year found at a shelter in Boston a 90 percent incidence of diagnosable mental illness: psychoses, chronic alcoholism and character disorders. The shelter selected for the study, which was under the direction of Alison Lauriat of the Massachusetts Association for Mental Health and Paul McGerigle of the United Community Planning Corporation, was considered demographically representative of Boston-area shelters.

The demographic data are themselves interesting. Men outnumbered women by four to one, although the number of women at Boston shelters seems to be increasing. The median age was 34 and apparently decreasing. One-third of the guests were either recent arrivals or only occasional users of the shelter, whereas the other two-thirds had been staying in shelters for more than six months. Some 20 percent had been on the streets and in shelters for more than two years.

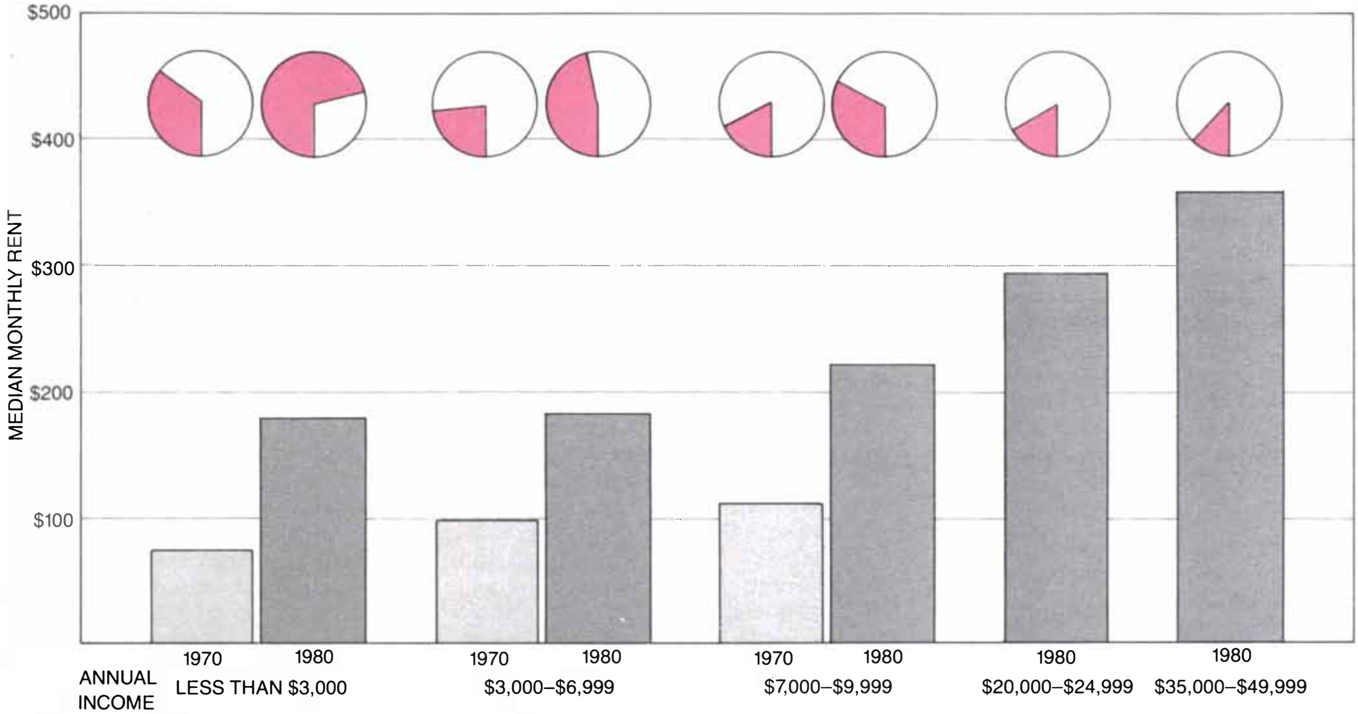
My colleagues (eight psychiatrists, psychologists and social workers) and I interviewed 78 guests at the shelter over the course of five nights. We diagnosed 40 percent as suffering from some form of psychosis: a generic term for major mental illnesses whose victims have difficulty distinguishing external reality from their own thoughts and feelings. The psychoses include some manic and depressive states and some organic brain syndromes, but most of the psychotics at the shelter were schizophrenic. Often subject to delusions and



SHORTAGE OF AFFORDABLE HOUSING for renter households with low incomes intensified in the 1970's and probably contributes to the growing homelessness problem. Colored bars indicate the number of renter households in a particular income range; gray bars indicate the number of rental units available in a rent range equal to 30 percent of the income range. As a result of inflation the total number of renter households with incomes below $7,000 fell from 13 million to about 9.2 million between 1970 and 1980, but the number of units available at rents below $175 (including utilities) dropped much more, from 17.9 million to 6.4 million. Many households in low-income groups that face an affordable housing gap must spend more than 30 percent of their income on rent. This illustration and the one on the opposite page are based on estimates provided by the Low Income Housing Information Service, which analyzed data from the census and from the Federal Government's Annual Housing Survey.



DEINSTITUTIONALIZATION of the mentally ill since the late 1950's has reduced the inpatient population at state and county mental hospitals. Because alternative residences and treatment programs have not been provided in many communities, many former patients, as well as some younger disturbed people who were never hospitalized, may now be homeless.



MEDIAN RENT paid by low-income households increased between 1970 and 1980, as did the fraction of income they devoted to rent (*colored wedges*). **The 1980 figures indicate that the households with higher income generally spend a smaller fraction of it on housing.**

hallucinations, they have trouble coping with the demands of daily life.

A 42-year-old man, at one time a talented artist, is an extreme example. When he was 24, he killed his wife with a baseball bat because she had been unfaithful to him. At the time he believed he was Raskolnikoff, the protagonist in Dostoevski's *Crime and Punishment.* The court psychiatrist diagnosed him as schizophrenic, and he was hospitalized in an institution for the criminally insane for the next 16 years. Since being discharged more than two years ago, he has lived both in shelters and on the streets; not long before we saw him he had been arrested for trespassing in a cemetery, where he was living in a tomb he had hollowed out. He says he receives messages from spirits who speak to him through spiders.

The story of an 18-year-old shelter guest is less striking but no less tragic. Until he became psychotic he was enrolled in an Ivy League college. He was hospitalized briefly in a state institution, where he was given antipsychotic medication, but when we saw him, he was receiving no treatment. For a while after his discharge his mother cared for him; eventually, however, she became too depressed to continue. Frightened and too confused to care for himself, he now wanders the streets by day, muttering incoherently and responding to voices he alone hears. At night he goes to a shelter where the staff are too busy feeding and clothing people to devote themselves to individual problems.

Many of the people we interviewed—we estimated 29 percent—were chronic alcoholics. One 33-year-old man had lived on the streets of Boston for 20 years and like many homeless alcoholics had been in and out of hospitals, detoxification centers and various treatment programs. In the past year he had made several suicide attempts, and he had recently been treated for pulmonary tuberculosis. (About 45 percent of the study group reported serious physical problems, including heart disease and cancer, in addition to their psychological difficulties.) Finally, about 21 percent suffered from personality disorders that made it hard for them to form and maintain relationships or to hold a job.

Chronic mental illness, even when it is severe enough to impair the ability to function in society, does not by itself cause homelessness, any more than unemployment does. For the great majority of shelter guests lack of a home is symptomatic of total disconnection from supportive people and institutions. Consider for a moment what would happen if a crisis were to strike your life—if you were to lose your job, say, or contract a serious illness. Most likely you are surrounded by family and friends, by co-workers and even by professional caretakers at various social agencies whose help you could call on to prevent a downward slide. You are insured, both in the literal sense of having coverage against financial loss and in the figurative sense of having a reliable support network.

To talk with homeless people is to be struck by how alone most of them are. The isolation is most severe for the mentally ill. Family and friends grow exhausted or lack the ability to help; overburdened social workers may be less responsive; the homeless themselves may be unwilling or unable to communicate their needs and to make use of the support available. Some 74 percent of the shelter residents we interviewed said they had no family relationships, and 73 percent said they had no friends, even within the shelter community. Those who had been hospitalized before for psychiatric reasons (about one-third of the group) reported even less social contact: more than 90 percent of them had neither friends nor family. About 40 percent of all the guests said they had no relationship with anyone or with any social institution; although only 6 percent worked steadily, only 22 percent received any financial assistance.

There is usually no single, simple reason for an individual's becoming homeless; rather, homelessness is often the final stage in a lifelong series of crises and missed opportunities, the culmination of a gradual disengagement from supportive relationships and institutions. A final example illustrates the point. A 45-year-old man whom I shall call Johnny M. has lived on the streets and in the shelters of Boston for four years. The youngest of four siblings in a lower-middle-class family, Johnny spent most of his adolescent years in an institution for the mentally retarded. He remembers washing dishes, going to classes and looking forward to the visits of his mother and older sister. When he turned 16, he moved back home and spent time watching television and puttering in the garden. Ten years later his older sister died suddenly and Johnny had a "nervous breakdown." He became terrified of dying, he cried constantly and his thoughts became confused. Because he was unable to care for himself, he was involuntarily committed to a state hospital, where he remained for the next eight years. He became very attached to a social worker whom he saw twice a week for therapy.

Although the hospital had become Johnny's home, he was discharged at the height of deinstitutionalization into a single-room-occupancy hotel. His father had died, his mother was in a nursing home and neither his remaining sister nor his brother could afford to support him. Within six months he had lost contact with the hospital. Johnny was forced out of the hotel when it was converted into condominiums; unable to find a room he could afford, he roamed the streets for several months until an elderly woman and her daughter took him into their rooming house.

When the daughter died unexpectedly of a stroke, Johnny became depressed, thought the other residents were trying to harm him and grew increasingly belligerent. His landlady evicted him. Without resources or supports and with an incipient psychosis, he ended up homeless. Resigned to street life, he now spends his days walking endlessly, foraging in dumpsters. Occasionally he collects bottles, sells his blood for transfusion or takes part in medical experiments to make pocket money. Itching from lice, wearing tattered clothes and suffering from cellulitis of one leg, he feels lucky that he can depend on an evening meal at the shelter and that on most nights he has access to a bed.

Shelters help to keep Johnny M. and his companions in misfortune alive. That is a shelter's function: to provide food, clothing and a bed. At a typical shelter guests line up outside until the doors open in the early evening. A security guard checks each person for alcohol, drugs and weapons. New guests are also checked for lice. At some shelters volunteers cook hot meals; at others dinner consists of soup, sandwiches and coffee. Some guests spend the evening socializing and playing cards, but most are too weary or too detached and go directly to sleep. The dormitory is typically a barren auditorium-size room with rows of cots or beds and one or two cribs. Sometimes groups of six or more beds are separated by partitions. Shelter guests usually have few opportunities to wash during the day, and so at night the bathrooms at the shelter are generally overcrowded. By 10:00 P.M. the lights are turned out, and the next morning the guests are awakened early, given coffee and a doughnut and sent out, even if the temperature is below zero.

The atmosphere in a shelter is sometimes volatile, and occasionally violent fights erupt that have to be broken up by the staff or the police. On the other hand, the anonymity and invisibility fostered by shelters is comforting to many of the guests, who spend their days as highly visible social outcasts. Shelter providers try to treat their guests with dignity and respect, asking no questions and attaching no strings to the help they offer.

Do they offer enough? In my view they do not. Shelters would be the appropriate solution if the homeless were simply the victims of unemployment, or of disasters such as floods or fires. Although these factors undoubtedly contribute to the problem, the overriding fact about the homeless is that most are mentally disabled and isolated from the support that might help to reintegrate them into society. Moreover, many are chronically, permanently ill and will never be able to live independently.

Although various innovative model programs exist, including one sponsored by St. Vincent's Hospital in New York City, shelters as a rule offer only minimal medical, psychological and social services. They are generally understaffed and have few personnel specifically trained to care for the severely disabled. Because they are open only

JOHNNY M. (a pseudonym) was one of the shelter guests interviewed by the author. He is seen here in line waiting to check into a Boston shelter (*left*) and waiting for dinner to start at the shelter on Easter Sunday (*right*). Treated in an institution for the mentally re-

at night, they cannot offer the continuing support and supervision that many chronically ill people need. People whose condition might improve with properly supervised treatment (for example the 18-year-old student I mentioned above) do not get it at the shelters. And it hardly needs saying that shelters are not a humane solution to the problem of providing a place to live for those who suffer from permanent mental disabilities.

The precise extent to which mental illnesses are prevalent among the homeless remains a matter of controversy. Recent clinical studies at shelters in Los Angeles, New York and Philadelphia support my contention that a majority of the homeless suffer from psychiatric disorders, but other estimates have put the incidence of mental illness among shelter populations as low as 20 percent. All these studies, including our own, have been largely descriptive and have been plagued by methodological problems. Differences in results can be attributed to the different theoretical biases of the various investigators, to the use of different standardized scales as the basis for psychiatric evaluation and most of all to the difficulty of obtaining a representative sample of a constantly shifting population. In addition, there is no reason to expect the characteristics of the homeless population to be constant throughout the country when mental health policies and economic conditions vary regionally.

The public debate on homelessness would undoubtedly be enlightened by more rigorous research into the causes of the problem. It can already be said, however, that at the very least a significant fraction of the people who frequent shelters have diagnosable mental disturbances. Public servants of all ideologies have failed to recognize the implications of this fact. Many political conservatives seem to believe the Government has little obligation to care for the homeless; this attitude is perhaps best exemplified by President Reagan's often quoted remark that "the homeless are homeless, you might say, by choice." For political liberals the plight of the homeless serves as ammunition in their attack on the Administration's economic policies, but the solution they tend to support is the expansion of emergency shelters: simply putting a temporary dressing on what has become a large, festering wound in the social body.

There is no mystery about the nature of a more appropriate solution. Essentially it would call for carrying out the aborted plans of the 1963 community mental health law by providing a spectrum of housing options and related health-care and social services for the mentally ill. These would entail living arrangements with varying degrees of supervision, from 24-hour care at therapeutic residences for patients with severe psychoses to more independent living at halfway houses for patients with less severe disorders. Some patients would receive counseling and therapy with the goal of rehabilitating them and even getting them jobs in the community. The one major change needed in the community mental health program, however, is a greater recognition of the limitations of psychiatry: given the current state of the art many chronically disturbed people simply cannot be rehabilitated, and the goal in these cases would be to provide the patient with comfortable and friendly asylum.

The community mental health movement failed primarily because the Federal and state governments never allocated the money needed to fulfill its promise. American society is currently trying to solve the problem cheaply, giving the mentally ill homeless at best emergency refuge and at worst no refuge at all. The question raised by the increasing number of homeless people is a very basic one: Are Americans willing to consign a broad class of disabled people to a life of degradation, or will they make the commitment to give such people the care they need? In a civilized society the answer should be clear.

tarded as a child, Johnny M. was later committed to a mental hospital when he had a nervous breakdown. He was discharged at the height of deinstitutionalization and lost his room when the building was converted into condominiums. He has lived on the streets for four years.

Undersea Volcanoes

New techniques for exploring the ocean floor are yielding a remarkably detailed picture of the volcanic processes responsible for generating the bulk of the earth's crust

by Roger Hékinian

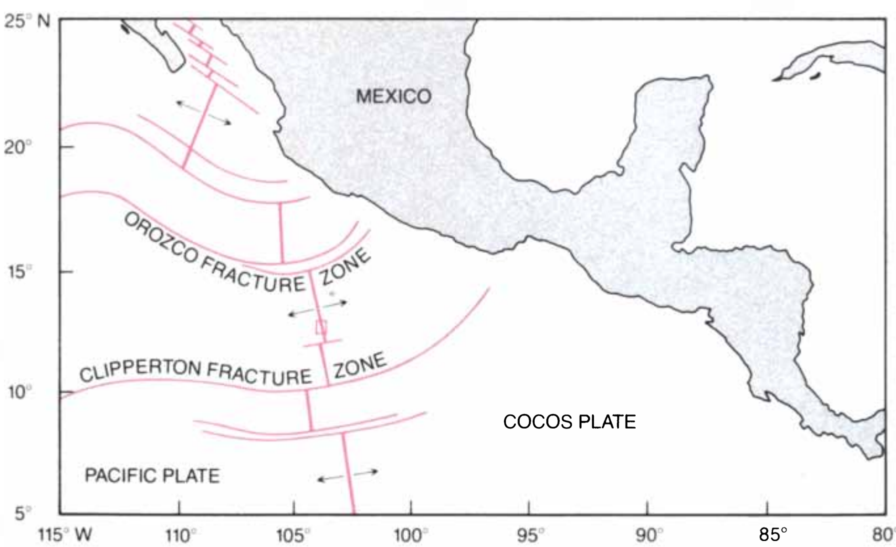
The volcanic activity that continually remakes the surface of the earth takes place almost entirely at the bottom of the ocean, far beyond the reach of the traditional investigative tools of the volcanologist. Magma, or molten rock, wells up from the mantle and spews out onto the ocean floor in two characteristic localities: along the active midocean ridges, where the spreading tectonic plates grow by the steady accretion of solidifying mantle material, and at the isolated volcanic structures called seamounts, which are typically strung out in chains across the interior of the plates.

It is only in the past decade or so that advances in the technology of undersea exploration have opened the way to a detailed study of the volcanic processes involved in the creation of fresh oceanic crust at these sites. The new methods of submarine volcanology range from remote sensing by instruments towed from surface ships to direct "field" observations made with the aid of manned submersibles. The picture that is emerging from this international research effort is full of surprises. The new findings promise to lead not only to a better understanding of some of the earth's most remarkable features but also to the discovery of valuable mineral resources and to their ultimate exploitation.

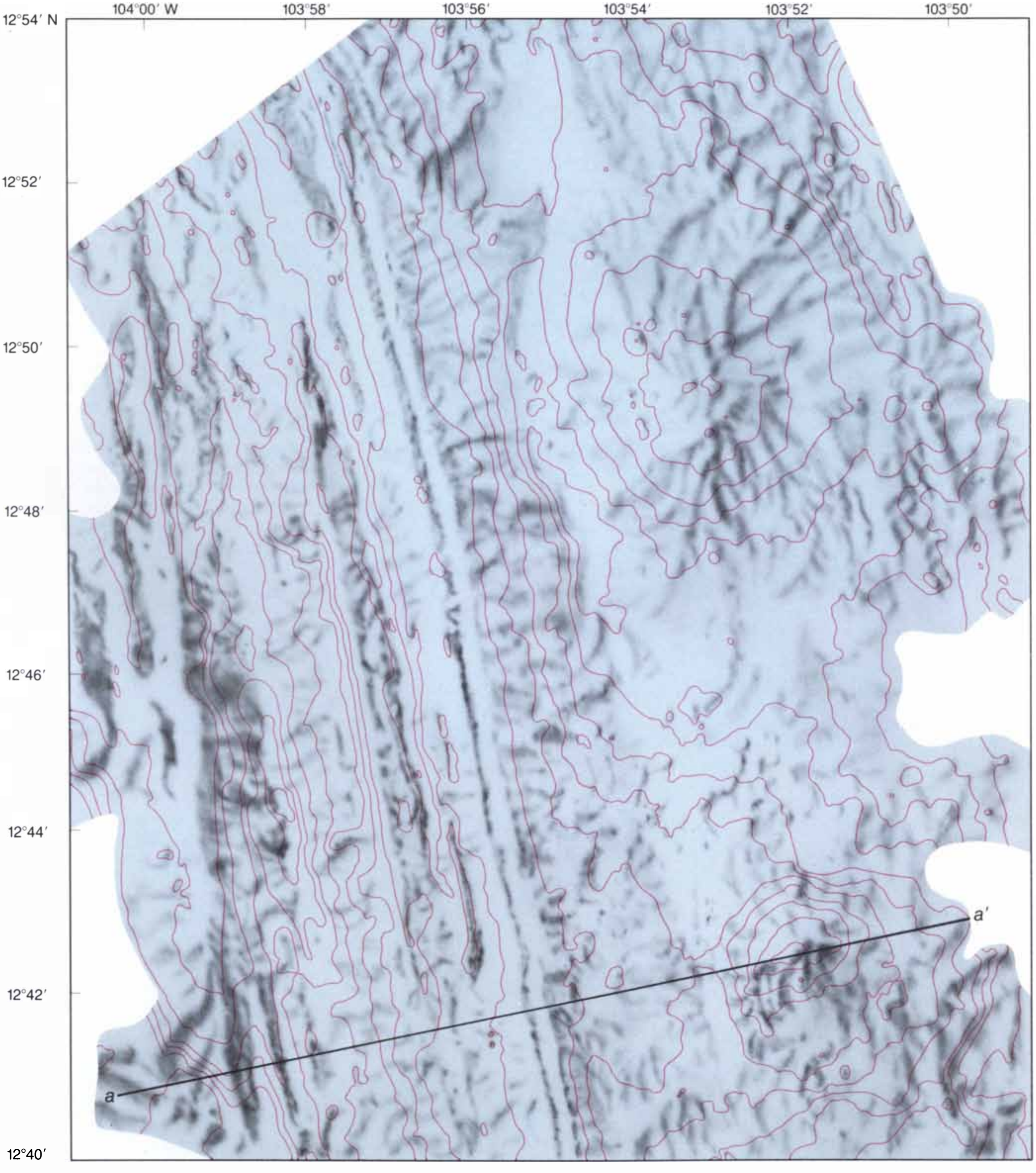
The products of volcanism most commonly found at the earth's surface are basalts: igneous rocks composed principally of silicates of iron, magnesium, aluminum and calcium. It has long been suspected, largely on the basis of seismic evidence, that the basaltic magma from which these rocks are formed originates in a zone of partial melting in the earth's upper mantle, more than 100 kilometers below the surface. The semimolten rock at this depth is less dense than the surrounding mantle material, and so it tends to rise slowly toward the surface in the form of giant blobs called diapirs. As a diapir ascends, the pressure on it decreases and consequently more mantle material becomes molten. The rising diapir contributes to the formation of the comparatively shallow magma reservoirs or feeder columns that are the immediate source of volcanic activity.

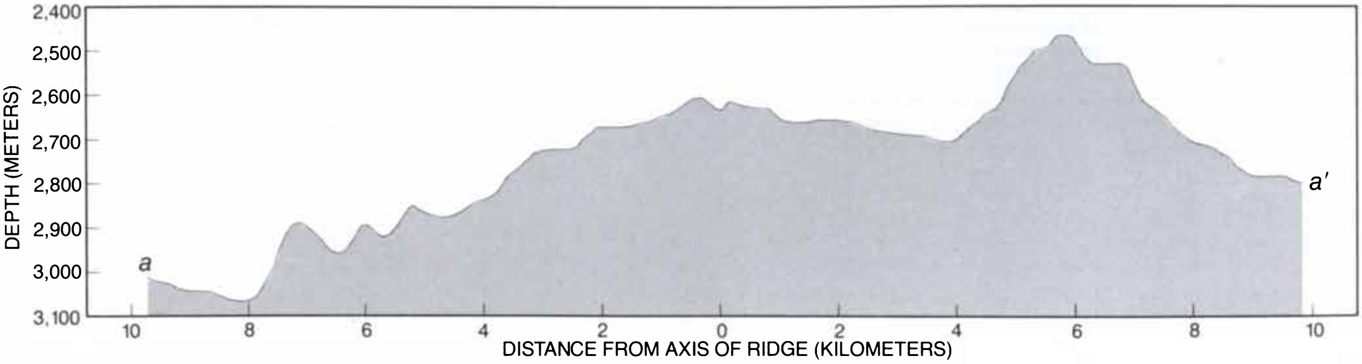
The magma chambers closest to the surface are those under the midocean ridges, where the thickness of the crust may be only 10 kilometers or less. If the molten rock is not extruded immediately, it slowly cools while it is still trapped by the crust. As the temperature falls, high-density minerals solidify first and thereby change the composition of the residual melt. This process of differentiation, or fractional crystallization, leads to the creation of heterogeneous crustal material. The high-density minerals, known collectively as peridotites, sink to the bottom of the reservoir, whereas the lower-density ones, called gabbros, line the walls and roof of the magma chamber. Some of the magma that is not extruded onto the ocean floor solidifies within the conduits above the magma chamber and forms the characteristic elongated structures known as dikes.

As long as the magma remains in the reservoir its composition continues to change. The ongoing process of fractional crystallization eventually results in the formation of more evolved products, that is, rocks that are extremely



SEGMENT OF A FAST-SPREADING MIDOCEAN RIDGE is depicted in unprecedented detail in the relief map on the opposite page. The map is based on high-resolution depth measurements made with a deep-towed multibeam sonar system called SeaBeam. The prominent diagonal structure running from the upper left to the lower right is the crest of the East Pacific Rise, an active volcanic zone where upwelling magma from the earth's mantle oozes onto the sea floor through fissures in the crust. The fissures are created by the separation of two tectonic plates: the Pacific plate (*left*) and the Cocos plate (*right*). The two isolated volcanoes on the eastern flank of the ridge were originally formed at the axis of the ridge and became detached from it in the course of sea-floor spreading. The 30-kilometer-long ridge segment lies between two major transform faults, the Orozco Fracture Zone and the Clipperton Fracture Zone, which cut across the East Pacific Rise more than 1,000 kilometers off the coast of Mexico (*see large-scale map above*). The data were obtained over a three-day period during the 1981 cruise of the French research vessel *Jean Charcot*. The site was subsequently explored at close range by the author and his colleagues in a series of dives in the manned submersible *Cyana*. The contours in the relief map are at 50-meter intervals. The vertical scale is exaggerated in the profile at the bottom, which corresponds to the line labeled *a—a'* on the relief map.


104°00′ W
103°58′
103°56′
103°54′
103°52′
103°50′
12°54′ N
12°52′
12°50′
12°48′
12°46′
12°44′
12°42′
12°40′
a
a′


DEPTH (METERS)
2,400
2,500
2,600
2,700
2,800
2,900
3,000
3,100
a
a′
10
8
6
4
2
0
2
4
6
8
10
DISTANCE FROM AXIS OF RIDGE (KILOMETERS)

differentiated from the original composition of the melt. The solidification of the high-density minerals removes these components from the melt, leaving behind a fluid with an increased concentration of low-density minerals and condensed gases. Most of the basalts that formed after magmatic differentiation and subsequently erupted onto the ocean floor are comparatively acidic rocks classified as andesites and trachytes. The large quantities of various gases (chiefly carbon dioxide, water and hydrogen sulfide) trapped in the melt after differentiation are responsible for the violent volcanic explosions occasionally witnessed in shallow marine environments as well as on land.

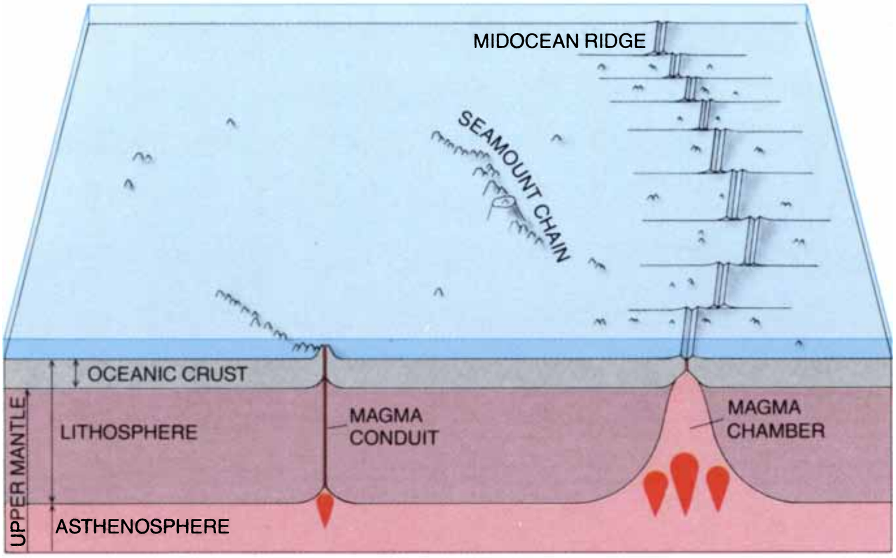
An important observation regarding the origin of the rocks produced by undersea volcanoes was made in 1963 by Albert E. J. Engel and Celeste G. Engel of the U.S. Geological Survey. They reported that the basalts found at the summit of volcanic islands and seamounts scattered across the Pacific basin tend to have a higher alkali content than the basalts recovered from the crest of the midocean ridges. Presumably the more alkaline basaltic magmas were formed at greater depth in the mantle and hence were subjected to less melting than those formed in the shallow reservoirs underlying the midocean ridges.

Two types of volcanic eruption are evident in the vicinity of the midocean ridges. In one type the magma oozes onto the ocean floor in the form of lava through long fissures. In the other type the magma rises in a central conduit, and the ensuing radial lava flows tend to build a conical structure.

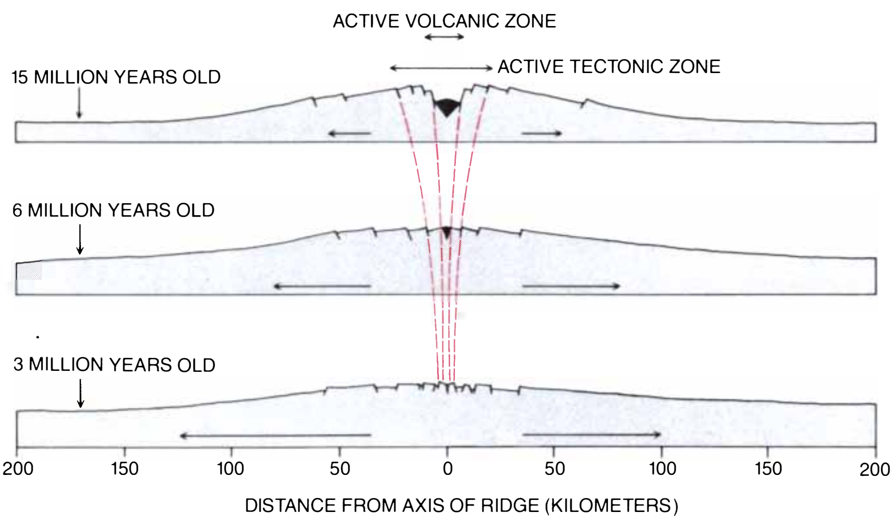
Fissure-type eruptions usually occur at a boundary between tectonic plates, where the brittle crust is split apart by the separation of the plates. Magma welling up along the entire length of a fissure tends to form large lava pools similar to those formed on land by "shield" volcanoes, such as the ones in the Hawaiian Islands; a volcano of this type generally has gentle slopes and a large crater flooded with lava.

The two main lava formations observed under such conditions are termed sheet flows and pillow flows. It was recently noted by Robert D. Ballard of the Woods Hole Oceanographic Institution that sheet flows are more prevalent in the active volcanic zone of fast-spreading ridge segments, such as those parts of the East Pacific Rise off the coast of Mexico and near the Galápagos Islands that have been intensively explored in the past few years by several joint U.S., French and Mexican expeditions. The sheet flows at these sites consist of flat slabs less than 20 centimeters thick and lobed forms with smooth surfaces. The flows resemble the pahoehoe lava associated with the Hawaiian eruptions. Viewed in cross section such formations are often seen to be hollow and to have a layered structure parallel to the surface.

In contrast, pillow lava typically exhibits a radial jointing pattern in cross section, caused by the contraction of the rock during cooling. Pillow lavas are often found in the active volcanic zones of slow-spreading ridges such as the one that runs down the middle of the Atlantic Ocean. There the pillow lavas typically form small hills that tend to be elongated in the downslope direction. The surface of the pillows often has corrugations, or small ridges transverse to the direction of flow, that are thought to correspond to successive cooling stages during which the melt continued to flow intermittently.



TWO PLACES ON THE OCEAN FLOOR where volcanic eruptions take place are shown in this idealized view. Basaltic magma originating in the earth's upper mantle, more than 100 kilometers below the surface, rises slowly in the form of giant blobs called diapirs, which supply both the comparatively shallow magma chambers underlying active midocean ridges (*right*) and the feeder columns responsible for isolated volcanic structures known as seamounts (*left*).



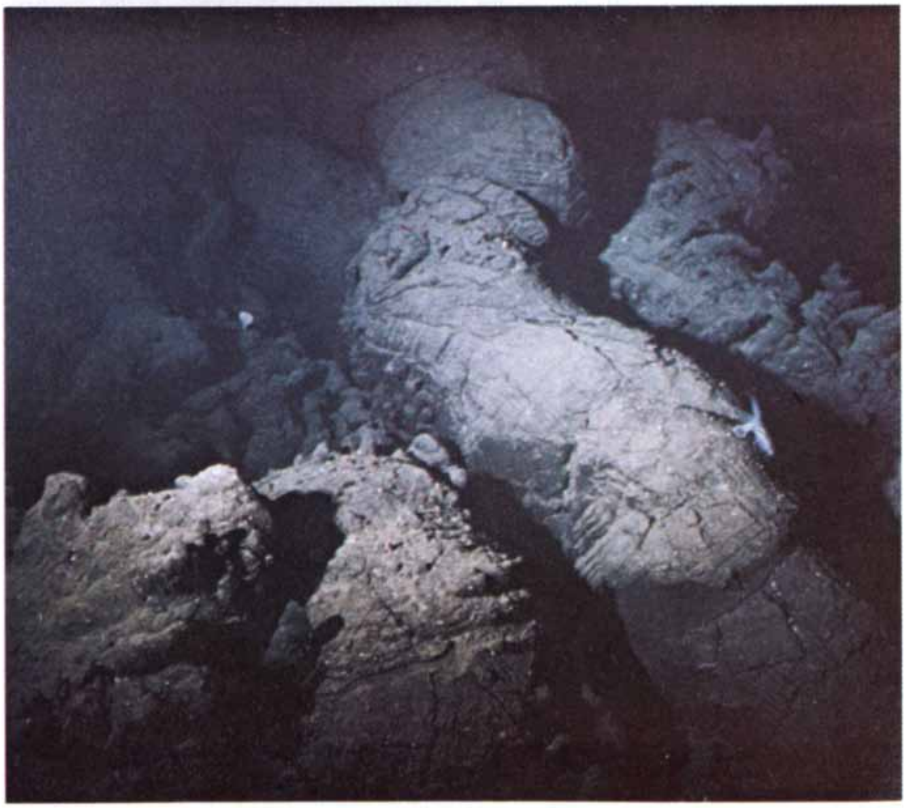
RATE OF SPREADING of a midocean ridge is reflected both in its topography and in the width of its active volcanic and tectonic zones. This schematic representation of three midocean ridges with different spreading rates is based on a study done by the author in collaboration with two of his colleagues, Jean Francheteau of the Institut de Physique du Globe of the University of Paris and Pierre Choukroune of the University of Rennes. The three profiles correspond to a slow-spreading segment of the Mid-Atlantic Ridge (*top*) and to two segments of the East Pacific Rise, one an intermediate-spreading segment (*middle*) and the other a fast-spreading segment (*bottom*). The actual spreading rates at the three sites are estimated to be respectively less than three centimeters per year, between three and seven centimeters per year and more than seven centimeters per year. The ridges respond to the stretching of the crust at the axis in two ways, depending on how a state of isostatic equilibrium is attained: through widespread fissuring near the axis of the fast-spreading ridge and through large-scale movements of the crust along major faults facing the axis of the intermediate- and slow-spreading ridges.

By comparing the undersea lava flows with those studied in Hawaii, Ballard has shown that the undersea sheet formations result from the eruption of a more fluid lava than that responsible for the pillow formations. The fluidity of the lava depends on the degree of crystallinity of the melt: the more solids carried along by the melt, the less fluid the lava. Both sheet flows and pillow flows have a glassy crust formed by the quenching of the hot lava when it comes in contact with the seawater. The crust is thicker on the sheet flows than it is on the pillow flows because the sheet flows have fewer crystals to impede the formation of the glassy layer. Ballard's observations also suggest that pillow flows could be a later eruptive stage than sheet flows. Thus a sheet flow produced during a widespread eruption of hot, highly fluid lava could be obscured by subsequent volcanic flows, giving rise to an overlying pillow formation.

The lavas emitted in sheet flows and pillow flows have no known compositional differences, which suggests they must have a common origin. The differences in their outward appearance and internal structure seem to be the result of their different eruption mechanisms. The most recently active volcanic zone along the well-explored part of the East Pacific Rise is made up almost entirely of sheet flows. Older volcanic events represented by hilly topographic features at the margins of such recent volcanic zones have given rise predominantly to pillow flows with only scattered remnants of sheet flows. The fact that the slow-spreading segments of the Mid-Atlantic Ridge have a higher proportion of pillow flows may be attributable to two causes. Because of the lower rate of spreading there, the upwelling of magma is slower and so less magma is available for extrusion; as a result the eruptions are less widespread. Alternatively, an initial widespread flow of hot, fluid sheet lava may have been covered by subsequent pillow flows.

A typical feature of terrain invaded by very fluid lava is the presence of collapsed structures. This phenomenon was first observed directly on the ocean floor in 1978 by Jean Francheteau of the Institut de Physique du Globe of the University of Paris in the course of a dive in the French research submarine *Cyana*. The features seen by Francheteau and others vary in size from pits a few centimeters in diameter to lava ponds as much as 50 meters across and 20 meters deep. The view from the edge of one of the larger collapsed structures includes a complex assemblage of cave-like recesses and slabs of rubble strewn on the ocean floor. Such scenes are thought to be the result of several eruptions that left a succession of empty conduits where the lava drained away along

FISSURE IN THE OCEAN FLOOR at the crest of the East Pacific Rise near 13 degrees north latitude is caused by the rapid spreading of the two tectonic plates in this region. It is through such fissures that volcanic eruptions take place. The color photographs accompanying this article were made by members of the French expedition while diving in the *Cyana*.



PILLOW LAVA, usually associated with less fluid volcanic eruptions, frequently takes the shape of elongated tubes through which the molten rock is channeled across the ocean floor.

a slightly inclined plane; the thin roofs of the empty conduits later collapsed. The overhanging walls and edges of the lobed conduits are often perceived as pillars, some of which are hollow, suggesting that seawater trapped between two large lava conduits forced its way to the surface, leaving chilled pathways in its wake. Alternatively seawater may have been trapped under a lava conduit; after being heated the water may have forced its way up through the molten lava, creating its own vertical pathways, the remnants of which are seen today as pillars.

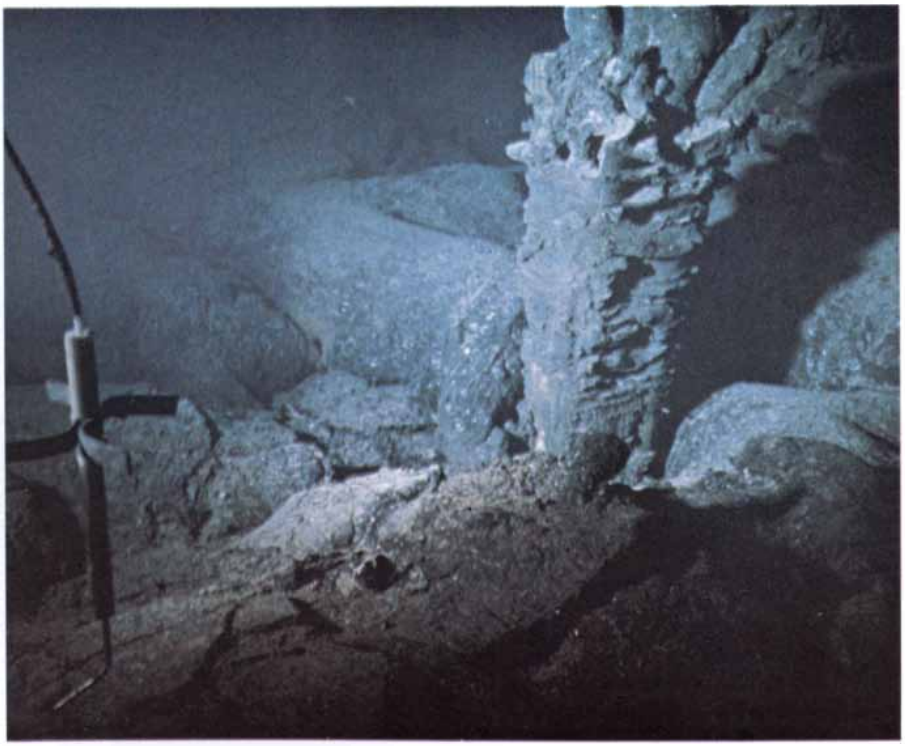
The more centralized type of volcanic eruption may or may not be associated with a widespread fissuring of the oceanic crust. Such an event takes place when the upwelling magma is concentrated in comparatively narrow conduits that lead to the main feeder columns of elevated volcanic structures. Some seamounts, for example, may be associated with extended fissures along which magma welled up through a main conduit, piling successive lava flows on one another. (The term seamount has been broadly defined—without reference to the question of origin—by Henry W. Menard, Jr., of the Scripps Institution of Oceanography as any isolated elevation of the ocean floor with a circular or elliptical horizontal cross section and a slope ranging between five and 35 degrees of arc with respect to the surrounding terrain.)

The summit of a seamount sometimes has a crater, or depressed area, within which lava is extruded. When a crater is larger than about two kilometers in diameter, it is considered a caldera. Depressions of this type vary between 50 and 300 meters in depth and are usually bounded by vertical fault zones. The depressions are formed when the magma reservoir becomes empty and there is no longer enough support to prevent the collapse of the top of the volcanic cone. Thomas E. Simkin of the Smithsonian Institution has suggested that undersea eruptions could take place from feeder vents along the vertical faults on the circumference of such a caldera. The lava flows from these "ring conduits" would eventually fill the caldera and flatten the top of the volcano. The flat-topped seamounts known as guyots, which are commonly found far from the midocean ridges, may have been formed in this way. Other undersea volcanoes do not have a collapsed caldera; instead the summit consists of a number of isolated volcanic peaks ranging in height from about 20 to 100 meters and usually having a slope steeper than that of the main volcano. (These generalizations are based primarily on submarine observations made near the axis of the East Pacific Rise in the vicinity of the Galápagos Islands.)

Sheet flows have often been observed as glassy slabs formed during the flow of very fluid lava that broke its crust during cooling. This type of flow is comparable to the "pressure ridge" phenomenon familiar in the case of the Hawaiian volcanoes. Other common undersea volcanic features are small lava domes made up of an accumulation of spiny basaltic glasses; the domes, which resemble the "spatter cones" formed during volcanic eruptions on land, were apparently formed when hot lava rose through cracks along the main path of an underlying lava flow.

The presence of superposed structures on the summit of many undersea volcanoes indicates that several eruptions contribute to the formation of a

HOLLOW LOBED SHAPES indicative of the eruption of a very fluid lava typically have a smooth top, a thick glassy crust and a layered structure parallel to the surface. In this case part of the volcanic formation collapsed after the lava drained away, revealing the inner cavity.



PILLARLIKE STRUCTURE (*right*) is thought to be the remnant of a collapsed lava pond that was partially buried by later lava flows. The object at the left is a temperature probe.

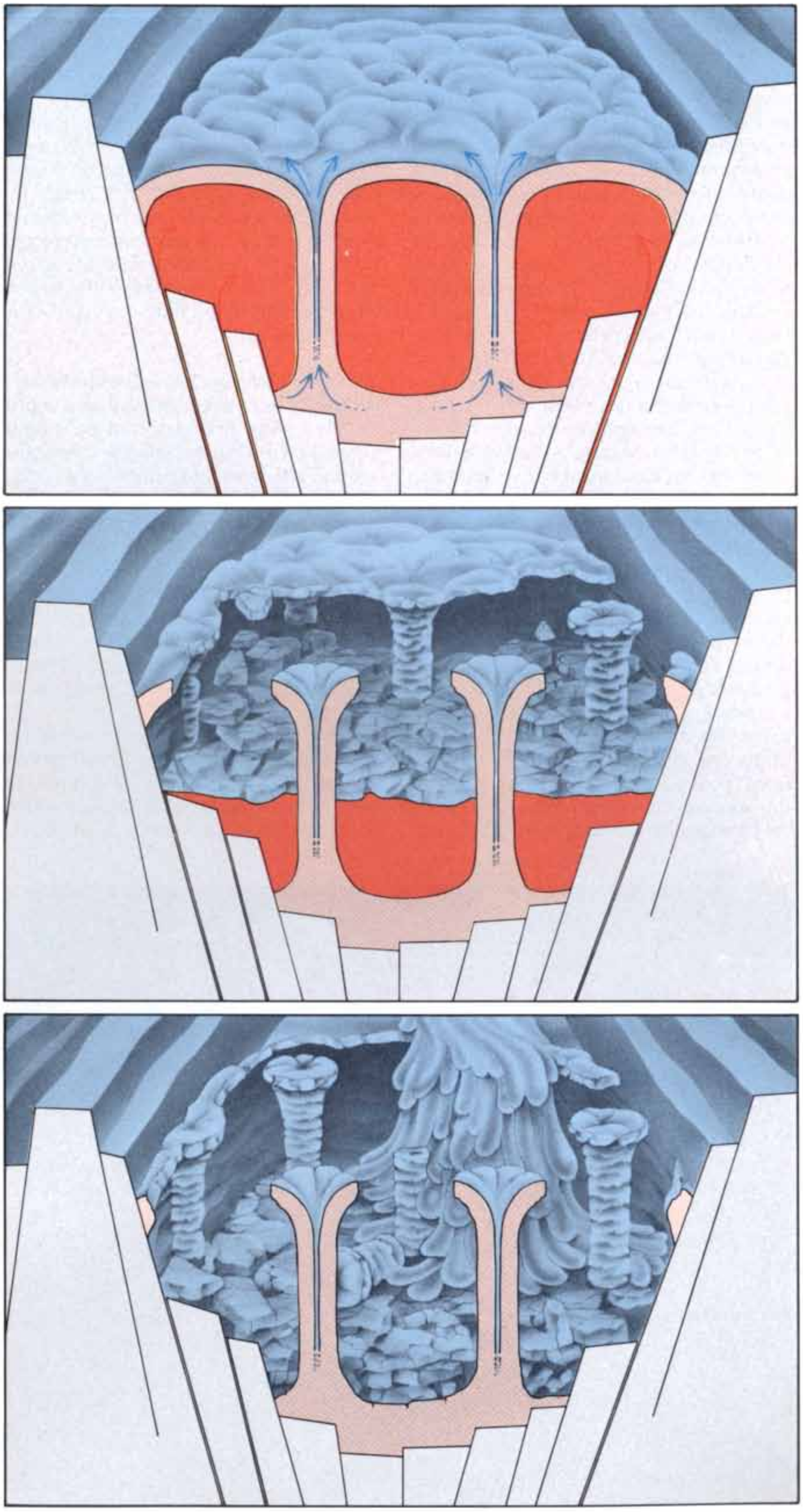
single volcano. A variety of lava flows can be found on seamounts, of which the most prevalent—sheet flows and pillow flows—are similar to those observed on the midocean ridges. Elongated tube-shaped pillows are also commonly found downslope from volcanic peaks, whereas sheet flows and large horizontal flows with slightly domed shapes tend to fill depressions. In addition pyroclastic flows, consisting of the fragmented debris of volcanic explosions at depths of less than 200 meters, are sometimes observed at the summit of seamounts.

Extraordinarily detailed topographic studies of the midocean ridges and other undersea volcanic structures have been made possible in recent years by devices such as the deep-towed multibeam sonar system called SeaBeam, a product of the General Instrument Corporation. SeaBeam generates a high-resolution contour map of the ocean floor along the track of the ship from which it is towed. The technique is capable of mapping large areas of the ocean floor in a comparatively short time. For example, during the 1981 cruise of the French research vessel *Jean Charcot* it was possible to map a 600-square-kilometer segment of the sea floor along the East Pacific Rise in just three days [*see illustration on page 47*].

Considerable progress has also been made in recent years in the use of manned submersibles, in particular the *Cyana* and the U.S. *Alvin,* to facilitate first-hand studies of the ocean floor at midocean ridges. The French group, of which I am a member, has relied on observations made with the aid of the *Cyana* to compare tectonic features of ridges with three different spreading rates: slow (less than three centimeters per year), medium (between three and seven centimeters per year) and fast (more than seven centimeters per year). We have visited ridges of all three types in the submersible in order to make direct visual observations of the tectonic and volcanic features associated with the process of sea-floor spreading under diverse conditions.

The active tectonic zone of a fast-spreading ridge is usually quite narrow: less than six kilometers across on the average. This fact may be attributable to the effect the heat from the shallow magma chamber has on the overlying crust. The axial region of a fast-spreading ridge tends to be warm and to have a "horst and graben" topography, in which the crust is broken into alternating high blocks (horsts) and low blocks (grabens) by repeated movements along vertical faults. Slow-spreading ridges tend to be cool and to be bounded by opposing pairs of steep, inward-facing scarps, or cliffs.

The difference can be attributed to the



PILLARS MAY HAVE BEEN FORMED when the highly fluid lava of a lava pond drained away, causing the thin roof of the formation to collapse and creating a large pit. According to this reconstruction of the possible sequence of events, the pillars, some of which stand as high as 15 meters, are the remains of chilled conduits through which trapped seawater escaped from under the lava. Formations of this kind were first observed in 1978 by Francheteau and his co-workers on the crest of the East Pacific Rise near 21 degrees north. The illustration is based on sketches made by Claude Rangin of the Pierre and Marie Curie University in Paris.

way each kind of ridge reaches a state of isostatic equilibrium, in which structures sink or float according to their density. Isostatic equilibrium is achieved everywhere across and along a fast-spreading ridge system through frequent fissuring of the warm crust, but a slow-spreading ridge does not initially come into equilibrium in this way. Instead equilibrium is restored near the axis of a slow-spreading ridge through sporadic, large-scale vertical motions along major faults. The presence of a central rift valley, a depressed region more than 1,000 meters deep on a slow-spreading ridge such as the one in the mid-Atlantic, is the result of a large-scale vertical adjustment of the crust caused by the need to restore isostatic equilibrium. Along a fast-spreading ridge there is no central rift valley since isostatic equilibrium is achieved initially in smaller increments.

The long fissures through which magma is extruded at a midocean ridge tend to be oriented at right angles to the direction of spreading. The magma under the fissures is channeled along the axis of the ridge, and when it is extruded through the cracked crust, it gives rise to a landscape with rather low relief. Volcanic structures formed on or near the midocean ridges can develop into isolated peaks as they move away from the axis of the ridge during sea-floor spreading. The oceanic crust cools and thickens as it leaves the accreting plate boundary. This process can influence the height attained by such volcanoes as they move away from the axis, because the thicker crust can support a greater mass on its surface. A volcano formed at the ridge cannot increase its mass, however, unless it continues to have a source of magma after it has left the upwelling part of the axial zone. Sometimes a volcano formed on a midocean ridge can even develop into an island, but only if it is supplied with magma from an underlying magma chamber or feeder column.

Isolated undersea volcanoes have been known to exist for more than a century; they were first detected by soundings taken during the H.M.S. *Challenger* around-the-world expedition of 1872–76. It was not until 1964, however, that Menard completed the first comprehensive survey of the Pacific seamounts. The fact that the Pacific seems to have a much higher density of seamounts than the Atlantic or Indian oceans is consistent with the view that the oceanic crust under the Pacific has had more volcanic activity than the crust under the other oceans.

The available bathymetric, or ocean-depth, data regarding the distribution of undersea volcanoes in the Pacific basin were recently analyzed by Rodey Batiza of Washington University in St. Louis. Basing his study on the geologic record of reversals in the earth's magnetic field, Batiza divided the Pacific into strips dating from the present (at the axis of the East Pacific Rise) to the Cretaceous period (between 65 and 136 million years ago). He showed that the number of volcanoes per unit area increases with increasing crustal age, reaching a peak in the Eocene epoch (between 37 and 55 million years ago). Batiza's analysis indicates that the volume of lava present in the form of seamounts also increases with the age of the crust; by this standard the volcanic activity was most pronounced early in the Cretaceous, more than 100 million years ago. This finding is in line with the observation that the tallest seamounts, standing more than four kilometers above the ocean floor, are in the western Pacific near the Philippine Trench, where the crust is more than 100 million years old.

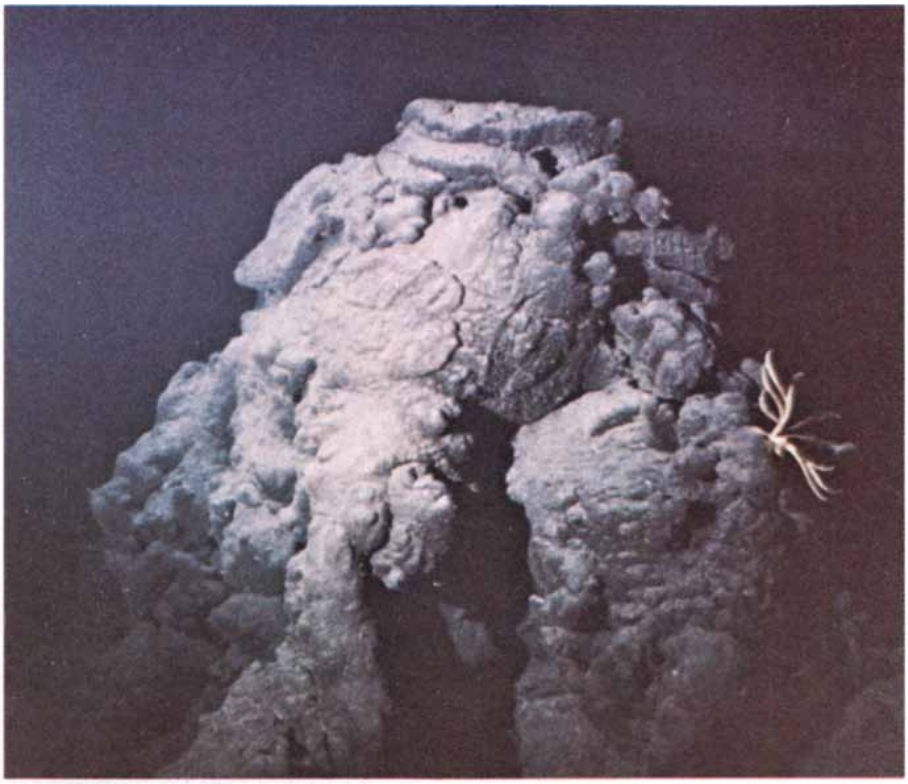
A large number of undersea volcanoes have been discovered in recent years with the aid of a side-looking sonar system named GLORIA, which was developed at the Institute of Oceanographic Sciences at Wormley in England. GLORIA is towed behind a research vessel quite near the surface; it provides extremely large-scale coverage of the ocean floor, surveying a swath between 36 and 50 kilometers wide in a single pass. In 1980 a group of British oceanographers led by Roger C. Searle,

SHEET LAVA was photographed from the *Cyana* at the summit of the Clipperton seamount 18 kilometers west of the axis of the East Pacific Rise. The lava formation is similar to the ones formed by the eruption of very fluid lava from "shield" volcanoes on land.

making only two passes with GLORIA in the southeastern Pacific between the East Pacific Rise and the coast of South America, discovered some 200 volcanoes with a diameter of more than one kilometer at the base.

Along the 3,300-kilometer route traversed on the two passes it was noted that the density of the volcanoes varied markedly with the age of the oceanic crust. A maximum density of more than 50 volcanoes per 10,000 square kilometers was found on a strip of crust dating from the Miocene epoch (between 6.5 and 26 million years ago) that extends between 700 and 1,700 kilometers from the axis of the East Pacific Rise. A somewhat smaller peak in the density distribution was observed about 200 kilometers from the ridge. Searle estimates that the average density of volcanoes in the South Pacific is about eight per 10,000 square kilometers. This is somewhat higher than Batiza's estimate for the North Pacific: between two and five volcanoes per 10,000 square kilometers.

The variation in the density of volcanoes from place to place suggests that in addition to the isolated volcanoes originating at the midocean ridge there are a number of truly intraplate volcanoes, that is, volcanoes arising within a tectonic plate itself instead of at the boundary between two plates. One of the mechanisms thought to be responsible for this kind of volcanism is based on the "hot spot" theory advanced by W. Jason Morgan of Princeton University. According to Morgan, intraplate volcanism can develop where an upwelling of hot mantle material rises to the surface as the tectonic plate is moving over a fixed zone of partial melting in the upper mantle.



SMALL HILL OF LAVA reminiscent of the "spatter cones" associated with the shield volcanoes of the Hawaiian Islands was presumably formed when hot lava rose through a crack in the roof of an underlying sheet flow. This structure is also on top of the Clipperton seamount.

It is difficult to distinguish a volcano formed in the interior of a tectonic plate from one formed at a midocean ridge because there is little detailed information for most undersea volcanoes. In order to discriminate between volcanoes built on the ridge and those created by intraplate volcanism, Anthony B. Watts and John H. Bodine of the Lamont-Doherty Geological Observatory have recently measured the effect of plate loading in the vicinity of islands and seamounts in the Pacific basin. The oceanic crust undergoes stress when a topographic load is added and so the outer layer of the earth must find a new equilibrium by moving up or down. In response to long-term loading the crust bends; thus when a volume of lava is added to a particular area of the oceanic crust, it deforms the crust. The redistribution of mass due to loading gives rise to a gravity anomaly, which can be measured by means of instruments carried on ships, airplanes or satellites. Based on an analysis of gravity-anomaly measurements Watts and N. M. Ribe of the University of Chicago have concluded that intraplate volcanism was very intense in the Cretaceous period; this of course corresponds to the time when the largest volume of lava erupted. Watts and Ribe also inferred from their bathymetric and gravity-anomaly data that volcanism due to midocean ridge activity has been more intense than intraplate volcanism on crust that is less than 100 million years old.

Peter F. Lonsdale and Fred N. Spiess of Scripps have proposed a model of the growth of a volcano as it moves away from the axis of a midocean ridge. They have suggested that a centrally fed volcano starts to develop at the ridge axis during central volcanism initiated above main feeder conduits supplied by magma from the upper mantle. As the crust spreads, the volcano moves away from the axial zone; when it has moved far enough (more than 15 kilometers), it is no longer over the main feeder column. In this situation a tall volcanic cone can develop only when the volcano is fed from a distinct magma source rising directly from the upper mantle. Lonsdale and Spiess have identified two symmetrical volcanic structures rising more than 1,000 meters above the ocean floor 35 kilometers on each side of the East Pacific Rise that they suggest were originally formed at the ridge axis and split during sea-floor spreading into two distinct structures.

Split seamounts more than 600 meters high were recently found right on the axis of the Juan de Fuca ridge system in the northeastern Pacific. The divided seamount showed up on a bathymetric contour map made by a team of American and Canadian investigators headed by H. Paul Johnson of the University of Washington. The volcano was evidently split by large fissures cutting across it. Small, centrally fed volcanoes immediately adjacent to the ridge axis have also been observed along the East Pacific Rise.

Elevated volcanic structures forming groups aligned in a direction apparently oblique to that of the adjacent midocean ridge system were probably created by intraplate volcanism. Examples of linear structures with centrally fed volcanic features have been found on the flanks of the East Pacific Rise near nine degrees north and the Juan de Fuca Ridge near 46 degrees north. These seamount chains start on or near the ridge axis and extend for at least a few hundred kilometers in a generally oblique northwesterly direction in contrast to the northerly orientation of the main axis of the midocean ridge.

The mechanism that causes these oblique seamount chains to extend from an accreting plate boundary is not fully understood. In 1977 Richard N. Hey of Princeton proposed that such oblique structures are formed in a region where the ridge axis is offset laterally by a small fracture zone; one ridge segment then propagates while the other recedes. Hey was led to his conclusion by the presence of an obliquely oriented pattern of magnetic reversals recorded in the oceanic crust in the Galápagos area. His model may also be applicable to the seamount chains observed near nine degrees north and 46 degrees north, which were apparently formed when a zone of weakness was affected by widespread fissuring that extended away from the main orientation of the spreading ridge axis. The alignment of the seamount chains may show the preferential direction of the moving plate, if one can assume that the hot spots feeding the successive volcanoes are fixed in the upper mantle as the tectonic plate slides over them.

Once the oceanic crust has been created either through central volcanism giving rise to isolated volcanoes or during the accretion of normal midocean ridges, it is subjected to alteration. One of the major processes altering the oceanic crust is related to the mineralogical and chemical changes that take place as a result of the interaction of seawater with the crust. The main factors are the deep circulation of seawater through fissures and faults and the heat released during magmatic solidification. Recently John M. Edmond and Karen Von Damm of the Massachusetts Institute of Technology summarized the compositional changes caused by seawater circulating through rocks in the oceanic crust [see "Hot Springs on the Ocean Floor," by John M. Edmond and Karen Von Damm; SCIENTIFIC AMERICAN, April, 1983].

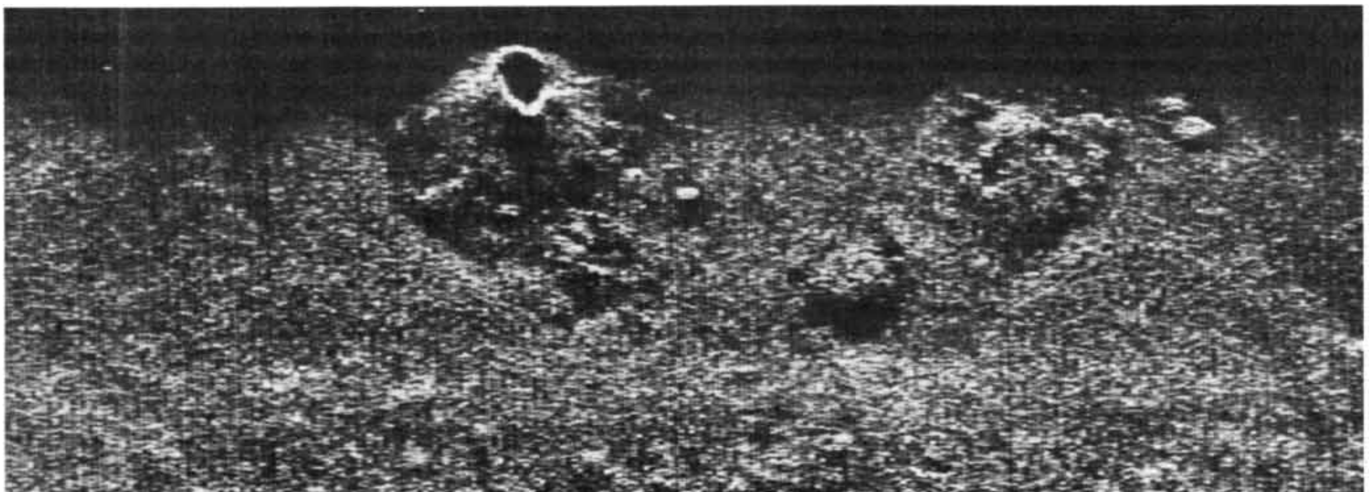
Direct evidence of young oceanic crust altered by hydrothermal circulation was recently obtained during *Cyana* dives on the axis of the East Pacific Rise. Green rocks made up of chlorite (a hydrated silicate) and clays associated with fresh plagioclase (a sodium-calcium silicate) were collected at the foot of a fault scarp on top of which a dead hydrothermal vent was observed. Similar discoveries were made along the crest of the East Pacific Rise in the South Pacific during an oceanographic expedition conducted by Preussag AG, a West German mining company.

As the alteration of the oceanic crust proceeds with the transformation of the major mineral phases, the subsequent release of hydrogen, hydrochloric acid and silicic acid changes the composition of the seawater, making it more corrosive. Positively charged ions of metals such as manganese, zinc, copper and cobalt, which are trace constituents of the basaltic rocks, are also leached and transported by the hydrothermal fluid during convection at high temperature. The trace metals are deposited on the sea floor as sulfide precipitates. Whereas the copper and zinc content of fresh oceanic basalt is respectively less than 100 and 150 parts per million, when the metals are transported and precipitated as sulfides their concentration has increased by more than 10,000 times.

It is estimated that to produce a sulfide deposit of three cubic meters containing 50 percent zinc it would be necessary to process about 1,500 cubic meters of basalt with a zinc content of 100 parts per million. Other minor constituents of the sulfide deposits are cobalt and silver, which are found in minerals such as pyrite, chalcopyrite and sphalerite. The bulk content of silver in some of the sulfides analyzed is as high as 385 parts per million, which corresponds to an enrichment of 5,000 times more than its concentration in the original basalt.

Some hydrothermal deposits on land, such as those in Cyprus, were mined up to the middle of this century for silver and gold. The economic importance of ocean-floor hydrothermalism cannot be overlooked. The known hydrothermal deposits of the world are probably as ancient as the oldest extant parts of the earth's crust, and some of them have been exploited since the dawn of civilization. Mining tools dating from before 2500 B.C. have been found in underground galleries of the Cyprus copper mines. Most of the minable deposits of the Canadian Shield were formed in the Paleozoic era about 500 million years ago and in the Tertiary period between 20 and 25 million years ago. Many of the deposits are thought to have been formed on old midocean ridges. Sulfide formations in Oman represent hydrothermal deposits that are thought to have been formed on elevated undersea structures associated with volcanic eruptions in the Tertiary period.

The exploration of fast-spreading ridge systems with the SeaBeam system, usually in conjunction with a deep-towed camera, enabled our group in 1981 to obtain valuable information about the distribution of hydrothermal deposits at two sites along the axis of the East Pacific Rise. In one 20-kilometer-long segment of the central depression, which is less than 600 meters wide and less than 50 meters deep, there are more than 84 hydrothermal sites. Six months later investigators diving at the same site with the *Cyana* found that out of the 84 hydrothermal sites at least 24 were extremely active. This segment of the East Pacific Rise is considered to be the most active hydrothermal area so far discovered on the ocean floor. The hydrothermal deposits are, on the average, less than 50 meters in diameter. They consist of several columnar structures reaching up to 27 meters in height and capped by several smaller active hydrothermal



LARGE UNDERSEA VOLCANO was discovered in the course of a sonar survey of the ocean floor between the East Pacific Rise and the coast of Peru conducted during the 1980 cruise of the British research vessel *Discovery*. The sonar image was made by means of a side-looking sonar system named GLORIA, which was developed at the Institute of Oceanographic Sciences at Wormley in England. The volcano, one of the largest of some 200 such seamounts found in the survey, is 1.5 kilometers high and 9.5 kilometers wide at the base.

chimneys. The hydrothermal deposits consist mainly of copper, zinc and iron sulfides with a small admixture of hydrated silica and anhydrite.

Temperature measurements of the hot exiting fluid gave values close to 330 degrees Celsius. The black fluids are spewing from the chimneys at a velocity of between .5 meter and two meters per second. The flow rate for a chimney with a diameter of three centimeters was calculated to be between 3.5 and 14 liters per second. The highest flow rates are those of the black fluids charged with abundant metallic ions. It was calculated that their concentration is more than .1 gram per liter. For a single chimney with a flow rate of 10 liters per second this concentration yields a total mass of 100 kilograms of metal per day. This considerable discharge of metal in seawater is not a continuous process; it is thought that the individual chimneys have a comparatively short life span. Most of the hydrothermal material discharged at the axis of the ridge does not precipitate nearby but instead is dispersed elsewhere, contributing to the formation of the metalliferous sediments and iron-manganese nodules common on old oceanic crust at some distance from the spreading axis.

METAL-RICH SULFIDE DEPOSITS flank an active hydrothermal vent of the "black smoker" type in this photograph made from the *Cyana* during a 1982 dive to the axial zone of the East Pacific Rise near 13 degrees north. Hydrothermal activity has also been observed on elevated volcanic structures. Indeed, the findings made by the author's group suggest that seamounts may be favored sites for the formation of mineral deposits by hydrothermal leaching.

Hydrothermal activity has been observed not only on normal mid-ocean ridges but also at elevated volcanic structures. In 1983 Canadian and U.S. geologists led by Richard L. Chase of the University of British Columbia dived in the U.S. submersible *Pisces IV* inside the caldera of a seamount on the spreading axis of the Juan de Fuca Ridge at a depth of almost 1,600 meters; they discovered warm hydrothermal vents and large sulfide structures about 10 meters high and three or four meters in diameter. Other off-axial seamounts were explored in 1982 by Lonsdale, Batiza and Simkin; they reported finding active sulfide deposits in the caldera of a volcano 18 kilometers from the axis of the East Pacific Rise.

At the same time that the American scientists were diving to the off-axis seamounts on one part of the East Pacific Rise, French scientists were exploring seamounts and the ridge crest at a site some 900 kilometers to the south. Large hydrothermal deposits were seen on a small seamount about six kilometers east of the ridge axis near 13 degrees north. (The seamount is the one at the lower right in the map on page 47.) The summit and the southern flank of the 400-meter volcano were surveyed with a deep-towed camera and with the *Cyana*. The base of the seamount is 2,675 meters deep and has a diameter of six kilometers; the structure terminates at a summit less than 500 meters in diameter at a depth of 2,440 meters. During one dive, going from the base of the seamount to its summit, my colleagues and I saw evidence of sedimentary, volcanic and hydrothermal structures. The base of the seamount is covered by a thin blanket of sediment for the first 500 meters or so. As the dive continued up the southern flank, large hydrothermal deposits covering about 80 percent of the surface were seen to extend almost continuously to the top of the seamount. Sporadic occurrences of tubular pillow flows were sometimes observed to have partially buried the hydrothermal products. This was usually manifested by localized abundances of pillow lava over sulfides and sediments near the top of the seamount. The hydrothermal deposit is an altered product of mainly iron hydroxide capping a massive iron-ore deposit that crops out in a few places. The lack of sediment on and near the summit and the freshness of the lava indicate that volcanic and hydrothermal activity has occurred here during a recent discharge of hot fluids.

Direct field observations have shown that the hydrothermal deposits on this small seamount are at least 800 meters long and about 200 meters wide along the track followed by the submersible. A first approximation indicates that the deposit has a volume of material 10 times as great as all the combined sulfide deposits distributed along the 20-kilometer segment of the adjacent axis of the East Pacific Rise. Further exploration of the area in February of this year has tended to confirm the conclusion that seamounts are favored sites for the formation of mineral deposits.

Immunotoxins

The idea is to link a toxic agent to a monoclonal antibody binding to a particular tumor antigen, thus fashioning a "magic bullet" that will destroy targeted cancer cells but leave normal cells unharmed

by R. John Collier and Donald A. Kaplan

How is it possible to destroy a particular subset of a patient's own cells and leave the rest of the cells unaffected? This is the difficult objective in the chemotherapy of cancer and certain other diseases. A cancer cell is a normal cell gone wrong. Freed from the usual constraints on growth, it multiplies rapidly; the cancer invades adjacent tissue and may metastasize to distant tissues. Most of the chemotherapeutic agents now available to the oncologist are drugs that are taken up by, or that primarily affect, cells that are multiplying rapidly. Unfortunately, this is a marginal basis for selectivity. Normal cells are not unaffected by these drugs. Doses large enough to eradicate cancer cells with high efficiency can be lethal to the patient, and even moderate doses can cause a variety of harmful side effects.

There is an alternative approach: a "magic bullet" that destroys its designated targets without significantly affecting any other cells. Antibiotics are one kind of magic bullet. An antibiotic can kill bacteria or prevent their proliferation without harming human cells because it inhibits metabolic processes peculiar to the prokaryotic bacterial cell. It is harder to develop comparable selectively toxic agents for fungal or parasitic diseases because the metabolism of the eukaryotic cells that cause them is too much like that of the infected mammalian host's cells.

A different kind of magic bullet would be one whose magic resides in its ability to home on a designated target such as a particular kind of cancer cell. That is, the toxic agent itself would be toxic for most cells, but it is allowed to reach only a defined population of cells. Within the past decade advances in cell-surface immunology, the advent of monoclonal antibodies and new understanding of certain highly toxic natural substances have begun to make such an approach seem feasible. The strategy is simple, at least in principle: Develop a monoclonal antibody that binds specifically to a target cell and not to other cells, and couple the antibody to a toxic molecule. The antibody-toxin conjugate, or immunotoxin, should kill target cells, and no other cells, with high efficiency. In practice, as we shall explain, there are difficulties in this approach, and there is a long way to go before immunotoxins can become standard therapeutic agents. A number of investigators, however, have been able to demonstrate the capacity of immunotoxins for targeted killing of cells growing in a laboratory culture, and the method is being tested in animals.

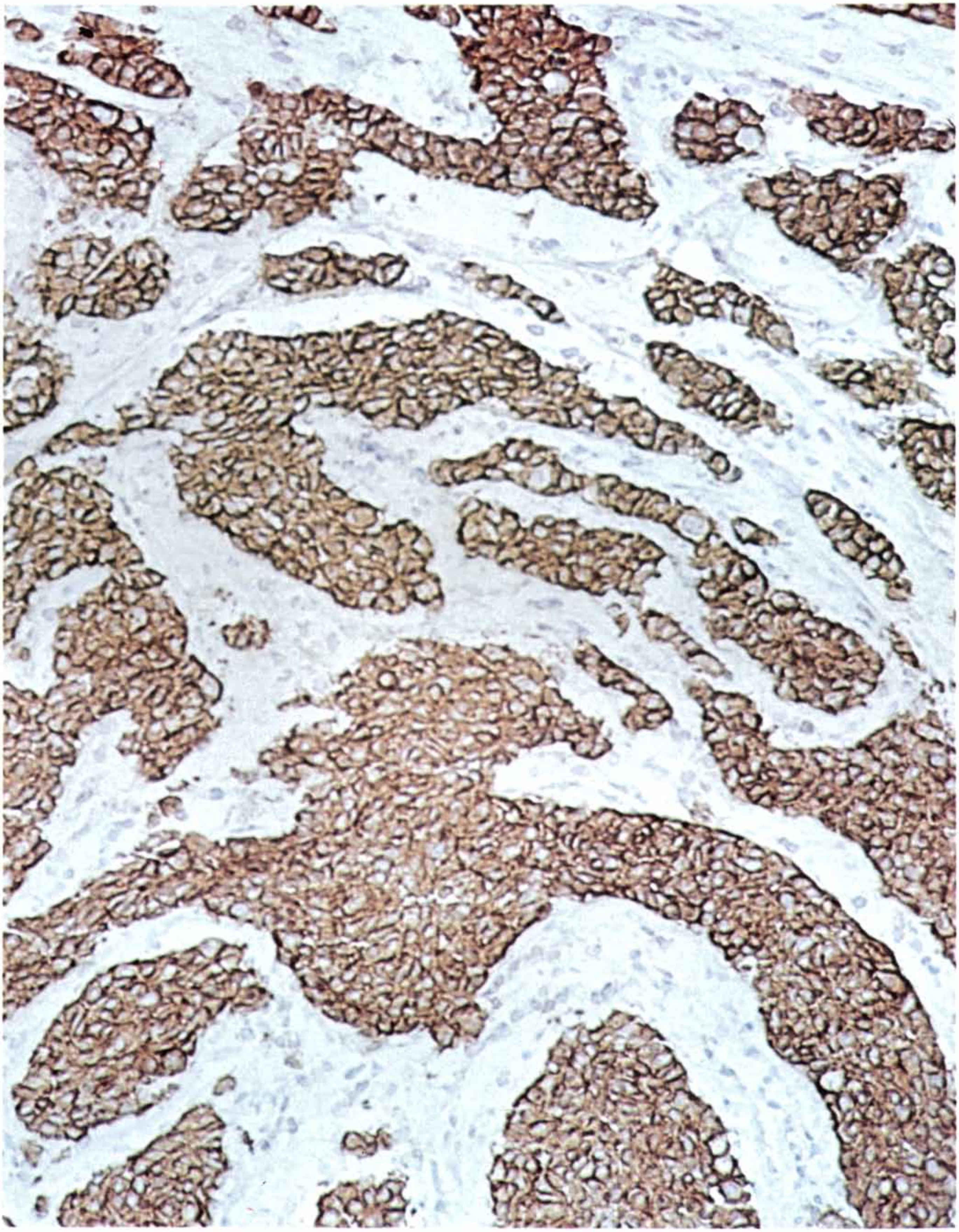
The lipid surface membrane of a living cell is studded with hundreds of chemical structures (mostly proteins, some of them with carbohydrate chains attached) that have various roles in cell communication and metabolism. Many of the structures differ from species to species and from individual to individual of the same species; in an individual many of them differ from one cell type to another. Significantly, some cell-surface structures are peculiar to certain malignant cells, and they distinguish tumor cells even from normal cells of the same tissue.

When human cells are introduced into the body of an animal, cell-surface markers that are different from the recipient's own are recognized as being foreign. They therefore serve as antigens, and the animal's immune system responds to their presence by making antibodies against them. Antibodies are highly specific protein molecules that are able to recognize and bind tightly to the particular antigens that induced their synthesis. The natural function of such antibodies is to initiate the defensive processes that inactivate or destroy harmful foreign substances, but for some time investigators have exploited the specificity of antibodies to find, identify, label and separate particular cells or molecules. Antibodies elicited by particular human cells should therefore provide a means of distinguishing those cells from others. Indeed, early in this century the German bacteriologist and immunologist Paul Ehrlich proposed that antibodies might somehow serve to deliver chemically coupled toxic agents to particular cells.

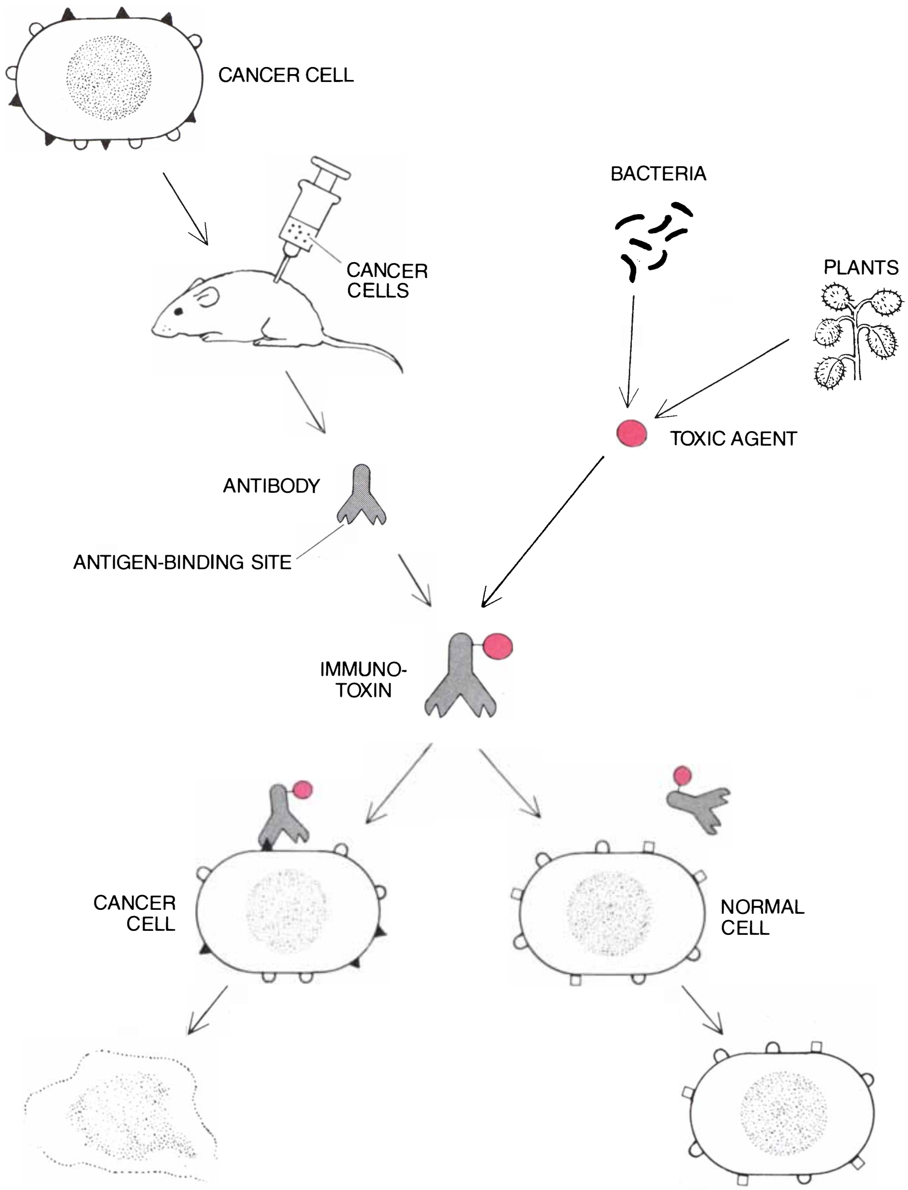
For many years that was not possible. Until recently antibodies reacting specifically with only a single antigen, and therefore binding dependably to a single class of cells, could be obtained only in minute quantities because it was difficult to purify them from the mixture of antibodies in the serum of an immunized animal. Each specific antibody is synthesized and secreted by plasma cells derived from a particular clone of *B* lymphocytes dedicated to the synthesis of that antibody, but there was no way to propagate such a clone in the laboratory because the antibody-secreting cells could not be grown in a laboratory culture medium.

That problem was overcome by the development of hybridoma technology in 1975 by Cesar Milstein of the Medical Research Council Laboratory of Molecular Biology in Cambridge. Milstein found a way to fuse *B* lymphocytes with the related but malignant myeloma cells. The resulting hybridoma cells are like *B* cells in that each produces a single antibody; they are like myeloma cells in that they can be grown indefinitely in culture. A single hybridoma cell can be grown into a clone of identical cells, which serves as a continuing source of a "monoclonal" antibody against a specific antigen.

To isolate a monoclonal antibody recognizing a certain type of cell one inoculates laboratory mice with a pure culture of the cells of interest or with a preparation of their cell membrane (which carries the cell-surface antigens). The animal's *B* cells are isolated and then fused with myeloma cells, and the resulting hybridomas are screened for the production of monoclonal antibodies that bind to the target cells in culture or to their membranes but do not bind effectively to a panel of control cells or membranes. The specificity of binding of monoclonal antibodies is rarely absolute, but it can be extremely high, in part



MONOCLONAL ANTIBODIES home in on cancer cells in a frozen section taken from a breast carcinoma. The antibodies bind specifically to a tumor-associated antigen, a cell-surface molecule peculiar to the cancer cells. The antibodies are made visible by the indirect immunoperoxidase technique, by which the antibody-covered surface of each cancer cell is stained brown; the nuclei of both cancer cells and normal cells are counterstained light blue. The antibodies were prepared and the photomicrograph was made in Arthur E. Frankel's laboratory at the Cetus Corporation. Monoclonal antibodies such as the ones illustrated here are linked to toxic agents to make immunotoxins: composite molecules that bind to cancer cells and kill them but fail to bind to normal cells, leaving them unaffected.



CONCEPT OF AN IMMUNOTOXIN is simple. Cancer cells, which include among their surface molecules a tumor-associated antigen (*black*), are injected into a mouse. The mouse synthesizes antibodies against the antigen. A toxic substance derived from a bacterial or other source is linked to the antibody to make an immunotoxin. The immunotoxin binds to the tumor-associated antigen on the cancer cells and kills the cells (*bottom left*). It does not bind to antigens on normal cells (*bottom right*), and so normal cells are not affected by the toxin.

because the antigen recognized by the antibody is much more prevalent on the surface of the target cells than it is on the surface of nontarget cells.

Even before monoclonal antibodies against cell-surface antigens became available, investigators had been studying the properties of toxic agents that might be delivered by antibodies to kill cancer cells. There are several candidates. There are toxic drugs, such as the ones we alluded to above that are currently administered for the chemotherapy of cancer. There are radioactive isotopes of various elements, some of which have an affinity for particular tissues (as iodine has for the thyroid gland). And there are naturally occurring toxic proteins produced by certain bacteria, plants and animals.

All these candidates have been coupled to antibodies and tested, but most of the work on immunotoxins has been done with naturally occurring toxic proteins, some of which are among the most potent cytocidal (cell-killing) substances known. Under appropriate conditions a human cell can be killed by a single molecule of the toxin secreted by *Corynebacterium diphtheriae,* the agent that causes diphtheria, or of the protein ricin, which is found in castor beans. We have worked mainly with diphtheria toxin.

The toxin is an enzyme that inactivates an essential component of a nonbacterial cell's protein-synthesis machinery. It catalyzes the transfer of adenosine diphosphate ribose, a part of the electron-carrier nicotinamide adenine dinucleotide (NAD), to a protein called elongation factor 2 (EF-2). The factor is required for the synthesis of protein on the cellular organelles called ribosomes, and it is inactivated by the attachment of the ADP-ribose group. In about a day a single diphtheria toxin molecule can inactivate most (perhaps all) of the two million EF-2 molecules in a typical animal cell. Unable to make protein, the cell dies.

Enzymatic action of this kind, such that one toxin molecule functions repeatedly to inactivate a very large number of target molecules, is probably characteristic of most highly toxic proteins. For example, Sjur Olsnes, Alexander A. Pihl and their colleagues at Norsk Hydro's Institute for Cancer Research in Oslo have shown that ricin and some related plant toxins also depend on an enzymatic process (although it is different from diphtheria toxin's) to inactivate ribosomes.

Powerful toxins such as the ones we have described would seem to be ideal for coupling to antibodies to make an effective immunotoxin. There is a major problem, however. A toxin molecule has its own binding sites, by means of which it can bind to most mammalian cells. An immunotoxin made by simply linking a toxin molecule to an antibody retains that nonspecific binding ability. It will attach itself not only to the target cells the antibody recognizes but also to almost any other cell, and so it will kill normal cells about as efficiently as it kills cancer cells. Clearly, then, one needs somehow to eliminate the toxin's own ability to bind to cells and make it rely for binding on the antibody to which it is linked.

A way to do that emerged from detailed studies of the structure of toxic proteins done in the early 1970's. One of us (Collier), working with colleagues at the University of California at Los Angeles, and D. Michael Gill, Alwin M. Pappenheimer, Jr., and their co-workers at Harvard University found that the intact diphtheria toxin molecule would not catalyze the ADP-ribosylation of EF-2. Enzymatic activity was observed only if the toxin was first cleaved into two unequal parts. First the long polypeptide chain of the toxin had to be cut into two smaller chains by a protease (an enzyme that attacks proteins) and then a disulfide bond linking the two chains had to be cut. (A disulfide bond is a chemical link between two sulfur atoms each of which is attached to a polypeptide chain.)

The shorter of the two resulting chains, designated the *A* chain, turned out to be responsible for the toxin's enzymatic activity. An *A* chain, separated from the longer *B* chain, was able to inactivate EF-2 in cell extracts. The *B* chain was found to be responsible for binding the toxin to receptors on the cell surface. Neither chain by itself was toxic for intact cells, implying that both binding to receptors and ADP-ribosylation are essential for the normal toxic process. Ricin and some other toxins derived from plants were found to have similar characteristics. Each such toxin is composed of two disulfide-linked chains: an *A* chain with enzymatic activity and a *B* chain serving to bind the toxin to the cell surface.

These findings suggested to several workers that if one could couple only the *A* chain to a cell-specific antibody, the toxin's own receptor-binding ability should be eliminated; the antibody alone should then mediate cell-surface binding and only the target cells should be killed. Moreover, the fact that the *A* chain alone is virtually nontoxic (being unable to bind to cells) meant that any adventitious breakdown of the immunotoxin into *A*-chain and antibody components within the body would fail to yield nonspecifically toxic agents.

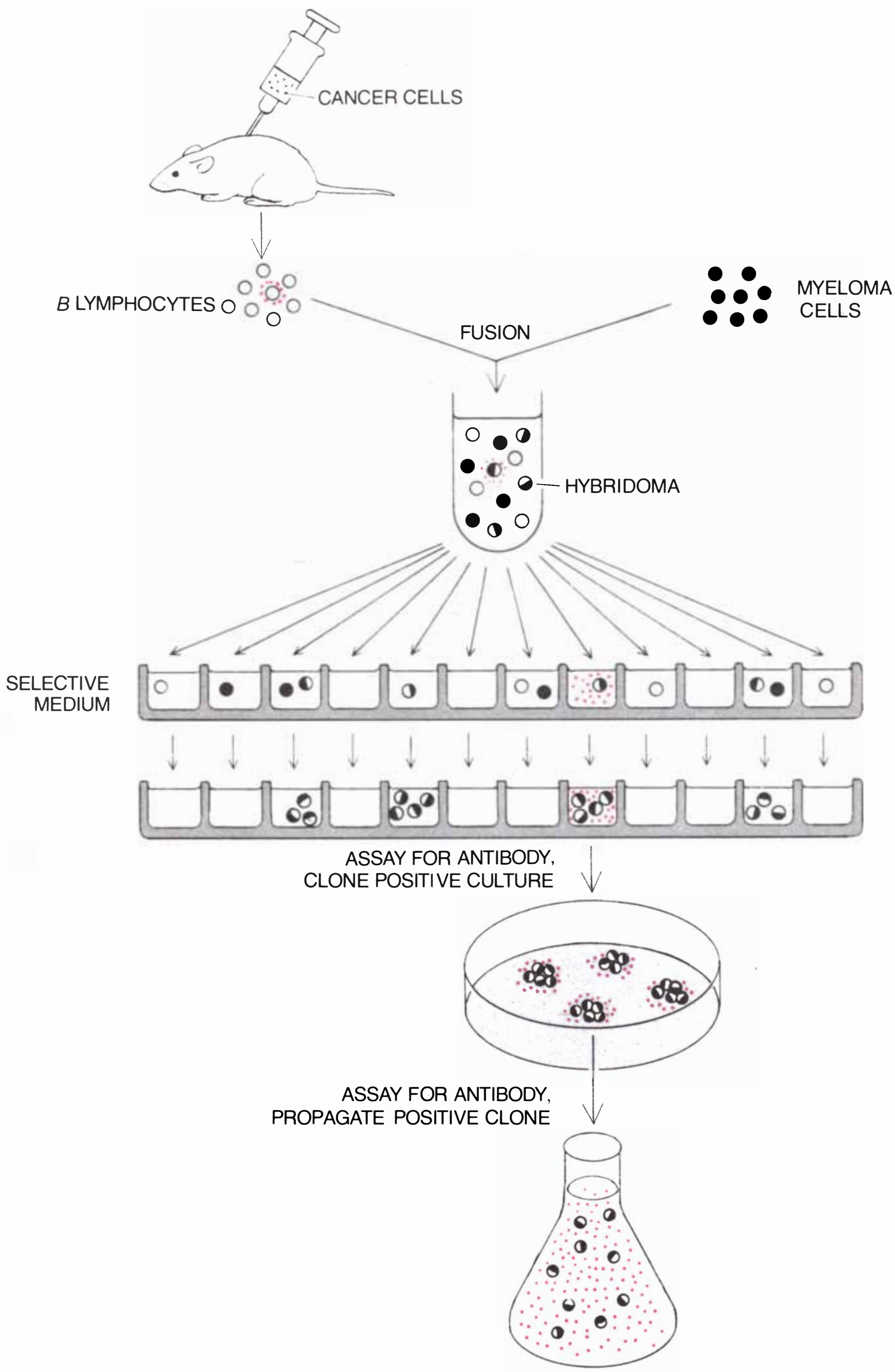
The fact that the two chains of the toxins we worked with are joined to each other by disulfide bonds suggested one should couple the *A* chain to an antibody by a similar link. Disulfide bonds are known to be ruptured easily within a cell and should therefore make for easy release of the *A* chain, and thus for the chain's activation through exposure of its enzymatic site, in a target cell.

Monoclonal antibodies directed against specific cells were not then readily available, and so several groups began to test the concept by coupling *A* chains to hormones or lectins. Hormones bind to specific cell-surface receptors; the lectins, which are nontoxic plant proteins, bind avidly to various glycoproteins and glycolipids on cell surfaces. David M. Neville, Jr., and his colleagues at the National Institute of Mental Health were the first to link a toxin *A* chain to a hormone.

At U.C.L.A., D. Gary Gilliland and one of us (Collier) linked the diphtheria toxin *A* chain to the lectin concanavalin *A*. The resulting conjugate was found to be toxic at low concentrations for human cells in culture. Its toxicity was inhibited when an excess of uncoupled concanavalin *A* was added to the culture medium. Apparently the free lectin was binding to the surface of cells, blocking the conjugate's attachment and thereby interfering with its toxic effect. This was a demonstration that the coupled lectin was indeed in control of the conjugate's binding, as we had hoped it would be. Tsuyoshi Uchida and Yoshio Okada of the University of Osaka got similar results by linking the diphtheria toxin *A* chain to a different lectin.

By the time these results had been achieved other laboratories were reporting the successful isolation of monoclonal antibodies against cell-surface antigens. Some of the antibodies were directed against tumor-associated antigens: antigens found to be present rarely, if at all, on the surface of normal cells. Zenon Steplewski and Hilary Koprowski of the Wistar Institute in Philadelphia developed a monoclonal antibody that appeared to bind to human colorectal carcinoma (CRC) cells but not to a number of other cells. The U.C.L.A. and Wistar groups decided to collaborate to prepare immunotoxins with the anti-CRC antibody.

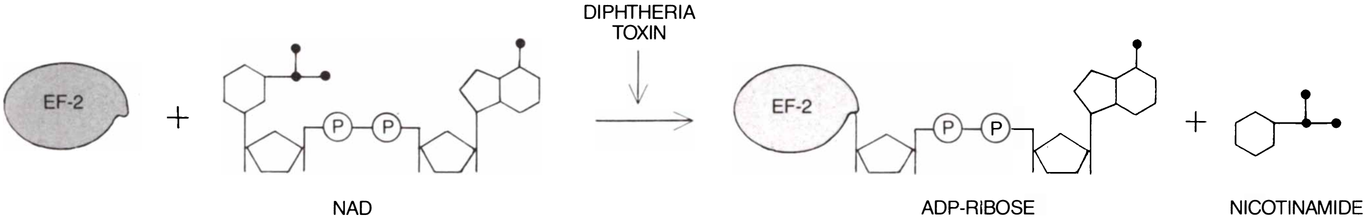
The immunotoxins were constructed at U.C.L.A. by linking the *A* chain of either diphtheria toxin or ricin to the antibody. They were tested at Wistar by assaying cells for their protein-synthesizing activity, the function inhibited by



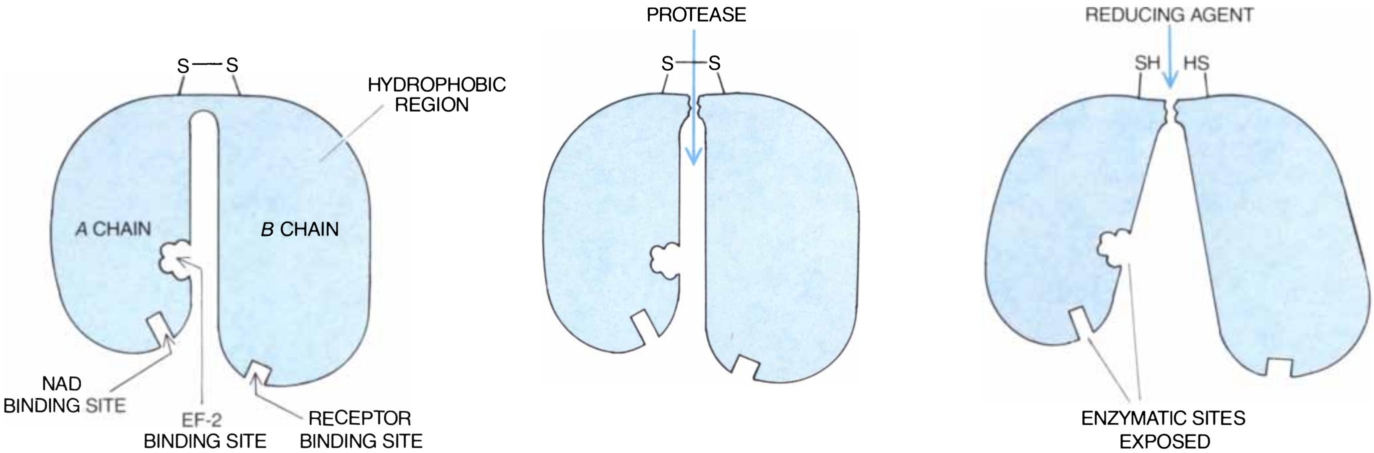
MONOCLONAL ANTIBODY against a specific cell-surface antigen is made by injecting a mouse with the cells, harvesting the immunized animal's *B* lymphocytes and fusing them with malignant myeloma cells. Successfully fused cells, called hybridomas, are recognized by their ability to survive and multiply in a selective medium. They have the myeloma's property of immortality, and so they can be maintained indefinitely in a laboratory culture. Some of the hybridomas secrete the wanted antibody (*color*), which is detected in the medium. A positive culture is plated to isolate a single colony: a clone of cells that all make the antibody. The clone is propagated to make a large supply of cells, a continuing source of the monoclonal antibody.

the toxin. Cultured CRC cells were incubated for 24 hours with the immunotoxin, with unconjugated antibodies or *A* chains or with intact diphtheria toxin; then the cells were assayed for their ability to incorporate amino acids into proteins. Melanoma cells were tested in the same way.

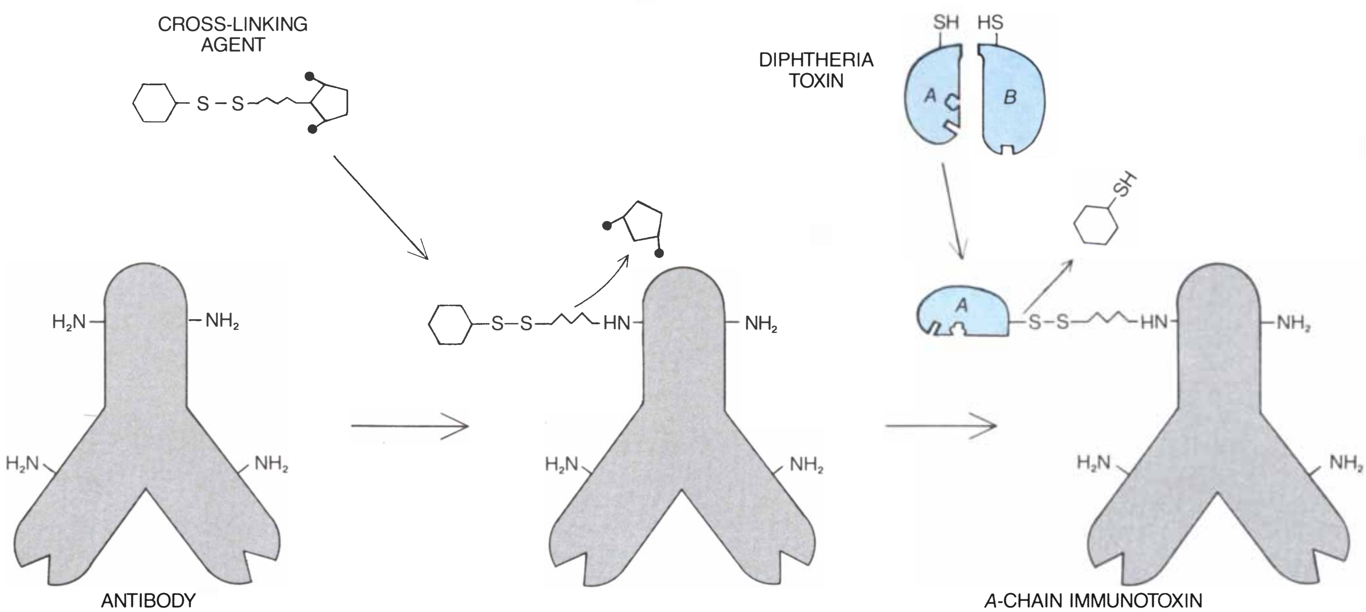
The results were encouraging. Even at rather low concentrations both *A*-chain immunotoxins blocked protein synthesis effectively in the CRC cells but not in the melanoma cells, whereas the intact diphtheria toxin had a powerful effect on the melanoma cells as well as on the CRC cells. The unconjugated antibodies had no effect, and the unconjugated *A* chains had little effect, on either kind of cell. In other words, the antibody part of the immunotoxin did dictate specificity for CRC cells and the toxin part did have an effect on those cells. This was the first demonstration that a monoclonal antibody against a tumor-associated



DIPHTHERIA TOXIN acts as an enzyme. Its effect is to inactivate elongation factor 2 (EF-2), an essential component of the cell's protein-synthesizing machinery. The toxin does this by catalyzing the transfer to EF-2 of adenosine diphosphate ribose, part of the electron carrier nicotinamide adenine dinucleotide (NAD). A single molecule of diphtheria toxin can inactivate a cell's EF-2 in a single day.



ENZYMATIC ACTIVITY is observed only after the toxin molecule (*left*) has been cleaved into two chains. The smaller one, the *A* chain, carries enzymatic sites catalyzing the transfer of ADP-ribose to EF-2. The *B* chain has a receptor-binding site and a hydrophobic region able to insert itself into a biological membrane. One way to make an immunotoxin that does not bind to nontarget cells is to use only the *A* chain. The chains are cleaved by cutting protein bridge between them (*center*) and disulfide bond (*S–S*) linking them (*right*).



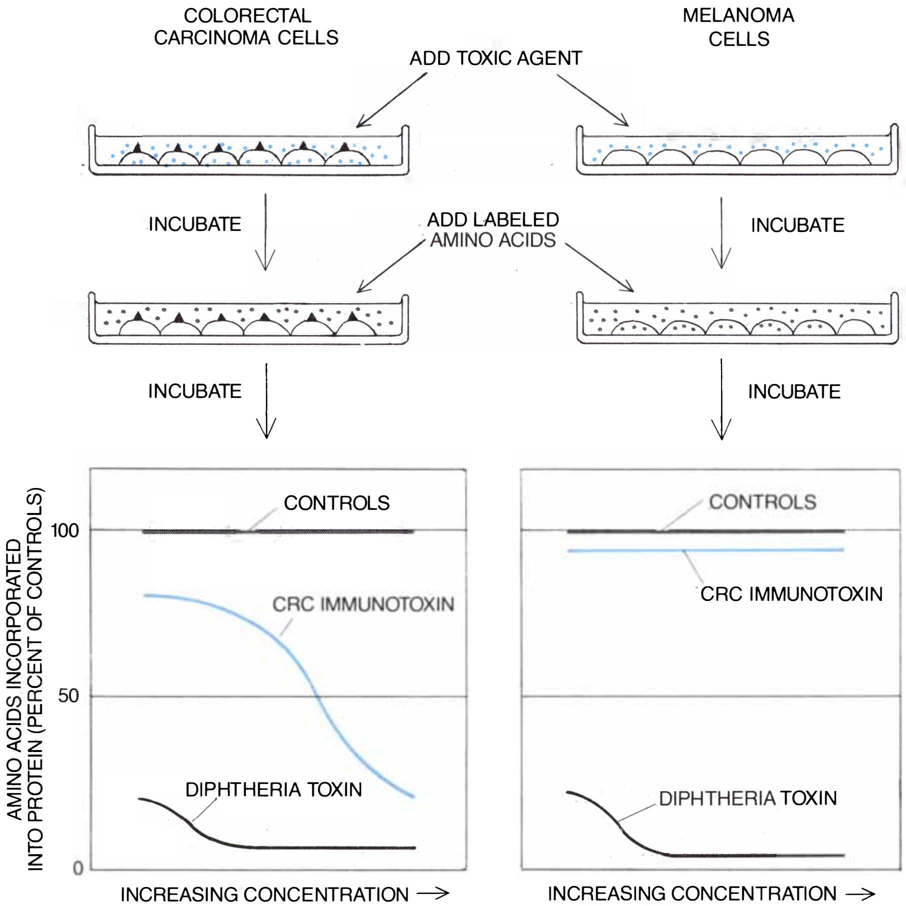
***A*-CHAIN IMMUNOTOXIN is made by conjugating an antibody against a tumor-associated antigen (*left*) with the *A* chain of diphtheria toxin. One end of a cross-linking agent carrying a disulfide bond is attached to an amino (NH_2) group on the antibody (*center*). The other end of the cross-linker is attached to the *A* chain of a toxin molecule (*right*); as a result the antibody replaces the toxin's *B* chain.**

antigen could direct the action of a potent toxin against specific cells. At about the same time Keith A. Krolick, Ellen S. Vitetta and Jonathan W. Uhr of the University of Texas Health Science Center at Dallas reported that antibodies against a mouse leukemia-cell antigen can make the leukemia cells specific targets of the ricin *A* chain.

These first-generation immunotoxins made with the *A* chain alone represented only a promising initial step. Their toxicity for target cells varied widely; in some cases it was many orders of magnitude lower than that of the intact parent toxin. At about the same time as these puzzling differences in toxicity were being reported, some possible reasons for the variation were beginning to emerge. Clearly the enzymatically active part of a toxin must do its work in the cell's cytoplasm, which is where EF-2 (the diphtheria toxin's target) and the ribosomes (ricin's target) reside. Not much had been known about how a large protein gets into a cell. In our early work we hoped our lectin conjugates and then our immunotoxins would make their way to the cytoplasm, but we did not know just how that would happen, or with what efficiency.

It is now known that many proteins are brought across the cell's outer membrane, the plasma membrane, by a process called receptor-mediated endocytosis [see "How Receptors Bring Proteins and Particles into Cells," by Alice Dautry-Varsat and Harvey F. Lodish; SCIENTIFIC AMERICAN, May]. The plasma membrane is studded with receptors, each of them specific for a particular protein or small particle. A diphtheria toxin molecule is thought to bind to such a receptor (which probably has a benign role in the ordinary life of the cell and is simply preempted by the toxin). The receptor either moves to or is already at one of many "coated pits" on the cell surface. At these sites the plasma membrane invaginates, or folds inward, forming a membrane-bounded vesicle called an endosome, which carries receptors and any ligand bound to them (such as a toxin molecule) into the cell.

A toxin molecule in an endosome is inside the cell, but it remains shielded from its substrates, EF-2 and NAD, by the endosomal membrane. How an *A* chain escapes from the endosome to get at its target has still not been clearly established. In the case of diphtheria toxin there is evidence that an increasingly acidic environment within the endosome causes the toxin molecule to insert itself into the endosomal membrane. The disulfide bond linking the *A* and *B* chains is apparently cleaved by substances called reducing compounds that are present in the cytoplasm, allowing the *A* chain to be released into the cytoplasm and leaving the *B* chain in the endosomal membrane.



IMMUNOTOXINS made with a monoclonal antibody against colorectal carcinoma (CRC) cells were tested for their ability to inhibit protein synthesis. The immunotoxins, *A* chains or antibodies alone or intact diphtheria toxin were incubated with CRC cells and with melanoma cells. Amino acids labeled with a radioactive isotope were added and the cells' ability to incorporate them into protein was assayed. *A* chains and antibodies alone (*controls*) had little or no effect. Diphtheria toxin inhibited protein synthesis in both CRC cells (*left*) and melanoma cells (*right*). The immunotoxins were effective in CRC cells but not in melanoma cells.

Apparently, then, diphtheria toxin is an unusual, highly specialized enzyme. It must perform at least three distinct functions: binding to a receptor, insertion into and traversal of the endosomal membrane, and the transfer of ADP-ribose to EF-2. As we have explained, the first function is accomplished by the *B* chain and the last function by the *A* chain. It has recently been established that membrane insertion is also a function of the *B* chain. It depends, in fact, on a particular region of the chain that is rich in hydrophobic amino acids, which are suited for insertion into the hydrophobic lipid membrane.

Less is known about how ricin gets into the cytoplasm. Acidity within the endosome does not seem to play a role. Nevertheless, the presence of the ricin *B* chain greatly increases the toxic activity of the ricin *A* chain, suggesting there may be a functional region of the *B* chain that (as in the case of diphtheria toxin) is operative during the *A* chain's transport into the cytoplasm.

The data on endocytosis point to two possible reasons for the reduced toxicity of *A*-chain immunotoxins compared with the toxicity of their parent toxin molecules. If, as would appear to be the case, the *A* chain can get to its substrates only by way of an endosome, the toxic activity of an immunotoxin must depend first of all on the efficiency with which it is incorporated into endosomes. That depends in turn on the ability of the cell-surface antigen to which the immunotoxin binds to transport it to a coated pit for receptor-mediated endocytosis. One reason for the reduced toxicity of certain immunotoxins may therefore be that the monoclonal antibodies with which *A* chains have been conjugated do not bind to antigens that frequent the coated pits. It should be possible, however, to find antibodies directed against such antigens.

There is a second possible reason, which may be more important. Having been carried into the cell in an endosome, the *A* chain needs to get out of the endosome to reach its substrate. At least in the case of diphtheria toxin, the membrane-insertion function required for efficient transport across the endosomal membrane is a property of the hydrophobic region of the *B* chain. By substituting an antibody for the toxin's *B* chain one therefore eliminates an activity that is critical for the efficient entry of the *A* chain into the cytoplasm. *A* chains linked to antibodies do seem to reach the cytoplasm, but they generally reach it with lower efficiency than intact toxin molecules do. Penetration of the endosomal membrane seems to be a limiting step in determining the toxic activity of an immunotoxin.

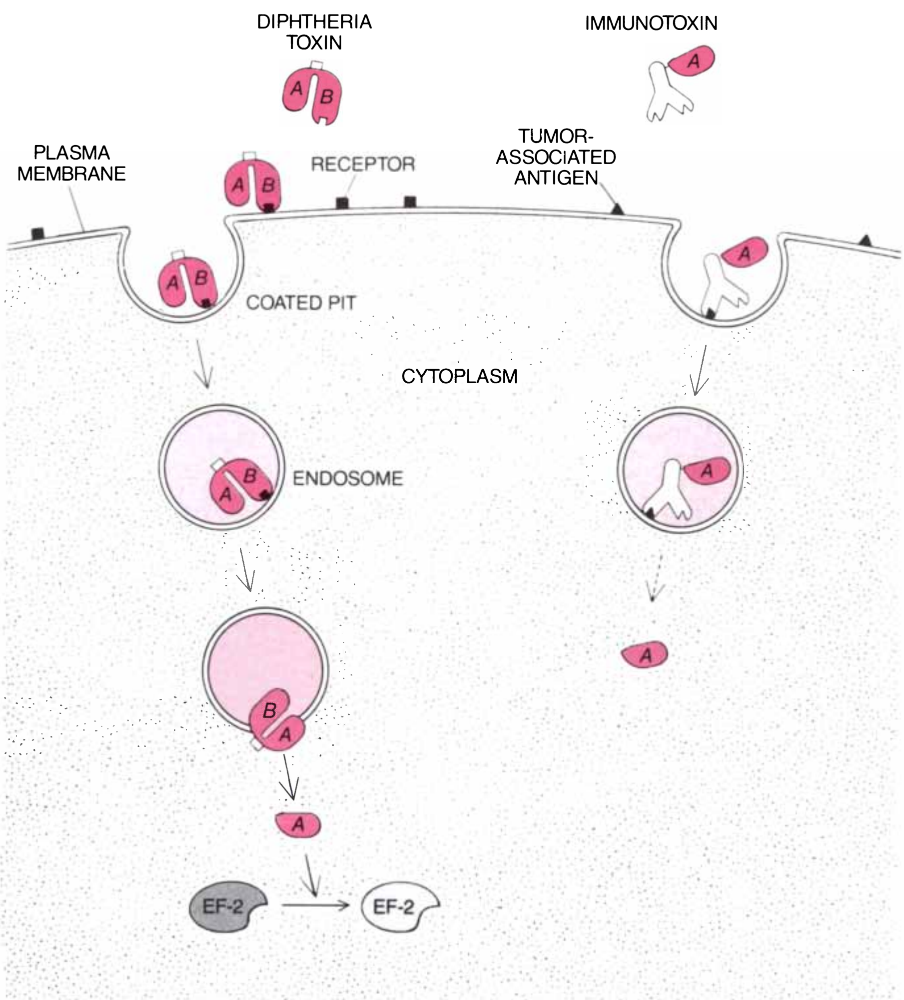
How is it possible to include in the immunotoxin the *B*-chain region required for membrane insertion without retaining the chain's receptor-binding function and thereby precluding cell-specific targeting? One approach would be to retain the entire *B* chain but block its binding to its receptors on the plasma membrane. It has long been known that the sugar lactose inhibits the binding of ricin to the surface of cells, presumably because the lactose occupies the ricin's receptor-binding site. Richard J. Youle and Neville have prepared immunotoxins by conjugating the intact ricin molecule with a monoclonal antibody. When cells in culture are treated with these whole-ricin immunotoxins in the presence of large concentrations of lactose, nonspecific toxicity is minimized. Such concentrations are toxic for animals, however, so that blocking binding sites with lactose, at least, does not appear to be feasible in human beings.

A more promising approach is to alter the *B* chain so that receptor-binding activity is eliminated but membrane-insertion activity is retained. One way to do this is to treat the intact toxin with a chemical that specifically disables only the receptor-binding part of the *B* chain and then to couple the modified toxin to an antibody. Ira H. Pastan of the National Cancer Institute and his colleagues were able to do this with *Pseudomonas* exotoxin *A*, a bacterial toxin resembling diphtheria toxin in its mode of action. No such method has yet been found for diphtheria toxin or ricin.

There is a powerful alternative way to alter the *B* chain of diphtheria toxin or ricin. Rather than modifying the toxin chemically, one can alter the gene encoding it. The new recombinant-DNA techniques make it possible to modify and manipulate genes almost at will. At U.C.L.A. we undertook the genetic engineering of toxins for the construction of immunotoxins. Our plan was to isolate the DNA coding for a toxin, to eliminate the nucleotide sequences that specify the receptor-binding site on the *B* chain and then to introduce the modified gene into *Escherichia coli* cells. The bacteria would translate the gene to make large quantities of a modified toxin: one that is unable to bind indiscriminately to cells but retains enzymatic and membrane-insertion activity, and that can be linked to a monoclonal antibody to make a cell-specific immunotoxin.

At U.C.L.A. Michael J. Bjorn and one of us (Kaplan) determined the nucleotide sequence of parts of the diphtheria-toxin gene. That enabled us to identify a fragment of the gene encoding the enzymatic and hydrophobic regions of the toxin but not the receptor-binding region. Because the protein product of the fragment would be essentially nontoxic, the Recombinant DNA Advisory Committee of the National Institutes of Health gave permission for the fragment to be cloned in *E. coli*. The fragment was cloned and then sequenced in full at the Cetus Corporation by Lawrence Greenfield, Bjorn and one of us (Kaplan). The cloned fragment is expressed in *E. coli*. The bacterial product retains the toxin's essential properties except for its ability to bind to cells.

As for the potential medical applications of immunotoxins, perhaps the most promising one for the near future is the treatment of bone marrow in the course of transplantation. Patients with leukemia are sometimes treated by total-body irradiation or chemotherapy in an effort to kill the leukemic cells. The treatment, however, also destroys normal stem cells in the bone marrow, which are the precursors of all blood cells. The patient therefore needs a bone-marrow transplant to provide a new population of stem cells. The trouble is that the transplantation can give rise to what is called graft-versus-host disease, in which *T* lymphocytes in the donated marrow recognize the recipi-



INTACT DIPHTHERIA TOXIN is thought to be taken into a cell by receptor-mediated endocytosis (*left*). The toxin's *B* chain binds to a receptor on the cell surface and the toxin is carried to one of many "coated pits," where the plasma membrane invaginates and pinches off to form a vesicle called an endosome. The acidic environment within the endosome (*light color*) causes the *B* chain's hydrophobic region to insert itself into the endosomal membrane; the *A* chain apparently crosses the membrane and is released into the cytoplasm, where it inactivates EF-2. An *A*-chain immunotoxin, on the other hand, presumably binds to a tumor-associated antigen (*right*). It may get to a coated pit and be internalized in an endosome, from which the *A* chain somehow emerges into the cytoplasm. *A* chains reach the cytoplasm at low efficiency, however; *A*-chain immunotoxins are less effective than their parent toxin molecules.

GTG AGC AGA AAA CTG TTT GCG TCA ATC TTA ATA GGG GCG CTA CTG GGG ATA GGG GCC CCA CCT TCA GCC CAT GCA
(f Met) Ser Arg Lys Leu Phe Ala Ser Ile Leu Ile Gly Ala Leu Leu Gly Ile Gly Ala Pro Pro Ser Ala His Ala
-21 -11 -1

GGC GCT GAT GAT GTT GTT GAT TCT TCT AAA TCT TTT GTG ATG GAA AAC TTT TCT TCG TAC CAC GGG ACT AAA CCT GGT TAT GTA GAT TCC
Gly Ala Asp Asp Val Val Asp Ser Ser Lys Ser Phe Val Met Glu Asn Phe Ser Ser Tyr His Gly Thr Lys Pro Gly Tyr Val Asp Ser
1 10 20 30

ATT CAA AAA GGT ATA CAA AAG CCA AAA TCT GGT ACA CAA GGA AAT TAT GAC GAT GAT TGG AAA GGG TTT TAT AGT ACC GAC AAT AAA TAC
Ile Gln Lys Gly Ile Gln Lys Pro Lys Ser Gly Thr Gln Gly Asn Tyr Asp Asp Asp Trp Lys Gly Phe Tyr Ser Thr Asp Asn Lys Tyr
31 40 50 60

GAC GCT GCG GGA TAC TCT GTA GAT AAT GAA AAC CCG CTC TCT GGA AAA GCT GGA GGC GTG GTC AAA GTG ACG TAT CCA GGA CTG ACG AAG
Asp Ala Ala Gly Tyr Ser Val Asp Asn Glu Asn Pro Leu Ser Gly Lys Ala Gly Gly Val Val Lys Val Thr Tyr Pro Gly Leu Thr Lys
61 70 80 90

GTT CTC GCA CTA AAA GTG GAT AAT GCC GAA ACT ATT AAG AAA GAG TTA GGT TTA AGT CTC ACT GAA CCG TTG ATG GAG CAA GTC GGA ACG
Val Leu Ala Leu Lys Val Asp Asn Ala Glu Thr Ile Lys Lys Glu Leu Gly Leu Ser Leu Thr Glu Pro Leu Met Glu Gln Val Gly Thr
91 100 110 120

GAA GAG TTT ATC AAA AGG TTC GGT GAT GGT GCT TCG CGT GTA GTG CTC AGC CTT CCC TTC GCT GAG GGG AGT TCT AGC GTT GAA TAT ATT
Glu Glu Phe Ile Lys Arg Phe Gly Asp Gly Ala Ser Arg Val Val Leu Ser Leu Pro Phe Ala Glu Gly Ser Ser Ser Val Glu Tyr Ile
121 130 140 150

AAT AAC TGG GAA CAG GCG AAA GCG TTA AGC GTA GAA CTT GAG ATT AAT TTT GAA ACC CGT GGA AAA CGT GGC CAA GAT GCG ATG TAT GAG
Asn Asn Trp Glu Gln Ala Lys Ala Leu Ser Val Glu Leu Glu Ile Asn Phe Glu Thr Arg Gly Lys Arg Gly Gln Asp Ala Met Tyr Glu
151 160 170 180

TAT ATG GCT CAA GCC TGT GCA GGA AAT CGT GTC AGG CGA TCA GTA GGT AGC TCA TTG TCA TGC ATA AAT CTT GAT TGG GAT GTC ATA AGG
Tyr Met Ala Gln Ala Cys Ala Gly Asn Arg Val Arg Arg Ser Val Gly Ser Ser Leu Ser Cys Ile Asn Leu Asp Trp Asp Val Ile Arg
181 190 S-S 200 210

GAT AAA ACT AAG ACA AAG ATA GAG TCT TTG AAA GAG CAT GGC CCT ATC AAA AAT AAA ATG AGC GAA AGT CCC AAT AAA ACA GTA TCT GAG
Asp Lys Thr Lys Thr Lys Ile Glu Ser Leu Lys Glu His Gly Pro Ile Lys Asn Lys Met Ser Glu Ser Pro Asn Lys Thr Val Ser Glu
211 220 230 240

GAA AAA GCT AAA CAA TAC CTA GAA GAA TTT CAT CAA ACG GCA TTA GAG CAT CCT GAA TTG TCA GAA CTT AAA ACC GTT ACT GGG ACC AAT
Glu Lys Ala Lys Gln Tyr Leu Glu Glu Phe His Gln Thr Ala Leu Glu His Pro Glu Leu Ser Glu Leu Lys Thr Val Thr Gly Thr Asn
241 250 260 270

CCT GTA TTC GCT GGG GCT AAC TAT GCG GCG TGG GCA GTA AAC GTT GCG CAA GTT ATC GAT AGC GAA ACA GCT GAT AAT TTG GAA AAG ACA
Pro Val Phe Ala Gly Ala Asn Tyr Ala Ala Trp Ala Val Asn Val Ala Gln Val Ile Asp Ser Glu Thr Ala Asp Asn Leu Glu Lys Thr
271 280 290 300

ACT GCT GCT CTT TCG ATA CTT CCT GGT ATC GGT AGC GTA ATG GGC ATT GCA GAC GGT GCC GTT CAC CAC AAT ACA GAA GAG ATA GTG GCA
Thr Ala Ala Leu Ser Ile Leu Pro Gly Ile Gly Ser Val Met Gly Ile Ala Asp Gly Ala Val His His Asn Thr Glu Glu Ile Val Ala
301 310 320 330

CAA TCA ATA GCT TTA TCG TCT TTA ATG GTT GCT CAA GCT ATT CCA TTG GTA GGA GAG CTA GTT GAT ATT GGT TTC GCT GCA TAT AAT TTT
Gln Ser Ile Ala Leu Ser Ser Leu Met Val Ala Gln Ala Ile Pro Leu Val Gly Glu Leu Val Asp Ile Gly Phe Ala Ala Tyr Asn Phe
331 340 350 360

GTA GAG AGT ATT ATC AAT TTA TTT CAA GTA GTT CAT AAT TCG TAT AAT CGT CCC GCG TAT TCT CCG GGG CAT AAA ACG CAA CCA TTT CTT
Val Glu Ser Ile Ile Asn Leu Phe Gln Val Val His Asn Ser Tyr Asn Arg Pro Ala Tyr Ser Pro Gly His Lys Thr Gln Pro Phe Leu
361 370 380 390

CAT GAC GGG TAT GCT GTC AGT TGG AAC ACT GTT GAA GAT TCG ATA ATC CGA ACT GGT TTT CAA GGG GAG AGT GGG CAC GAC ATA AAA ATT
His Asp Gly Tyr Ala Val Ser Trp Asn Thr Val Glu Asp Ser Ile Ile Arg Thr Gly Phe Gln Gly Glu Ser Gly His Asp Ile Lys Ile
391 400 410 420

ACT GCT GAA AAT ACC CCG CTT CCA ATC GCG GGT GTC CTA CTA CCG ACT ATT CCT GGA AAG CTG GAC GTT AAT AAG TCC AAG ACT CAT ATT
Thr Ala Glu Asn Thr Pro Leu Pro Ile Ala Gly Val Leu Leu Pro Thr Ile Pro Gly Lys Leu Asp Val Asn Lys Ser Lys Thr His Ile
421 430 440 450

TCC GTA AAT GGT CGG AAA ATA AGG ATG CGT TGC AGA GCT ATA GAC GGT GAT GTA ACT TTT TGT CGC CCT AAA TCT CCT GTT TAT GTT GGT
Ser Val Asn Gly Arg Lys Ile Arg Met Arg Cys Arg Ala Ile Asp Gly Asp Val Thr Phe Cys Arg Pro Lys Ser Pro Val Tyr Val Gly
451 460 S-S 470 480

AAT GGT GTG CAT GCG AAT CTT CAC GTG GCA TTT CAC AGA AGC AGC TCG GAG AAA ATT CAT TCT AAT GAA ATT TCG TCG GAT TCC ATA GGC
Asn Gly Val His Ala Asn Leu His Val Ala Phe His Arg Ser Ser Ser Glu Lys Ile His Ser Asn Glu Ile Ser Ser Asp Ser Ile Gly
481 490 500 510

GTT CTT GGG TAC CAG AAA ACA GTA GAT CAC ACC AAG GTT AAT TCT AAG CTA TCG CTA TTT TTT GAA ATC AAA AGC TGA
Val Leu Gly Tyr Gln Lys Thr Val Asp His Thr Lys Val Asn Ser Lys Leu Ser Leu Phe Phe Glu Ile Lys Ser STOP
511 520 530

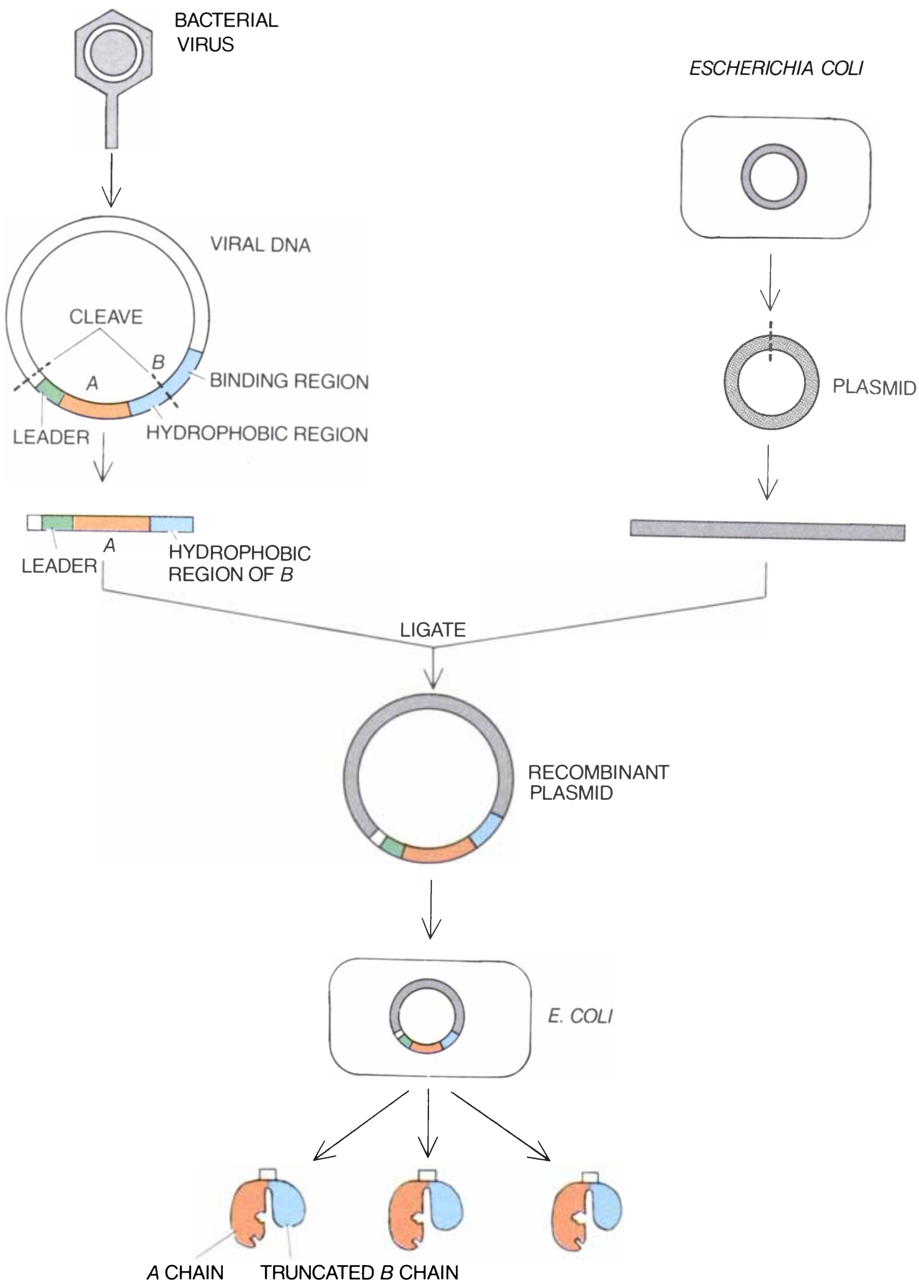
NUCLEOTIDE SEQUENCE of the diphtheria toxin gene was determined and then was translated according to the genetic code to establish the amino acid sequence of the toxin molecule. The protein has a leader, or signal, peptide (*green*) involved in secretion from the bacterium, an *A* chain (*orange*) and a *B* chain (*blue*). The *B* chain has a region rich in hydrophobic amino acids (*dark blue*) that is responsible for the toxin's insertion into the endosomal membrane and therefore for the *A* chain's emergence from an endosome into the cytoplasm. The toxin's receptor-binding site is somewhere in the *B* chain beyond the hydrophobic region. Two disulfide bonds (*S–S*), one of which links the two chains, are indicated. Amino acid 148 (*boxed*) is thought to be in the active site catalyzing ADP-ribosylation of EF-2.

ent's cells as foreign and attack and destroy them.

To avoid this complication one would like to eliminate the *T* cells before introducing the donor's marrow into the patient. Daniel A. Vallera of the University of Minnesota Medical School and his colleagues have treated marrow cells in culture with immunotoxins made by linking intact ricin to antibodies against *T* cells. They added lactose to minimize nonspecific toxicity. The procedure reduced the marrow's *T*-cell population by about 99 percent, with minimal effect on the stem cells. The efficacy of marrow treated in this way is now being tested in patients.

Applications of this kind, in which tissues are treated by immunotoxins outside the body, are attractive as an initial step in demonstrating the effectiveness of immunotoxins. Treating cells outside the body is implicitly safer than administering the immunotoxin directly to a patient, because one can remove any excess immunotoxin by rinsing the cells before injecting them into the recipient.

The long-term challenge, however, is to develop immunotoxins into a new family of chemotherapeutic agents with which to treat patients directly. Tests of immunotoxins have been done in animals, with some reports of tumor regression, or reduction in size. For example, Michael I. Bernhard of the National Cancer Institute and his colleagues treated liver carcinoma in guinea pigs with an immunotoxin consisting of the diphtheria toxin *A* chain linked by a disulfide bond to a monoclonal antibody directed against an antigen on the cancer cells. A single dose of the immunotoxin led to tumor regression but did not completely eradicate the tumor.



MODIFIED DIPHTHERIA TOXIN GENE has been cloned in bacteria and translated into protein. The source of the toxin gene is a bacterial virus that infects *Corynebacterium diphtheriae.* The viral DNA is extracted (*top left*) and a fragment encoding the *A* chain and the hydrophobic region of the *B* chain, but not the receptor-binding site, is isolated. The selected fragment is inserted into a plasmid: a small circular piece of nonchromosomal DNA taken from the bacterium *Escherichia coli.* The recombinant plasmid is introduced into *E. coli* cells. The bacteria make large quantities of the foreign DNA (which can be sequenced) and also translate the gene fragment to make a modified toxin molecule lacking the *B* chain's receptor-binding site. John R. Murphy of the Harvard Medical School and his colleagues recently linked a mutant toxin molecule that is similarly truncated to a small hormone and found it had high toxic activity.

Much research still needs to be done: on antibodies, on toxic agents and on methods of treatment. Many laboratories are now at work isolating monoclonal antibodies against various human cancers. Almost certainly a large number of different immunotoxins will have to be developed whose antibody portion is specific for various tumor-associated antigens on different cancer cells. As we mentioned above, a radioactive isotope or one of the conventional chemotherapeutic agents could be coupled to these antibodies instead of a modified natural toxin. It is also possible that antibodies will be found that by themselves can initiate an attack by the patient's own immune system against cancer cells.

It is likely that genetic engineering will play an increasing role in the development of immunotoxins. Just as one can eliminate the receptor-binding region of the *B* chain by manipulating the diphtheria toxin gene, so one can modify toxin genes in other ways to improve the efficacy and safety of toxin molecules. Eventually it may be possible to make the entire immunotoxin by genetic-engineering techniques: to isolate the gene for a wanted monoclonal antibody, perhaps modify it to improve its affinity for a particular antigen, link it to an appropriate toxin gene and let bacteria or such eukaryotic cells as yeasts synthesize the immunotoxin as a single construct.

In the human body an immunotoxin will be subjected to an environment much more complex than that of a cell culture. To be effective an immunotoxin must remain stable as it moves through the circulatory system, must be able to gain access to target cells in many parts of the body and of course must not damage normal tissues that are crucial to survival. The degree to which various immunotoxins satisfy these requirements is currently under study. The task of alleviating or curing cancer by chemotherapy is a formidable one. Probably no single approach will suffice. Immunotoxins may become one of many weapons in the armamentarium.

Science under Wraps

In the past year several Federal agencies have proposed policies aimed at restricting the publication of research results from U.S. universities that could be of military benefit to the U.S.S.R. The Department of Commerce has drafted regulations that could require a scientific worker to obtain an export license before presenting a paper at a meeting where citizens of countries other than the U.S. are present. The Department of Defense is considering the adoption of rules that would enlarge the power of Government officials to block publication of the results of research done under contract to the department.

The argument offered to justify the proposed restrictions is that depriving the Russians of advances in fields such as computer technology and materials science strengthens the capacity of the U.S. to defend itself and undermines the defensive capacity of the U.S.S.R. The costs to the U.S. and its allies of restricting scientific communication are rarely considered in detail. Recent appraisals suggest that whereas the efficacy of restrictive policies in hindering Russian military development is not well established, the potential for damage to science in the West is considerable.

Under the current Export Administration Regulations, which are administered by the Department of Commerce, scientific data are largely exempt from export controls. The department is considering changes in the regulations that would eliminate the exemption and require a license for the export of "critically technical data." The definition of export in the draft regulations includes the presentation of papers at scientific meetings attended by foreign nationals. The definition also includes the hiring of a person who is not a U.S. citizen as a research worker on a project that could yield "critical technical data."

The Department of Defense Steering Committee on National Security and Technology Transfer has recommended that the department adopt new review procedures for contract research. Papers describing work in some areas would have to be submitted to the contract officer 90 days before being submitted to a journal for publication. Officials of the Department of Defense would be empowered to demand changes in the manuscript or to forbid its publication. The regulations would apply not to classified research (which in general cannot be published) but to a new category of nonclassified "sensitive" work. In April the presidents of Stanford University, the California Institute of Technology and the Massachusetts Institute of Technology told the department that their institutions could not accept research contracts under the proposed review procedure.

Neither the Department of Commerce regulations nor the Department of Defense review procedure has been put into effect; the agencies are currently operating under older or interim policies. According to Mitchel B. Wallerstein of the National Research Council, the recommended changes have nonetheless had a chilling effect on the exchange of research results.

As a member of the staff of the council Wallerstein helped to prepare a report issued in 1982 on the military consequences of exporting technical information. The report was that of a panel sponsored by the National Academy of Sciences, the National Academy of Engineering and the Institute of Medicine and chaired by Dale R. Corson, president emeritus of Cornell University. Wallerstein said recently that because of the current political climate, "many investigators are not sure of their rights and obligations and so they are being cautious about publication."

In addition there has been a rapid increase in the number of U.S. scientific meetings that are closed to citizens of other countries. Jacques Bodelle, head of the French scientific mission to the U.S., said recently in *Science & Government Report:* "There were very few of these restricted meetings four or five years ago. Now there are more—meetings in electronics and materials science. We have difficulties attending some of the meetings on the frontiers of these subjects."

The chilling of scientific communication is the most immediate consequence of the attempt to restrict publication. According to Wallerstein, however, subtler, long-term effects could be even more damaging. In an update of the Corson report published in *Science* Wallerstein notes that the fraction of graduate students and research workers in the U.S. who are not U.S. citizens has increased rapidly in many scientific fields. Excluding them from work in those fields would cause large perturbations in the supply of scientific manpower. Indeed, the proportion of foreign-born workers is greatest in the disciplines on which military technology relies the most. Wallerstein states: "Perhaps the largest risk in this regard is the long-term changes that controls are likely to cause in the demographic distribution of scientists and engineers among the various disciplines and subfields."

An additional consequence of the cool ideological climate is the reduction of contact with Russian science. A study done by Andrew M. Sessler and Rita La Brie of the Lawrence Berkeley Laboratory shows that between 1975 and 1981 there was a sharp decrease in the frequency with which Russian scientific publications were cited in Western European and U.S. journals. Sessler and La Brie suggest that one factor in the decrease could be the decline in face-to-face contact between American and Russian scientific workers.

It should be noted that not all of the reduction in contact is due to actions taken by the Reagan Administration. In 1982 the exchange agreement under which thousands of Russian scientists had visited the U.S. every year was allowed to lapse by the National Academy of Sciences. The academy's action was in part a protest against the treatment of politically dissident scientists in the U.S.S.R. Efforts are under way to revive the agreement, but in the meantime the number of exchanges has fallen to a few hundred per year.

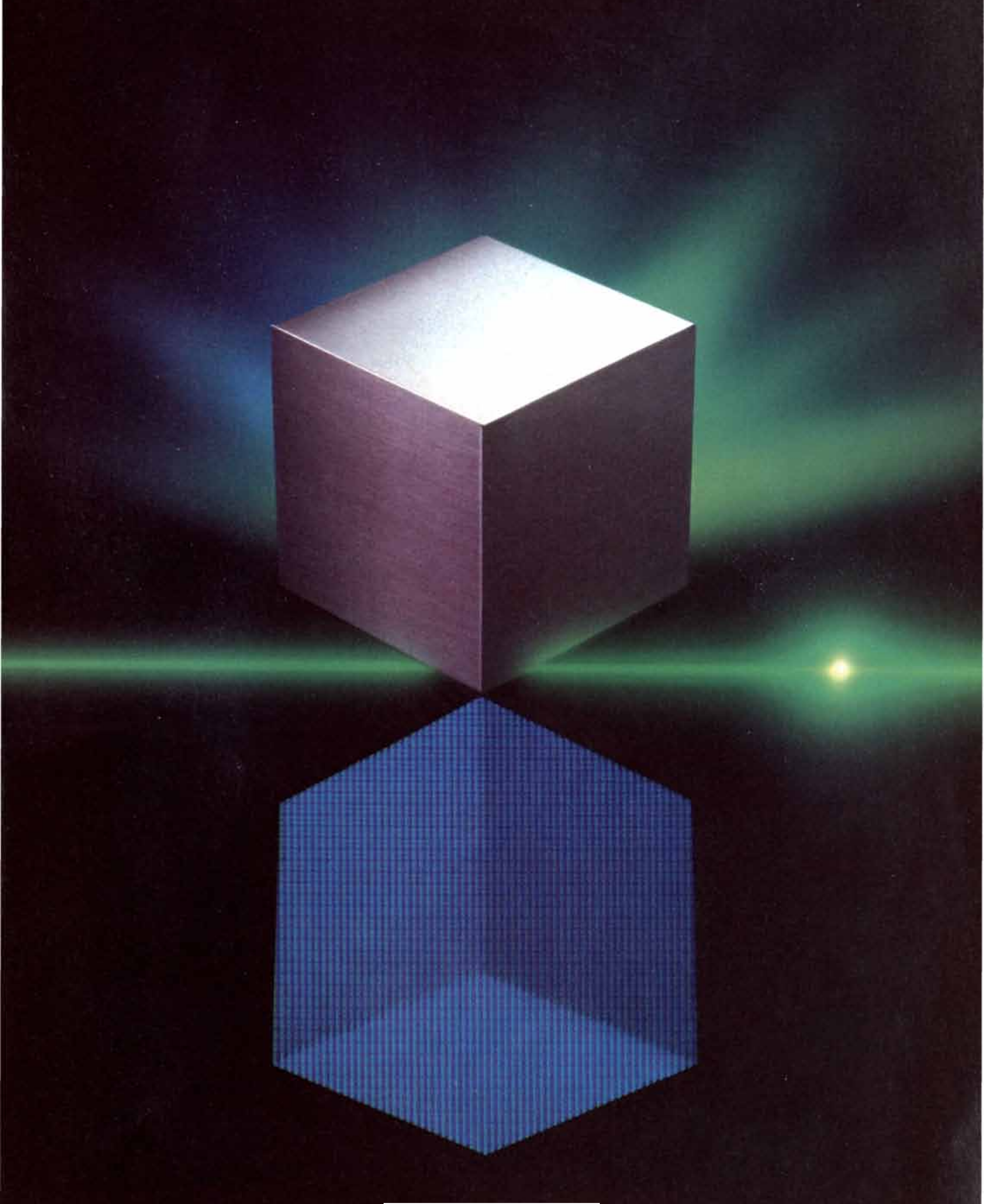
One of the ironies of the current situation is that in spite of the potentially serious consequences of interrupting scientific exchange, there is little evidence that scientific or technical information reaching the U.S.S.R. by such routes has helped the Russian military effort. The Corson panel "found no case of significant damage to security associated with research dissemination." In the update of the panel's report Wallerstein concludes that since the original report was published "no major initiative has been undertaken to better characterize the relative importance of sources, channels or types of information that leaks out or the relative significance of scientific communication within the large picture. The intelligence community reports no cases during this period in which loss through the U.S. scientific community has led to identifiable damage to national security."

Double Agent

The recent announcement by workers at the National Cancer Institute that the virus responsible for AIDS (acquired immunodeficiency syndrome) has been identified was attended by indications of the announcement's social and political importance as well as its scientific significance. Conflicting rumors about the finding circulated before the formal announcement; the Secretary of Health and Human Services and her lieutenants turned up at the news conference to claim credit for the Government's scientific establishment and an international controversy over scientific precedence was fanned by the press.

Publication of the substantive reports

The Robot Abstraction



The Robot Abstraction

Until now, matching robots to specific industrial tasks has been done by trial and error, requiring the creation of expensive prototypes. Recent advances at the General Motors Research Laboratories have produced a computer system that can be used not only to select the right robot, but also to program it to perform the task in the most efficient way.

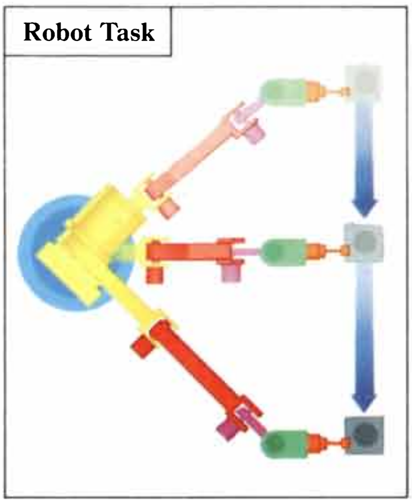


Figure 1: Two-dimensional overhead view of a robot task—the straight path trajectory of a solid.

Figure 2: Three-dimensional illustration of the robot work cell layout, showing reach capability for the task in Figure 1. Areas of color show total reach as well as the joint limits stored in the robot model.

THE DECISION to use robots to automate a manufacturing facility introduces the need for more decisions. There are several dozen kinds of robots, each with different capabilities. Thus far, choosing the right robot for a given set of tasks has been largely a manual process, involving great expenditures of time and money. By combining previously separate disciplines in a single computer system, two General Motors researchers have made the introduction of robots to the factory floor a more rational, less costly undertaking.

RoboTeach is the first computer system which integrates robotics, solid modeling, and simulation. It was designed and developed by Dr. Robert Tilove and Mary Pickett, both members of the Computer Science Department.

The use of powerful programming languages for manipulating robots is a major new development in the discipline of robotics. The languages specify desired robot motions, but they have no way of describing the robot's environment. Hence, they cannot automatically take into account physical obstacles or anticipate collisions. With only robot programming languages at one's disposal, assuring proper interaction with the environment requires testing with actual robots and parts.

Solid modeling, on the other hand, provides geometrically complete representations of environmental components and their spatial relations. But solid modeling cannot represent processes, because it has no way of representing temporal relations. Traditional solid modeling deals only with static relationships. While robot programming is without physical context, solid modeling is nothing but physical context. Neither by itself is adequate.

Nor are they satisfactory together. Only by simulation of both the robot and its environment can the sequence of discrete steps in a robot task be converted into the continuous motion of a process. Also without simulation, there is no way to represent accurately the robotic process as it unfolds in its environment.

RoboTeach, by combining all three disciplines, provides computer representations of the environment, the robot, and the task. Consequently, it helps users reach high-level decisions about the real world without the investment of time or money in actual robots, actual parts, or the factory setting.

One key RoboTeach abstraction is a mathematical robot model. Solid modeling techniques represent the geometric form of each link of the robot. Then constraints are imposed on the relative positions of mating links to produce a mathematical abstraction of a mechanical joint. By insisting that the joint constraints always be satisfied, RoboTeach insures that the abstract robot model corresponds to a physically realizable geometric configuration.

OTHER representational facilities in RoboTeach handle robot task definitions. The representation of any task can be matched with the representation of any robot. In this way, RoboTeach helps users to determine the optimal robot for the task. Once a robot has been selected, RoboTeach can be used to program the robot off-line.

Not only are robots proliferating, but the tasks assigned to them are becoming more complex, making the need for off-line programming more urgent. When there are only a half dozen robots in a factory, the prospect of reprogramming them all by conventional show-and-teach methods for every new task is not overwhelming. But when there are hundreds of robots, the value of being able to reprogram without interaction with each robot becomes more apparent. Without off-line programming, the savings which justified the initial robot investment may quickly vanish.

RoboTeach distinguishes between two kinds of off-line programming: at the task level (what to do) and at the robot level (how to do it). For example, in the creation of a mechanical assembly, task-level instructions would include what components to assemble, the alignment of the components for the assembly process, and criteria for verifying that the final assembly is correct. Typically, there is a one-to-many relationship between task-level instructions and robot-level instructions.

"RoboTeach is currently in use," says Robert Tilove, "to study robot reach capabilities and to simulate simple robot-level tasks."

"Future research," adds Mary Pickett, "will explore the possibility of using RoboTeach to approach problems from the more abstract task level, with the user defining the task at a high level and RoboTeach filling in the details."

General Motors

GM MARK OF EXCELLENCE

THE PEOPLE BEHIND THE WORK

Dr. Robert Tilove and Mary Pickett are Staff Research Scientists in the Computer Science Department at the General Motors Research Laboratories.

Mary Pickett received her B.S. in mathematics from Iowa State University and her Master's in computer science from Purdue University. She was a member of the team that developed GMSOLID, an interactive geometric modeling system. Her research at GM has also included the design of real-time programming languages. She joined GM in 1971.

Robert Tilove received his undergraduate and graduate degrees in electrical engineering from the University of Rochester. His Ph.D. thesis concerned the design and analysis of geometric algorithms for solid modeling. His current research interests also include the application of geometric modeling to computer vision and robot control. He joined GM in 1981.

in *Science* showed that an NCI team headed by Robert C. Gallo has indeed brought to successful culmination a painstaking series of investigations by isolating a new virus called HTLV-III and linking it convincingly with the causation of the disease. The fact that investigators at the Pasteur Institute in Paris had apparently isolated the same virus a year earlier but had not assembled the same kind of evidence that it causes AIDS did not detract from the NCI achievement—or, indeed, from the French one.

The sudden outbreak of a devastating disease that has not been reported before and whose cause is unknown is a rarity in an advanced 20th-century society. AIDS is such a disease. In the U.S. it first appeared early in 1981, and since then some 5,000 cases have been reported, almost 45 percent of which have ended in death. The disease came into view as a syndrome, or group of symptoms, whose common source is a major defect in the immune response. Patients show severe depletion of *T* lymphocytes and specifically of *T* helper cells, which promote the secretion of antibodies by *B* lymphocytes in response to foreign antigens. Victims are afflicted by multiple infections, often from low-grade pathogens that would not be dangerous for people with an intact immune system.

At first AIDS seemed to strike only promiscuous homosexual or bisexual men. They continue to be by far the most grievously affected group, but others at risk include their female sexual partners, Haitian immigrants to the U.S. and immigrants from equatorial Africa to Europe, users of intravenously administered drugs, hemophiliacs treated with preparations made from pooled blood, recipients of repeated blood transfusions and infants born of parents in a high-risk group. AIDS itself is often preceded by pre-AIDS, a "prodrome" characterized by fatigue and lymph-gland disease.

The culprit virus was pinpointed by a line of investigation undertaken 14 years ago. It had just been shown that the viruses responsible for certain animal leukemias and other cancers are retroviruses: their genetic material is RNA, which is "reverse-transcribed" into DNA in an infected animal cell by a viral enzyme, reverse transcriptase. At the NCI, Gallo and his colleagues began to develop techniques for detecting retroviruses that might be associated with human leukemias and lymphomas. Their discovery of *T*-cell growth factor, a protein promoting long-term growth of *T* lymphocytes in a culture medium, and their development of a sensitive assay for reverse transcriptase made it possible to identify a family of human T-cell leukemia/lymphoma viruses (HTLV), the first human retroviruses and the first viruses apparently responsible for a human cancer. HTLV-I turned out to be associated with a *T*-cell malignancy endemic in parts of southern Japan, the Caribbean and Africa. A related virus was designated HTLV-II.

Gallo and his colleagues thought a member of the HTLV family might well be a causative agent of AIDS. There were a number of good reasons. The viruses tend to infect *T* cells, and in particular the helper cells. Another retrovirus, feline leukemia virus, suppresses the immune response in addition to causing leukemia of *T*-cell origin. HTLV-I was known to be prevalent in the Caribbean (particularly in Haiti) and in Africa and sometimes to be spread by intimate contact or by blood products.

Last year Gallo's group reported evidence for the presence of an HTLV in a few patients with AIDS; so did Myron Essex and his colleagues at the Harvard School of Public Health. The evidence was not strong enough to suggest causation, however. It seemed to Gallo and his colleagues that some HTLV variant whose main effect is to damage *T* cells instead of making them immortal might be more likely than HTLV-I or II to cause AIDS. They got glimpses of such a variant, but the glimpses were inevitably transient because the affected cells died before a cytopathic virus could be isolated. Then Mikulas Popovic and his co-workers in Gallo's group found a line of malignant *T* cells that could be infected with the new cytopathic variant but would survive and proliferate, providing an immortal cell population in which the cytopathic virus could be propagated in quantity.

The new virus was shown to be a distinguishably different member of the HTLV family, and so it was designated HTLV-III. Like HTLV-I and II it infects *T* helper cells, but it kills them instead of causing them to proliferate. In micrographs it looks much like the other two HTLV's, but its dense core is distinctively cylindrical.

The investigators went on to determine the prevalence of HTLV-III in AIDS patients. The virus was found in 18 of 21 blood-cell samples from patients with pre-AIDS and in 26 out of 72 samples from AIDS patients. The prevalence of the virus may be understated because many samples were received in poor condition and because the *T*-cell population was depleted in samples from late-stage patients. The virus was detected in only one of 22 samples from healthy, nonpromiscuous homosexual males; six months later the one man who tested positive developed AIDS.

Even when a virus cannot be isolated, a person can be shown to have been exposed to it by the presence in his blood of antibodies elicited by viral antigens. The Gallo group found antibodies to HTLV-III in blood serum from 87.8 percent of a group of patients with AIDS and 78.6 percent of pre-AIDS patients.

From their data Gallo and his colleagues conclude that "HTLV-III is the primary cause of AIDS." The French data seem to point to the same conclusion, because it is very likely that the virus they found and HTLV-III are the same, or at least are members of closely related subgroups. The issue of precedence has apparently been overstated. The French workers, headed by Luc Montagnier and Jean-Claude Chermann, were the first to identify the new virus but could not lay claim to having established its causative role; the Americans published a year later, but they had propagated the virus, characterized it in detail and found the virus and antibodies to it in many more patients.

In 1983 the Pasteur Institute group reported in *Science* on the isolation, from a patient with pre-AIDS, of what looked like a new virus and seemed unrelated to HTLV-I. The French called it lymphadenopathy-associated virus (LAV). They have since found LAV in most AIDS patients tested; other French investigators have found antibodies to the virus in a high proportion of the pre-AIDS patients they tested, in about a third of AIDS patients and in only one out of 130 healthy blood donors.

The discovery of an infectious agent is only one step toward the control of a disease. The first major benefit to flow from the identification of HTLV and its propagation in quantity will be the ability to test blood for antibodies to the virus. There is also the prospect of developing a prophylactic vaccine to stimulate the formation of antibodies against the virus. It is not yet clear, however, whether a strong enough antibody response can be induced to cope with a virus whose primary target is the immune system itself.

Bound Bosons

A proposed new accelerator, the Superconducting Super Collider (SSC), is the topic of much discussion among physicists who study elementary particles. The machine would be a ring roughly 100 kilometers in circumference where physical processes could be investigated at 40 times the energy supplied by the largest existing accelerators. It would also take at least 10 years to build and would consume a sizable fraction of the nation's budget for basic research. Is the payoff worth the cost? In recent months several phenomena that cannot be explained by accepted theory have been observed at the European Laboratory for Particle Physics (CERN). The phenomena suggest that the accelerators currently in operation are already beginning to glimpse an intricate new structure in the fundamental forces of nature that could be explored by more powerful accelerators.

The suggestive experimental data

from CERN are still quite sparse: only about 15 unexplained events have been recorded so far. The events arise from high-energy collisions of protons and antiprotons circulating in opposite directions in a ring. It was in this way that the long-sought W^+, W^- and Z^0 particles were found at CERN last year.

The anomalous events are of at least four kinds. At the UA1 detector, which records the products of collisions at one point on the CERN ring, a reaction has been observed in which a highly collimated jet of particles carries energy away from the collision site. Because momentum is conserved in the experiment, the jet must be balanced by the emission of an equal amount of energy in the opposite direction, presumably carried by neutrinos or other neutral particles that are not detected. The total energy released is as high as 120 GeV (billion electron volts), which is much more than would be expected from the decay of any particle discovered so far.

A second reaction has been seen at both the UA1 detector and the UA2 detector, installed at another point on the ring at CERN. The reaction leads to the emission of a high-energy photon and a pair of leptons, either an electron and a positron or a positive muon and a negative muon. The emission of the two leptons is one of the accepted signatures of the decay of the W and Z particles, but in the standard theory photon emission should be quite rare. Actually, in two events observed at the UA1 detector a single photon with an energy of about 50 GeV was emitted without an accompanying pair of leptons. The events are difficult to interpret in the standard theory, but the energy released is large enough for the photon to have been emitted through some unforeseen decay channel of the Z^0.

The third kind of reaction is also notable because of its frequency. The decay of the Z^0 can give rise to jets of particles other than electrons and positrons. The theory predicts that with each additional jet the probability of the reaction decreases by a factor of 10. Although fewer than two dozen Z^0 decays have been observed, in some of them the decay products form as many as five jets.

The fourth kind of event has been seen only at the UA2 detector. The decay gives rise to an electron, a neutrino and as many as three jets of particles. The total energy of the decay products is as much as 160 GeV, which is too great for the decay of a single W or Z but too little for the decay of two such particles.

Could the four kinds of interaction reflect a single underlying phenomenon? There is no all-encompassing explanation so far, but theorists have made several preliminary calculations suggesting links between the new results and earlier experimental anomalies. The ideas go to the heart of the unified theory of the electromagnetic force and the weak force. Both forces are mediated by particles called bosons: the electromagnetic force is carried by the photon, which has no mass, and the weak force is carried by the very massive W^+, W^- and Z^0 particles. According to the standard version of the theory, the fact that the bosons differ greatly in mass is only an apparent asymmetry between the two forces, and it should be manifest only at relatively "low" energy. At extremely high energy the two forces should be indistinguishable.

The predicted symmetry of the unified force at high energy requires a symmetry-breaking mechanism whereby the weak bosons can acquire a mass. The standard theory postulates an undiscovered particle called the Higgs boson to account for the mass. It has long been thought that if the Higgs boson itself is extremely massive, there could be an interaction between the W and the Z particles that is strong enough to bind two or more of them together. The nature of such a force is almost totally unknown, but it is argued that the more massive the Higgs boson is, the stronger the force must be.

Martinus Veltman of the University of Michigan has now suggested that many of the anomalous events at CERN mark the decay of a bound state of a pair of W or Z particles. For example, the 160-GeV events recorded by the UA2 detector could be caused by the decay of such a bound pair, because the energy of the bound state is lower than that of two unbound particles. Veltman proposes that the decay of such a bound system could also be responsible for the unexpectedly high rate at which pairs of leptons (electrons, positrons and muons) in which both members of a pair have the same electric charge are generated in high-energy experiments. The lepton pairs have been observed for several years but have never been explained. David B. Cline of the University of Wisconsin at Madison has also suggested that a pair of leptons with the same charge signals the existence of a new particle.

According to Veltman, if the anomalous events do represent evidence of a bound pair of weak particles, the Higgs boson must be much more massive than 1,000 GeV. In fact, Veltman argues, the Higgs particle must be so massive that it can be effectively dispensed with in the unified theory.

Other theorists are not yet ready to grant that the recent observations have any direct bearing on the Higgs boson. Chris Quigg of the Fermi National Accelerator Laboratory and Thomas Appelquist of Yale University both stress that the case for a bound state of weak particles is far from proved. They point out that other explanations are possible and might be explored with existing accelerators or those now being built. If the answer does lie in the Higgs boson, Quigg and Appelquist suggest the mass could be so large that Higgs effects do not become manifest at energies much below 1,000 GeV. A definitive experimental investigation of Higgs particles would therefore require energies available only at the proposed SSC accelerator.

Hot Rocks

The pebble-bed reactor is a standard device of industrial chemistry in which fluids percolate through a loosely packed matrix of solid material. A pebble-bed reactor of another kind entirely is about to go into commercial operation in West Germany. It is a nuclear power reactor whose novel design addresses some of the questions of safety and economy that have virtually becalmed the U.S. nuclear power industry.

The new reactor is known formally as the thorium high-temperature reactor, or THTR 300. The concept was developed by Hochtemperatur-Reaktorbau (HRB) in Mannheim. The reactor is gas-cooled, employing helium to transfer heat from the reactor core to a steam generator. The fuel is a mixture of uranium oxide and thorium oxide formed into small spherical particles coated with carbon; the particles in turn are packed into 60-millimeter spheres (the pebbles). A pebble contains about 30,000 fuel particles embedded in a graphite matrix and surrounded by a five-millimeter shell of graphite. A reactor core contains about 675,000 pebbles.

The designers see safety as the principal contribution of the combination of pebble fuel and gas coolant. The graphite in which the fuel is sealed has a high heat capacity. In a mishap such as a failure of the cooling system the pebbles' high capacity for absorbing heat would slow the temperature rise in the reactor core, giving the operators time to decide how to deal with the problem. Moreover, because the circulation of the coolant gas is entirely within the reactor vessel, the reactor cannot lose its coolant owing to a rupture of pipes outside the vessel.

The pebble configuration also allows for continuous refueling of the reactor. Fuel elements can be discharged by gravity through the conical bottom of the reactor vessel while the reactor is operating. Spent or damaged pebbles are removed; the rest are loaded back into the core along with new ones.

The pebble design was tested in a small experimental reactor at the West German nuclear research facility in Jülich some years ago. Now the design has been scaled up to a prototype commercial plant, a 300-megawatt facility at Schmehausen, near Dortmund. The plant is in the final stages of testing and is scheduled to begin delivering electricity to the national grid next spring.

Under a contract with 16 public utilities in West Germany, HRB has designed a 500-megawatt pebble-bed plant. If it is approved for construction, it could be in operation within a decade.

Almost a Star

An extremely faint object thought to be traveling in the disk of our galaxy only 28 light-years from the sun may be the first recorded example of a "brown dwarf": a planet-size ball of gas whose gravitational contraction generates enough heat to make it glow dimly but whose mass is insufficient to ignite the thermonuclear fire that causes all true stars to shine. A borderline object of this kind would be expected to emit radiation predominantly in the infrared part of the spectrum, which is precisely what is observed in the present case. The brown-dwarf candidate, designated LHS 2924, is the coolest and least luminous condensed object ever detected outside the solar system. Its tentative identification as a substellar brown dwarf has important implications not only for theories of star formation but also for the still unanswered cosmological question of whether or not the universe will expand forever.

The "peculiar spectrum" of LHS 2924 was first reported last year in a joint paper in *The Astrophysical Journal* by Ronald G. Probst of the Kitt Peak National Observatory and James Liebert of the University of Arizona's Steward Observatory. The two investigators had been engaged with their respective colleagues in separate searches for unusual dwarf stars among the thousands of low-luminosity white and red dwarfs listed in published surveys of stellar objects with a large proper motion (that is, motion across the line of sight). Because of their rapid angular displacement, all such objects are presumed to be nearby. The attention of both groups of astronomers was drawn to LHS 2924, a "uniquely cool" dwarf. The designation LHS 2924 refers to the object's listing in the Luyten Half Second catalogue, compiled by William J. Luyten of the University of Minnesota, which includes only objects with a proper motion of at least .5 second of arc per year.

According to Probst and Liebert, their initial spectroscopic and photometric studies of LHS 2924 showed it to be "a very red object" with an effective surface temperature of approximately 1,950 degrees Kelvin. This preliminary temperature reading, they noted, indicated that LHS 2924 is "significantly cooler" than any previously known red dwarf. The coolest objects in this class (called late M dwarfs) are thought to be low-mass stars that consume their hydrogen fuel very slowly. Although a later measurement gave a new estimate of 2,450 degrees, the object "remains... the coolest red dwarf known."

The mystery deepened when a comparison of certain atomic and molecular absorption lines in the spectrum of LHS 2924 with those in the spectra of other low-luminosity red dwarfs revealed "pronounced" differences. After eliminating several alternative explanations for the "confusing, contradictory and unprecedented" features of the spectrum, Probst and Liebert concluded that two hypotheses remained: the object could be an extraordinarily cool M dwarf nearing the end of the hydrogen-burning phase of its lifetime, or it could be a young brown dwarf of substellar mass that is destined never to reach the temperature needed to initiate thermonuclear reactions.

The potential cosmological significance of LHS 2924 arises from the fact that the critical density of matter needed to reverse the current expansion of the universe is much greater than the total mass of the known luminous bodies. Hence the question of whether or not the universe will expand forever turns on estimates of how much matter is distributed throughout space in a nonluminous or extremely dim form. The discovery of a nearby object with the peculiar spectral properties of LHS 2924 raises the possibility that interstellar space may be populated with countless similar objects, thus accounting for much of the universe's missing mass.

The Sign of the Beast

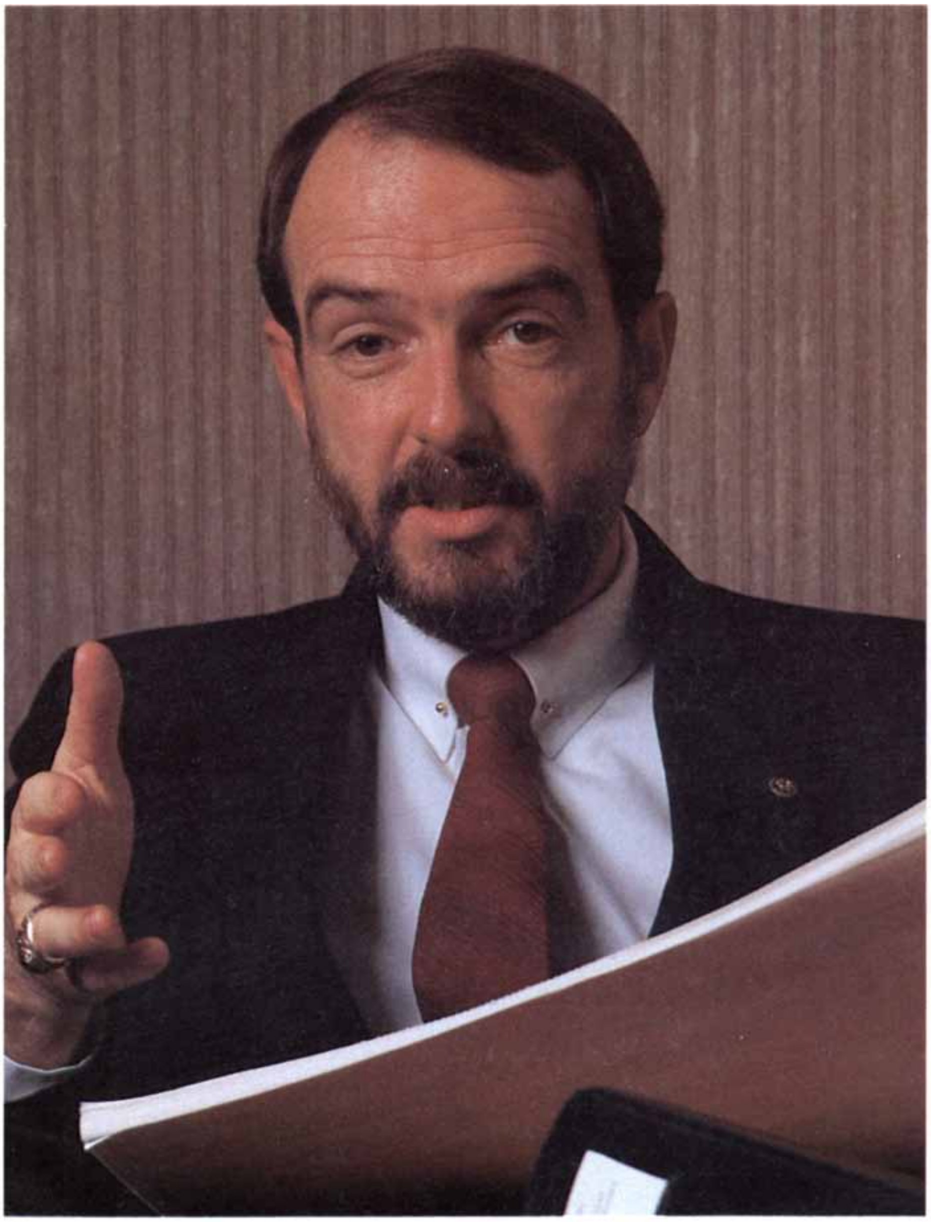
The modern classification of animals has been in progress for roughly two centuries; it goes under the name of zoology, and its findings derive in large measure from the collection and study of specimens. In 1955 a contrasting endeavor emerged. Called cryptozoology, it takes as its realm the study of hidden animals, which distinguishes it both from zoology, the study of manifest animals, and from paleontology, the study of extinct ones. By hidden animals the cryptozoologist means not only animals thought to be extinct that are later found alive but also animals known only from peoples' claims that the creatures have been sighted. The exemplars of cryptozoology range, therefore, from *Latimeria* (a hidden deep-sea fish first netted in 1938 off Africa) to the Loch Ness monster (a hidden dinosaur?) and the Himalayan yeti (a hidden Neanderthal man?). The term cryptozoology was coined by Bernard Heuvelmans in his book *On the Track of Unknown Animals*. There is also an International Society of Cryptozoology (its president is Heuvelmans) and a journal, *Cryptozoology*.

Writing in *Proceedings of the American Philosophical Society,* the paleontologist George Gaylord Simpson examines the cryptozoological premise that animals can hide in the world today. Simpson begins with an analysis of mammals discovered in this century. He consults two compilations "of all known living orders, families, genera and species of mammals." He notes that "any mammal not obviously related to some previously known taxon would be at least a new genus." He finds that 126 mammalian genera have been discovered in this century. Two-thirds of them (84) were discovered before 1930. Moreover, three-fourths of them (91) belong to just two orders: Chiroptera, or bats (23), and Rodentia, or rodents (68). Within the order Primates, where cryptozoological hopes run high, a mere two genera have emerged in the 20th century: the South American marmoset *Callimico* in 1904 and the African monkey *Allenopithecus* in 1907.

Simpson turns next to mammals first known from fossil evidence and later found to be living. They belong to only four genera. Fossil evidence of *Burramys* was found in Australia at the end of the 19th century; it was a mouse-size marsupial. In 1966 a living *Burramys* was found in a ski hut high in the Australian Alps. *Burramys* is now assigned to the family Burramyidae. Fossil evidence of *Catagonus* was found in Argentina early in the 20th century. In 1975 the animal itself was found alive on the Gran Chaco plain of South America. It is a type of peccary "well known to native hunters for uncounted generations." Fossil evidence of *Aproteles,* a fruit bat, was found in New Guinea in 1974. In 1975 the search of a cave in New Guinea revealed "remains of living bats shot by natives with bow and arrow." The genus may be extinct now. Fossil evidence of *Speothos* was found in Brazil early in the 19th century. By the end of the century the animal was known to be alive. It belongs to the family Canidae, or dogs.

In every case where an animal thought to be extinct has turned up among the living the fossil evidence has been no older than the late Pleistocene, which is to say no older than a few tens of thousands of years. Moreover, in every case the fossils, and then the living animals, proved to be from families with which zoologists are familiar. In short, Simpson writes, "there is no significant time gap and no discovery of a truly unknown species, genus or family." Simpson finds it not surprising that most of the living mammals discovered in recent years have been "small animals in unusual and remote environments." He assigns "some reasonable probability" to the future discovery of living mammalian species. The chance of finding new mammalian genera he thinks is slighter; the chance of finding new mammalian families he thinks is "very slight."

WHAT SHOULD YOU WORKING PARTNERSHIP



"At Pattie A. Clay Hospital an HP 3000 computer system improves cash flow and cuts data processing costs by 50 percent.

"With HP's Patient Accounting program we handle patient registration six times faster than before. And HP's error checking feature prevents major errors from entering the system."

Pattie A. Clay Hospital in Richmond, Kentucky, is using an HP 3000 computer system and HP's Hospital Financial Management software. Administrator Richard M. Thomas says: "By eliminating outside data processing services, the HP system enables us to control information management while cutting data processing costs by 50 percent.

"We no longer have to wait three weeks for monthly financial reports. With general ledger online, we have an accurate picture of our financial status at all times. Because the HP system bills patients and insurance companies simultaneously, we reduce paperwork and overhead costs by $25,000 a year.

"HP's support team provided solutions to every obstacle we encountered. They trained us and saved us the cost of hiring applications specialists. The system works for us; we don't work for the system."

EXPECT FROM A
WITH HEWLETT-PACKARD?

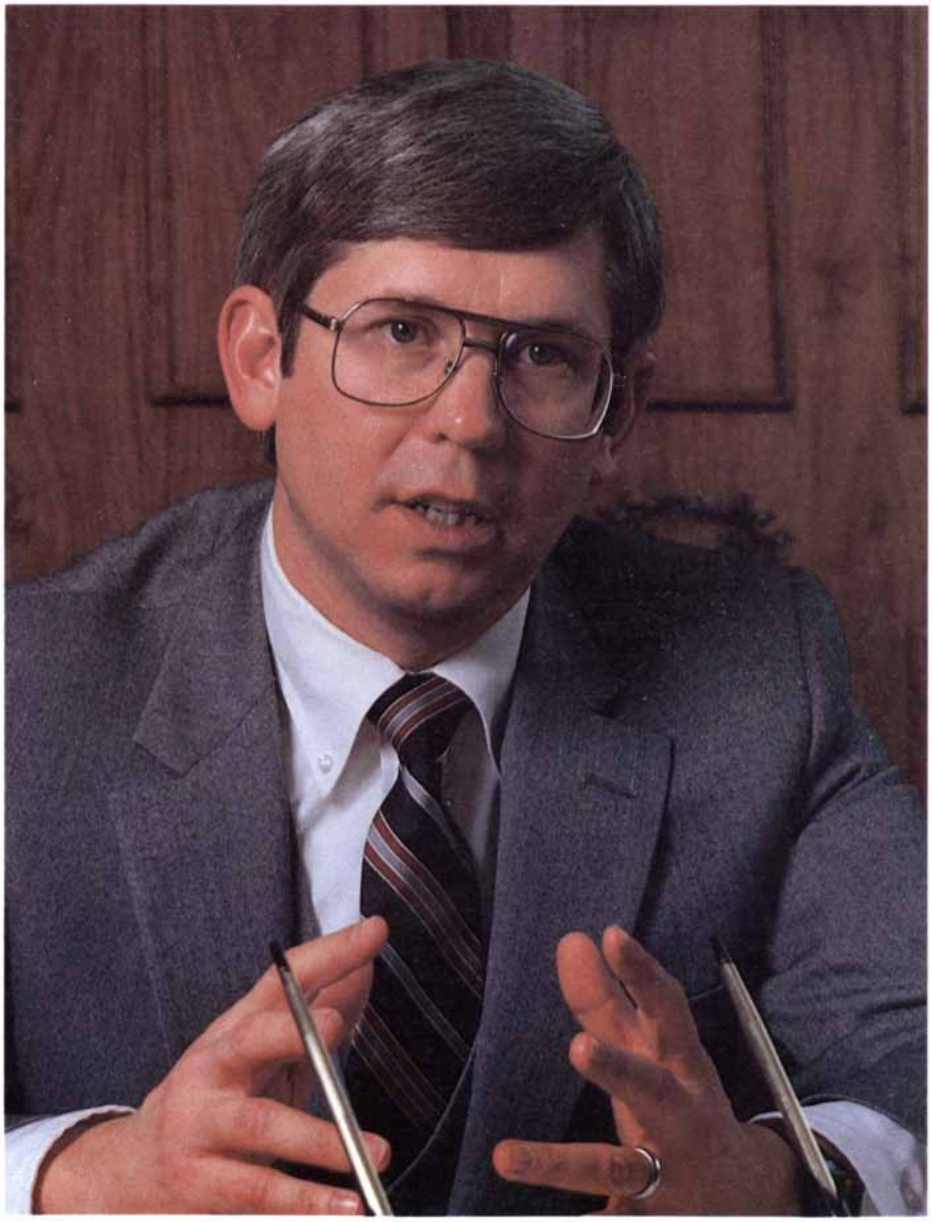
At Fairview Hospital,
an HP CareNet system facilitates critical care and saves $54,000 a year in monitoring costs.

"Because the HP computer-based system delivers extremely accurate vital sign data, we no longer have to repeat procedures or revert to manual methods. On-line monitoring eliminates the need for outside hemodynamic monitoring services, so our staff accesses critical data immediately."

At Fairview Hospital in Minneapolis, Bill Maxwell, Hospital Administrator, says: "HP offered us a total solution to intensive care monitoring—a means to deliver cost-effective quality care. HP CareNet gives us control over our monitoring strategy. And we can continuously monitor all patients from any location in the network.

"The flexibility of HP hardware will enable us to expand our network threefold without significantly increasing our hardware costs. And through HP's upgrade program, we can put tomorrow's technology on today's system without increasing our capital investment.

"Our HP system is maintaining close to 100% uptime. The reliability of our HP network and HP's support team has convinced us to equip another Fairview intensive care unit with an HP CareNet system."



Productivity. Not promises.

HEWLETT PACKARD

0002406

Form and Function in Fish Swimming

Some fishes are specialized for cruising, accelerating or maneuvering, but most are generalists with a combination of locomotor qualities. Recent work correlates each fish's form with its habit of swimming

by Paul W. Webb

Human beings have been reflecting on how fishes propel themselves through water for a remarkably long time. Some of the earliest thoughts on how fishes swim were recorded in India 2,500 years ago. In the Western world the hypotheses of Aristotle, although they were not very close to the mark, were generally accepted until late in the 19th century. Aristotle thought all vertebrate animals moved by pushing at a certain number of points of contact with the world. He hypothesized that fishes pushed against the water with two pairs of fins, four bends in the body or a combination of fins and bends. One reason his views prevailed for so long is that the study of fish swimming presents difficult problems of observation and analysis. The movements of the fish are quick and the fish can travel a long way in a short time. Moreover, the propulsive forces are complex and are exerted in a fluid medium, which makes them hard to measure.

By the end of the 19th century, however, much progress had been made in cinematography, which made it possible to begin examining in detail how fishes swim. At the same time advances were being made in the construction of wind tunnels and experimental water tanks and in the concepts of hydrodynamics. Such developments encouraged new work on the physical forces that have a role in fish swimming. As is true of any solid body that is being propelled, the forward motion of the fish is the net result of the forces that tend to advance it and the forces that tend to retard it. In the case of the fish the impelling force is the thrust generated by the swimming movements. The retarding forces come from inertial resistance or from the drag of the water.

By the 1930's Charles M. Breder of the American Museum of Natural History and Sir James Gray of the University of Cambridge had described swimming kinematics (the movements entailed in swimming). By the 1960's Gray, Richard Bainbridge and John R. Brett of the Canadian Fisheries Research Board had made great contributions to the understanding of swimming energetics (the relations of force and energy generated by the propulsive movements). In the past 15 years particularly rapid progress has been made by building on the work of these pioneers. Quantitative theories formulated by Sir James Lighthill of Cambridge, T. Y. Wu of the California Institute of Technology and Daniel Weihs of the Israel Institute of Technology show how fishes generate thrust and how to calculate the magnitude of the thrust. The new theories also make clear how the form of the fish affects the thrust and the drag.

Thus the subject of how fishes swim presents a curious spectacle: it has been a source of interest for two and a half millenniums but only in the past two decades has quantitative work begun to illuminate how a fish's form is related to its mode of swimming and its way of life. The interplay of thrust and resistance makes some body shapes much better than others for accelerating, some better for cruising and some better for maneuvering. These are the three main swimming functions. Some fishes have bodies specialized for one function. For example, the long, slender body of the pike is well suited to accelerating, which benefits the pike in preying on small fishes. In return for specialized abilities, however, the pike sacrifices peak performance in cruising and maneuvering.

Most fishes are not specialists, but generalists with bodies that give them moderately good performance in all three functions and superior performance in none. The bass can cruise, maneuver and accelerate fairly well but cannot accelerate as well as a pike, cruise as well as a tuna or maneuver as well as an angelfish. The work on fish swimming done in the past 15 years has made it possible to determine which features of the fish's body make a significant contribution to its performance. Since the fish's behavioral options are largely determined by the animal's swimming ability, it has become possible for the first time to correlate in a precise way the design of a particular fish with its mode of life.

Undulation and Oscillation

Fishes clearly have a remarkable variety of shapes and sizes. At one extreme is the knifefish, a long, narrow, tapered animal resembling an eel. At another extreme is the triggerfish, a tropical fish with a body that is roughly diamond-shaped when it is viewed from the side and extremely narrow when it is viewed from the front or the rear. Not only fishes' bodies but also fishes' appendages are diverse. The knifefish has a fin that extends down the median line of the ventral (lower) surface of its body. The triggerfish, on the other hand, has many small fins with short bases on the dorsal (upper) and ventral surfaces of the body and also on the sides.

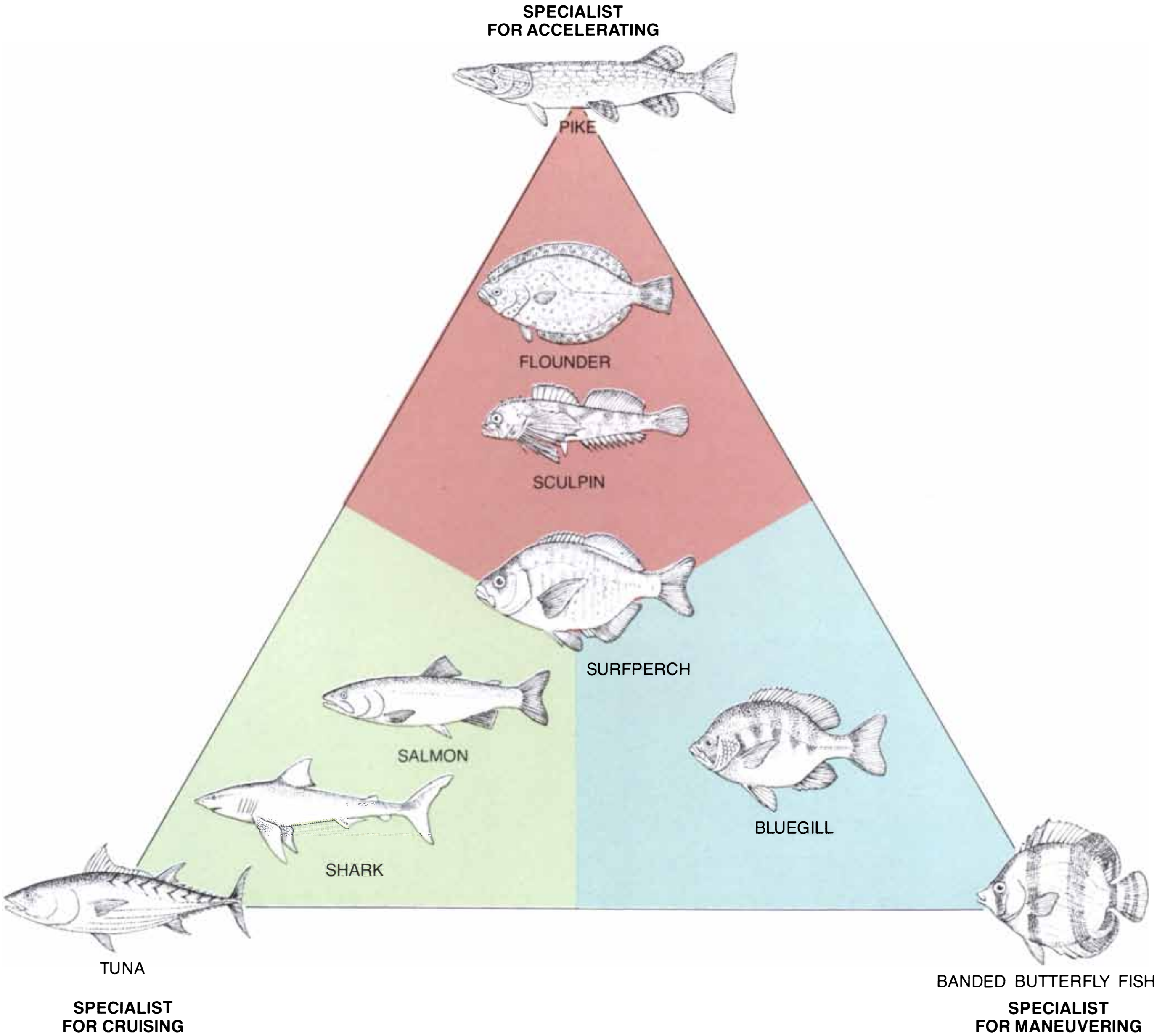
To make matters more complicated, every part of the body and every appendage is employed for swimming by some fish or other, including the body itself, the caudal fin (the tail), the paired fins on the sides of the body and the median fins on the dorsal and ventral surfaces. Such diversity would appear to make the task of understanding the functional morphology of fishes—the connection between the form of the animal and its way of swimming—a daunting one.

Fortunately for the ichthyologist the job is made easier by the fact that locomotion is underlain by unifying principles. Knowledge of these principles makes it possible to divide the fishes into large groups according to their mode of swimming. The first principle is the distinction between undulatory motion and oscillatory motion. In undulatory motion a wave passes along the propulsor, the structure that provides the propulsive force. Two types of structure can undulate: (1) the body and the caudal fin acting as a single unit and (2) fins attached to the body by a long base. Such long-based fins can be on the median line of the dorsal and ventral surfaces or in pairs on the sides. The mackerel, the shark, the salmon, the pike, the bass, the trout and the eel are propelled by the undulatory motion of the body and the caudal fin. The knifefish is propelled by the long-based undulatory fin on its ventral surface.

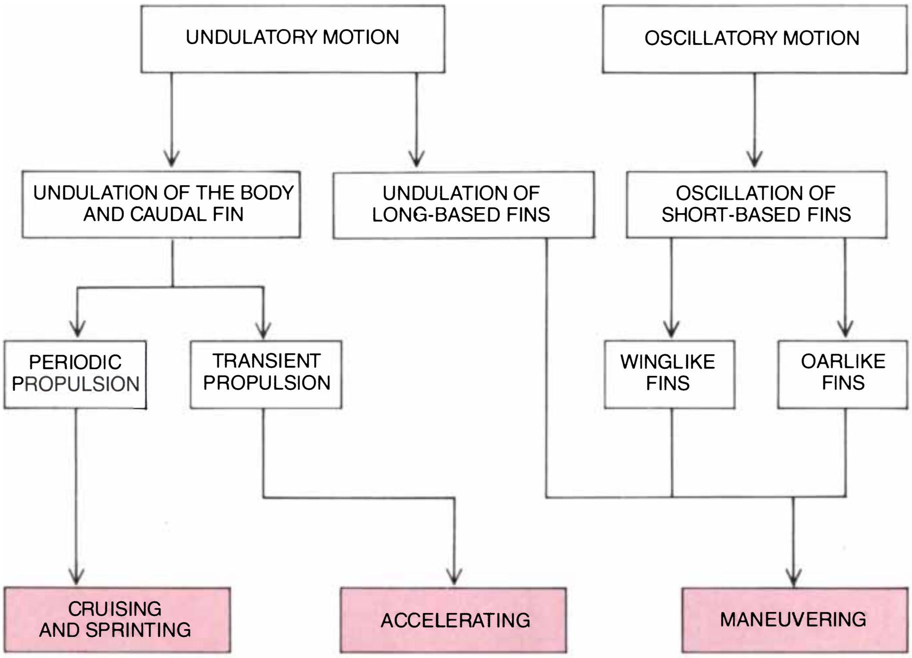
In oscillatory motion the part of the body that provides the propulsive force moves back and forth by pivoting on its base, without the wave motion shown by undulatory structures. Fins attached to the body by a short base generally oscillate. As we shall see, there are two quite different physical principles by which oscillatory fins can generate a propulsive force. The surfperch and the mandarin fish are examples of fishes that rely on oscillatory fins for their main propulsive force. In general, oscillatory and undulatory fins unassisted by body movements are utilized at low speeds in situations where precise maneuvering makes a substantial contribution to the fish's feeding or to its survival. Faster swimming requires more power; therefore the myotomal muscle, the large mass of tissue on each side of the body, is recruited for accelerating and high-speed swimming.

How does the undulation or oscillation of a part of the fish's body result in the thrust that propels the fish through the water? Much of the work that has been done in response to this question concerns undulatory movements, and the best-understood of these movements is the joint undulation of the body and the caudal fin. When the fish's body undulates, one or more half wavelengths travel from its head to its tail at a speed greater than the fish's speed with respect to the water.

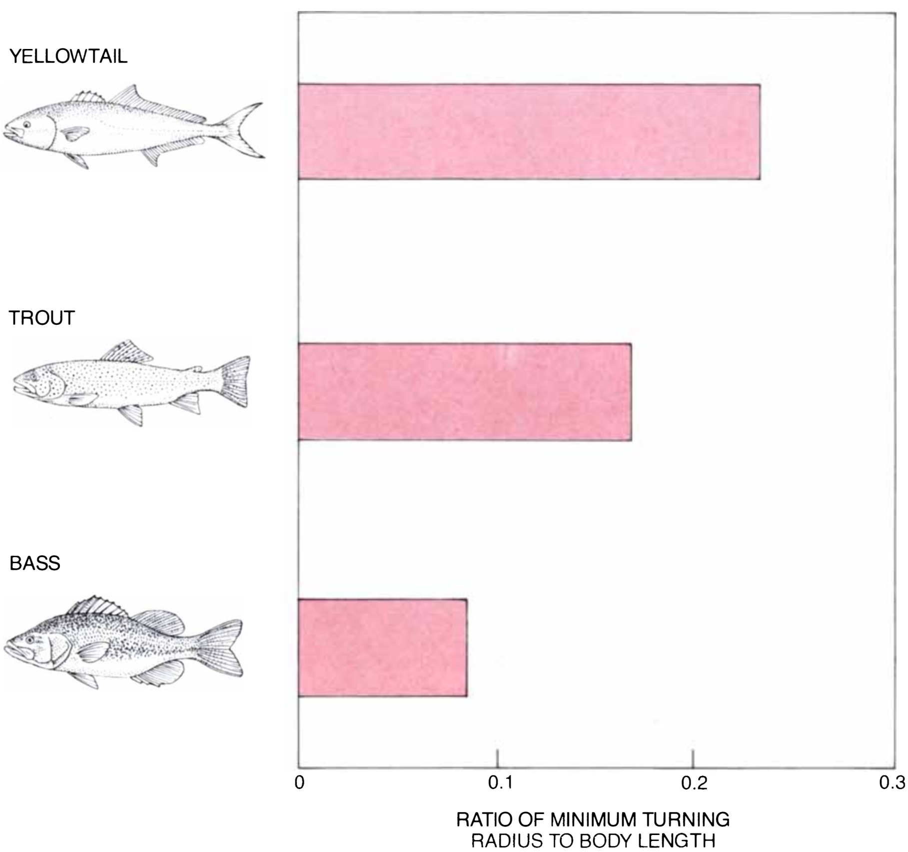
In the undulation each propulsive element, or small segment of the body, moves laterally with respect to the head.



FUNCTIONAL-MORPHOLOGY PLANE for locomotion in fishes includes specialists and generalists. The fishes at the corners of the triangle are specialized for one swimming function: the tuna for sustained cruising, the pike for accelerating in quick strikes at prey and the banded butterfly fish for low-speed maneuvering in and around the coral reef. Each specialist performs poorly in the other two functions. For example, the tuna does not accelerate or maneuver well. The surfperch, at the center of the triangle, is a generalist. It can accelerate, cruise and maneuver fairly well but does not perform any function as well as a specialist. Between the center of the triangle and the corners are fishes that are more specialized than the surfperch but are less specialized than the tuna, the pike or the butterfly fish.



TYPES OF PROPULSION utilized by fishes are shown in the boxes. In undulatory motion a wave passes down the length of the body part or the fin that provides the propulsive force. In oscillatory motion the anatomical structure moves back and forth. Cruising, sprinting and accelerating call for the coupled undulatory motion of the body and the caudal fin (the tail). In periodic propulsion, which is employed in cruising and sprinting, the swimming movements are repeated cyclically. In transient propulsion, which is employed in accelerating, the movements are not cyclical. Fishes that maneuver slowly utilize the oscillatory motion of fins with short bases (which operate like wings or like oars) or the undulatory motion of fins with long bases.



TRANSIENT SWIMMING PERFORMANCE can be measured by the ratio of the minimum turning radius of the center of body mass in a fast start to the length of the body. The smaller the ratio, the better the performance. Bass perform better in a quick start than trout or yellowtail. The reason for the superior performance of bass is that they have the largest body area in relation to their mass; the area is utilized for thrust whereas the mass resists the turn.

As the wave passes, the propulsive element accelerates the water nearby. The force generated by the muscles accelerates the water; an equal opposing force, called the reaction force, is exerted on the propulsive element by the water. The magnitude of the reaction force is equal to the product of the acceleration given to the water and the mass of the water accelerated. The force is perpendicular to the propulsive element and is inclined toward the head of the fish.

The force exerted on a propulsive element has two components. Consider a fish swimming in a straight line (meaning without turning). One component of the force on the propulsive element is parallel to the fish's overall direction of motion; this is the longitudinal force. The other component is perpendicular to the overall direction of motion; this is the side force. When the fish is swimming without turning, only the longitudinal component contributes to thrust.

Contributions to Thrust

The longitudinal force generated by the propulsive elements near the tail is greater than the force generated by the elements near the head. There are two reasons the caudal elements make a greater contribution to thrust. The first reason is that the caudal inclination of the segments near the tail is large.

To understand what caudal inclination means, imagine a line drawn along the axis of the propulsive element, that is, along the median line of the fish's body [*see illustration on page 78*]. When the propulsive wave passes down the body, the line corresponding to the axis of the element curves so that the side pushing against the water faces toward the tail. Now imagine a second line perpendicular to the first; the perpendicular line represents the reaction force that acts on each element and is directed toward the head. As the propulsive wave passes, the elements near the tail curve farther toward the rear than the elements near the head. As a result the reaction force on the rear elements has a direction closer to that of the overall motion of the fish, and more of the force exerted by the water on the element is parallel to the direction of motion.

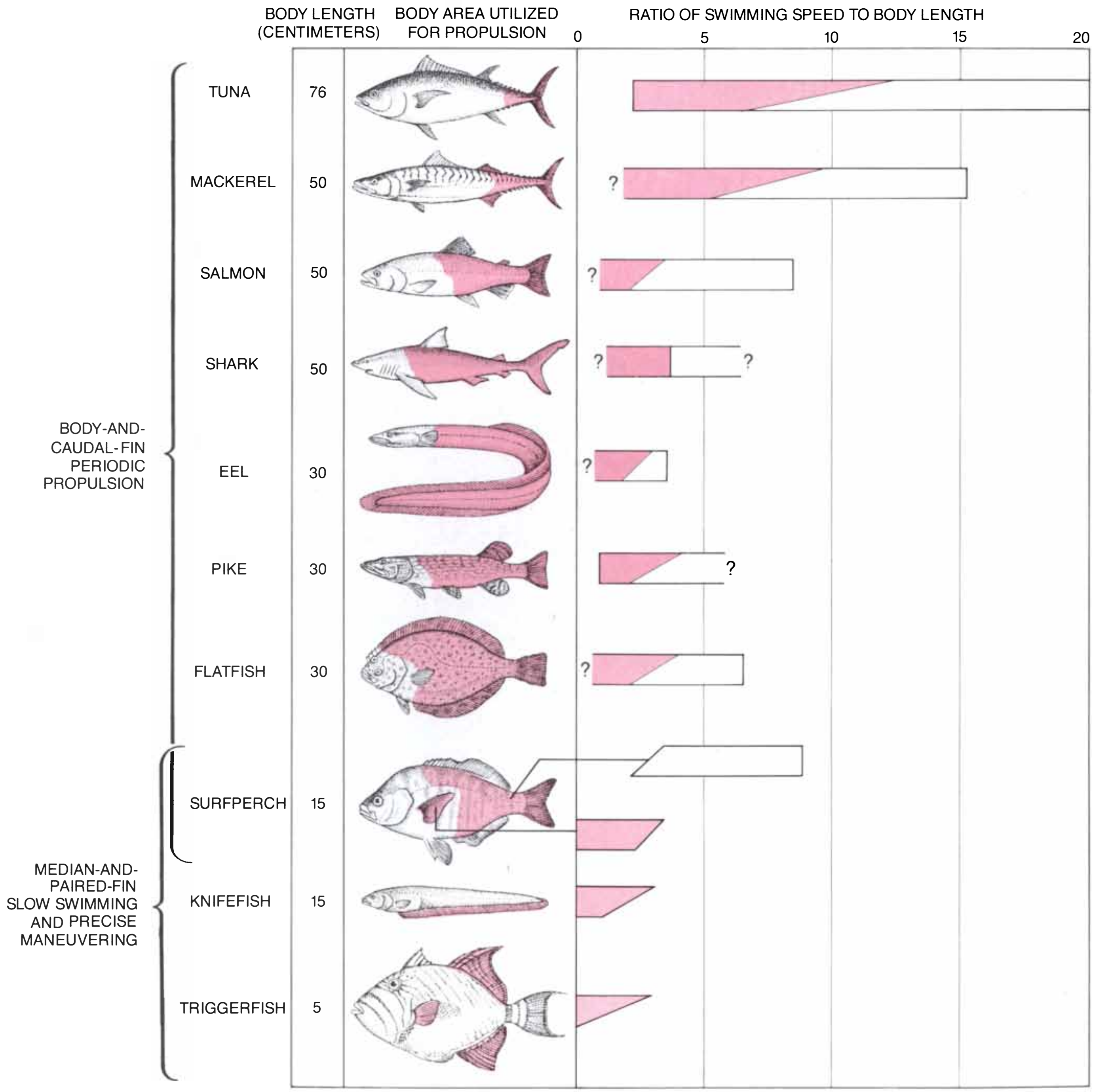
The second reason the elements near the tail make a greater contribution to the thrust than the elements near the head is that as the wave passes, the rear elements traverse a greater distance than the forward ones. Hence the speed of the elements near the tail is greater and they accelerate the water more than the elements near the head of the fish.

The way thrust is generated by the propulsive elements as the wave moves down the body and the caudal fin has significant implications for the shape that most efficiently meets the needs of a particular fish. Swimming based on the coupled motion of the body and caudal fin can be divided into two categories: transient swimming and sustained swimming. Transient swimming, which is the type of movement observed in fast starts and powered turns, entails propulsive movements that are brief and have a large amplitude. The movements can be as brief as tens of milliseconds and in them the tail can traverse a distance equal to half of the body length.

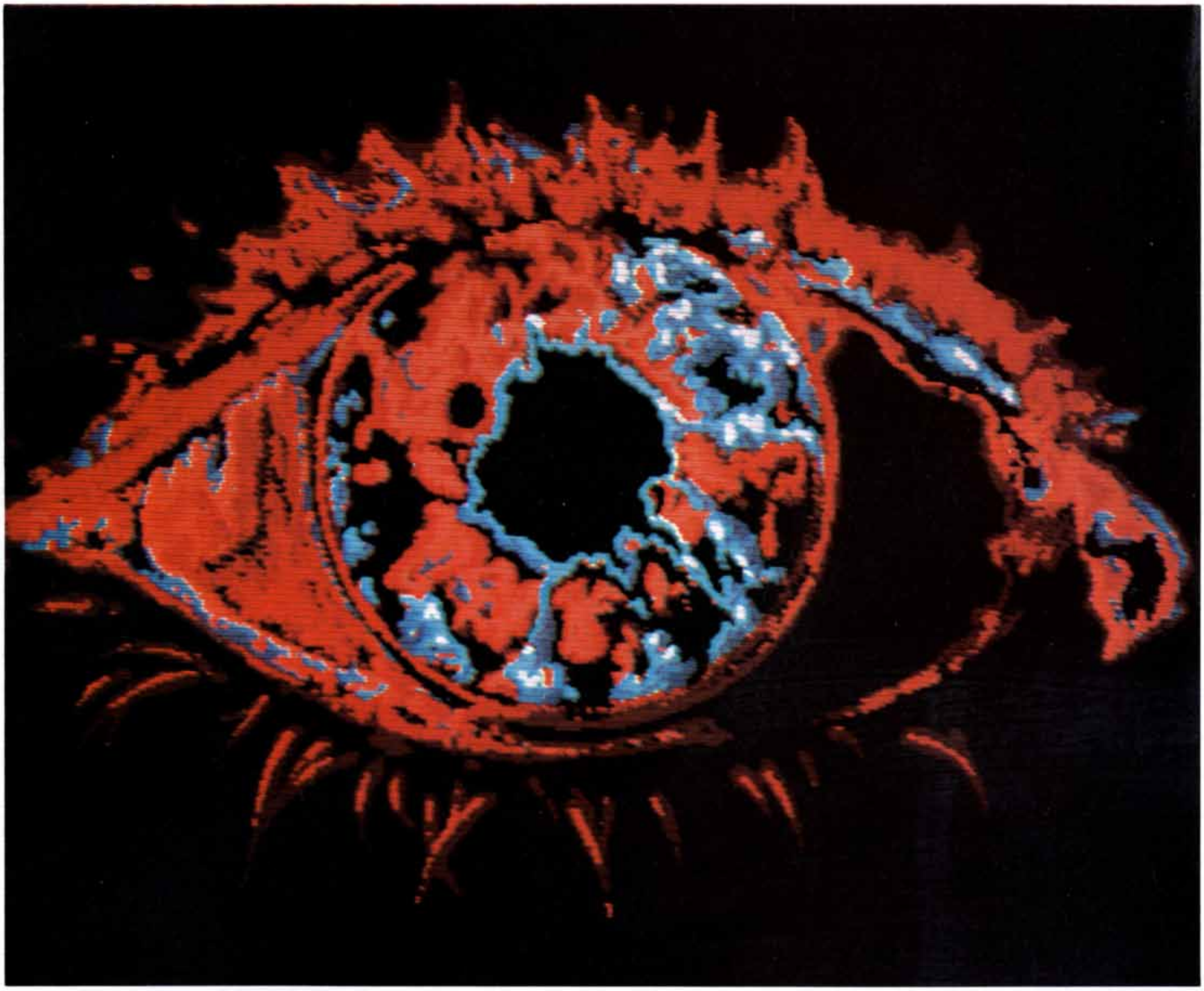
Sustained swimming, on the other hand, is any propulsion with cyclically repeated swimming movements; such movements generally include several beats of the tail. Sustained swimming can itself be divided into several types according to the duration of the propulsion. For example, cruising is a type of swimming that can be sustained for an hour or more. Sprinting, another type of sustained swimming, may last for only a few seconds. The mechanism whereby thrust results from the wavelike motion of the body has implications for transient swimming different from those it has for sustained swimming, and as a result fishes that rely mainly on quick acceleration have a shape different from that of fishes that cruise large areas of the ocean at a steady pace.

Transient Swimming

For transient swimming one of the most significant implications of the method of thrust generation is the shape of the silhouette seen from the side. When the propulsive element moves, it



PERIODIC SWIMMING AND MANEUVERING can be accomplished with almost any part of the body. This illustration shows the structures utilized for propulsion (*color*) by some periodic swimmers and maneuverers. Also shown are the range of speed in cruising (*colored part of bar*) and the range in sprinting (*open part of bar*); maneuverers tend to cruise at low speeds. Question marks indicate that the limits of performance are not known with precision. The surfperch can be either a periodic swimmer or a maneuverer: it employs the motion of the body and the caudal fin for sprinting and the oscillatory motion of the paired pectoral fins for cruising and maneuvering.

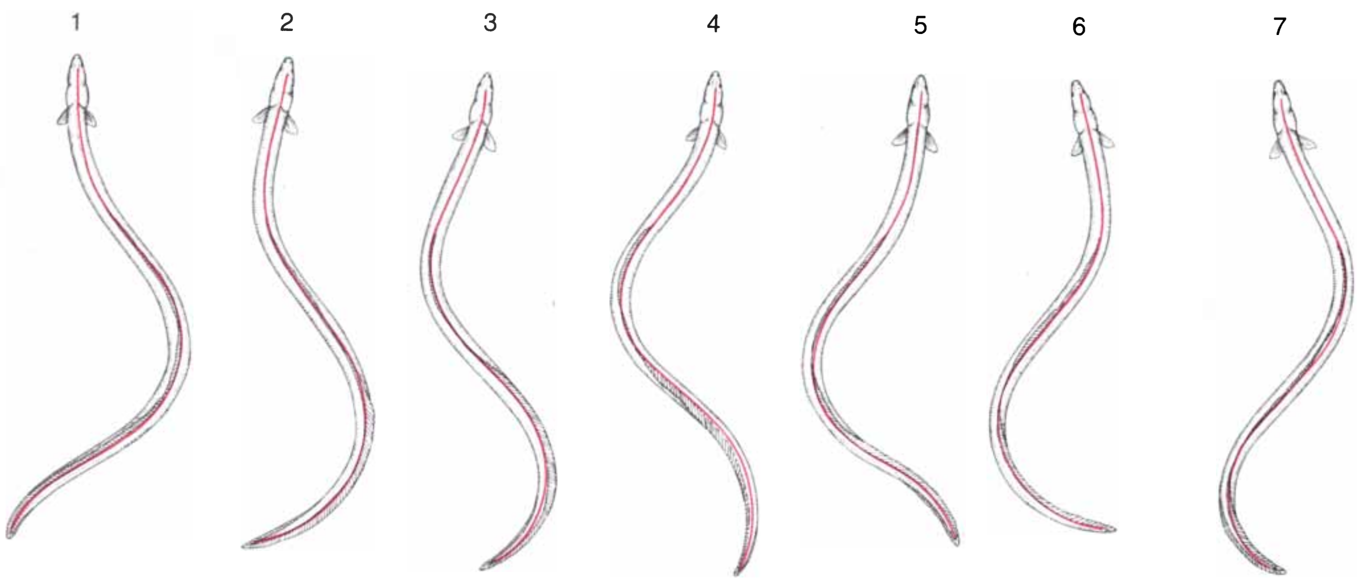


The human eye, greatest of natural sensors, is a key model for Hitachi electronic devices that translate environmental conditions into signals machines can act on.

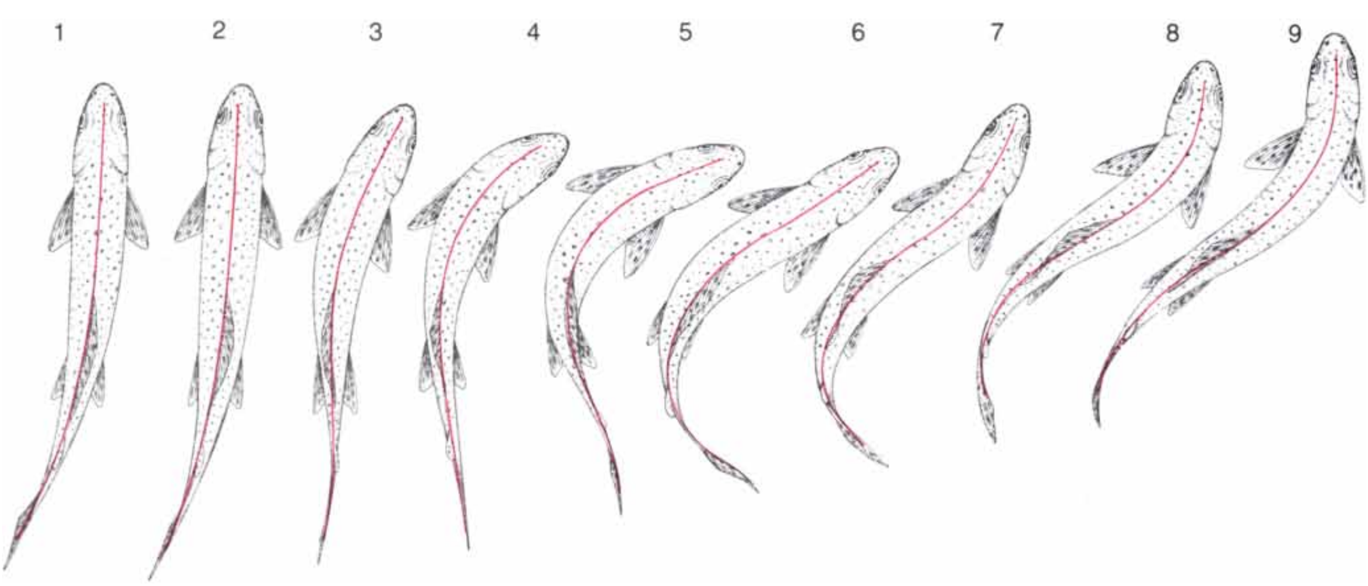
accelerates a quantity of water that is referred to as the added mass. The magnitude of the added mass is approximated by the mass of the water contained in a cylinder with a diameter equal to the depth of the propulsive element (measured from the upper edge of the body or fin to the lower edge) and a small length.

It follows that to maximize the thrust each propulsive element should be as deep as possible, which would yield a very deep silhouette for any fish that relies on the body and the caudal fin for propulsion. As it happens, a deep silhouette maximizes thrust only when each propulsive element makes an independent contribution to the thrust, that is, when there is no significant interaction among the elements. In sustained swimming the action of one element has a considerable effect on the action of the other elements; the interaction has a profound influence on the silhouette that yields the best performance. In transient swimming, however, there is little interaction among elements, and a deep silhouette leads to an increase in thrust. The result is a very deep body and median fin, such as those of flatfishes and sculpins, fishes that are frequent transient swimmers.

The level of performance made possible by a particular body form does not depend solely on the magnitude of the thrust. On the contrary, performance depends on the net balance between the thrust and the resistance. As is the case with thrust, the factors that determine resistance in transient swimming are different from those that determine resistance in sustained swimming. In transient swimming the rate of acceleration is high and resistance comes mainly from inertia. The effect of inertia can be minimized by making the nonmuscle dead weight (the body weight contributed by tissues other than skeletal muscle) as small as possible. Indeed, the myotomal muscle in a fish specialized for transient swimming can contribute 60 percent of the total body mass.



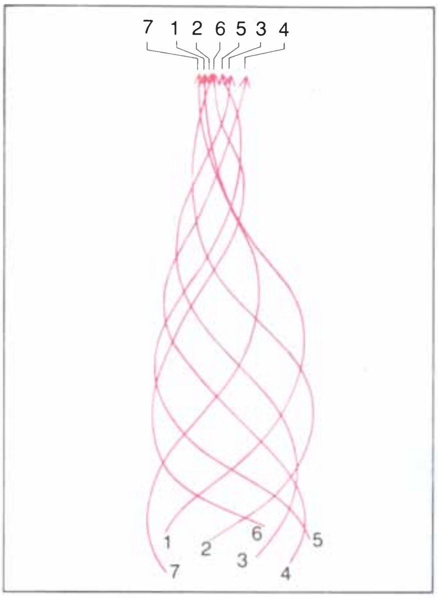
CRUISING EEL bends into more than one complete wavelength as it moves through the water. In the two illustrations on this page the numbered figures at the left are based on frames of a motion picture made at short intervals. The panel at the right shows the position of the centerline of the fish during the propulsive movements. In every type of body-and-caudal-fin swimming the propulsive wave travels along the body at a speed higher than the net speed of the fish with respect to the water. In periodic propulsion, such as that of the cruising



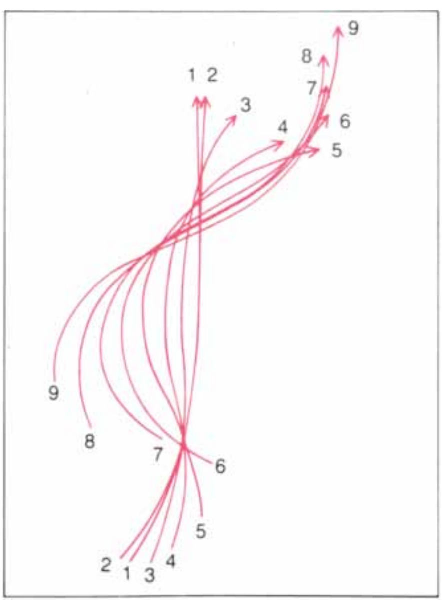
ACCELERATING TROUT bends its body into only part of a complete wavelength as it increases speed from a standing start. In such transient swimming the overall propulsive movement is not repeated cyclically. Like most transient swimmers, the trout combines a fast

The analysis of transient swimming in the preceding paragraphs suggests there is an optimum design for fishes with a mode of life that relies heavily on quick starts. The design includes a body flexible enough to bend into a wave of large amplitude, a body and fins with a large area and a body that is mostly muscle. Although this would appear to be an advantageous combination, no fish has all these features. Nonlocomotor factors, in particular factors related to feeding, exert an influence. As a result the optimum design for transient swimming is not necessarily the best design for the overall welfare of the fish.

Rather than showing all the features of the optimum design, many transient swimmers have a form in which one feature predominates. For example, the sculpin is a transient swimmer with a design that maximizes thrust. It has large dorsal and ventral fins that give the fish a deep silhouette along the length of its body. The design of the sculpin, however, does not yield the maximum proportion of myotomal muscle. The reason is that the sculpin's head is large and heavy, with a geometry that is valuable in feeding on bottom animals. As a result the nonmuscle dead weight cannot be reduced to the minimum and only 30 percent of the body mass of the sculpin is myotomal muscle.



eel, the movements are cyclical. The undulatory movements of the eel were first described by Sir James Gray in the 1930's; much of the current work draws on Gray's observations.



start with a power turn. In transient swimming the propulsive movements are small and brief.

The Pike

The design specifications of the pike, another acceleration specialist, result in low resistance. The body of the pike is from 55 to 60 percent muscle; the fish even has a thin skin. The minimizing of nonmuscle dead weight is not surprising in a predator that catches its prey by rapid acceleration. Yet the pike does not conform to the optimum design for transient swimming because its silhouette is deep only near the tail; near the head the silhouette is shallow.

Since a deep silhouette all along the body offers maximum thrust in a quick start, the shallowness of the pike's silhouette near the head does not appear to make sense for catching small prey. To ascertain what other factors influence the design of the pike several of my students at the University of Michigan and I compared the strike of the tiger muskie, a hybrid relative of the pike with a shape similar to that of the pike, with the strike of the trout, the bass and the rock bass.

It was found that prey fish were much slower in responding to the muskie than they were in responding to the other predators, with the result that in 70 to 80 percent of the strikes the muskie caught its prey before the smaller fish responded by moving away. The prey were much quicker to evade the other predators: only about 30 percent of the strikes of the trout and the two kinds of bass were successful.

The disparity in the response threshold of the prey is due largely to the shape of the lateral cross section of the predator's body, which is the outline the prey responds to when the predator strikes. Trout and bass have a lateral cross section that resembles a vertically aligned ellipse. Fishes are extremely sensitive to that shape. The pike, however, has a lateral cross section more like a disk, because its dorsal fins end well behind the head. Thus in the body of the pike the optimum design for locomotion is compromised to accommodate sensory factors in the prey.

In sustained swimming the interaction

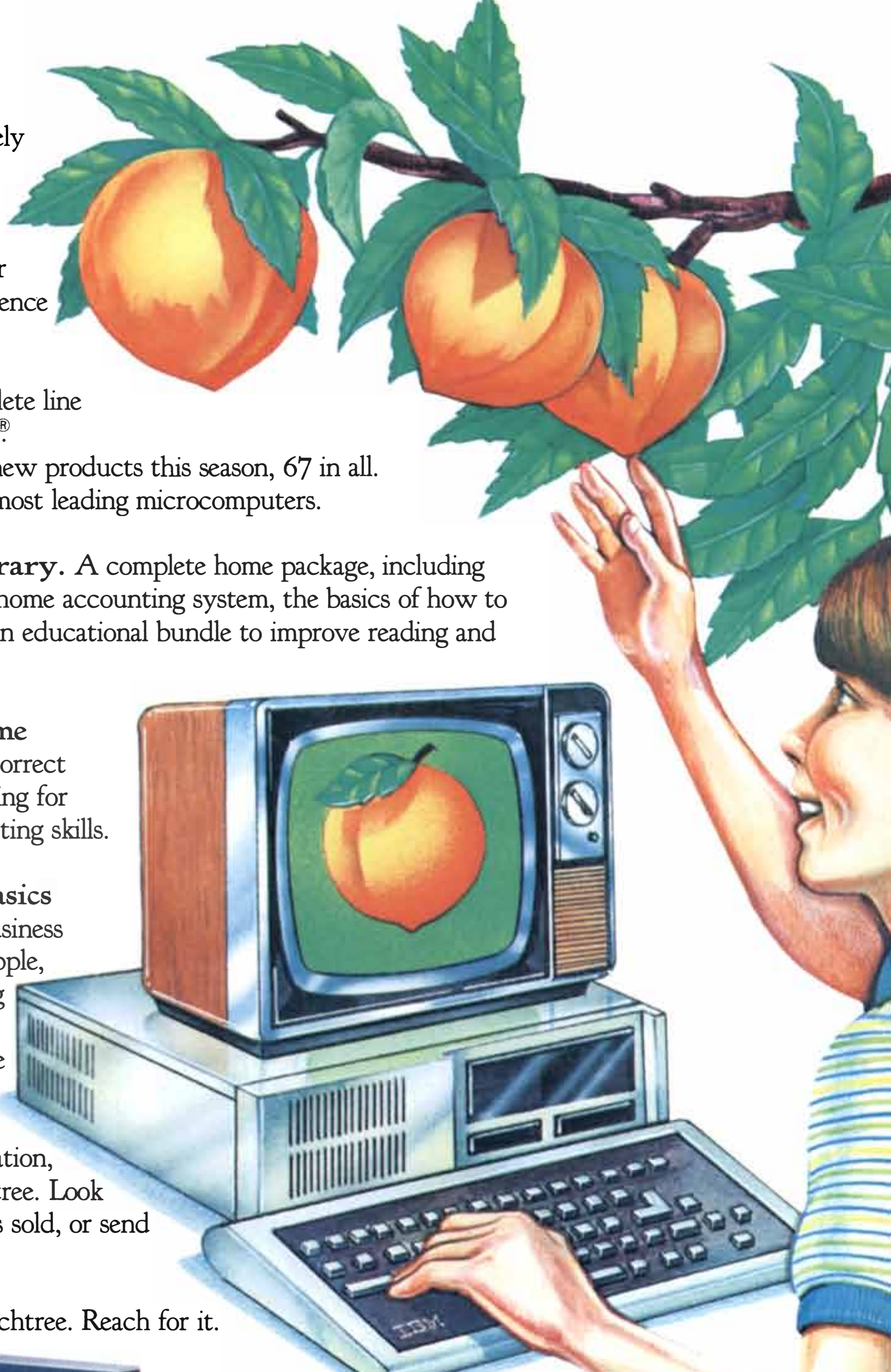
And it couldn't be in better hands.

Only one software company so completely covers home, education and business.

Only one has behind it the resources of the world's largest independent producer of mainframe software, Management Science America, Inc. (MSA).

And only one offers you the most complete line for microcomputers. Peachtree Software®. The leader that's still leading. With 25 new products this season, 67 in all. With applications for IBM, Apple and most leading microcomputers.

For instance: our new Home Library. A complete home package, including word processing, financial spread sheet, home accounting system, the basics of how to write your own programs, a game, and an educational bundle to improve reading and writing skills for all ages.



For instance: our new five-volume Writing Skills System. Created to correct common writing problems; teaches writing for all levels through the development of editing skills.

For instance: our new Back to Basics Accounting System. Perfect for a business conducted at home. For your PCjr™, Apple, Atari, Commodore or almost any leading microcomputer. Lets you set up a basic accounting system even if you have little or no previous accounting knowledge.

Whether it's for home, business or education, America's software grows on the Peachtree. Look for all our products wherever software is sold, or send us the coupon for a free brochure.

Your family's software grows on the Peachtree. Reach for it.

America's Software grows on the Peachtree.

PCjr is a trademark of International Business Machines Corporation.

Peachtree and Peachtree Software are registered trademarks of Peachtree Software Incorporated, an MSA Company.® 1984

Peachtree Software
3445 Peachtree Road N.E.
Atlanta, GA 30326 • 1-800-554-8900

Please send me more information about Peachtree Software.
I'm interested in software for ( ) business, ( ) home, ( ) education.

Name____________________

Address____________________

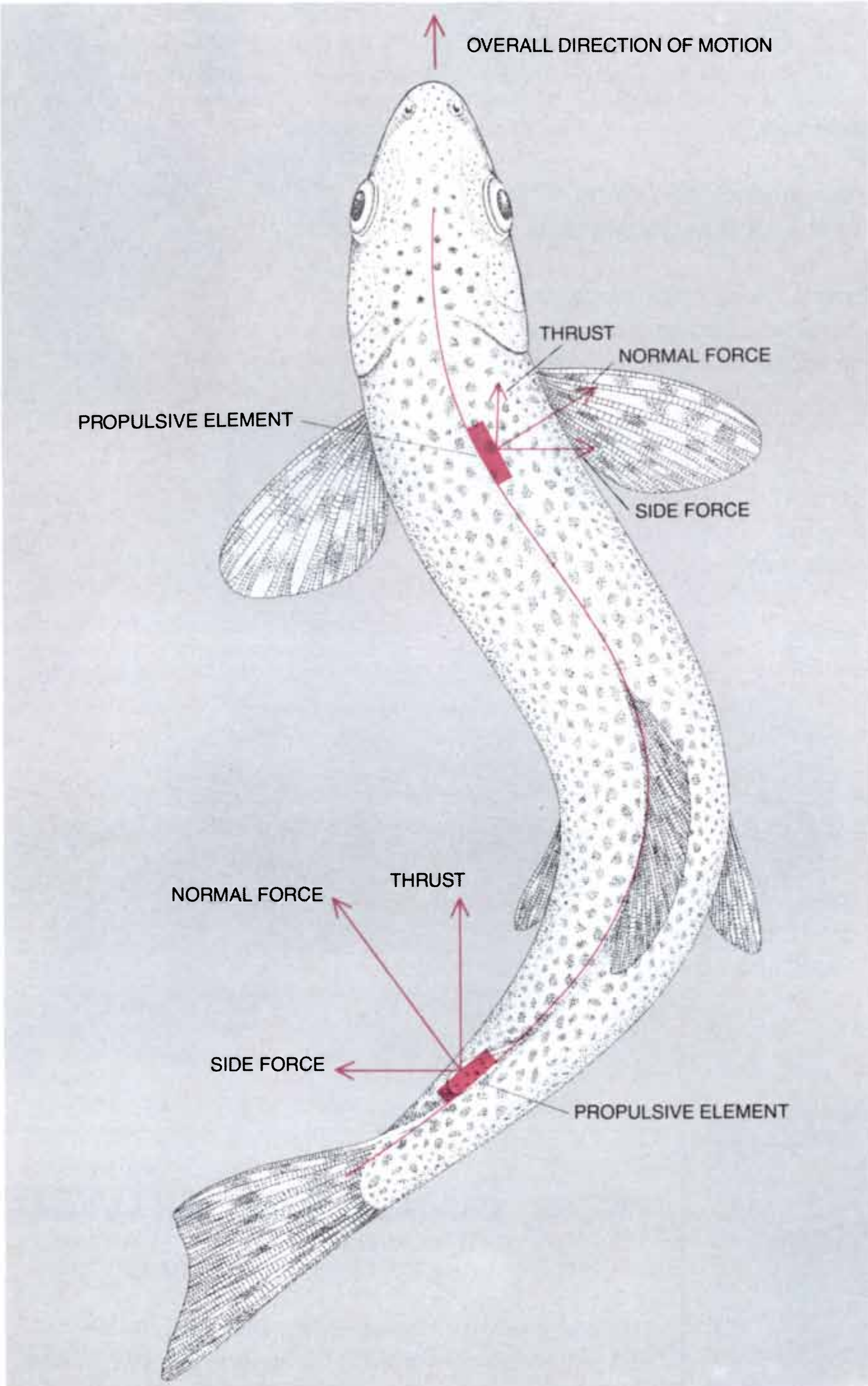
City__________ State______ Zip______

Phone______________ CA4

of the fish's body and the surrounding medium leads to optimum body and fin shapes quite different from those of the sculpin and the pike. In cruising and sprinting the propulsive movements are repeated cyclically. Because of the repeated movements, sustained swimming is also referred to as periodic propulsion. In cruising and sprinting the amplitude of the tail beat is smaller than it is in transient propulsion. The tail moves a distance that rarely exceeds 20 percent of the body length, and so it is possible for the body to be bent into a larger number of wavelengths.

In periodic propulsion the water accelerated by a propulsive element is immediately affected by the action of the next element toward the rear of the fish. As we have seen, the rear elements travel faster than the forward ones, and the caudal inclination of the elements toward the rear is greater than the inclination of the elements toward the head. Therefore the effect of the rear propulsive element is to increase the magnitude of the acceleration given to the water by the element in front of it. In periodic propulsion the propulsive elements interact, and the last element at the tail (corresponding to the trailing edge of the fish) determines the net acceleration imparted to the water. In most fishes the movement of the trailing edge also determines the final added mass. Hence in thrust generation the trailing edge of the tail is the most crucial element. It follows that it is valuable for the tip of the tail to be as deep as possible.

Because the action of the rear propulsive elements is of great significance in sustained swimming, it would appear that the optimum design for a fish that spends much time cruising would include a deep rear silhouette, perhaps something like the rear silhouette of the pike. There are several reasons, however, for this not being so. Indeed, it can be shown that among the optimum design features for periodic swimming is the scooping out of the body in front of the tail to form a narrow caudal peduncle: a slender stem to which the tail is attached. Lighthill calls such a tail design narrow necking and the design is a characteristic feature of fishes that spend their lives cruising.



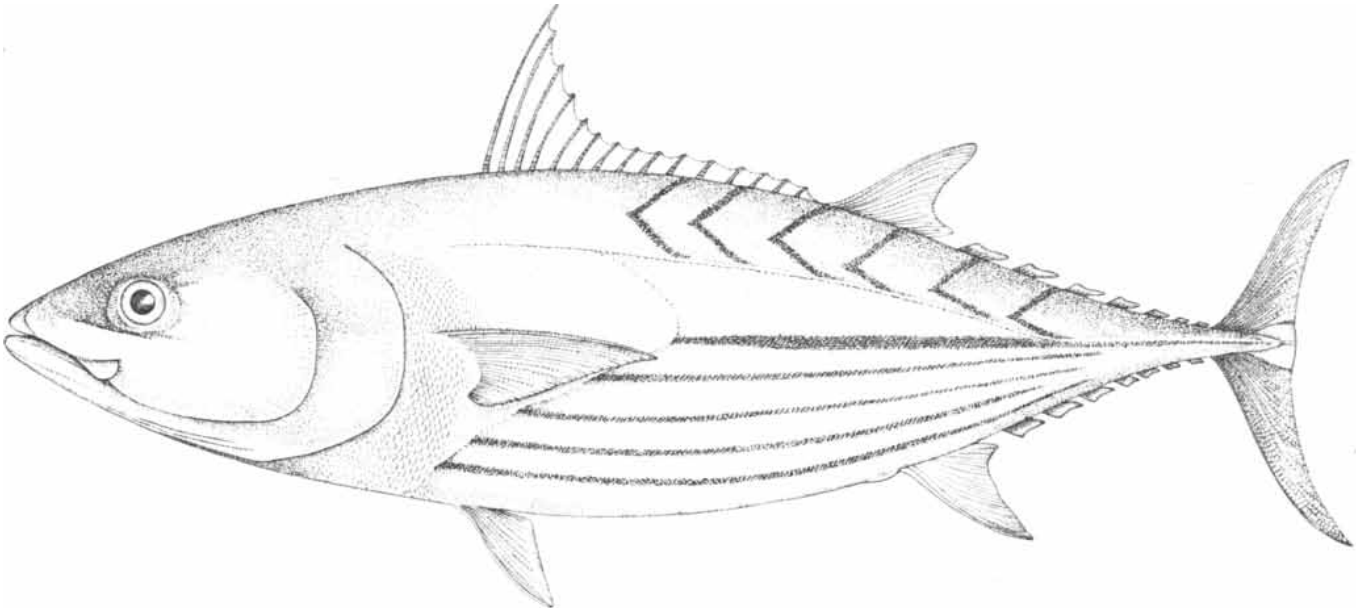
THRUST IS GENERATED by propulsive elements: small segments of the body. The forces operating on two propulsive elements are shown. As the propulsive wave passes, each element moves laterally and accelerates the water nearby. The element is inclined toward the tail, and so the water is accelerated toward the tail. The normal force, which is equal and opposite to the force exerted on the water, presses against the propulsive element. The normal force has two components: the side force is perpendicular to the overall motion of the fish and the thrust is parallel to the overall motion. The combined thrust from all the propulsive elements propels the fish. The magnitude of the thrust contributed by each element increases toward the tail.

Narrow Necking

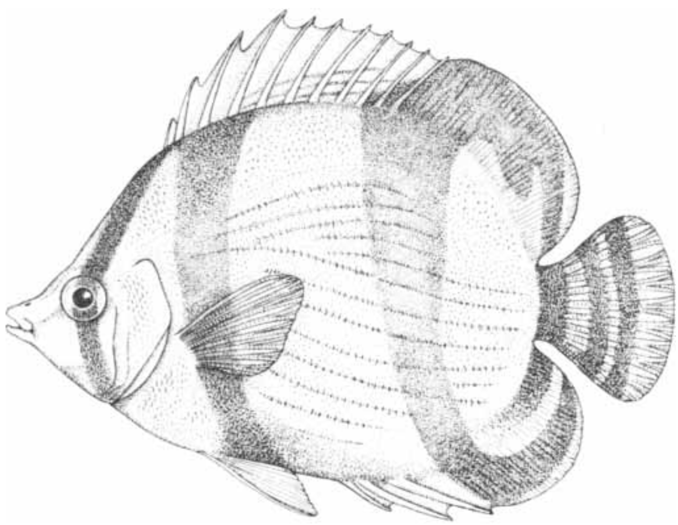
Why is narrow necking so valuable? One reason is related to the side force, the component of the reaction force that does not contribute to the forward motion of the fish but turns the body to the side. The side force tends to make the part of the body near the head recoil, or oscillate laterally. In periodic swimming the side force can lead to the waste of considerable energy. If all the propulsive elements were of equal depth, the side force would be greatest in the rear elements. With a decrease in the depth of the body in front of the tail the wasteful side force is reduced.

The lateral oscillation is reduced further by an increase in the mass of the body near the head. The resistance of the forward part of the body to the effect of the side force is also increased by a fin on the upper surface of the body, which helps to explain the presence of forward median fins in many aquatic animals.

Narrow necking also contributes to the lowering of resistance. The major source of resistance in sustained swimming, in contrast to the resistance in transient swimming, is drag that arises



SKIPJACK TUNA is a superior cruiser with a design that approximates the optimum for periodic swimming. It has a stiff body with a "streamlined" shape that is deepest about halfway between the head and the tail. The deep, narrow caudal fin generates a strong thrust. The caudal fin is connected to the body by a slender peduncle. Such a design serves to maximize the thrust while reducing the drag force.



BANDED BUTTERFLY FISH has a design that approximates the optimum for low-speed maneuvering. The disk-shaped body is quite short, which facilitates rotating movements in the median vertical plane. The oscillating fins that provide the propulsive force are distributed all around the center of the body mass, and so small, precise thrusts can be delivered in any plane. Such a form is of considerable advantage to the fish in geometrically complex habitats such as the coral reef, where many specialists for low-speed maneuvering live.

from the viscosity of the water. The viscous drag force depends on the rate at which the fluid is distorted. The rate of distortion is greatest in the boundary layer: the thin layer of fluid immediately adjacent to the surface of any body moving through a fluid.

The viscous drag force is proportional to the square of the velocity and the surface area of the body. Propulsion that depends on flexing the body increases the viscous drag force because the motion of the propulsive elements increases their velocity with respect to the surrounding fluid (in comparison with the corresponding elements of a rigid body). In addition, the propulsive movements modify the local flow patterns and increase the distortion of the fluid, which increases the frictional force.

Reducing Viscous Drag

As a result of such factors the drag on a self-propelled flexing object such as a fish can be greater at the tail by an order of magnitude than the drag on a rigid body in the corresponding position. If the drag is summed along the entire body, it is found that the average drag on the fish is from two to four times greater than the drag on a rigid body. Viscous drag is reduced if the area of the moving body is minimized. Moreover, the reduction is particularly great if the area is reduced in the region in front of the tail, because it is there that the effect of body movement leads to the greatest increase in drag. The result is narrow necking. In addition a narrow tail section makes it possible for the forward part of the body to be more rigid and for the body to be fusiform ("streamlined"), which also decreases the drag.

The interaction of the propulsive elements in periodic swimming and the benefits associated with decreased viscous drag suggest that there is an optimum design for fishes that rely on sustained swimming. The design includes a stiff body attached by a slim caudal peduncle to a deep, narrow tail. The body itself has a fusiform shape: it is deepest at a point that is between a third and a half of the distance from the head to the tail and narrows gently to the tail. Perhaps the best approximation of the optimum sustained-swimming design is the form of the tuna. A similar morphology is found in sharks, dolphins, whales and even in ichthyosaurs, a group of extinct swimming reptiles. All these animals show a specialization for fast, efficient cruising.

As with transient swimmers, there are deviations from the optimum locomotor design among fishes that live by cruising. The most notable deviations are observed in the sharks. Although most sharks depend heavily on cruising for their livelihood, only a few have a body shaped like the body of a tuna. The reason for the divergence is that in spite of the fact that both the shark and the tuna are cruising fishes, most sharks have a mode of swimming quite different from the tuna's. Indeed, it appears that the swimming movements common among sharks are more like those of the eel than they are like those of the tuna.

The body of the shark, however, is more fusiform than the body of the eel. In addition, whereas the eel has no significant median fins, the shark has several and they are sizable. The median fins on the dorsal surface of the body are particularly pronounced. There is generally a large forward dorsal fin and a somewhat smaller rear dorsal fin. The upper extension of the tail is also large. Between the three dorsal appendages there are substantial spaces.

The combination of the shark's swimming movements and the large gaps between the median fins is of great mechanical significance. The effect arises from the fact that any fin with a sharp rear edge trails a wake behind it. Furthermore, because of the periodic motion of sustained swimming the wake follows a sinusoidal path.

The Fins of the Shark

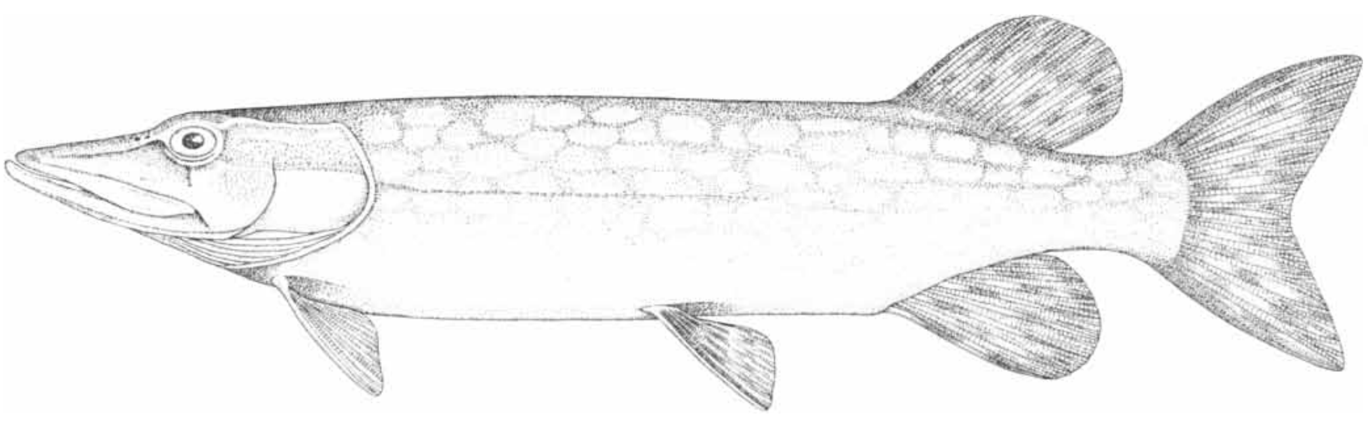
Consider the wake of the shark's fin in a frame of reference that has the same velocity as the fish and is moving in the direction the fish moves. In such a frame the wake will appear to be traveling downstream, or toward the tail of the fish. The velocity of the wake with respect to the fish will be equal to that of the fish when the fish is viewed by an observer in the normal frame of reference. The propulsive wave is also traveling toward the tail but, as we have seen, the velocity of the propulsive wave is greater than that of the fish. Since the propulsive wave travels faster than the wake, there is a phase difference between the two sinusoidal paths. If the fish is swimming at a steady pace, then the phase difference will be constant at any point that is downstream from the fin.

Next consider a second fin closer to the tail than the first. If at the second fin the phase difference between the propulsive wave and the wake has the right magnitude, the wake begins to move inward just as the second fin is moving outward. If that is the case, the thrust resulting from the motion of the second fin can be considerably increased. The effect is roughly analogous to reaching out to push a swinging door just as the door is pushed from the other side.

As it happens, the minimum phase difference, or the difference in the position of the sinusoidal paths, needed to yield the thrust-enhancing effect is 90 degrees. With Raymond Keyes of Sea World in San Diego I studied six species of shark at the shark facility of Sea World to find out whether the phase difference between the trailing wake and the downstream fin is large enough for the thrust-enhancing interaction to take place. We found that all six species swim in such a way that the forward fin and the rear fin can interact to increase the thrust.

One reason sharks rely on thrust-enhancing mechanisms rather than on the specialized refinement of the tuna design is the material of which the shark's skeleton is composed. The skeleton of sharks is not bony but cartilaginous. Although the skeleton is in many instances reinforced with stronger, calcium-containing materials, the cartilaginous members cannot sustain as much bending as the bone found in most other fishes. This is one reason the few sharks that look like tunas cannot swim as fast as tunas can.

The differences in functional morphology between sharks and tunas reflects far-reaching differences in the ecology of the two groups of fishes. James F. Kitchell of the University of Wisconsin at Madison analyzed the pattern of energy utilization by tunas and concluded that they are "energy speculators." Tunas expend considerable energy swimming continuously at high



NORTHERN PIKE is a specialist for acceleration with a design that yields the minimum resistance. In transient swimming the resistance comes mostly from inertia. The inertia with respect to the magnitude of the propulsive force can be reduced by having the nonmuscle component of the body mass as small as possible: the pike has a body mass that is 60 percent muscle. The pike's long, slender body can bend into a large-amplitude wave, which is advantageous in accelerating. The silhouette of a pike viewed from the side is deep near the tail, yielding considerable thrust in darting strikes at small prey fish. Near the head, however, the silhouette is shallow. The author has found that the rounded lateral cross section near the head aids the pike because prey fishes are comparatively insensitive to that shape.

speed over great distances searching for prey and breeding sites. The continuous high-speed swimming is facilitated by endothermy: internal regulation of body temperature. Conversely, continuous swimming and endothermy call for much energy in the form of food, hence tunas must search the ocean at a steady high speed.

Sharks: Energy Conservatives

Sharks, on the other hand, are much more conservative in the management of their energy accounts. Sharks are rarely endothermic. Indeed, their metabolic rate is low compared with the rates of most other fishes. Sharks often stop swimming and rest on the bottom. Their low energy requirements enable them to be opportunistic predators: by stopping and waiting they can take advantage of food sources that are unpredictable. Alternative patterns of energy acquisition and consumption such as those shown by tunas and sharks influence the form and behavior of organisms throughout the animal kingdom.

Undulatory swimming with the body and caudal fin together has been subjected to more intensive scrutiny than any other mode of fish swimming. Many fishes, however, rely on the oscillatory motion of short-based fins for propulsion, without any body motion. Such a mode of swimming is characteristic of fishes that rely on precise maneuvers at low speed to get food and to hide in small spaces. Many of the tropical fishes living in coral reefs show specialization for maneuvering.

Oscillatory swimming is not as well understood as undulatory swimming but it is currently being examined by Robert W. Blake of the University of British Columbia. Much of the work concerns the two distinct mechanisms that underlie the operation of oscillatory fins. The principles could be called the oar principle and the wing principle. Both types of short-based fin are found in pairs on the sides of the body. The paired fins generally operate symmetrically. The oscillatory motions of the two types, however, are different.

Fins that are like oars move back and forth horizontally much as the oars of a small boat do, alternating the power stroke and the recovery stroke. During the power stroke the fin moves in the direction opposite to the direction of overall motion. The fin blade is broadside to the fish's direction of motion and moves faster than the fish's body. During the recovery stroke the fin moves in the same direction as the body. The power stroke generates a substantial drag force that can be oriented to thrust the fish in the direction it chooses. In the recovery stroke no thrust is generated and the fin blade adds to the resistance that retards the motion of the fish through the water. The resistance contributed by the fin in the recovery is reduced by collapsing the fin blade and twisting it parallel to the water flow.

The shape of the fin has considerable influence on how much thrust the blade can be made to yield. Imagine the fin divided into narrow strips like the propulsive elements employed in the analysis of undulatory propulsion. The contribution of each blade element to the thrust is proportional to the area of the element and the square of its relative velocity (the velocity measured with respect to the body of the fish).

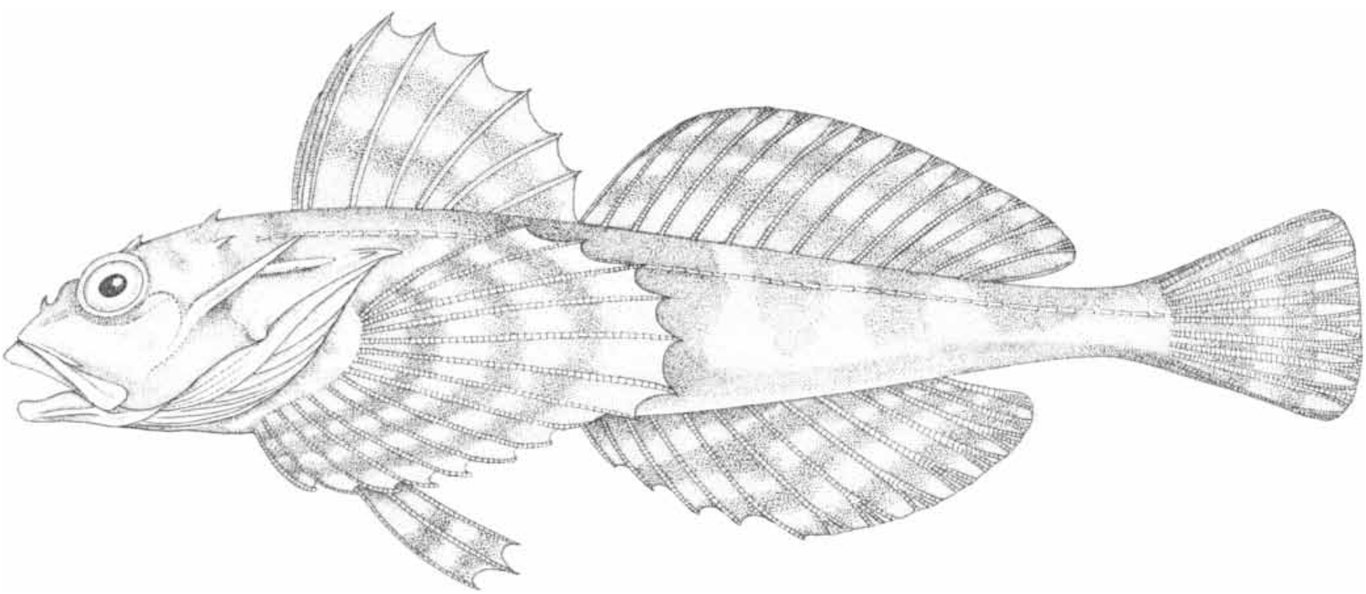
The blade elements farthest from the body (the ones at the end of the fin blade) traverse the greatest distance during the power stroke. Therefore such elements make the greatest contribution to the thrust. Indeed, elements close to the base of the fin sometimes move so slowly that they actually generate resistance rather than thrust. Hence the optimum shape for a fin that operates like an oar is a triangle with the apex at the fin base.

Fins that Are like Wings

Whereas a rowboat is propelled by a drag force, the wing of an airplane raises the craft by means of a lifting force. Because of the difference in the propulsive force, there are differences in design between fins that operate as oars do and fins that operate as wings do. When an oarlike fin moves, thrust is generated in the same plane as the plane of motion of the fin blade. Lift, on the other hand, is generated in a plane perpendicular to the direction of motion: when an airplane travels through the air, the lift is vertical although the wing is moving horizontally.

Therefore in order to propel the fish forward a lifting fin must move up and down in a plane that is roughly perpendicular to the long axis of the fish's body. Since the lift is generated at right angles to the plane of the fin blade, no recovery stroke is necessary. The fins generate lift on the upstroke and on the downstroke.

As is the case for drag, lift is propor-



SCULPIN is a specialist for acceleration with a design that yields the maximum thrust. The silhouette of a sculpin viewed from the side is quite deep. Fins on the upper and lower surfaces increase the area of the silhouette, thereby yielding the greatest thrust in transient swimming. The body is long and flexible, so that the fish can bend into a large-amplitude wave. The nonmuscle mass of the sculpin, however, is not small. The fish's body is only about 30 percent muscle, which implies that in fast starts the propulsive force is small compared with the inertial resistance. One reason for the increased inertia is the form of the head. The head is large and heavy, with a geometry that is advantageous in feeding on bottom animals. Thus the optimum design for transient swimming is compromised by nonlocomotor factors.

tional to the area of the fin blade and the square of the velocity of the blade. The lift force, however, can be an order of magnitude greater than the drag force generated by a fin blade with the same area. As a result of such factors the thrust yielded by lift-based fins is larger and more continuous than the thrust yielded by drag-based fins. Hence the fastest oscillatory swimmers have fins that operate like wings; these include the surfperch, the wrasses and the ocean sunfish. In contrast to oarlike fins, winglike fins are not limited to placement on the sides of the body; they can also be on the dorsal and ventral surfaces.

The shape of a fin that generates lift tends to be different from the shape of a fin that generates drag. One reason is the need to minimize crossflow, which is the flow around the tip of the fin. The same pattern is observed in air flowing around the tip of an airplane wing. In both instances crossflow leads to a decrease in lift and an increase in drag.

Airplane designers minimize crossflow by making the tip of the wing as narrow as possible. This is generally done by tapering the wing from the base to the tip. A fin cannot be tapered along its entire length because the fin must be narrow at the base in order to oscillate freely. Therefore lifting fins tend to be diamond-shaped, with a taper at the inboard end and at the outboard end.

Oscillating fins and long-based undulatory fins serve for slow swimming and precise maneuvering in structurally complex habitats such as weed beds and coral reefs. In such environments fishes hide in small spaces and pluck food from surfaces where it is growing or has settled. The surfaces are oriented in all directions, and in order to be able to put its mouth in the right place the fish must be capable of delivering small propulsive thrusts with considerable precision.

This implies that the fins should be flexible and should have the capacity to bend and rotate precisely in several directions. In addition, for slow swimming and delicate maneuvering some overall body forms appear to be better than others. Fishes that are successful in geometrically complex environments tend to have laterally flattened bodies with the outline of a disk or a diamond.

Such shapes offer the least resistance to rotation in the median vertical plane of the body. Moreover, many successful reef fishes have top and bottom fins that extend over the tapering rear segment of the body. There are paired pectoral fins high on the body and paired pelvic fins low on the body. Thus the propulsive appendages are distributed around the center of mass of the body and thrust can be directed in any plane. The butterfly fish is an example of a reef fish with such an optimum design for maneuvering. Fishes with that morphology tend to displace other species in habitats where precise maneuvering is at a premium.

My analysis of the three basic locomotor designs—for transient swimming, for sustained swimming and for maneuvering—suggests a fundamental conclusion: that it is not possible to combine all the optimum features for the different types of swimming in one fish. I call this the principle of the mutual exclusion of optimum designs. Theoretical and experimental work shows that the design elements resulting in the best transient-swimming performance and the elements resulting in the best periodic-swimming performance are mutually exclusive. There are also good reasons to think the optimum design for maneuvering excludes the elements that favor high performance in accelerating or in sprinting and cruising.

The butterfly fish, a maneuvering specialist, could not perform optimally in transient swimming and periodic swimming. During periodic propulsion the large area of its body and fins exacts a high price in increased drag. The shortness of the body leaves little room for muscle and therefore there is considerable dead weight, which impairs performance in a quick start. Moreover, with this type of body the large-amplitude movements needed for rapid acceleration are difficult to execute.

Conversely, fishes that are specialists in body-and-caudal-fin swimming are not particularly adept at slow swimming and maneuvering. Periodic swimmers have stiff fins that serve as keels for stability and hydroplanes for regulating the depth of the fish in the water; such stiff fins are of little use for delivering the small, precise thrusts needed in low-speed maneuvering. The elongated body of the transient-swimming specialist reduces its ability to turn in habitats with a complex geometry.

The Functional-Morphology Plane

Therefore it appears that there are three basic optimum designs for swimming and that the designs are mutually exclusive. It would be foolish, however, to conclude that all fishes should conform to one of the three basic designs. Even the most perfunctory examination of the shapes of fishes shows that most of them have locomotor designs with a combination of the design elements of the specialists. Hence the majority of fishes are locomotor generalists rather than locomotor specialists.

Such generalists clearly perform better than any specialist in the two broad types of swimming the specialist sacrifices to achieve high performance. No generalist, however, can perform as well as a specialist in the specialist's domain. Therefore in spite of the fact that few fishes approximate the optimum design for a particular swimming function the principle of the three basic designs is not invalidated. Instead the three fundamental designs should be regarded as defining a locomotor functional-morphology plane with the shape of a triangle.

Most fishes have a position on the functional-morphology plane, with the position depending on the particular combination of specialized design elements each fish shows. Some fishes, however, cannot be put on the plane. Among them are fishes with modes of life in which swimming has a secondary role, such as the anglerfish, which lives deep in the ocean and relies on mouth suction to catch its prey.

The Qualities of the Generalists

Straight line segments drawn across the triangular plane, from the specialists at the corners to the generalists in the middle, pick out groups of fishes with swimming qualities arranged in a continuum that is based on differences in the resources exploited by the fishes. One such pattern is shown in the success rate of the strike at prey. The tuna, at one corner of the triangle, catches from 10 to 15 percent of the fish it strikes at. The reason is that the specialized design of the tuna, which is excellent for cruising, limits acceleration and maneuverability. By cruising widely at high speed, however, the tuna can increase the number of prey it encounters and thereby increase the number of strikes it makes.

As we have seen, fishes shaped like the pike, which are at another corner of the triangle, catch from 70 to 80 percent of the fish they strike at. Because the pike is not able to cruise, however, it must wait for prey, which limits the number of prey it encounters and the number of strikes it can make. Fishes such as the trout and the bass, which are intermediate in form between the pike and the tuna, are also intermediate in the rate of successful strikes. Such generalists eventually catch from 40 to 50 percent of the fish they attack, and because they are generalists, they can also cruise looking for prey. Their compromises in design open up feeding opportunities that are closed to the specialists.

The biomechanical approach to the biology of fishes has thrown light on aspects of the animals' lives that could hardly have been illuminated in any other way. As a result it is becoming possible to correlate precise measurements of performance in the water with the fish's way of swimming and its way of life. One hallmark of a successful scientific model is that it leads to new knowledge that would have been overlooked in the absence of the model. The significance of hydromechanical models in understanding the lives of aquatic animals is an example of such success. In a general way it shows the role biomechanics can have in elucidating fundamental biological puzzles.

Alliance 2-door

Symbiotic Stars

Satellite observations have revealed that certain celestial objects with a peculiar spectrum consist of a red-giant star surrounded by a small dense nebula heated by a compact hot companion star

by Minas Kafatos and Andrew G. Michalitsianos

The process that makes new stars out of clouds of dust and gas seems to yield single stars and double stars in roughly equal numbers. Stars born in pairs circle each other in gravitationally bound orbits throughout their evolution. If the distance between them is large, their interaction may be negligible. The closer the two stars get to each other, however, the greater the effect on their evolution is and the more likely it is that the radiation they emit will exhibit features setting them apart from the common run of stars.

Among the several hundred million binary systems estimated to lie within 3,000 light-years of the solar system, and thus to be theoretically detectable on sky-survey photographs, a tiny fraction, no more than a few hundred, belong to a curious subclass whose radiation has a wavelength distribution so peculiar that it long defied explanation. Such systems radiate strongly in the visible region of the spectrum, but some of them do so even more strongly at both shorter and longer wavelengths: in the ultraviolet region and in the infrared and radio regions.

This odd distribution of radiation is best explained by the pairing of a cool red-giant star and an intensely hot small star that is virtually in contact with its larger companion as the two travel around a common center. Such objects have become known as symbiotic stars. On photographic plates only the giant star can be discerned, but evidence for the existence of the hot companion has now been supplied by satellite-borne instruments capable of detecting ultraviolet radiation at wavelengths that are absorbed by the earth's atmosphere (and therefore cannot be detected by instruments on the ground). Recently two symbiotic-star systems, the first to be detected outside our galaxy, have been observed in the Large Cloud of Magellan, one of the satellite galaxies associated with ours.

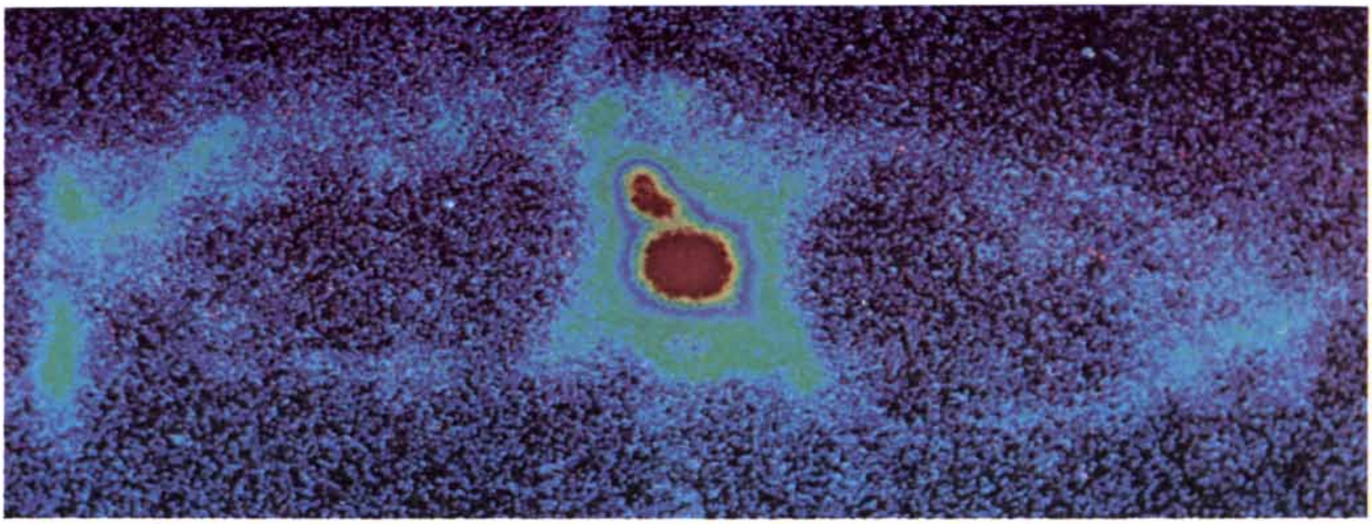
The spectra of symbiotic stars indicate that the cool red giant is surrounded by a very hot ionized gas. The existence of the ionized gas marked such objects as being peculiar several decades before satellite observations finally identified the ionizing source as the radiation from an invisible hot companion. Symbiotic stars also flared up in outbursts indicating the ejection of material in the form of a shell or a ring, reminiscent of the recurrent outbursts of a nova. Symbiotic stars may therefore represent a transitory phase in the evolution of certain types of binary systems in which there is a substantial transfer of matter from the larger partner to the smaller. It seems likely that in the course of the transfer the material often forms a disk around the smaller partner. In at least one instance, however, a symbiotic star has evidently expelled matter in the form of a highly directional jet, resembling the much larger jets associated with active galaxies and quasars.

The term symbiotic was applied to these peculiar star systems by Paul W. Merrill, who with Milton L. Humason studied them at the Mount Wilson Observatory in the 1930's and 1940's. The term was prescient because at the time there was no convincing evidence for the existence of a hot small companion star. Merrill was applying it to objects of a particular class whose spectrum at visible and near-infrared wavelengths suggests the presence of a cool red-giant star, about 200 times the radius of the sun, that is surrounded by a rarefied hot gas. Because a red giant with a surface temperature of 2,500 degrees Kelvin would be incapable of raising the temperature of the surrounding gas to a much higher temperature, Merrill suspected the presence of an unseen hot companion.

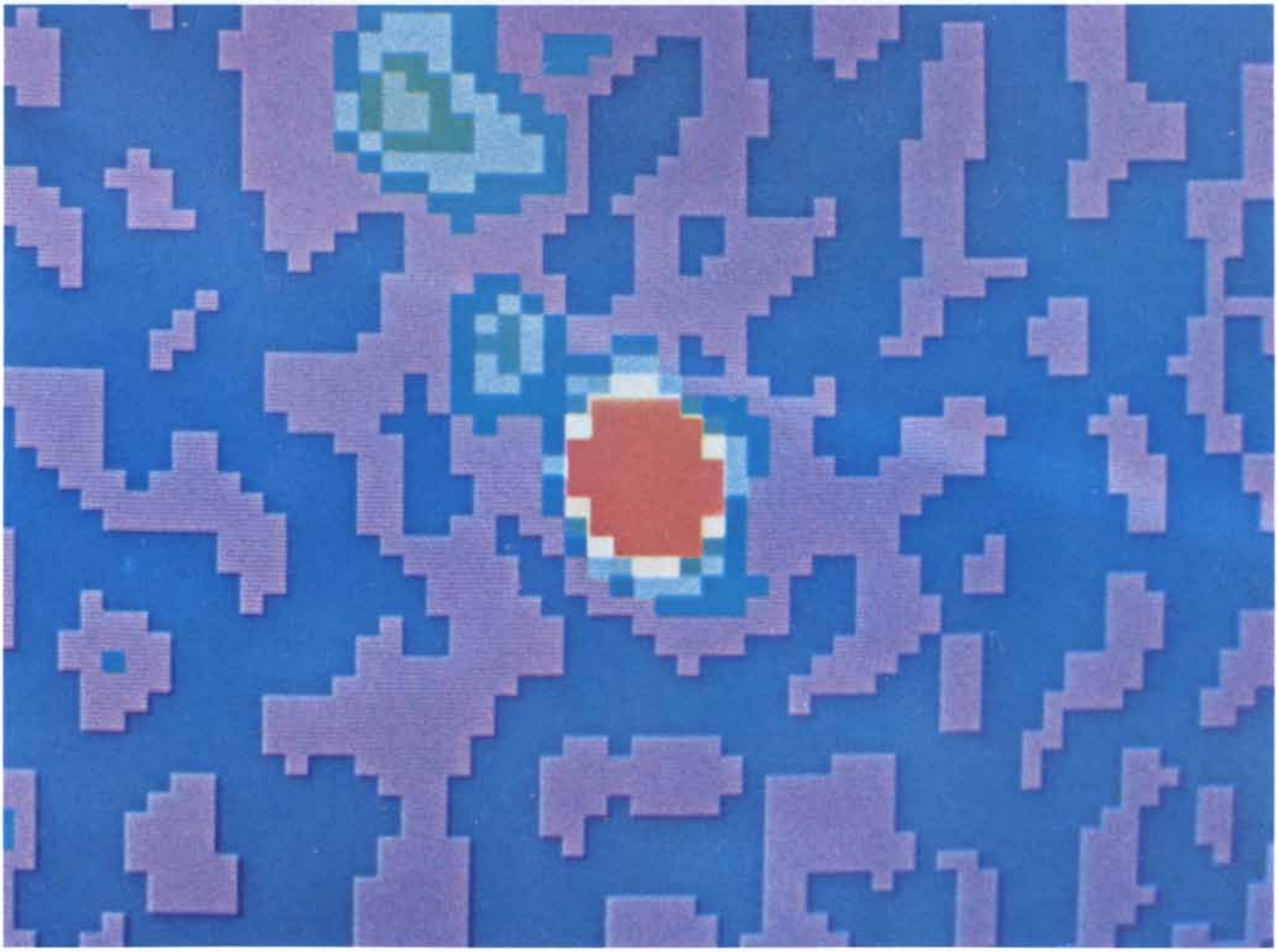
A formal model of symbiotic stars as binary systems was proposed independently 15 years ago by Alexander A. Boyarchuk of the Crimean Astrophysical Observatory and by Jorge Sahade of the Argentine Institute of Radio Astronomy. In this model a cool red giant and a hot compact star coexist in a binary system that is surrounded by a dense small nebula of ionized gas. The spectra recorded by instruments aboard the International Ultraviolet Explorer satellite (IUE) support the binary model. A few astronomers, however, are not quite convinced. They speculate that the observed states of ionization in the surrounding envelope of gas arise not from a binary object but from a single one. In what follows we shall adopt the binary interpretation.

On the basis of spectral surveys made with ground-based telescopes by Nicholas Sanduleak and C. Bruce Stephenson of Case Western Reserve University, Lloyd R. Wackerling of Northwestern University and Karl G. Heinze of the Johnson Space Flight Center of the National Aeronautics and Space Administration there are perhaps 20,000 symbiotic stars in our galaxy, of which a few hundred—those within a distance of 3,000 light-years—are accessible to observation. Since spectroscopists disagree somewhat on the criteria for classification, a more exact estimate cannot be given. As the satellite instruments yield more observations in the deep-ultraviolet region of the spectrum, however, the spectra of true symbiotic systems are rather easily distinguished from those of peculiar stars that seem similar in observations made from the ground.

Cool giant stars are so luminous that their emission completely dominates the output of a symbiotic system in the visible part of the spectrum. A red giant expands to prodigious size when the star, initially the size of the sun, exhausts the supply of hydrogen available in its core for thermonuclear burning. At that point the interior of the star has become so hot that nuclear burning extends to the shell surrounding the nuclear core, causing the outer layers of the star to expand. As the star swells by a factor of more than 50 its luminosity increases by a factor of several hundred. In the course of the expansion the surface temperature finally reaches 2,500 degrees K. and the star glows orange-red. The energy that supplies heat to the extended outer envelope originates in a thin hy-



SYMBIOTIC SYSTEM R AQUARII and its peculiar jet are shown in this false-color image derived from a photographic plate made with the three-meter telescope at the Lick Observatory by George H. Herbig of the University of California at Santa Cruz. R Aquarii is believed to consist of a cool red giant and an invisible hot companion whose radiation ionizes the red giant's surrounding compact nebula, causing it to glow. The two stars evidently circle each other in an elliptical orbit with a period of 44 years. At the point of closest approach between them there is presumably a sudden transfer of mass from the larger body to the smaller one. The interaction of that mass with the nebular gas may account for the jet, which is visible as a blob extending to the upper left. The entire system is enclosed within a lens-shaped diffuse nebula that may be the remnant of a nova-like outburst recorded by Japanese observers in A.D. 930. The image processing was done by Daniel A. Klinglesmith III of the Goddard Space Flight Center of the National Aeronautics and Space Administration.



RADIO IMAGE OF R AQUARII, which resolves the jet extending from the red giant into two blobs, was made with the Very Large Array radio telescope near Socorro, N.Mex., by the authors and Jan M. Hollis of the Goddard Space Flight Center. The color-coded map depicts the radio emission from R Aquarii at a wavelength of six centimeters. The scale of the radio map is 5.5 times larger than the scale of the photographic image. The separation between the central star and the more distant blob is about .01 light-year. If the blob was ejected from the central star between 1970 and 1977, as is suspected, its velocity lies between 800 and 1,800 kilometers per second.

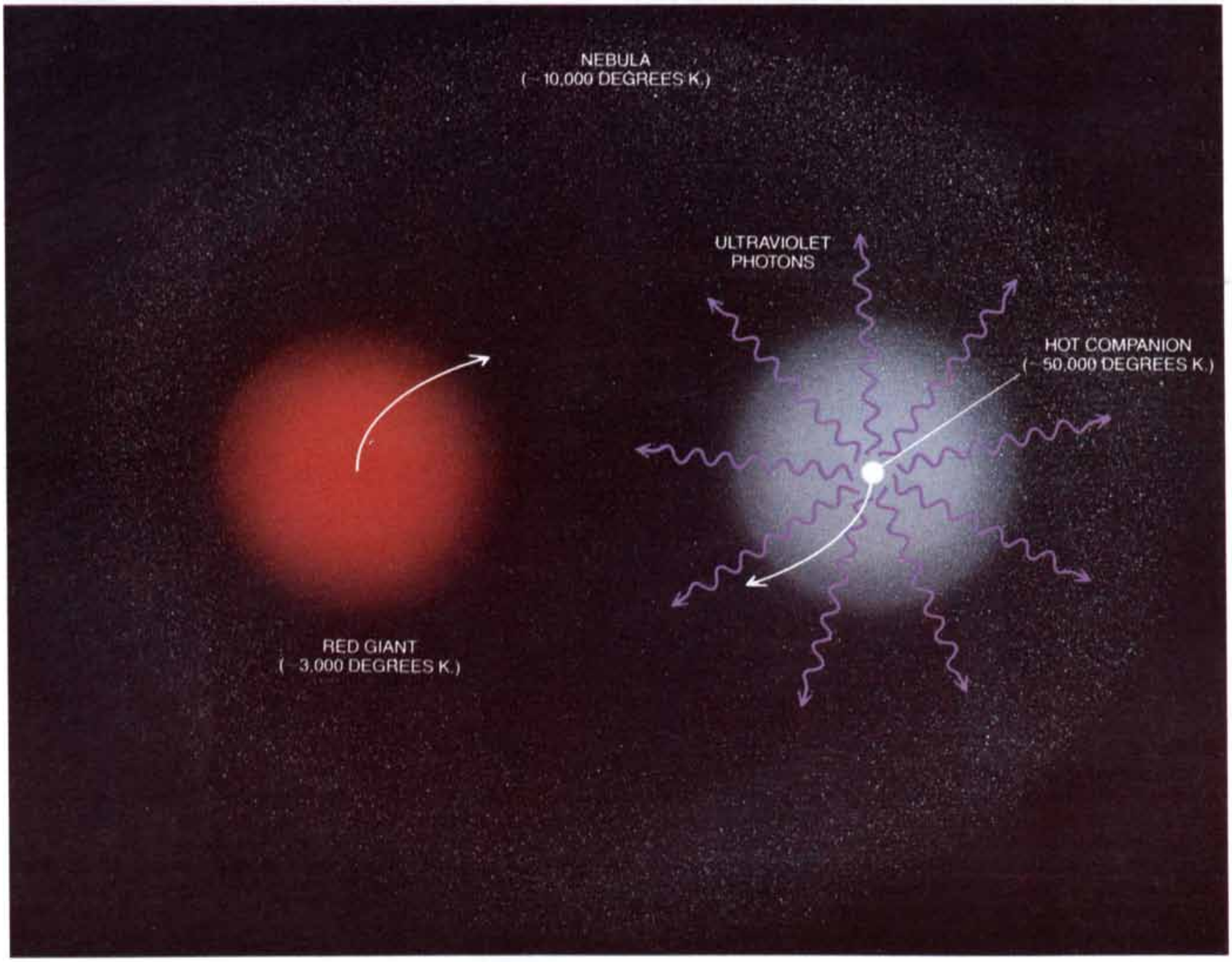
drogen-burning shell at the edge of the core. The star's high luminosity is explained by its vast size. Whereas the luminosity falls as the fourth power of the surface temperature, it rises as the square of the radius, so that the decrease in temperature is more than offset by the increase in radius.

The tenuous cool envelope of the giant star emits most of its energy in the visible and infrared regions of the spectrum, in accordance with the curve corresponding to the emission of a black body (a perfect emitter and absorber of energy) at a temperature of 2,500 degrees K. The emission curve declines exponentially with decreasing wavelength until in the region below a wavelength of 3,500 angstrom units at the blue end of the spectrum the emission from the red giant is barely detectable.

The postulated compact companion star of the red giant has a surface temperature of between 50,000 and 100,000 degrees K. The black-body emission from such an object rises continuously with decreasing wavelength, culminating with a maximum output in the far ultraviolet at wavelengths below 1,100 angstroms. Satellite instruments capable of observations in the far ultraviolet have now made it possible to probe the emission properties of the hot companion. The radiation from the luminous cool star that overwhelms the radiation from the hot companion in observations from the ground is itself overwhelmed in the ultraviolet.

The characteristic property that distinguished the visible-light spectra of symbiotic stars from those of other stars is the emission of excited hydrogen atoms at wavelengths ranging from the hydrogen-alpha line at 6,563 angstroms in the red to the hydrogen-delta line at 4,102 angstroms in the blue. In many symbiotic-star spectra there are also the emission lines of heavier atoms that are multiply ionized: stripped of two or more electrons. Such emission lines, which are not present in the spectra of normal red giants, also characterize the spectra of planetary nebulas: a hot star surrounded by a shell or ring of luminous gas. In addition many spectra of symbiotic stars exhibit dark bands indicating absorption by molecules in the extended outer atmosphere of the cool star.

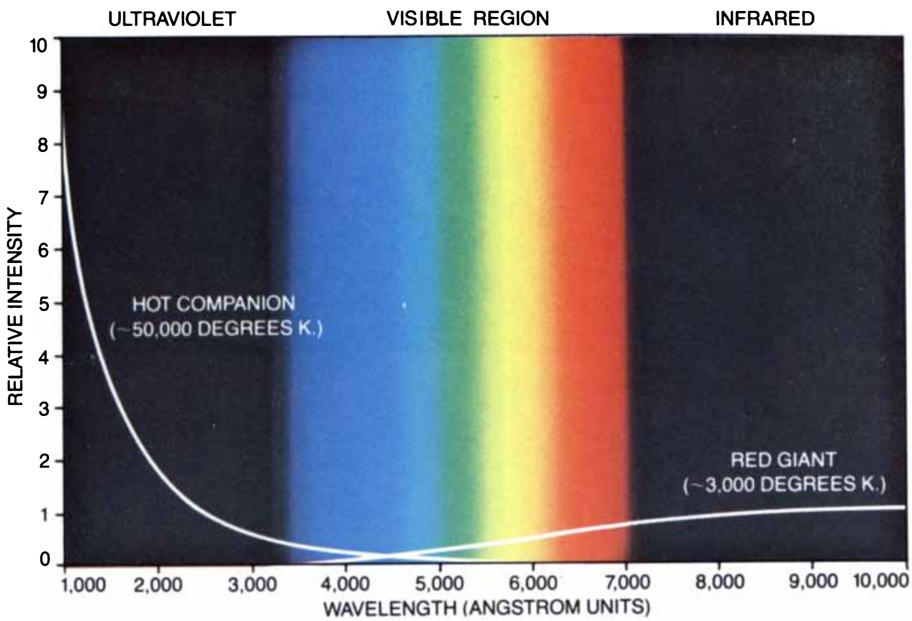
In early spectra of symbiotic stars recorded by Merrill the molecular absorption bands have a saw-toothed appearance. This can be attributed to unresolved transitions between energy states of compounds such as titanium oxide, vanadium oxide, carbon monoxide and in some spectra even water vapor. The compounds form in the atmosphere of the giant star and absorb or scatter photons in the continuous spectrum (as distinct from the line spectrum) of radia-



MODEL OF A SYMBIOTIC SYSTEM illustrates how the cool red-giant star is embedded in a nebula whose atoms have been ionized (stripped of one or more electrons) by the energetic radiation from the compact hot companion star. The nebula consists of gases expelled from the red giant in the form of a stellar wind. The radius of the giant star is at least 50 to 100 times the radius of the sun, so that its tenuous outer envelope is close to its small circling companion. The ionization of the nebula accounts for the many emission lines that characterize the spectra of symbiotic systems in the visible and ultraviolet regions of the spectrum. This version of the model depicts the symbiotic system in a quiescent state, a condition in which the emission lines in the spectrum have narrow, sharply defined profiles.

tion emitted by the star. Molecular absorption is notably responsible for the many saw-toothed features observed in the continuous spectrum of the symbiotic star Z Andromedae, regarded as the prototype of its class; this star has recently been closely examined in the ultraviolet by Roberto Viotti of the Istituto di Astrofisica Spaziale at Frascati in Italy and his collaborators.

Because a red giant has such a huge radius, the pull of gravity at its surface is low, about 10,000 times less than that at the surface of the sun. As a result the pressure exerted by radiation at the surface of the star can inflate the star's atmosphere, generating a wind of particles many times stronger than the sun's. As the wind of a red giant expands and cools, its constituents condense to form an envelope of dust consisting chiefly of compounds of silicon and carbon. The dusty envelope is revealed by enhanced emission in the far infrared at wavelengths between 10 and 20 micrometers, where the cooler envelope with its much larger volume outshines the red star itself. In a symbiotic system, however, a fraction of the diffuse envelope can be ionized by energetic photons from the hot companion star. Emission lines from excited atoms in the visible region of the spectrum were therefore an early clue to the presence of the hot star. In order to confirm this conclusion and to establish the nature of the companion it was necessary to explore other regions of the spectrum.



WAVELENGTH DISTRIBUTION of energy emitted by the two stars in a symbiotic system, shown schematically, is puzzling if observations are limited to the visible region of the spectrum at wavelengths between 3,400 and 7,000 angstrom units. If the emitting object were solely a red-giant star with a typical surface temperature of 3,000 degrees Kelvin, its emission should fall essentially to zero at wavelengths shorter than 4,000 angstroms. The slight upturn below 4,000 angstroms is the only clue to the existence of a hot companion with a temperature of 50,000 degrees K. To investigate the ultraviolet region in which the hot star's emission rises steeply to a maximum below 1,000 angstroms it is necessary to conduct observations from a satellite outside the earth's atmosphere. Such observations have been made with the International Ultraviolet Explorer satellite (IUE). The curves assume the two stars are equally luminous.

Satellite instruments for probing the ultraviolet region of the spectrum are optimized to detect wavelengths below 3,300 angstroms, the cutoff point for the transmission of ultraviolet radiation by the earth's atmosphere. Information about symbiotic stars has expanded at a furious rate since the International Ultraviolet Explorer went into operation in the spring of 1978. The IUE satellite is operated as a joint project by the European Space Agency, the British Science and Engineering Research Council and NASA. Controlled from stations at NASA's Goddard Space Flight Center in Maryland and at Villafranca del Castillo near Madrid, the satellite has gathered some 30,000 ultraviolet spectra of a diverse sample of objects, including several dozen symbiotic stars.

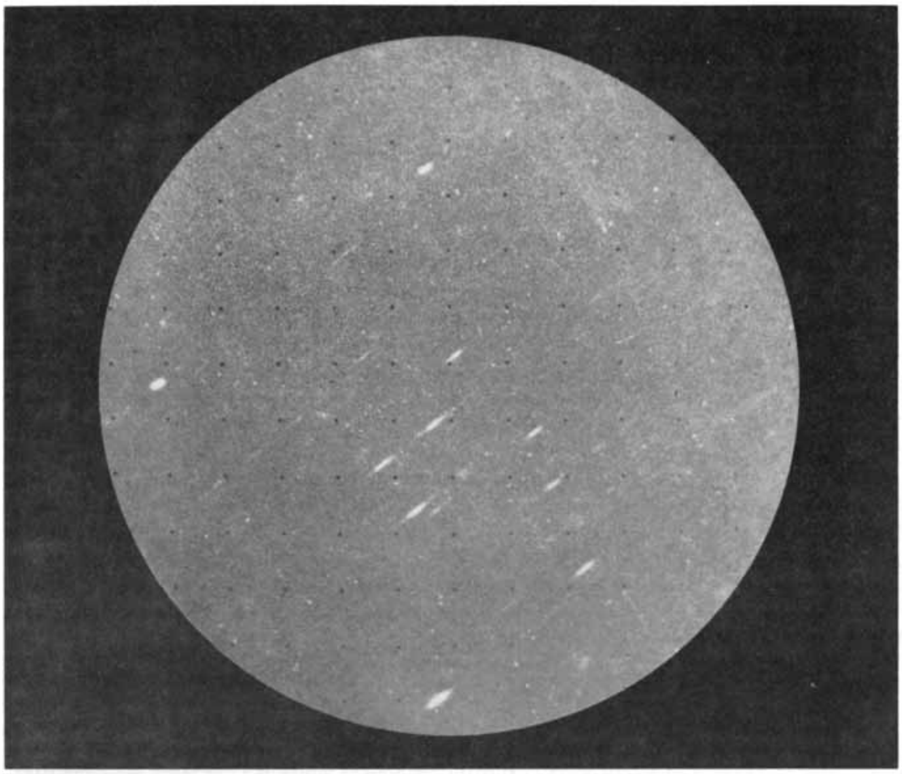
The satellite has a telescope with an aperture of 45 centimeters (18 inches) that gathers light for two spectrographs. Four Vidicon television cameras capture images of spectra in the wavelength range between 1,100 and 3,200 angstroms. The instruments can be operated at two levels of resolution. After selecting his target the astronomer can request ultraviolet spectra with a limiting resolution of six angstroms or one of .1 angstrom.

When the satellite spectra in the ultraviolet region are taken together with ground spectra in the visible, infrared and radio regions, there seems to be little doubt that symbiotic systems are binary. For example, the continuous spectrum of the symbiotic star AG Pegasi at visible wavelengths greater than 4,000 angstroms corresponds to the output of a cool red giant whose surface temperature is 3,300 degrees K. Observations from the IUE satellite analyzed by Mirek Plavec and Charles D. Keyes of the University of California at Los Angeles suggest that the continuous spectrum, which rises steadily with decreasing wavelength at wavelengths shorter than 4,000 angstroms, is nicely explained by the existence of a 30,000-degree companion star. The hydrogen lines that dominate the spectrum in the visible region evidently arise from a surrounding nebula of gas excited to 15,000 degrees. At that temperature a substantial fraction of the ionized hydrogen atoms in the nebula recombine with electrons and fall to lower levels of excitation, thereby yielding the emission spectrum of hydrogen.

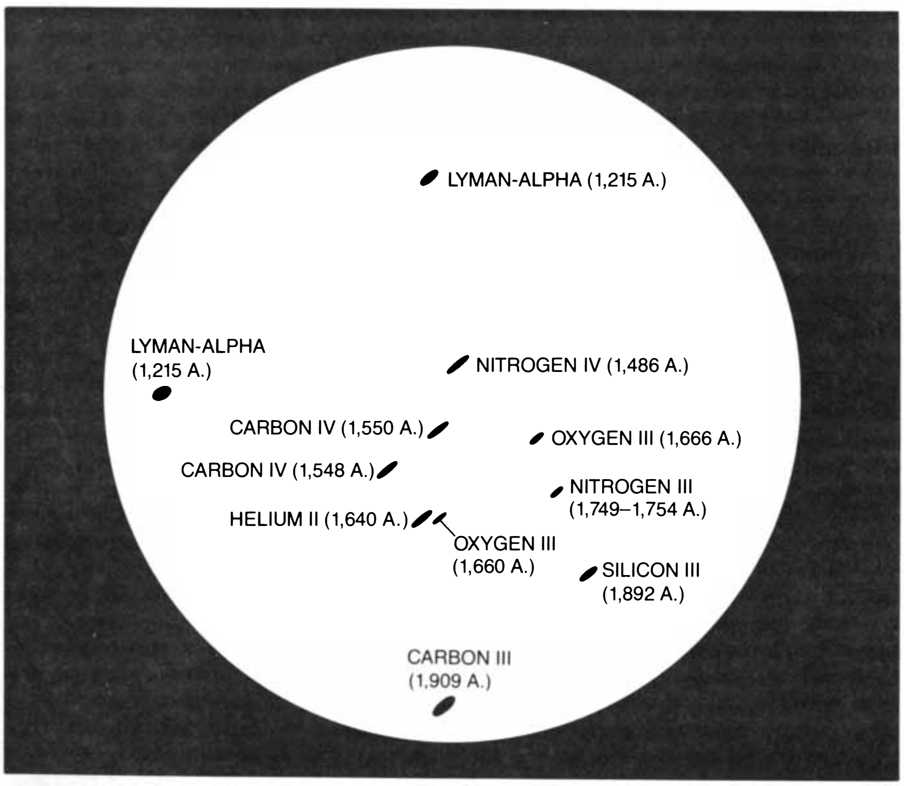
At wavelengths below 4,000 angstroms the ultraviolet emissions observed with the IUE satellite represent the radiation from multiply ionized atoms of helium, carbon, oxygen, silicon and nitrogen, exactly what one would expect from a highly excited gas. Triply ionized carbon yields the intensest lines at wavelengths of 1,548 and 1,550 angstroms. In other symbiotic systems these lines are rivaled in intensity by the emission of doubly ionized carbon at 1,907 and 1,909 angstroms. A system that exhibits the emission of once-ionized helium at 1,640 angstroms evidently harbors a very hot nebular gas, most plausibly excited by a hot companion star with a surface temperature of at least 65,000 degrees K.

The key to the nature of the companion star is the continuous spectrum in the ultraviolet. The satellite observations reveal continuous spectra of two distinct types. Some symbiotic stars have a continuous spectrum that rises steadily with decreasing wavelength in the range between 1,200 and 2,000 angstroms. This behavior is consistent with the presence of a compact hot star having a temperature higher than 20,000 degrees K. Good examples of this type of spectrum are the continuous spectra of the symbiotic stars RW Hydrae, AG Pegasi and Z Andromedae.

Other symbiotic stars exhibit a continuous spectrum that is essentially flat below 2,000 angstroms but that rises with increasing wavelength between 2,000 and 3,200 angstroms. Such behavior suggests the presence of a moderately hot companion star with a temperature of 10,000 degrees K., but this hypothesis runs into trouble. Mark H. Slovak of the University of Minnesota and we have independently calculated the



ULTRAVIOLET SPECTRUM of the symbiotic system RX Puppis was obtained by the authors with the International Ultraviolet Explorer. The spectrum covers wavelengths between 1,100 and 2,000 angstroms, divided into a series of spectral segments each about 20 angstroms wide. In each segment wavelength increases from the lower left to the upper right. The bright dashes are intense emission lines of multiply ionized atoms, identified in the illustration below.



PRINCIPAL EXCITED ATOMS in the spectrum of RX Puppis are those of carbon, nitrogen, oxygen, helium and silicon. The number of electrons stripped away in each case is one less than the number indicated by the roman numeral. The Lyman-alpha line of hydrogen at 1,215 angstroms is an emission feature created in the vicinity of the earth by the scattering of Lyman-alpha light emitted by the sun. The line appears twice in the spectrum because of a peculiarity in the way the IUE spectrometer disperses the ultraviolet radiation according to its wavelength.

spectra appropriate to a star of 10,000 degrees. We found the energy distribution did not match the spectra actually observed. Presumably the observed spectra arise from ionization processes in the nebula that are not fully understood or that are at least more complex than those postulated in our models. Examples of symbiotic systems with such anomalous spectra are CI Cygni, RX Puppis and R Aquarii.

Another revelation of the International Ultraviolet Explorer is that the ultraviolet emission of symbiotic stars can vary substantially over a period measured in months. IUE measurements of the ultraviolet continuous spectrum of the star SY Muscae suggested the presence of a hot dwarf companion star with a temperature of 50,000 degrees K. Nine months earlier, however, the spectrum indicated a much cooler object. There is no reason to think the brightening in the ultraviolet represents an outburst, because the emission lines superposed on the continuous spectrum did not change. If mass were being ejected violently from one of the stars, that fact would show up as a broadening of the emission lines. It now seems that the single depressed continuous spectrum of SY Muscae was recorded when the hot member of the system was briefly eclipsed by its cooler and more luminous companion. The hypothesis will be tested by future observations.

Actually two symbiotic systems are known to be eclipsing binaries: AR Pavonis and CI Cygni. On the basis of regular dips in the curve of the light they emit, the former system has an orbital period of 606 days and the latter a period of 855 days. If one assumes that each system has a total mass equal to twice the mass of the sun, it follows that the distance separating the two stars in each system is on the order of a few hundred million kilometers. That distance is comparable to the radius of the red giant and to the emitting region of the surrounding nebula. This implies that the component stars in symbiotic systems are almost in contact. The cool giant and its hot companion circle each other precariously, surrounded by a compact cloud of ionized gas whose emission is at times overwhelmed by the radiation from the hot gases released when there is a vast tidal transfer of mass between the two stars.

Infrared observations made by David Allen of the Anglo-Australian Observatory at Epping in Australia have revealed strong thermal emission from dust surrounding about a dozen symbiotic stars. Around other symbiotic stars, however, dust seems to be completely absent. The presence of dust dramatically affects both the overall brightness of the system and the general character of the ultraviolet continuous spectrum. Because small particles that absorb or scatter photons are comparable in size to the wavelength of ultraviolet radiation, they have their greatest effect in that region of the spectrum.

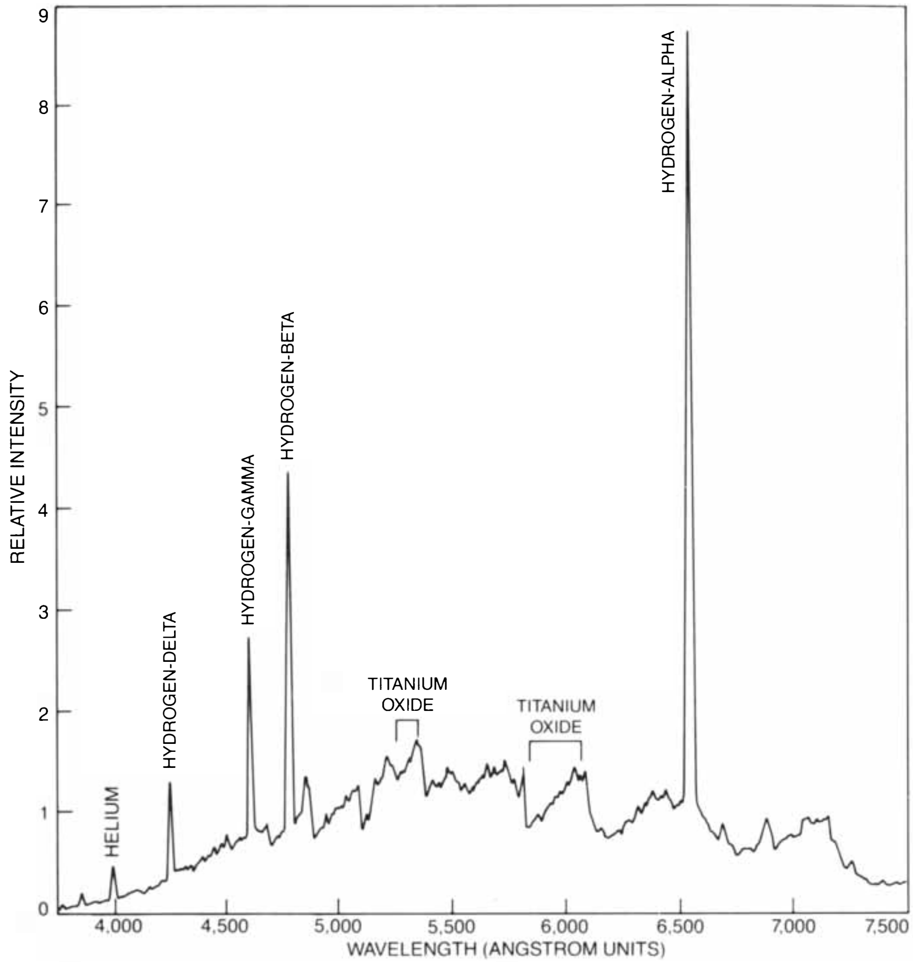
In interstellar space dust particles absorb or scatter not only ultraviolet photons but also photons of visible wavelengths, totally obscuring objects in or near the central plane of the galaxy at distances of more than a few thousand light-years. Astronomers refer to this extinction as reddening. For observations in the ultraviolet a conspicuous dip at 2,200 angstroms indicates the presence of graphite particles in the dust. In fact, astronomers commonly exploit the depth of the absorption at 2,200 angstroms to estimate the quantity of dust in the direction of ultraviolet-emitting stars.

Paradoxically a few symbiotic stars that exhibit a strong dust-reddening at visual wavelengths and a strong emission from dust at the infrared wavelengths between 10 and 20 micrometers show no sign of extinction by graphite particles at 2,200 angstroms. The absence of such absorption in symbiotic systems suggests that the extinction in the galactic plane may represent not dust particles widely diffused along the line of sight but much more localized concentrations of dust associated with the particular objects or regions being observed.

At the very least the absence of absorption at 2,200 angstroms in the direction of symbiotic stars otherwise showing clear evidence of dustiness argues that the physical properties of dust particles in symbiotic systems are quite different from the properties of particles thought to constitute dust in the interstellar medium. By the same token we are led to suspect that the stellar wind from ordinary red giants is not, as is commonly believed, the primary source of interstellar dust. If it were, in order to account for the absorption at 2,200 angstroms the dust particles being expelled from the red giant would have to change in size after entering interstellar space.

The chemical composition of the nebular material in symbiotic systems also presents puzzles. Analyses of ultraviolet emission lines from the hottest regions of the nebula indicate that certain heavy elements, notably silicon and carbon, are less abundant than they are in the universe at large. It is possible that the relative scarcity of such elements simply reflects peculiarities in the nuclear-burning processes of symbiotic systems.

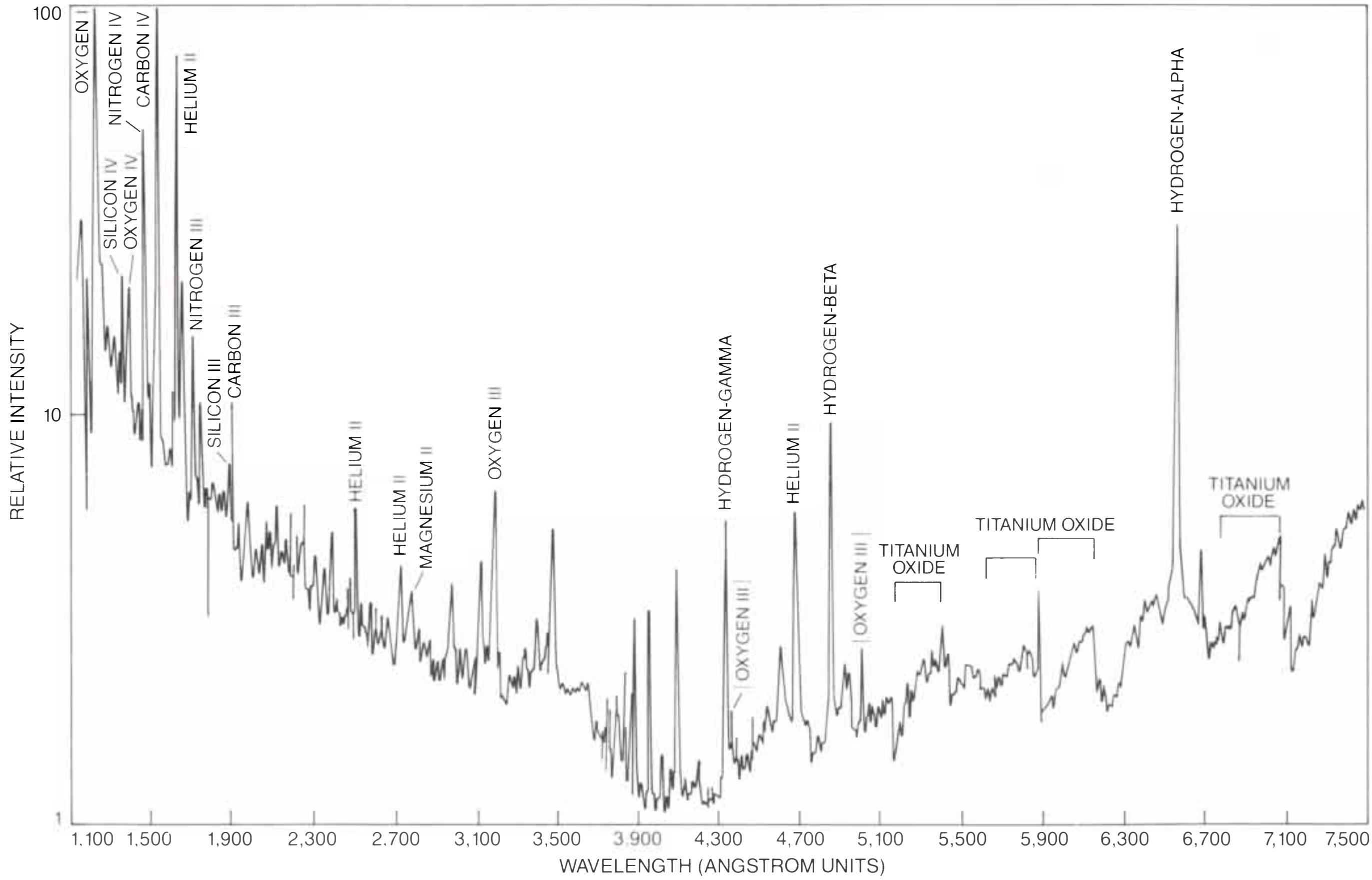
The scarcity is particularly evident in the spectra the International Ultraviolet Explorer has obtained of the two symbiotic stars in the Large Cloud of Magellan. One of the two, identified as LMC Anonymous for want of a better designation, exhibits strong emission lines of nitrogen and helium in the far ultraviolet but virtually no indication of the presence of carbon. The theory of the synthesis of elements by nuclear burning predicts that nitrogen and helium will be abundant with respect to carbon in the core of red giants in the phase of nuclear burning known as the CNO cycle (for carbon, nitrogen and oxygen). The excess of nitrogen with respect to carbon in LMC Anonymous could well be explained by the dredging up of nitrogen-rich material from the core of the red giant. Therefore strong nitrogen emission lines are evidence that the CNO cycle is operating in the red giant of the symbiotic system.



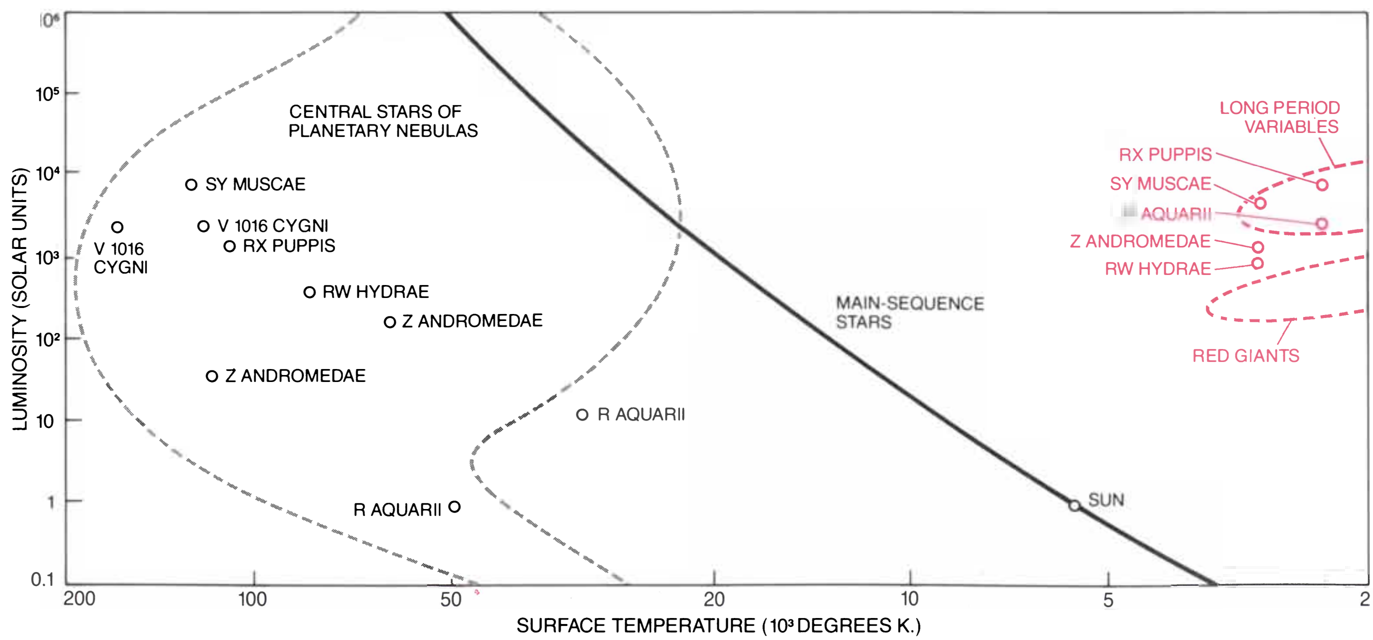
OPTICAL SPECTRUM of the symbiotic star Z Andromedae obtained by William P. Blair of the University of Michigan shows strong emissions of excited nebular gas rising above the continuous spectrum characteristic of a giant star with a surface temperature of about 3,000 degrees K. Z Andromedae is recognized as the prototypical symbiotic star. The strongest peaks are the emission lines of excited hydrogen atoms known as the Balmer series. They indicate that the nebula around the giant star is being ionized by a nearby hot source. The atmosphere of the red giant evidently also contains molecules of titanium oxide, which give the spectrum a saw-toothed appearance by absorbing some of the giant star's continuous-spectrum radiation.

The depletion of carbon observed in the spectra of symbiotic stars could be intensified by the precipitation of atomic carbon, along with silicon, into chemical compounds that end up in dust particles and so are no longer available for excitation and emission. Another suggestion, made by Allen, is that the luminous cool component of dust-rich symbiotic systems is not a typical red giant but rather a long-period variable star of the type known as a Mira variable (after the prototype star Mira Ceti). Mira variables typically pulsate with a period of several hundred days and a change in magnitude of five or more (a change in luminosity by a factor of 100 or more). Mira variables are known to be surrounded by grains of silicate dust at temperatures of between 800 and 1,000 degrees K. It is perhaps significant that symbiotic stars exhibiting a high level of dust-reddening also lack an ultraviolet continuous spectrum that rises with decreasing wavelength. We believe it is likely, therefore, that the dust effectively masks the expected rise in the ultraviolet continuous spectrum of the hot companion.

Important features of the ionized region in symbiotic systems can be deduced from the ultraviolet emission lines. A leader in such analysis is Harry Nussbaumer of the Swiss Federal Institute of Technology. The intensity of



ULTRAVIOLET AND VISIBLE SPECTRUM of AG Pegasi combines a visible spectrum recorded from the ground with an ultraviolet spectrum recorded from the IUE. It was prepared by Mirek Plavec and Charles D. Keyes of the University of California at Los Angeles. The continuous spectrum that rises with increasing wavelength to the right suggests the presence of a cool red-giant star with a surface temperature of 3,300 degrees K. The dips in the continuous spectrum are ascribed to titanium oxide, and the hydrogen Balmer lines indicate that the giant star is embedded in an excited nebula. The ultraviolet continuous spectrum rising steeply with decreasing wavelength to the left reveals the presence of a 30,000-degree hot companion star. The lines of twice-ionized oxygen in brackets are "forbidden" energy transitions, which can occur only in a gas so tenuous that collisions between the atoms are too infrequent for the excited species to decay spontaneously to their lowest energy state. Un-ionized oxygen I is normally found only in the atmosphere of very cool stars.



PLOT OF LUMINOSITY V. TEMPERATURE shows where the component stars of symbiotic systems fall on the classic Hertzsprung-Russell diagram. The large majority of the stars, still amply supplied with nuclear fuel, occupy the band called the main sequence, in which luminosity rises with increasing size and surface temperature. The red-giant members of symbiotic systems lie to the right near the region populated by typical red giants: stars that have become more luminous and cooler as their supply of nuclear fuel has dwindled. The hot members of symbiotic systems cluster to the left of the main sequence in a region populated by hot central stars of planetary nebulas. Because of uncertainties in measuring the spectrum in the ultraviolet, some hot members of symbiotic systems are plotted in two positions.

emission lines is proportional to the square of the electron density and to the volume of the emitting region, so that one can readily estimate the size of the ionized nebula. Nussbaumer finds this region is invariably only slightly larger than the diameter of the red giant. Certain emission lines of doubly ionized oxygen and of singly ionized helium are also a measure of the total amount of ionizing radiation needed to heat the system's gaseous nebula. If all the ionizing radiation is assumed to come from the hot companion star, the temperature and the luminosity of the star can be calculated and compared with those of more familiar stars. This is done on the classic Hertzsprung-Russell diagram, in which temperature is plotted against luminosity.

On such a diagram one finds that the hot secondary stars are roughly 10 times hotter for a given luminosity than typical stars in the galaxy, which occupy the band called the main sequence. It may be no coincidence that the central stars of planetary nebulas fall in the same area of the diagram as the hot members of symbiotic systems. The cool members of symbiotic systems generally occupy the area of the diagram where long-period variable stars are also found. Such stars have surface temperatures that are only about a tenth as high as those of main-sequence stars of similar luminosity. Their luminosity, in turn, is anywhere from 1,000 to 10,000 times greater than the sun's.

Strong ultraviolet emission lines can also reveal the density of matter in the symbiotic nebula. The typical density turns out to be high: about 10^9 particles per cubic centimeter, compared with about 10^4 particles in an equal volume of gas in planetary nebulas. This has led some astronomers to suggest that symbiotic systems may represent an early stage in the evolution of a planetary nebula. The argument against this idea is that most symbiotic systems are almost certainly binary whereas most planetary nebulas seem to harbor only a single star.

So far we have been considering only the ultraviolet emission of symbiotic stars. Do they also emit radiation of shorter wavelength, namely X rays? Operating between 1978 and 1981, the High Energy Astronomical Observatory 2, also known as the Einstein Observatory, observed about a dozen symbiotic stars in a search for X-ray emission between four and 12 angstroms. In reviewing the results Allen found no significant X-ray emission from typical symbiotic stars. In three instances, however, X rays were recorded from systems that exhibited peculiarities at other wavelengths: HM Sagittae, V 1016 Cygni and RR Telescopii.

X-ray emission is believed to arise when hot, swirling gases pouring out of a giant star form a disk as they fall onto a compact companion star. Such accretion disks are formed because the streaming gases leave the surface of the giant star with high angular momentum, a result of the high orbital velocity of the stars in the binary system. In such a tidal exchange the ejected material cannot travel to the surface of the compact star in a direct line but must circle toward it in an ever tightening spiral. If there is sufficient material, an accretion disk is formed. Although accretion disks probably exist in most symbiotic systems, they would be likely to be larger and to develop less heat than the tighter systems that generate X rays. As a result the symbiotic disks give rise to ultraviolet photons rather than X-ray ones.

One symbiotic system that shows direct evidence in the ultraviolet for a tidal exchange of material between the component stars is RX Puppis. The system was studied by the late Belgian astronomer Pol F. Swings and his son Jean Pierre Swings. They suggested that RX Puppis harbors a star somewhat resembling a "slow nova," a type of nova whose decrease in light emission following an outburst is slow compared with that of regular novas. Infrared observations of RX Puppis show the presence of dust and even water vapor, features characteristic of a cool Mira variable.

After a quiescent period lasting for roughly 40 years RX Puppis is now returning to a state of high excitation. High-resolution spectra obtained by us with the International Ultraviolet Explorer over a two-year period disclose significant changes in structure not seen in the spectra of any other symbiotic system. The multiple emission features in the spectra are displaced toward the red, evidence for the existence of streamers between the component stars of the system. Comparable spectra of other symbiotic systems exhibit narrow emission lines, indicating that the lines arise in compact, quiescent nebulas with no turbulent or streaming gas.

It will be difficult to observe accretion disks in symbiotic systems directly because the density of the gas in such a disk is probably too great for the formation of emission lines. Furthermore, the radiation from the luminous cool giant will overwhelm the continuous-spectrum radiation of the disk at visible wavelengths. Nevertheless, Geoffrey T. Bath of the University of Oxford and James E. Pringle of the University of Cambridge believe the energetic outbursts observed in some symbiotic systems are strong evidence that large accretion disks are formed whenever large volumes of material escape from the red giant and are pulled onto the hot companion. In their model a hot compact star, or perhaps a somewhat larger and cooler main-sequence star, exerts a strong gravitational pull that attracts the hot gases evolved in the system. As the volume of accreted material around the compact star increases, the gravitational attraction of the compact star is steadily diminished by the growing mass of the matter around it and finally becomes weaker than the radiation pressure acting on particles in the accretion disk. At that point the radiation pressure triggers a massive outflow of accreted gas in the form of a powerful stellar wind. The hypothesis is supported by a certain amount of spectroscopic evidence.

It has been calculated that the wind thus triggered has a much higher velocity than the wind normally flowing out of individual red-giant stars. Drawing on this model, Sun Kwok of the University of Calgary, George Wallerstein of the University of Washington and Lee Anne Willson of Iowa State University have proposed that the low-velocity wind of the red giant may collide with the higher-velocity wind from the accretion disk around the compact star, yielding a highly ionized sheet of material between the two stars. One way or another, if accretion disks are actually present in symbiotic systems, they could readily account for the high state of excitation observed in the nebulas of such systems.

So far it has not been possible to determine the exact nature of the compact star around which an accretion disk is likely to form. It may well be that a disk forms in symbiotic systems that happen to have as a junior member a fairly cool, sunlike star rather than a hot, compact one. In collaboration with Robert E. Stencel of NASA and Boyarchuk we have recently found evidence for a sunlike star in CI Cygni, the eclipsing binary whose orbital period is 855 days. As we noted above, the ultraviolet spectrum of CI Cygni is unusual in that its intensity does not increase with decreasing wavelength, indicating that the temperature of its secondary member is less than 10,000 degrees K. Our new evidence suggests the secondary may be a star as cool as 6,000 degrees.

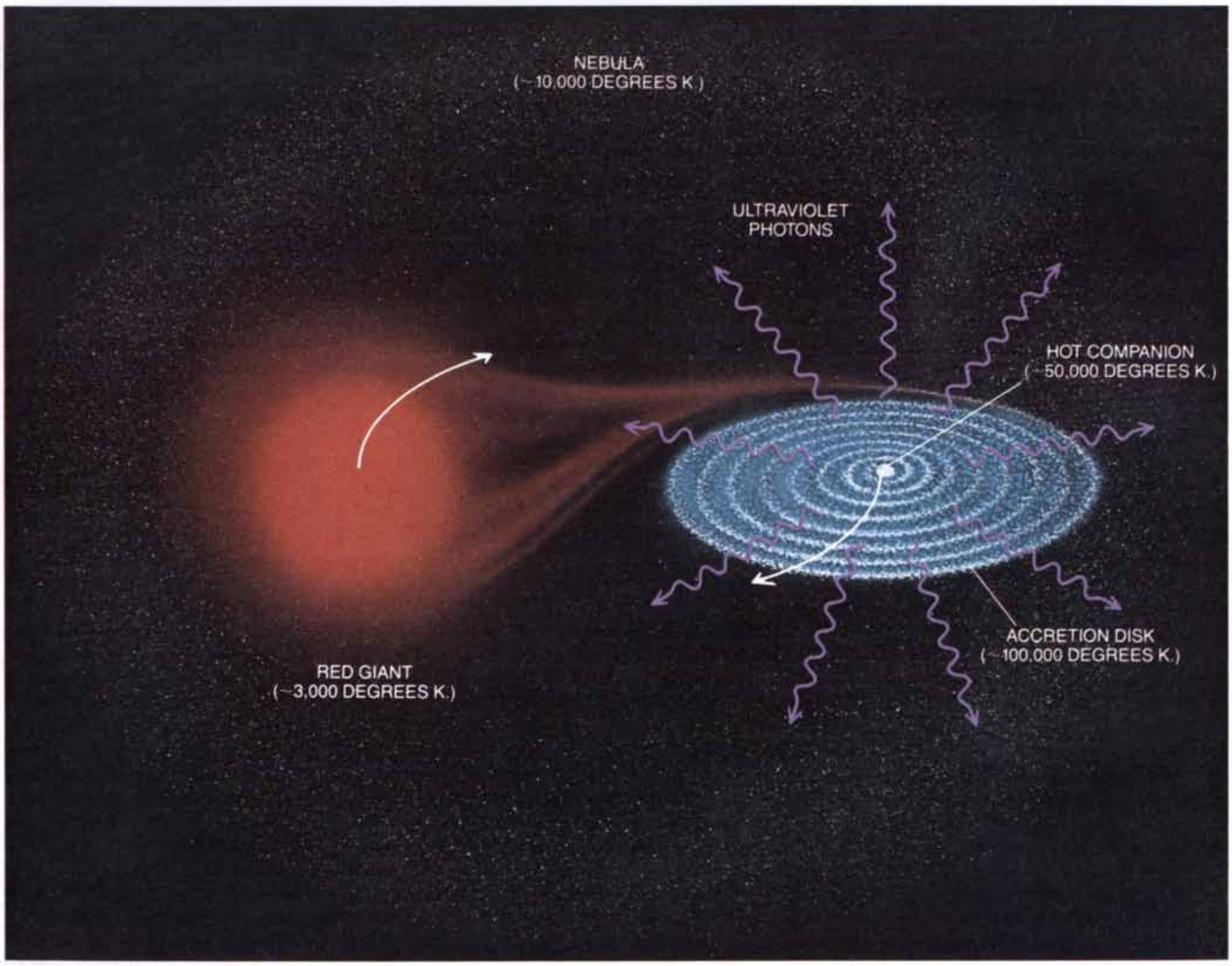
The nature and temperature of the compact secondary determines the characteristics of any outbursts that occur. Accordingly observations of outbursts in symbiotic stars should yield important clues to the entire system. The cause of such outbursts remains uncertain. Two quite different hypotheses have been advanced. In the model proposed by Bath outflowing matter from the cool giant star forms an accretion disk whose temperature may reach 100,000 degrees K. near the surface of the compact companion because of turbulence and friction in the swirling gases. This subjects the surrounding gas to intense radiation pressure that drives great volumes of material away from the system in the form of a stellar wind.

Because in Bath's model the main source of high-temperature radiation is the accretion disk, it matters little whether the compact secondary is a hot subdwarf star, the central star of a planetary nebula or a cooler main-sequence star. Recently Scott J. Kenyon of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory has calculated energy distributions for accretion disks in symbiotic stars. His work supports the existence of an accretion disk in some symbiotic systems. The intensity of the outflowing wind would be modulated by variations in the rate at which mass is transferred from the larger star to the smaller one. Such variations might reflect periodic upheavals in the outer envelope of the red giant or changes in the orbital distance between the two members of the system.

According to a second model, proposed by Bohdan Paczynski and Bronislaw Rudak of the Copernicus Astronomical Center in Warsaw, the periodic outbursts in symbiotic systems result from the sudden onset of nuclear burning at the surface of a white dwarf. In such a dense, hot star, with little hydrogen left in it, the nuclear fires have long since been banked. As material ejected from the red-giant companion accretes onto the surface of the dwarf star, however, pressure near the surface steadily increases until the remaining hydrogen is ignited. Variations in the rate of mass transfer modulate the intensity of the nuclear reaction, thereby accounting for the observed outbursts. Conceivably each model may be valid for some fraction of symbiotic systems. Only Bath's model, however, can explain the outbursts observed in CI Cygni, in which a white dwarf is clearly absent.

Symbiotic stars present still other puzzling phenomena. The most bizarre, perhaps, is associated with the peculiar star R Aquarii, the brightest symbiotic system known in the visible region of the spectrum. Issuing from R Aquarii is a large jet of material similar in appearance to the high-velocity jet that extends from the core of an active galaxy. Unlike such a jet, whose velocity approaches the speed of light, the jet from R Aquarii is traveling at only several hundred kilometers per second, or about a thousandth the velocity of light.

Nevertheless, the highly directional expulsion of matter from R Aquarii is both fascinating and unexpected. It was first noticed by Wallerstein and Jesse L. Greenstein of the California Institute of Technology and later studied by George H. Herbig of the University of California at Santa Cruz. The R Aquarii jet may be similar to the low-velocity blobs that have been observed in gas clouds where new stars seem to



OUTBURSTS IN SYMBIOTIC SYSTEMS have been explained by two principal models, which are depicted schematically on these two pages. The accretion-disk-outburst model, on this page, suggests that the enhanced ultraviolet and visible emission observed in some symbiotic systems arises from a disk that forms around the hot member of the pair following a tidal outflow of material from its red-giant companion. Because of the high-velocity orbital motion of the two stars, the material expelled from the tenuous envelope of the red giant is not deposited directly onto the surface of the small hot star but forms a disk as it spirals inward. Collisions among the swirling particles heat the accretion disk to approximately 100,000 degrees K., so that the disk itself becomes a strong source of ultraviolet radiation.

be forming. Ejected blobs are also reported near the planetary nebula known as Abell 30.

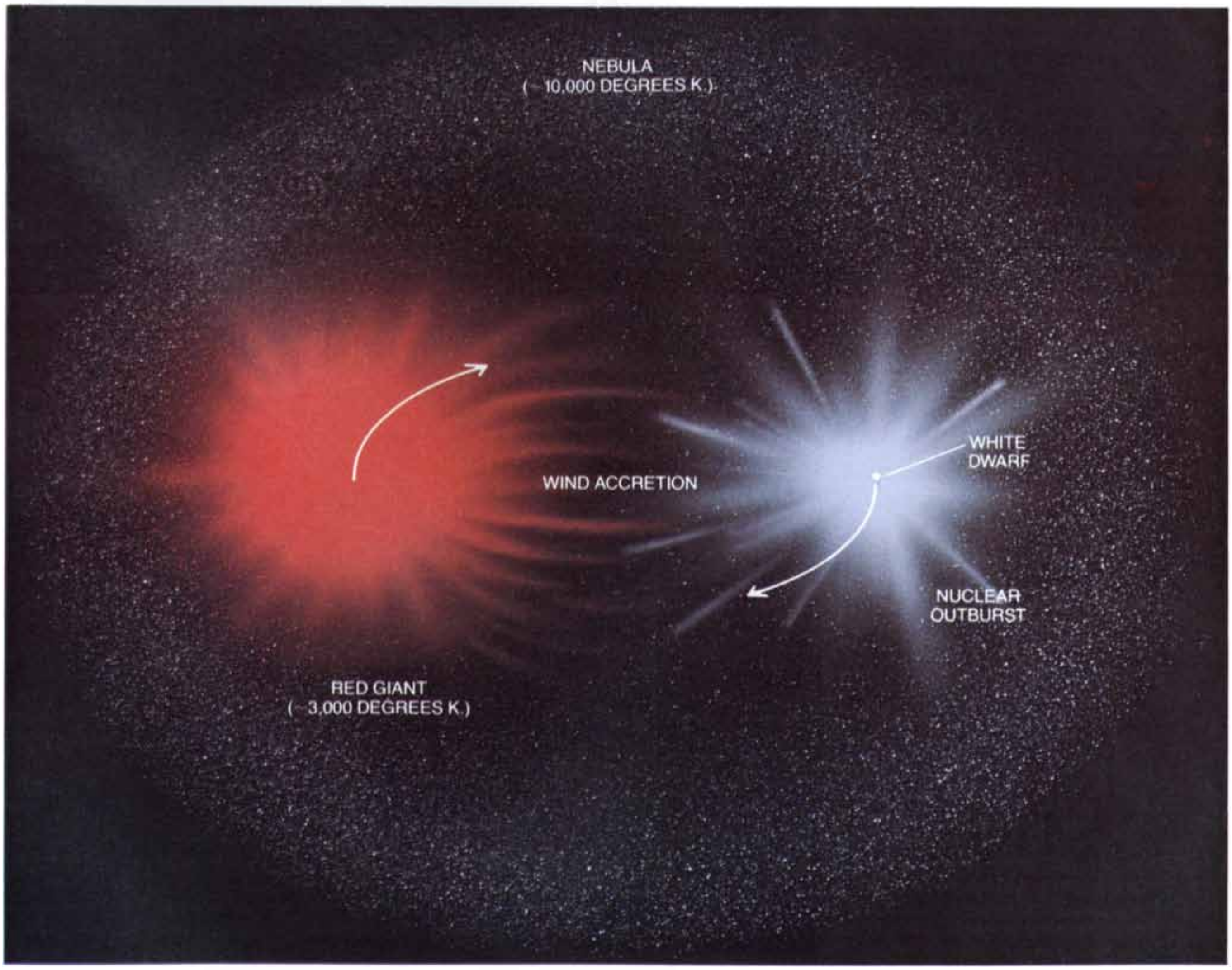
Recently we and others have observed the R Aquarii jet at radio frequencies with the Very Large Array radio telescope near Socorro, N.Mex. We have concluded that the jet forms as the two stars in the system make a close approach every 44 years in a highly elliptical orbit. At such times the envelope of the giant star, probably a Mira variable, comes almost in contact with its much hotter companion, and there is a rapid transfer of mass.

This model is supported by the most recent data we have obtained with the Very Large Array in collaboration with Jan M. Hollis of the Goddard Space Flight Center. At radio wavelengths between two and 20 centimeters the energy distribution of the jet remains constant with wavelength, which is unlike the slope exhibited by the central ionized region that engulfs the binary system. This finding suggests the jet is powered by thermal processes of the kind observed in gases heated by shock waves. Such shock heating could result if ejected material were to plow into the surrounding tenuous nebula, thereby raising its temperature sharply above that of the quiescent material.

The radio emission from the central ionized nebula is thermally generated by the ultraviolet radiation from the hot companion and its accretion disk. The accretion rate is so high when the two stars are closest together that radiation pressure drives excess material out of the system along an axis perpendicular to the plane of the orbit and the accretion disk. Whatever the detailed explanation, it seems remarkable that a symbiotic system can create a collimated stream of high-velocity material.

According to our model, the present jet formed during a close encounter in the mid-1970's. Presumably if it did, there should have been a similar jet 44 years earlier, in the 1930's. The emergence of the jets at each close encounter may be related to the strong variations in light emission that have also been observed in R Aquarii. In the mid-1930's R Aquarii flared up in an outburst Merrill compared to the behavior of a slow nova. The entire system is enclosed by a lens-shaped nebulosity. Merrill and other observers suggested the nebulosity is the remnant of a major outburst that may have taken place some 600 years ago. We now believe, however, that the outburst was a "guest star" described by Japanese astronomers in A.D. 930, hundreds of years earlier. Today we would call it a nova.

The exact evolutionary course that turns a binary system into a symbiotic one is a matter of conjecture. The comparatively small number of known

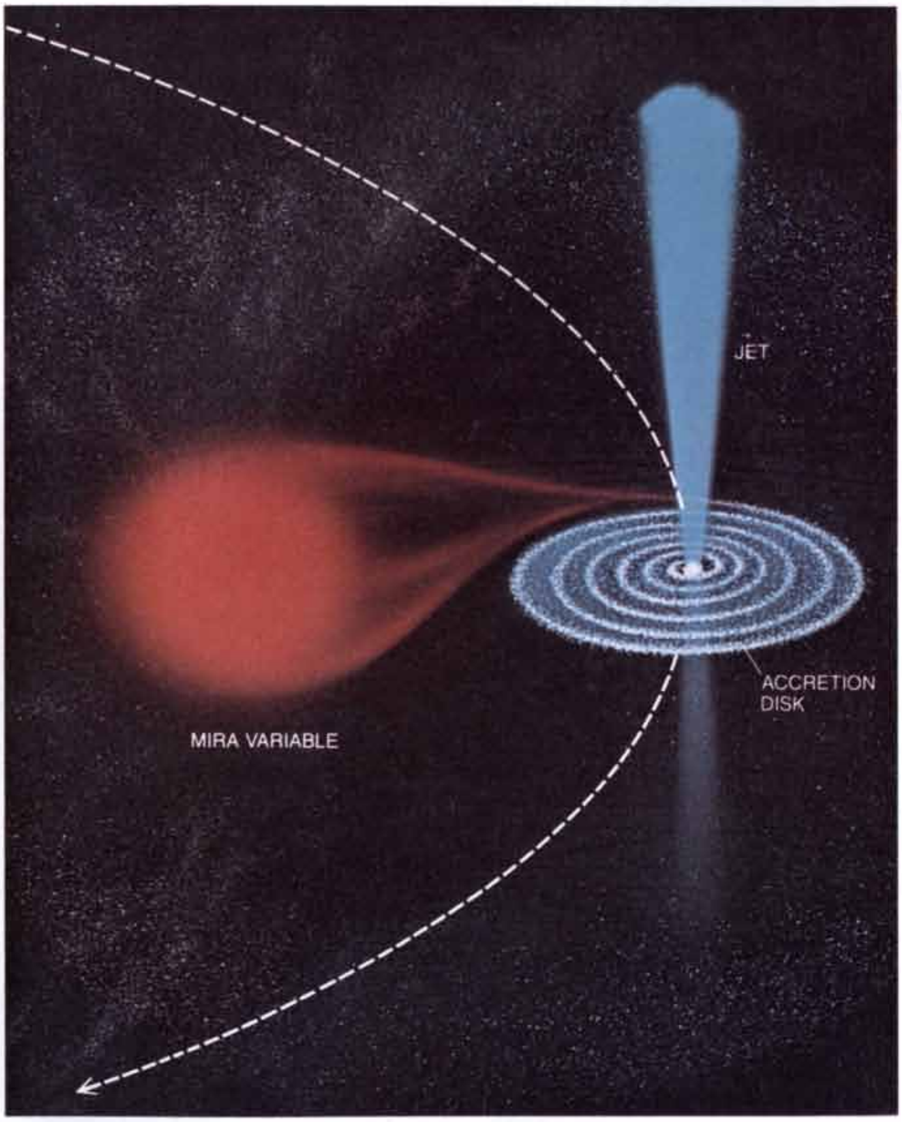


THERMONUCLEAR-OUTBURST MODEL also calls for the transfer of material from the red giant to its smaller and hotter companion, but the possible formation of an accretion disk is not critical to the argument. In this model the hot star is specifically a white dwarf, a star that has exhausted its supply of nuclear fuel and has collapsed to roughly the size of the earth. Such a star will still shine for billions of years as it slowly cools. The material ejected from the red giant accumulates on the surface of the white dwarf, forming a hot ionized envelope that impedes the flow of radiation from the dwarf's surface. Ultimately the temperature of the superheated shell of the dwarf reaches some 100 million degrees K., at which point thermonuclear reactions are rekindled and an outburst of radiation follows.

symbiotics in our galaxy suggests that if all binaries of modest mass normally pass through a symbiotic phase in their evolution, the phase must be extremely brief, perhaps as short as a million years. It is suspected that the evolutionary course of a binary system is predetermined by the initial mass and angular momentum of the gas cloud within which binary stars are born. Since red giants and Mira variables are thought to be stars with a mass of one or two suns, it seems plausible that the original cloud from which a symbiotic system is formed can consist of no more than a few solar masses of gas.

The distance separating the stars at the time the gas cloud condenses may be the critical factor in determining the final stages of binary evolution. Thus symbiotics seem to constitute a family of objects distinct from the binary systems that exhibit strong X-ray emission. X-ray binaries have extremely short orbital periods, ranging downward from days to hours. The typical orbital periods of symbiotics run to hundreds of days, indicating a much greater separation between the two components.

One must be cautious, however, in trying to infer the initial properties of gas clouds that may give rise to different kinds of binary systems, because the rate of loss of angular momentum and mass from such systems is unknown. Other complicating factors, such as the presence of stellar magnetic fields, created by the dynamo action of stellar rotation, and the way in which such fields may affect the evolutionary course of the system, have not yet been explored even in a preliminary way.



R AQUARII JET seems to be created when the system's hot secondary star, in its elliptical 44-year orbit, comes closest to its cool companion. At that time material ejected from the giant may rapidly form an accretion disk around the smaller star. Collisions among the swirling particles heat the disk to a high temperature with a corresponding high luminosity. By a mechanism that is still obscure material is driven away from the system along an axis perpendicular to both the disk and the orbital plane of the system. As the collimated stream leaves the system, it cools and forms the blobs of gas that are observed in photographs and radio images.

Multilingual Word Processing

The advantages of computerized typing and editing are now being extended to all the living languages of the world. Even a complex script such as Japanese or Arabic can be processed

by Joseph D. Becker

The personal computer has become a familiar fixture in the office and even in the home. It is useful in many ways, and yet for all its novel applications it is probably most useful to the greatest number of people when it serves the function of a typewriter. In that role it enables the user to see text displayed on a screen, so that words can be reviewed and revised before they are ever committed to paper. The kind of computer program underlying such an ability is known as word-processing software.

So far computers have largely been limited to the processing of words in the English language. That is not surprising: most computers have been developed in English-speaking countries, and English is the principal language of international commerce. Yet there is no technical reason for word processing to be confined to English. Indeed, it is possible for word-processing software to handle not only French, German, Italian, Russian, Spanish and other European languages but also more complex scripts such as Arabic, Chinese, Hebrew, Japanese and Korean. My colleagues and I at the Xerox Corporation have been developing multilingual word-processing software for a personal computer work station called Star, which is manufactured by Xerox. Our basic idea is that the computer should deal with a universal notion of "text" broad enough to include any of the world's living languages in any combination.

In effect, therefore, the fascinating diversity of mankind's written symbols must be made to coexist in the computer. At first it hardly seems possible. Arabic script, for example, flows from right to left in curlicues. Thai and other scripts, originally from ancient India, have letters that sometimes step around their neighbors and thus get out of phonetic order. Occasionally a letter even surrounds its neighbors. Korean groups its letters in syllabic clusters. (The Korean alphabet was designed from scratch by a group of scholars in 1443.) Chinese, the most ancient of living writing systems, consists of tens of thousands of ideographic characters. Each character is a miniature calligraphic composition inside its own square frame. It seems the developers of the computer and of word-processing software were coddled by the English language, which happens to have the simplest writing system of all: unadorned alphabetic letters laid out one after the other.

How can computer software originally designed to handle only English text be broadened to encompass the full diversity of the world's writing systems? The many challenges of the task can be divided into three basic realms. There must be a way for text to be represented in the memory of a computer; there must be a way for text to be typed at the keyboard of the computer; there must be a way for the computer to present text to the typist. I shall refer to these realms as encoding, typing and rendering. By rendering I mean both the display of text on the screen of a computer and the printing of text on paper.

Encoding is governed by a single, basic fact: the computer can store only numbers. Indeed, it can store only binary numbers, consisting of strings of 0's and 1's. Hence text is represented in a computer by storing a binary code number for each letter. In the case of the English language the American Standard Code for Information Interchange, abbreviated ASCII, assigns the binary code number 01000001 to the letter *A*, 01000010 to *B*, 01000011 to *C* and so on. Thus when you type an *A* on a computer keyboard, the computer is really being instructed to store the code number 01000001. When the computer comes to display or print a letter encoded as 01000001, its instructions cause it to draw a symbol you recognize as an *A*. As long as the input and output instructions are consistent, you have the illusion that the letter *A* itself was stored.

Computers generally store information in units of bytes, where each byte is a group of eight bits. It therefore seems a sensible strategy to store text as one byte per character. The trouble is, there are only 2^8, or 256, ways in which eight 0's and 1's can be combined in a byte. The living scripts of the world have far more letters than that. A two-byte coding scheme, in which each letter would be identified by two successive bytes, would yield 2^{16}, or 65,536, possible codes; a three-byte scheme would yield 2^{24}, or well over 16 million, codes. But employing two or three bytes per letter where only one byte is needed would waste space in the computer's memory. The answer is to arrange for the encoding to expand to two or three bytes per letter only when necessary. This can be done by setting aside a few bytes as signals to the computer and putting those signals into encoded text.

The first step in establishing a scheme for multilingual text encoding is to assign a binary code number to each of the alphabets of the world. The Roman alphabet is assigned 00000000, Greek 00100110, Russian 00100111, Arabic 11100000 and so on. (The particular choices, like the choices for individual letters, are based on international standards.) Next the code 11111111 is designated the shift-alphabet signal. The

MULTILINGUAL TEXT occupies much of this "screen dump," which shows the content of the display screen of a computer. At the upper right is a "virtual keyboard." It establishes that the Arabic alphabet has been assigned to the standard keyboard attached to the computer. Pressing the *g* key, for example, stores the Arabic letter *lam*. Below the virtual keyboard is a document in Arabic and English. The former reads from right to left, the latter from left to right; the computer enables the typist to embed one language in the other while preserving the directionality of each script. Below the bilingual text is a set of "icons" that represent facilities such as printers. The left is occupied by a document that shows symbols available to the typist.

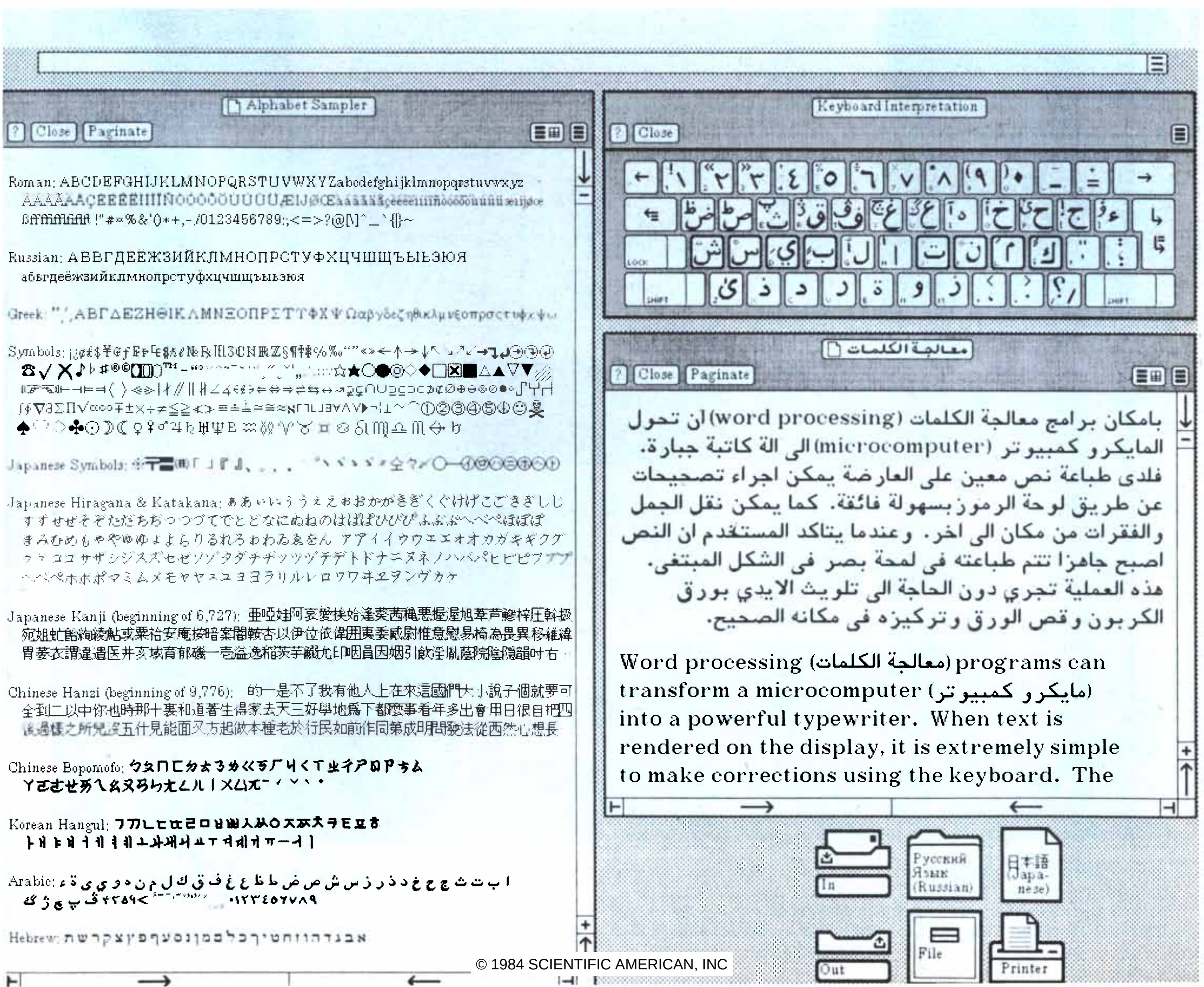
computer is to start by assuming the text is in alphabet 00000000—that is, in Roman script—but whenever it encounters 11111111, it is to interpret the next byte as the code specifying a new alphabet and the succeeding bytes as codes for letters in that alphabet [*see top illustration on page 100*].

That solves the text-encoding problem for the world's phonetic alphabets. The Chinese ideographs remain: they push the total number of letters above the 65,536 the scheme can encompass. To accommodate all the ideographs one must create another level in the encoding hierarchy. One can consider a group of 65,536 letters to be a "superalphabet" and specify two bytes of 11111111 in succession to be a shift-superalphabet signal. The signal causes the computer to interpret the following byte as the code number of a new superalphabet and succeeding pairs of bytes as codes specifying symbols in that superalphabet. The main superalphabet, designated 00000000, is all one needs except for very rare Chinese characters.

This strategy of "flexible encoding," which was devised by Gael Curry of Xerox's Office Systems Division, opens the computer to a range of more than 16 million characters, including all Chinese ideographs. At the same time it optimizes the storage of text by encoding ordinary (that is, phonetic) alphabets with a single byte per letter. It allows text in any mixture of living languages to be represented economically in the computer as a sequence of bytes.

The sequence of bytes is stored in the linear order in which the text would be spoken, and as such it is isolated from graphical complexities such as the variant forms of letters and the mixing of the directions in which multilingual text might have to be written. This accounts for a fact that may seem surprising: the internal computer processing of multilingual text is not affected by the presence of exotic scripts. Word-processing operations such as the editing of text, the search of text for particular characters or words and the electronic transmission of text depend not on the graphical form of the text but on the internal sequence of bytes that represents its information content. The only real complexities in multilingual word-processing software involve the typing and rendering of text.

I turn next to typing. For most languages the process can be quite simple, since almost any living alphabet will fit comfortably on a standard typewriter keyboard. Indeed, in computerized typing it is easy for the software to change the computer's "interpretation" of the keys so that the typing is in another alphabet. For example, pressing the *A* key can cause the computer to store the Russian letter Ф or the Arabic letter ش, depending on the keyboard interpretation. The computer can even display on its screen a small diagram called a virtual keyboard, which reminds the typ-



ist what alphabet is currently assigned to the keyboard and which keys correspond to which letters.

For some languages the computer can simplify typing itself. Arabic script, for example, includes a special combination character that replaces the letters *lam* and *alif* whenever the two appear in sequence. On an Arabic typewriter the special character occupies its own key. The computer can automatically render all *lam-alif* combinations as the special character, making the extra key superfluous. The computer can also handle letters that change their appearance depending on their context. With word-processing software for such a script the typist need only enter a natural (that is, phonetic) sequence of characters; the computer will take over the burden of handling the complexities of the script.

There is one living script whose complexity challenges even a computer's capabilities. It is the ideographic system employed in writing Chinese, Japanese and occasionally Korean. The Chinese ideographic characters, which can conveniently be called by the Japanese term *kanji,* originated as pictographs more than 4,000 years ago and assumed their

COMPUTER WORK STATION includes the screen whose content was shown in the illustration on the preceding page. The computer itself is part of a computer system called Star manufactured by the Xerox Corporation. The space bar on its keyboard is actually four space bars. In Japanese word processing they specify three Japanese "alphabets" (*katakana, hiragana* and *kanji*) or English letters. In front of the keyboard is a "mouse" connected to the computer. The mouse is moved about on a desktop to position a pointer on the video screen.

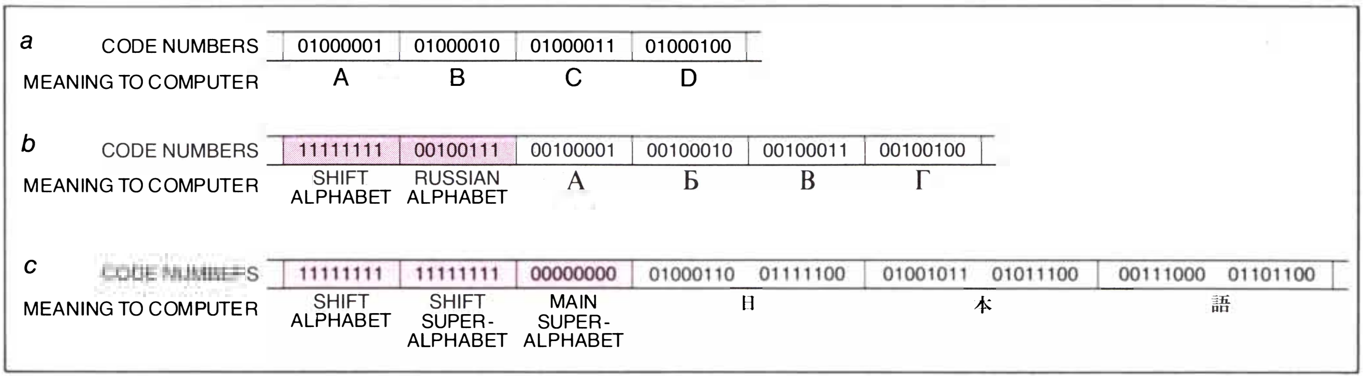
present form some 2,000 years later. About 500 years after that the Japanese and the Koreans adopted the Chinese language for official and scholarly discourse. For writing the native language they devised phonetic letters: the *kana* (*hiragana* and *katakana*) characters in Japan and the *hangul* in Korea. The Chinese characters were unsuitable from the beginning for writing the native language; now they were also unnecessary. Nevertheless, the prestige of Chinese culture led to hybrid scripts. Today the phonetic alphabets in Japan and Korea are used mainly to write the inflectional endings of words (endings somewhat similar to English inflections such as *-ed* and *-ing*). *Kanji* are used to write word roots, that is, the basic dictionary form of words. In Japan the number of *kanji* commonly seen in published text is declining: it has been cut to about 3,500, about half the number commonly seen in China.

As a result of its curious history the Japanese language has the most complex script in the world: it remains a mixture of *kanji* ideographs and *kana* phonetic letters. Because of this complexity, the Japanese have had no reasonable way to type their own language; more than 90 percent of all documents in Japan are handwritten, or rather handcrafted. A slip of the writer's hand, and a page must be torn up. Moreover, most documents are hard to read unless the writer happens to be an accomplished calligrapher. To be sure, there does exist a *kanji* typewriter, rather like a small typesetting machine, but the device is slow and tiring to use. Professional typists are comparatively rare, and their productivity is typically about 20 characters per minute, or only 10 pages per day.

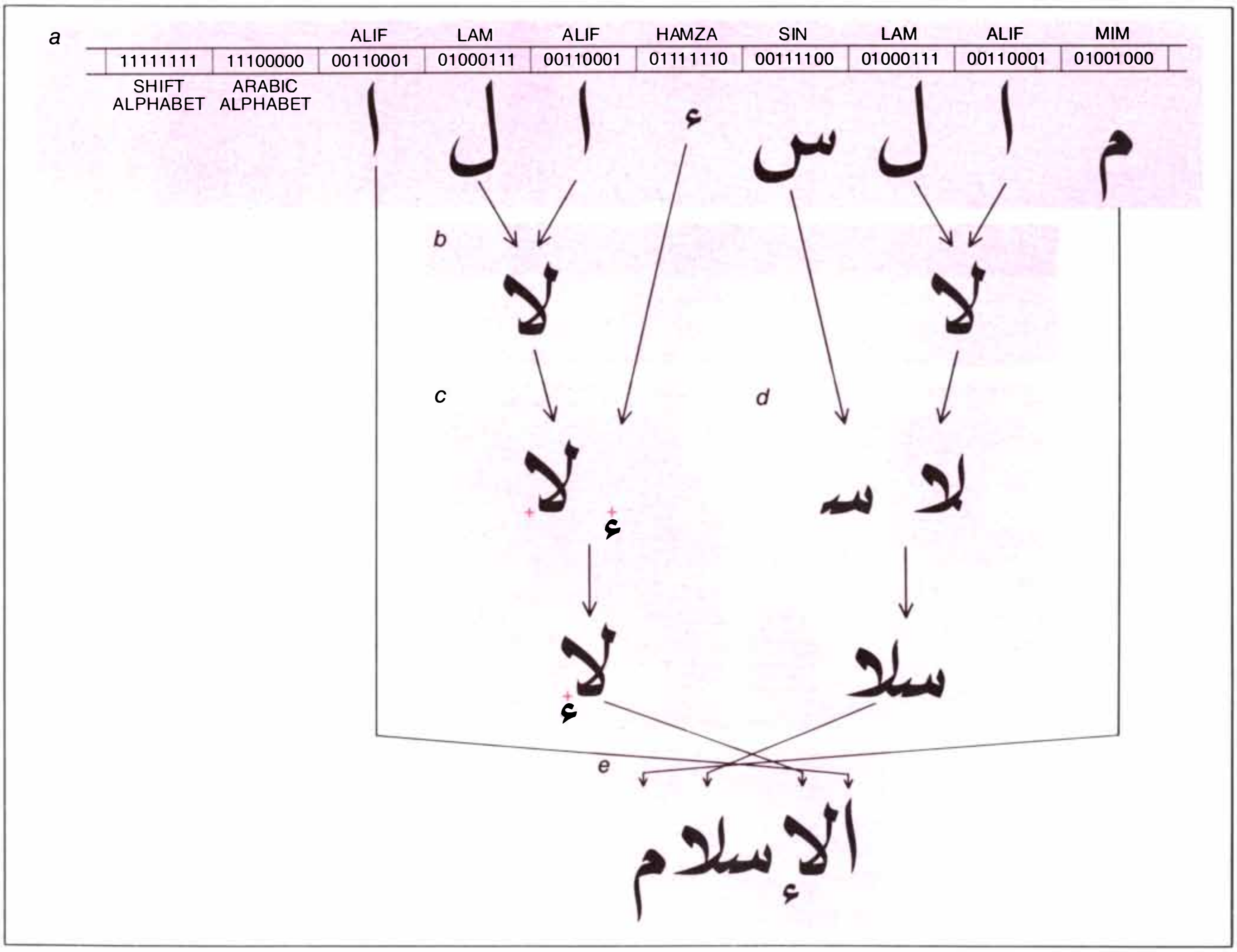
Hisao Yamada of the University of Tokyo, a scholar of the social history of typing, notes that it is not easy for a society to envision a usable typing system where none has existed. In the U.S. the possibility of high-speed typing never occurred to the inventor of the typewriter; touch-typing (with all 10 fingers potentially in contact with the keyboard) was not devised until 14 years after the typewriter was patented. In Asia the situation is much the same. The initial methods for entering *kanji* into a computer were not well conceived. Some methods employ a keyboard with several hundred keys, where each key has several characters on it. The right hand presses such a key while the left hand manipulates a bank of shift keys to specify a particular character. Other methods require the typist to analyze each *kanji* character and then enter some sequence of code symbols. This turns out to be even more tiresome than searching among hundreds of keys.

In recent years a new *kanji* typing

WORLD'S LIVING LANGUAGES raise problems for multilingual word processing. Mongolian (*a*) is in vertical columns. The specimen reads: *Erte togha tomshi ugei, nogchigsen galab-un urida anu*, or "Once upon a time, countless past ages ago. . . ." Arabic (*b*) is written from right to left in descending curlicues. The specimen reads: *najmu-l-fajri*, or "star of dawn." Hebrew (*c*) is written from right to left in letters that carry markings. The word is *ba'asher*, or "where." Greek (*d*) includes a letter, *sigma* (*colored arrows*), that has a special form at the end of a word. Here the word is *aschimos*, or "ugly." Arabic scripts (*e*) require the combining of certain letters. The word is *al-islam*, or "Islam." Hindi (*f*) has letters written out of phonetic order. The word is *hindî*; the spelling, in effect, is *ihndî*. Thai (*g*) has vowels surrounding consonants. The word is *deuan*, or "month"; the letter *eua* is in three parts. Korean (*h*) is characterized by syllabic clusters. The specimen reads: *han-gug-ŏ*, or "Korean language." Chinese (*i*) is written with thousands of ideographs. The specimens both read: *zhong-guo-hua*, or "Chinese language."



FLEXIBLE ENCODING accommodates all the world's writing systems while minimizing the length of the string of bits, or binary digits (0's and 1's), that represents a multilingual text in the memory of a computer. For English (*a*) a standard scheme, the American Standard Code for Information Interchange, is available. It assigns one byte, or eight bits of code, to each character. For other languages, such as Russian (*b*), the byte 11111111 is designated a "shift alphabet" signal: it instructs the computer to interpret the following byte as the code specifying a new alphabet, and the bytes after that as codes specifying characters in that alphabet. For Chinese and Japanese (*c*) a still more complex scheme is required. Two successive bytes of 11111111 are a "shift superalphabet" signal, the following byte specifies a particular superalphabet and the subsequent bytes, taken in pairs, specify characters in the superalphabet (here ideographs).



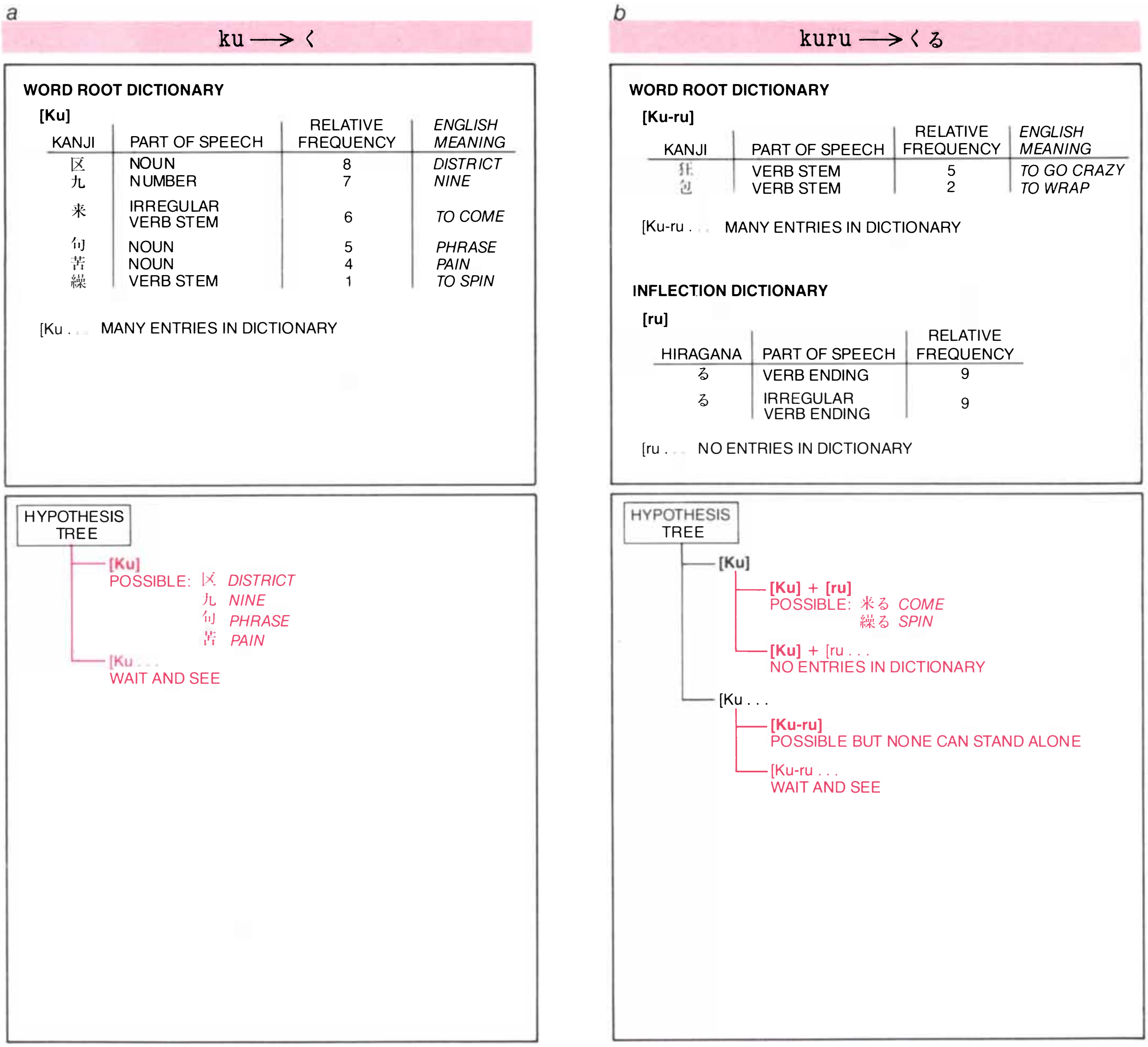
RENDERING OF TEXT means the process of displaying text on a screen or printing it on a page. The rendering of the Arabic word *al-islam* suggests some of the complexities. In storage in the computer's memory the word is simply a sequence of eight letters (*a*). Their proper rendering requires some special computations. First the computer finds two instances in which *lam* and *alif* are neighbors; Arabic script requires that this combination be replaced by a single ligature that looks much like the Greek letter *gamma* (*b*). The computer "notes" this in temporary memory; the stored code for the original text is unchanged. Next the computer searches for small markings such as *hamza,* which have to be positioned above or below other letters. The small crosses in the illustration are registration marks (*c*). Third the computer searches for letters that must be joined to their neighbors. The *sin* and its neighboring *lam-alif* ligature will join, and so they are given special joining forms (*d*). Finally, the computer displays the word or prints it on paper, in proper right-to-left order (*e*).

method has emerged; it suits the average typist so well that it is fast becoming the standard for Japanese manufacturers. Indeed, it makes possible touch-typing at a rate of about 50 characters per minute for the average typist and 150 characters per minute for the best speed typists. Called phonetic conversion, it has three steps that enable the Japanese to type their language phonetically. The typist types a word in phonetic *hiragana* symbols. (This phonetic spelling is learned in childhood by the Japanese.) To change the spelling to *kanji* the typist presses a special "look-up key" and the computer finds the *kanji* spelling in a dictionary it holds on a magnetic storage disk. If several words share the same phonetic spelling or if the word has several *kanji* spellings (this happens about 40 percent of the time), the computer assigns each alternative to a key on a virtual keyboard, and the typist chooses the one wanted.

Phonetic-conversion typing employs a standard typewriter keyboard. The *hiragana* characters fit neatly on such a keyboard; in fact, there is a standard arrangement for them. Yet many Japanese are familiar with the English typewriter keyboard and do not want to have to learn another arrangement. In consideration of their preferences many phonetic-conversion systems offer a *hiragana* typing method called *romaji* conversion. *Romaji* signifies the use of Roman letters to spell Japanese words. *Fuji, sayonara, samurai* and a host of other words are *romaji* spellings well known to Westerners. Most Japanese know them too. To type the word *sushi,* therefore, the typist can simply type *s u s h i.* The computer converts this first into *hiragana* すし and then into *kanji* 寿司. Offered a choice between typing *hiragana* on a standard Japanese keyboard and typing *romaji* on an English keyboard, nine out of 10 Japanese users prefer the latter.

The second step in phonetic-conversion typing—the computer's search in a dictionary—is greatly complicated by the inflectional endings of the Japanese language, which are somewhat similar to those of a European language. (For example, *aruku* means "walk," *aruita* means "walked" and *aruite* means "walking.") The dictionary may contain well over 100,000 word roots. The inflected forms of those words would number in the millions. Plainly the computer cannot store them all. The problem is solved by software that draws on a grammar of Japanese inflections to analyze the phonetic spellings supplied by the typist [*see illustrations on pages 102–104*]. In essence the typewriter is made to know the complete grammar of the language typed on it.

The chief advantage of phonetic-conversion typing is that the typing is analyzed by the computer word by word rather than one *kanji* at a time. Consider the nature of *kanji*. Although each *kanji* character signifies a separate concept, most words in Chinese, Japanese and Korean are compounded from two such concepts (and therefore from two *kanji*), much like English words such as *blackboard* and *railroad*. The compounding allows the creation of far more words than there are *kanji*. In addition it reduces the phonetic ambiguity of the language. For example, a Japanese dictionary lists 64 common *kanji* pronounced *tou* and 53 common *kanji* pronounced *kyou* but only one compound word pronounced *toukyou*, namely 東京, or "Eastern capital," the name of the city Tokyo. In the face of the alternatives each *kanji* in the compound word uniquely determines the other.

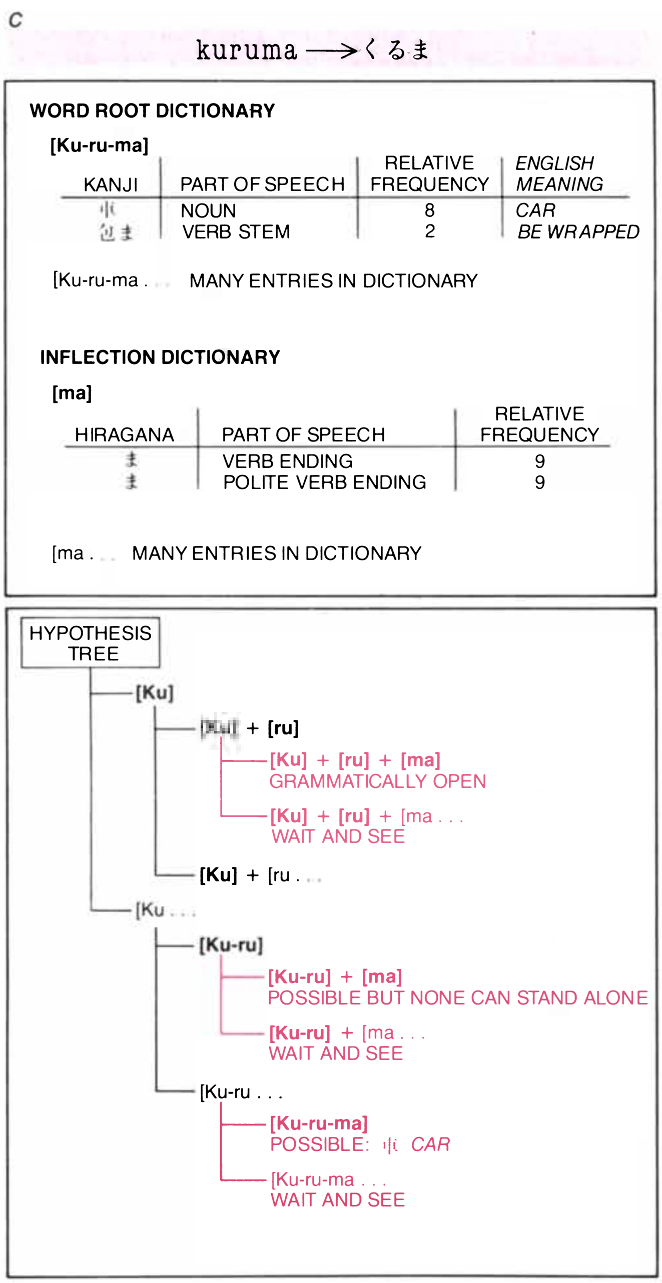


JAPANESE WORD PROCESSING challenges a computer because the Japanese language is a complex mixture of thousands of *kanji* (ideographic symbols borrowed from the Chinese) and dozens of *kana* (indigenous phonetic symbols). Here a typist types *kurumade*, which means "by car." The typing is in *romaji*, that is, it employs a standard English keyboard. Thus the typist begins (*a*) by typing an English *k* and *u*. The computer translates that immediately into a single *kana* character. Meanwhile the computer consults an internal dictionary to find entries consisting entirely of *ku* (*boldface type in square brackets*) and entries beginning with *ku* (*light type with a left bracket*). From the entries it constructs a "hypothesis tree." The processing continues (*b*) as the typist enters the English letters *r* and *u*, completing the second syllable of *kurumade*. Again the computer translates the *romaji* into *kana*. The result is a string of two *kana* characters. Then the computer consults the internal dictionary. The possibilities are more varied. *Kuru* could be a complete word. *Kuru* could be the start of a word. *Ku* could be the root of a word and *ru* could be an inflection, or modification of the root, somewhat like the English *-ed*, *-ing* and so on. Finally, *ku* could be a root and *ru* could be the start of an inflection. Some of the possibilities are ruled out because the dictionary has no such entries; others remain conceivable. Accordingly the hypothesis tree grows further branches (*color*). The third step in the processing (*c*) occurs when the typist enters the English letters *m* and *a*. The computer now has a string of three *kana*, from which it constructs the corresponding additions to its hypothesis tree. Some of the additions are conceivable but are not preferred. In particular, the root *ku* plus the inflection *ru* could in principle act as an adjective meaning "coming" or "spinning" (in the sense of making thread);

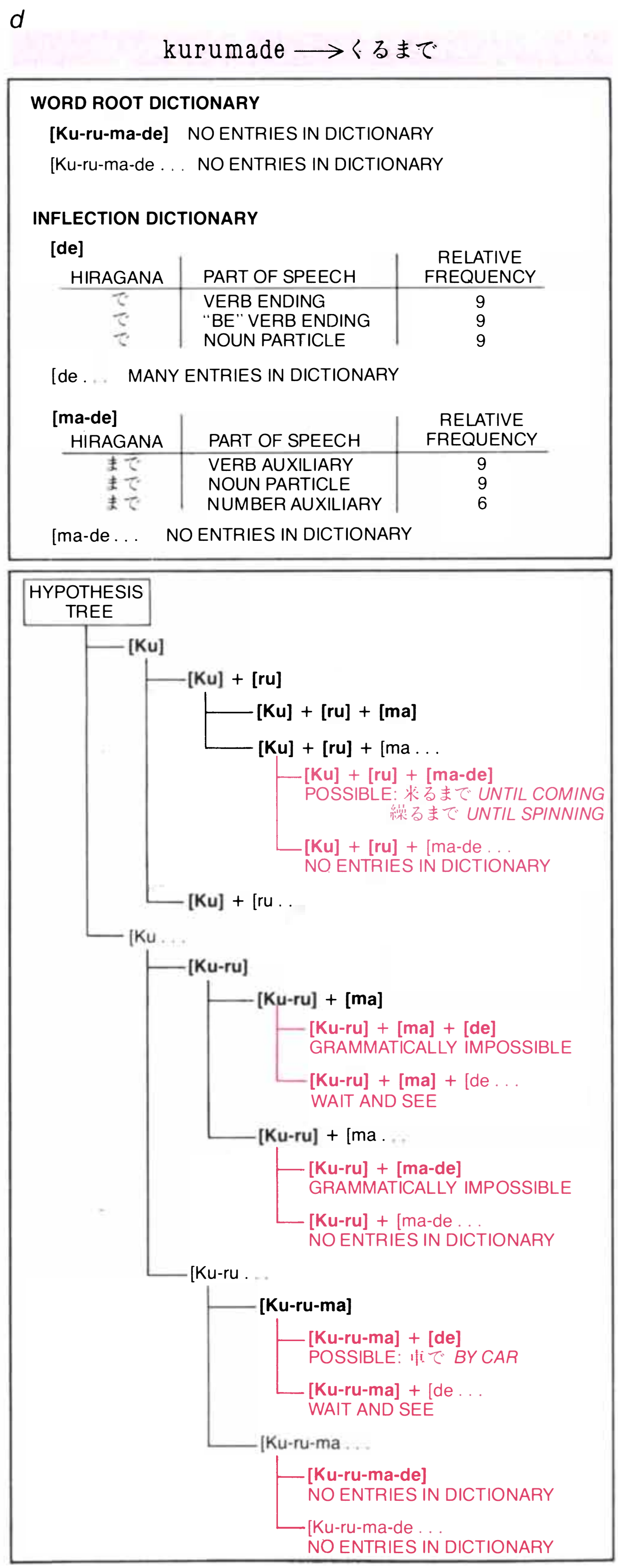
Consider, then, the typing of *toukyou*. When 東京 is typed a *kanji* at a time, the typist must somehow pick the desired *tou* from among 64 alternatives and the desired *kyou* from among 53 alternatives. In contrast, phonetic conversion requires merely that the typist enter *t o u k y o u*. The computer will respond with the unique, correct *kanji* pair. The creation of the computer's dictionary bestows on the computer a store of information about *kanji*. This greatly reduces the amount of information the typist must transmit to the computer.

I have treated phonetic-conversion typing in reference to Japanese, the language in which it was first applied, but it is equally effective in Korean and Chinese. For Korean a *hangul* keyboard is likely to be popular; for Chinese the romanized spelling called *pinyin* can serve

c



d



hence it could modify any noun. The computer will revert to this interpretation only if nothing more concrete emerges from further input. The tree is now at its most complex; the typing of further syllables will introduce additional hypotheses, but most of them will be eliminated by linguistic constraints, such as grammatical impossibility. The fourth step in the processing (*d*) occurs when the typist enters the English letters *d* and *e*, which complete the string *kurumade* and supply the computer with a sequence of four *kana*. The computer consults the internal dictionary one last time. The information retrieved from the dictionary constrains the possibilities. The root *ku* followed by the inflections *ru* and *made* is conceivable: it could mean "until coming" or "until spinning." Moreover, the root *kuruma* followed by the inflection *de* is conceivable: it means "by car." The last step in the processing of *kurumade* is shown in the illustration on the next page.

in the same role as *romaji*. For example, the name of the city Beijing can be typed by simply entering *b e i j i n g*. To be sure, each syllable in Chinese has a particular tonal pattern, and this input does not specify it. Still, that leaves only two dictionary alternatives: 背景, or "background," and 北京, or "Northern capital." It turns out to be far more efficient for a typist to choose among alternatives of this kind than it would be to enter an explicit encoding of the tone for each syllable. Chinese is further complicated by dialects that differ widely in pronunciation, but the standard Mandarin pronunciation is taught worldwide and already is native to 750 million potential typists.

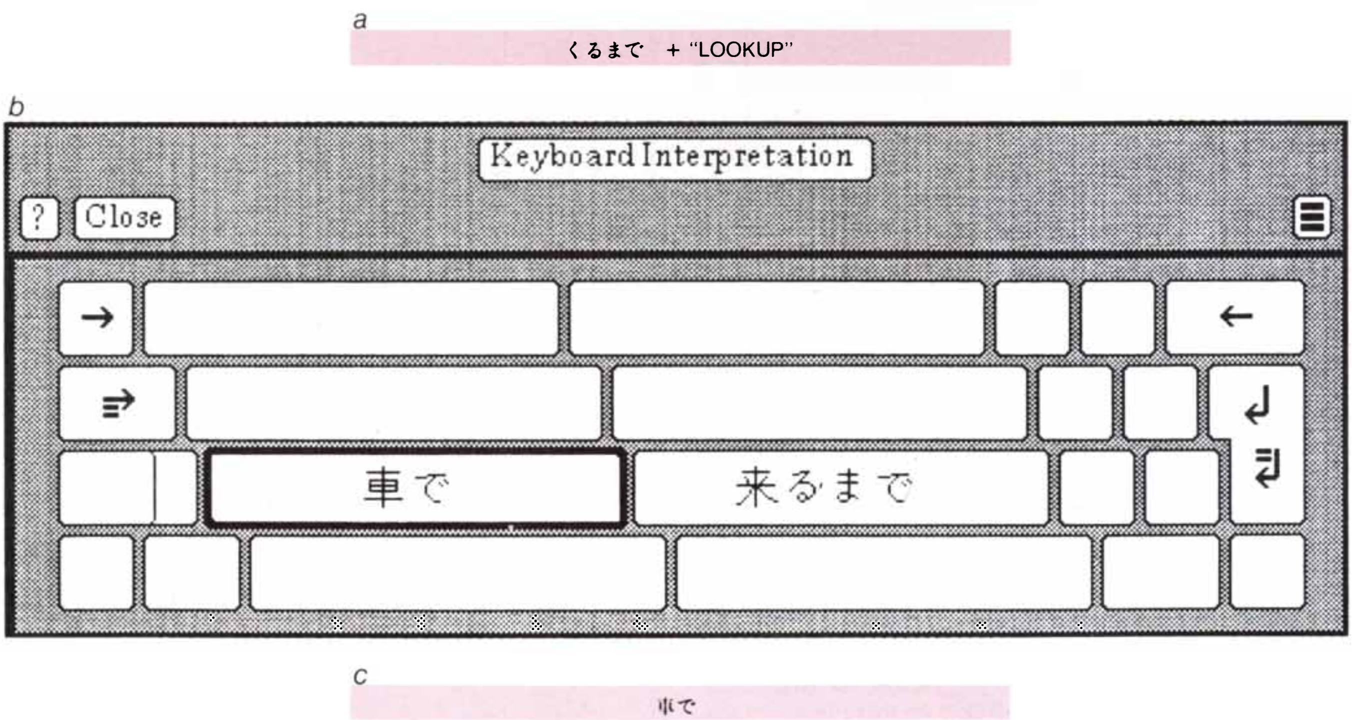
I have described how multilingual text can be typed into a computer and stored in its memory. I turn now to the third aspect of multilingual word processing: the rendering of text that is stored in the computer. For word processing in the English language there is a simple one-to-one correspondence between code numbers in computer storage and rendered characters on a display screen or a printed page. In fully multilingual software, however, that correspondence must be abandoned and replaced by a much richer scheme. When the computer processes a sequence of text codes for rendering, it must be empowered to examine any number of consecutive codes at a time, and it must be allowed to make any computation in order to choose the graphical forms and positions for the characters. In particular, the computer must be provided with variant character forms that it can use to represent the changeable letter shapes characteristic of many writing systems.

An example of the application of these ideas is provided by the handling of the Greek letter *sigma* (σ), which takes on a special shape (ς) when it appears at the end of a word. In the computer every *sigma* is stored as the code for an ordinary *sigma*, even if it comes at the end of a word. Whenever the computer is called on to render a *sigma*, however, it examines the character that follows the *sigma*. If the *sigma* is found to be at the end of a word, the computer renders the variant form instead of the ordinary form of the letter.

Arabic offers a more complex example: most Arabic letters have four forms, depending on whether the letter stands alone or is at the beginning, in the middle or at the end of a word. In addition the rules for joining Arabic letters to their neighbors must accommodate the presence of small markings placed above or below the letters. Nevertheless, the solution is much the same: the computer can be instructed to consider each letter's context before choosing its rendered shape. All the contingencies can be taken into account, given that the computer can make any computation in determining the final appearance of the text.

An important element in rendering is the ligature, in which two or more letters fuse to form a single rendered character. In most English typefaces (including the one in which this article has been typeset) there are ligatures for a few common letter combinations, chiefly *ff*, *ffi*, *fi*, *fl* and *ffl*. Software that takes each letter's context into account can instruct the computer to render the word *office* in four characters—*o*, *ffi*, *c* and *e*—while retaining in its memory the six-letter sequence *o f f i c e* required to process the word correctly.

A similar approach can be adopted for accents, or diacritical marks, such as the German umlaut (*ü*) or the French accent grave (*è*). The International Organization for Standards requires such marks to be represented in the computer as separate codes preceding the code for the letter to which the mark is applied. Hence a letter such as *ü* is represented by two bytes. The computer can render the two bytes ¨ and *u* as the single character *ü* by the same process that renders the two bytes *f* and *i* as the ligature *fi*. Some scripts require the application of multiple marks above or below a base character; the computer can handle this problem by a somewhat more complex rendering procedure.



FINAL STEP in the word processing of *kurumade* comes after the *romaji* typing of the word, when the typist presses a special "lookup key" (*a*). Inside the computer the dead branches of the hypothesis tree are ruled out. The wait-and-see branches, now superseded, are ruled out. The grammatically unpreferred branches, also superseded, are ruled out. The surviving hypotheses are ordered by the frequency of their appearance in Japanese; then the more likely ones (here the ones that mean "by car" and "until coming") are displayed on a virtual keyboard (*b*). Roots are displayed in *kanji*; inflections remain in *kana*. Extremely unlikely possibilities (here the one that means "until spinning") are not displayed unless the lookup key is pressed again. When the typist chooses an alternative, it replaces the string of four *kana* (*c*).

Even the unique syllabic clumps that characterize the Korean *hangul* script can be rendered if the computer is given appropriate software and rendering variants of the Korean letters, which it can build into square groupings. Similarly, the slanting descent to a baseline characteristic of many Arabic fonts can be rendered, freeing Arabic typography from the stricture of a flat horizontal baseline. In the Hindi language the word *hindî* itself has the first vowel (*i*) written out of phonetic order: it is placed before the initial consonant (*h*). When the rule for placing the vowel *i* is incorporated into the computer's instructions, however, the typist can enter the word in its normal phonetic order and the computer will automatically place the *i* before the *h* when it renders the word. The same approach can even rationalize the handling of scripts such as Thai, in which vowels can actually split into fragments that surround a neighboring letter.

Equipped with enough flexibility in the rendering process, the computer can handle any instance in which the letters of a script have a contextual effect on one another. The only remaining rendering problem is a broader one: How is the computer to mix scripts that run in different directions?

A computer cannot do the impossible: there is simply no sensible way to mix vertical text and horizontal text in a single paragraph. Chinese, Japanese and Korean are vertical by tradition. Asian printers, however, have developed the practice of printing these languages horizontally from left to right. Mongolian too is vertical by tradition. Mongolian printers rotate text by 90 degrees when it is necessary to combine it with horizontal text.

In sum, typographical usage calls for all languages to be typeset horizontally at need. The one circumstance that remains for the computer to handle is the mixing of a horizontal left-to-right script such as English with a horizontal right-to-left script such as Arabic. The typist wants to type Arabic in its proper phonetic order and have it appear in its proper right-to-left sequence, even if it is typed into the midst of English text. Conversely, the typist wants an English passage to appear in left-to-right sequence even in the midst of Arabic text.

The problem may seem perplexing, but it looks simpler from the standpoint of the computer. The computer works much faster than the typist; thus the computer spends most of its time waiting for the typist to enter the next character. Between keystrokes the text is simply displayed without change on the screen; the computer has done no more than produce a static text layout. The static arranging of mixed-direction text is handled routinely by printers of books and magazines in which Arabic or Hebrew script appears.

Mixed-direction text can go through remarkable transformations as it is entered into a computer. In most word processors the place on the screen where the next typed letter will appear is indicated by a blinking marker. The marker moves along as the text continues to grow. When a line fills up with type and the marker reaches a margin, the marker automatically drops to the beginning of the next line. For the typing of all-English text the marker moves from left to right. For the typing of all-Arabic text it moves from right to left. For mixed-direction text its activity is novel. Throughout the typing of English text embedded in Arabic the marker must stay put and the newly entered text must slide away from it. The stationary marker cannot, of course, reach a margin. All the same, the filling of a line means that the marker and subsequent words must drop to the next line. The typist has the strange impression that text is falling from the middle of a line. Yet the drop creates the correct layout of the mixed-direction text.

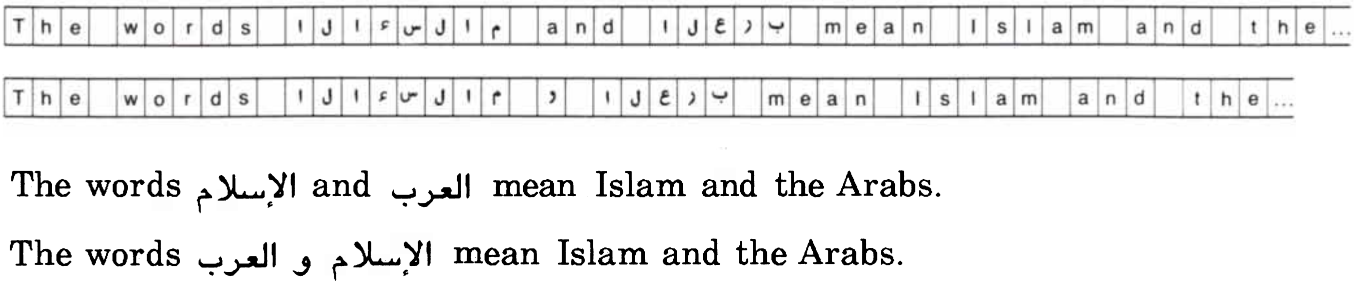
The editing of a mixed-direction text can cause changes that are even more remarkable. Consider the English-Arabic sentence "The words *al-islam* and *al-'arab* mean Islam and the Arabs," where *al-islam* and *al-'arab* are written in Arabic [*see illustration at right*]. Suppose in the course of editing the sentence the English word *and* is replaced by the Arabic equivalent; the Arabic words must then switch places, because they are now part of a phrase expressed in Arabic that should read from right to left. Yet the Arabic words themselves were not involved in the editing operation. Indeed, the text in computer storage changes only to accommodate the replacement of the word *and*. The visible permutation of the Arabic words is entirely a consequence of the rendering process.

The encoding, typing and rendering of text are the basic elements of multilingual word processing, but they do not exhaust the challenges that must be faced in designing a multilingual word-processing system. Suppose an oil company's Texas offices are made part of a worldwide network of word-processing computer work stations. An Arabic document arrives electronically from the Middle East, but the computers in Texas do not have the software needed to render Arabic text. The Texas computers must nonetheless handle the text as best they can without "crashing" and without mistaking the Arabic codes for English ones. They must at least render the English portions of a multilingual document. Furthermore, the computers in Texas should be capable of accepting an

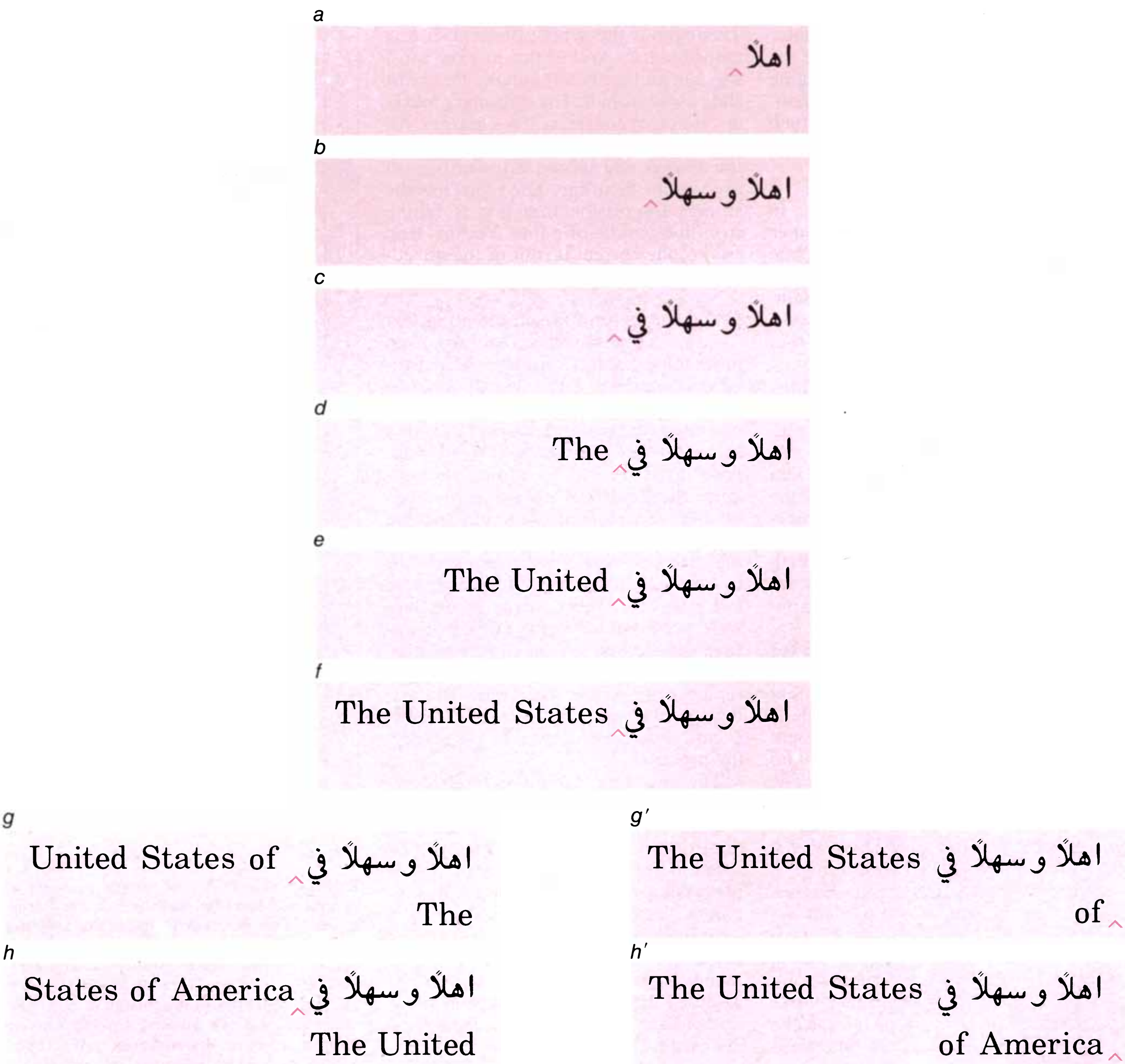
T	01010100
h	01101000
e	01100101
	00100000
w	01110111
o	01101111
r	01110010
d	01100100
s	01110011
	00100000
	11111111
	11100000
ا	00110001
ل	01000111
ا	00110001
ء	01111110
س	00111100
ل	01000111
ا	00110001
م	01001000
	11111111
	00000000
	00100000
a	01100001
n	01101110
d	01100100
	00100000
	11111111
	11100000
ا	00110001
ل	01000111
ع	01000010
ر	00111010
ب	00110010
	11111111
	00000000
	00100000
m	01101101
e	01100101
a	01100001
n	01101110
	00100000
I	01001001
s	01110011
l	01101100
a	01100001
m	01101101

T	01010100
h	01101000
e	01100101
	00100000
w	01110111
o	01101111
r	01110010
d	01100100
s	01110011
	00100000
	11111111
	11100000
ا	00110001
ل	01000111
ا	00110001
ء	01111110
س	00111100
ل	01000111
ا	00110001
م	01001000
	00110000
و	01001011
	00110000
ا	00110001
ل	01000111
ع	01000010
ر	00111010
ب	00110010
	11111111
	00000000
	00100000
m	01101101
e	01100101
a	01100001
n	01101110
	00100000
I	01001001
s	01110011
l	01101100
a	01100001
m	01101101

BILINGUAL TEXT in storage presents no special problems for multilingual word processing. Here the text includes the sentence "The words *al-islam* and *al-'arab* mean Islam and the Arabs," where Arabic words are expressed in Arabic. The sentence is stored in phonetic order (*left*), even though Arabic characters are rendered from right to left whereas English ones are rendered from left to right. Some bytes (*color*) are special codes that shift alphabet and specify a new alphabet. When the sentence is edited (*right*), so that the English word *and* is replaced by its equivalent in Arabic, the only change to the stored text is the removal of the bytes specifying *a*, *n* and *d*, along with some special shift codes, and the insertion of the code for the Arabic letter *wa*.



RENDERING OF THE STORED TEXT shown in the illustration on the preceding page yields a remarkable consequence. The change made in the text involves only the English word *and,* as shown at the top of this illustration. Yet in the rendering of the text the words *al-islam* and *al-'arab* must exchange places, because the entire phrase is now rendered in Arabic script, which must be read from right to left.



"WORD WRAP" is the ability of a computer to move a word to the next line of the display screen if the word will not fit on the line being typed. If the text combines a language that reads from left to right with one that reads from right to left, a special problem arises. Here *ahlan wa sahlan fi* is typed in Arabic (*a*). It means "Welcome to. . . ." The Arabic characters appear on the screen in right-to-left order; a marker (a caret) has moved to the left to show where the next new character will be entered (*b, c*). The typist then switches to English to type "The United States" (*d, e, f*). Now the caret is stationary; the newly typed characters slide to the left to make room for input. That input ("of America") will require a wrap, but if the computer continues to push text to the left, the words "The United" will be forced down a line (*g, h*), reversing the halves of the English. A more appropriate wrap (*g′, h′*) drops "of America" and the caret onto the next line.

Arabic-software module, which would add Arabic rendering instructions to the repertory they already have.

The oil company's situation suggests the broad-scale design goals for a multilingual word-processing system. They are compatibility (all the computers in the system must be able to exchange documents in any combination of scripts); open-endedness (each computer must be able to deal sensibly with scripts unknown to it), and modularity (it must be possible to add the capacity for new languages one by one). These goals are difficult, but they can be achieved. The solutions derive in large part from the treatment of text in ways that are broad enough to include any mixture of the world's living languages.

The initial applications of multilingual word processing surely lie in the creation of multilingual documents. After all, as worldwide commerce and politics grow, all kinds of multilingual documents become essential. In addition software that can manipulate multilingual text is certain to bring benefits to language teachers and translators. The automatic translation of documents from one language into another is a far more distant goal. Indeed, at the moment there is little cause to imagine that high-quality machine translation can be achieved. The faithful translation of a passage requires that the translator understand the passage both in its explicit content and in its implications. In a quarter century of intensive research there has been no significant progress in supplying a computer with such an ability. Low-quality translation of texts with circumscribed meaning (such as instruction manuals) is already a reality. Even there, however, the success of machine translation depends heavily on editing by a human proofreader.

The questionable prospects for high-quality machine translation are balanced, I think, by the brilliant outlook for electronic mail: the near-instantaneous transmission of messages or documents by way of private electronic networks or public communication lines. International electronic-mail systems are already in everyday use. They are similar to telephone and postal systems in that their usefulness increases with the extent of the network. It seems inevitable, therefore, that electronic mail services will eventually expand and join to form a single worldwide electronic-mail utility. If this is to come about, one prerequisite is inescapable. The telephone does not require its users to speak only English, nor does a postal system require its users to write only English. Electronic mail will not succeed as a global communication medium unless the text it carries is fully multilingual. To my mind that is the ultimate application for multilingual word processing.

The Mathematics of Three-dimensional Manifolds

Topological study of these higher-dimensional analogues of a surface suggests the universe may be as convoluted as a tangled loop of string. It now appears most of the manifolds can be analyzed geometrically

by William P. Thurston and Jeffrey R. Weeks

Thousands of years ago many people thought the earth was flat. The flatness of the earth's surface must have seemed self-evident to anyone who looked out across an ocean or a prairie, and it was argued, not entirely unreasonably, that the surface of the earth must either be infinite or have an edge. It is now understood, of course, how such a basic misconception could have come about: even from a few thousand meters above the ground a small part of the roughly spherical earth looks like a small part of a plane. What is less often appreciated is that an unlimited number of terrestrial shapes would give rise to the same local observations. For example, it would be consistent with such local observations for the earth to have the shape of an irregular blob or of a doughnut.

Investigations in the branch of mathematics called topology make it clear that we confront an analogous situation when we attempt to describe the overall form of the universe based on the limited view from our point in space. An observer on the earth cannot conclude that the universe retains the geometric structure of ordinary Euclidean space to indefinite distances, although there is still no evidence to the contrary. If the structure of the universe is not Euclidean, what are the alternatives? One familiar idea is that space may be "curved" in much the same way as a surface can be curved. The three-dimensional curvature of space and a closely related concept, the four-dimensional curvature of space and time, have become important ideas in astronomy and cosmology because of the key role they play in Einstein's general theory of relativity.

Nevertheless, the determination of curvature alone is not enough to specify what can be called, loosely speaking, the shape of the universe. Certain kinds of possible three-dimensional structure for the universe can be specified by analogy with the two-dimensional surfaces, but the analogy only begins to suggest the richness of form that is introduced by a third dimension. Indeed, since space and time are treated in the theory of relativity as a single entity called space-time, one might suppose that the appropriate mathematical structure of the universe must be a four-dimensional one. There is good reason to believe, however, that the structure of four-dimensional space-time is governed by the structure of three-dimensional space alone. Hence in order to investigate the overall structure of the universe without prejudice one must begin to understand the kinds of three-dimensional structure that could give rise to the observed universe. The structures are called three-dimensional manifolds, or three-manifolds for short.

The study of three-manifolds is, in a sense, a generalization of the study of two-manifolds, or surfaces. Topologists have known how to describe and classify all possible two-manifolds for more than a century, but the systematic classification of all three-manifolds remains an unsolved problem due to the exceedingly complex forms to which some three-manifolds give rise. A mathematical procedure called surgery suggests a measure of the complexity. Surgery makes it possible to construct a three-manifold from any tangled loop of string, no matter how knotted or convoluted the tangle. Imagine confronting two snarled masses of fishing line and trying to determine whether or not they are tangled in exactly the same way. Unless it is possible to classify such tangles of line in a systematic way, there is no hope that three-manifolds can be analyzed either. Until recently, therefore, mathematicians saw little reason for thinking a systematic theory of three-manifolds could be devised.

That pessimistic assessment must now be reconsidered. Investigations by one of us (Thurston) into the geometry of three-manifolds show there is a pattern that may lead to an understanding of all possible three-manifolds. All known three-manifolds fit the pattern, and as a result their twisting and winding can be described in geometric terms.

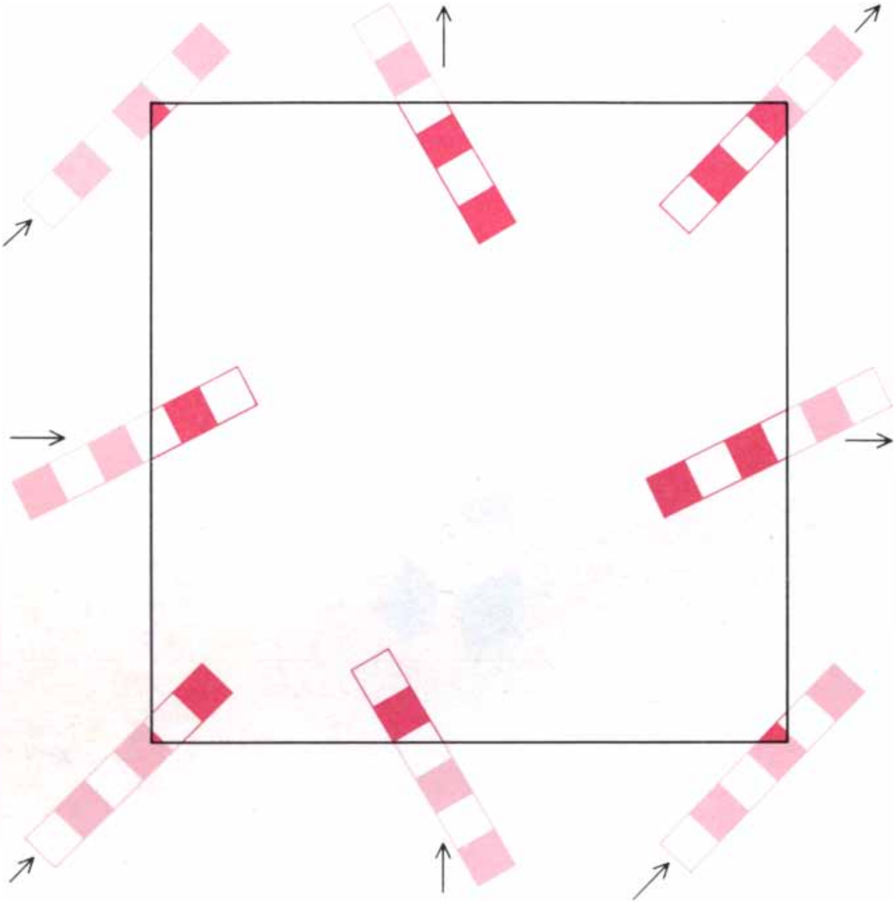
The theory of manifolds arose in the 19th century out of a need to understand quantitative relations geometrically. For example, the set of solutions to an equation that has two variables can be plotted as a set of points in the plane. Each point represents a pair of values for the variables that make the equation true; typically, the set of points is a curve or a set of curves. Similarly, the set of solutions to an equation that has three variables can often be plotted as a two-dimensional surface in three-dimensional space, such as the surface of a sphere. For equations with more than three variables one can describe the set of solutions geometrically in much the same way: it is a higher-dimensional manifold in a still higher-dimensional space. Although one cannot visualize such objects directly, mathematicians have developed conceptual tools for the study of equations that lead to higher-dimensional manifolds.

Topology cannot actually solve equations. What it provides is a mathematical vocabulary—adjectives and nouns—that allow a set of solutions to be discussed in a general way without actually being specified. Thus, although the manifold of points that makes up the set of solutions to an equation has a precise and unambiguous shape, the topology of the manifold is not constrained by the properties of that shape. Instead the topology encompasses whatever properties are retained when the manifold is deformed in an arbitrary way, as long as the deformation is done without cutting, tearing or puncturing.

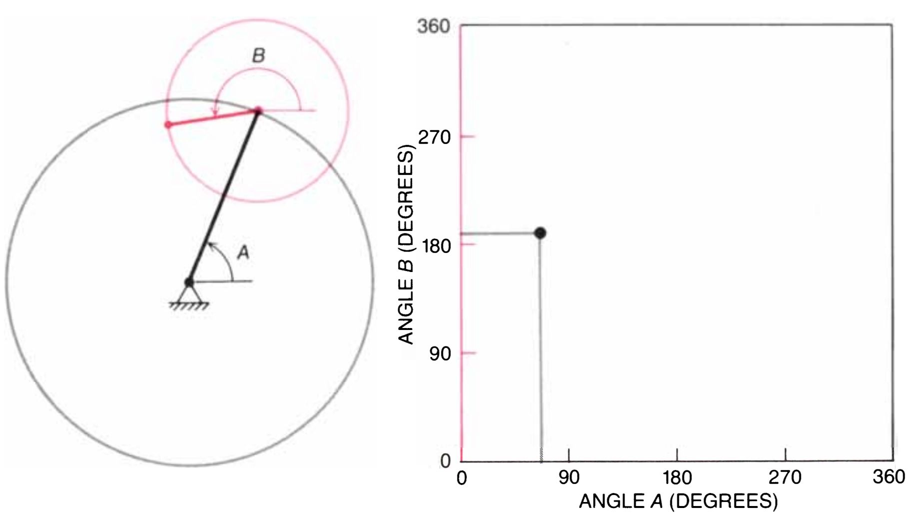
A doughnut can be deformed into a coffee cup by making a concave depression in the surface of the doughnut and then enlarging the depression while shrinking the rest of the doughnut. As the old joke goes, a topologist is a person who cannot tell a (one-hole) doughnut from a (one-handle) coffee cup. On the other hand, the topologist does distinguish the surface of a doughnut from the surface of a glass without handles, because there is no way one shape can be continuously deformed to yield the other. It may seem that by allowing arbitrary deformations topology discards most of the interesting features of a manifold. In many mathematical questions, however, topological information plays a significant role.

The first substantial contributions to the topological theory of three-manifolds were made around the turn of the century by Henri Poincaré, Max Dehn and Poul Heegaard. One difficulty with the study of three-manifolds is that direct visualization must partially give way to abstract representation. Many surfaces can be visualized because they

TOPOLOGICAL STRUCTURE of the universe need not conform to the structure of infinite, three-dimensional Euclidean space. The mathematical theory of three-dimensional manifolds, or three-manifolds, demonstrates that space may "curve back on itself" in an infinite variety of ways. One possible model for the topology of space is the three-manifold discovered by Herbert Seifert and C. Weber, who is now at the University of Geneva, in 1932. The manifold cannot be depicted from without because to do so one would have to view it from a fourth or higher dimension. Nevertheless, it can be visualized in a more limited sense as a dodecahedron whose opposite faces are mathematically glued together, or identified. The colored, ruled bars moving into and out of the faces of the dodecahedron indicate how the gluings are to be carried out: one member of each pair of faces is matched to its counterpart after a rotation of three-tenths of a turn about the axis perpendicular to the two faces. Although parts of the bars are shown as ghosted images outside the dodecahedron, the bars do not really exist there because it is assumed that only points inside the dodecahedron exist. When one of the bars moves toward one of the faces of the dodecahedron, it disappears at that face and reappears at the opposite face as if it were entering the dodecahedron from another direction. If the structure of the universe is that of the Seifert-Weber manifold, the universe is finite but will expand forever.



TWO-DIMENSIONAL MANIFOLD known as the two-torus can be represented as a square whose opposite edges are abstractly glued together. In other words, the top edge of the square is identified with the bottom edge and the left edge is identified with the right edge. If a ruled bar moves off the right edge, it reappears at the left edge; if the bar moves off the top edge, it reappears at the bottom. The motion is similar to that of objects in many video games. When the edges are abstractly glued together, all four vertexes of the square coincide in the manifold; when a ruled bar moves toward a vertex, pieces of it reappear at the other three vertexes.



DOUBLE-CRANK LINKAGE SYSTEM is made up of two rigid bars pinned together at one point; the end of one bar is held fixed. The bars are free to rotate about the pins as long as the movement is confined to the plane of the page (*left*). Every possible configuration of the two bars can be given by a point plotted on two perpendicular coordinate axes: one axis gives the angle between the direction of the first bar and a fixed direction, and the second axis gives the corresponding angle for the second bar. The set of all the plotted points, which represent all possible positions of the linkage, is called the configuration space of the linkage (*right*). Because the configuration of the double crank does not change if the angle of either bar is changed by 360 degrees, the configuration space is a square bounded by the lines that represent zero-degree and 360-degree rotations for each bar. Points on opposite sides of the square represent identical configurations of the linkage; in other words, the configuration space is a two-torus.

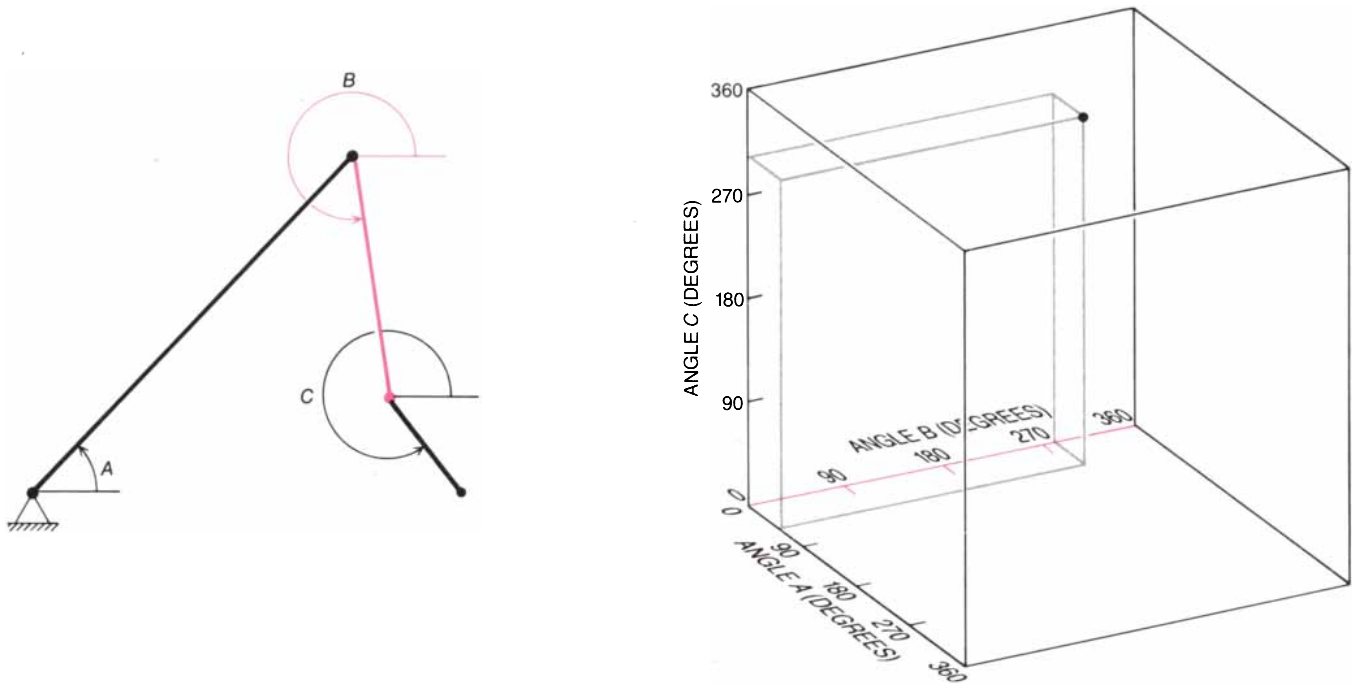
can be seen externally from the third dimension, a dimension that is one higher than the dimension of the surface. The extra dimension gives the surface enough room to bend around and close up with itself. One might try to visualize a three-manifold externally, as if one were viewing it from a space with four or more dimensions, but it turns out such contortions are not necessary.

In the 19th century mathematicians found that two-manifolds can be represented as polygons whose edges are to be glued together, or in other words identified with each other in a specified way. In the novel *Flatland,* published in 1884, Edwin A. Abbott describes a two-dimensional creature living entirely within the plane. Consider the movements of such a creature on a two-manifold with a more exotic topology, namely a square whose opposite edges are identified. When the creature moves off the top edge of the square, it reappears at the bottom; when it moves off the right edge, it reappears at the left. Intrinsically, therefore, the top of the square is glued to the bottom and the right edge is glued to the left. It is worth noting that many video games operate on the same principle: when a figure moves off the top edge of the screen, it reappears at the bottom, and so on.

For a square it is a simple matter to carry out the gluings. Attaching the top of the square to the bottom gives rise to a cylinder open at both ends, and gluing the open ends leads to a one-hole doughnut. After the edges have been glued the seams are erased; the Flatlander cannot tell where the gluings were made. The doughnut and the square (with edges properly identified) count topologically as the same abstract manifold, namely the two-torus.

As the video games demonstrate, however, it is not necessary to do the gluing to get an intuitive understanding of the two-torus. With a little practice it is just as easy to follow the motion of an object on the square, where the gluings are specified only in an abstract sense. Abstract gluing brings within the compass of geometric intuition a great many manifolds that would otherwise be difficult to visualize. What is most important for us is that the gluing trick can readily be generalized to bring geometric intuition to bear on the understanding of three-manifolds.

Consider the three-manifold generated from a rectangular block of space, such as the space inside a room. Abstractly glue the front wall of the room to the back wall, the left wall to the right wall and the floor to the ceiling. If the gluings were actually done, one would have to imagine the room bending around and joining itself in a fourth dimension. All that is needed for the description of the manifold, however, is



THREE-MANIFOLD analogous to the two-torus arises from the set of all possible configurations of a triple crank whose motion is confined to a plane (*left*). If the angles between each bar and a fixed direction are plotted on three mutually perpendicular axes, every possible position of the triple crank can be plotted as a point in a cube (*right*). The configuration of the linkage does not change with any complete rotation of a bar. Thus in the configuration space every face of the cube that corresponds to a rotation of 360 degrees is abstractly identified with its opposite face, which corresponds to a rotation of zero degrees. The resulting three-manifold is called the three-torus.

given by the procedure for abstract gluing. If an object within the manifold is moved toward the front wall, it disappears at that wall and reappears on the back wall; similarly, the object disappears at the right wall as it reappears on the left wall and disappears at the ceiling as it reappears on the floor. Evidently the motion is strikingly similar to the motion of an object within the two-torus; the manifold is the three-dimensional analogue of the two-torus, and so it is called the three-torus.

If ordinary concepts of space and physical reality are momentarily set aside, one can readily imagine living in a three-torus. Look at the back wall and the line of sight passes through that wall and returns from the opposite point on the front wall. What you see is a copy of yourself from behind. Look to the right and you see a copy of yourself from the left; look down at the floor and you see the top of your head. Indeed, because the line of sight continues crossing the room in all directions, you see what appear to be infinitely many copies of yourself and the room, all arranged in a rectangular lattice. The optical effect is similar to the one created by a room whose walls, floor and ceiling are covered with mirrors. The difference is that there are no reflections reversing the images of the room; instead all the images are direct copies of the original.

Does the fact that astronomers have not observed such peculiar visual effects imply the universe cannot be a three-torus? No. The universe is between 10 and 20 billion years old. If it were a three-torus, say, 60 billion light-years across, no light would have had enough time to complete a round trip. Another possibility is that observational astronomy has already recorded light that has traveled all the way around the universe: if the universe is a three-torus, one of the distant galaxies we observe may be our own. The possibility would be hard to verify because the image of our galaxy would be formed from light that left its source billions of years ago and spent the intervening time crossing the universe. What could be seen, given unlimited resolution of the image, would be the Milky Way in its earliest stages of evolution, as it looked when the light was emitted. Such a universe has a finite volume but no boundary of any kind.

Similar models of the possible spatial structure of the universe can be derived from other polyhedrons as well as from the cube. In each case the best way to understand the manifold is to imagine certain faces of the polyhedrons abstractly glued together. Two such manifolds are readily constructed from the regular dodecahedron. The 12 faces of the dodecahedron are regular pentagons arranged in pairs in such a way that the members of each pair are parallel and on diametrically opposite sides of the dodecahedron.

In the first dodecahedral three-manifold one member of each pair of pentagons is identified with the opposite pentagon by rotating the first member one-tenth of a turn counterclockwise about the axis perpendicular to its surface. The manifold is called the Poincaré manifold because it is equivalent to a three-manifold discovered by Poincaré in 1902. (Poincaré was unaware, however, that the manifold could be made from a dodecahedron.) The second dodecahedral manifold arises when each pentagon is glued to its opposite counterpart after a counterclockwise rotation of three-tenths of a turn. The resulting manifold is called the Seifert-Weber dodecahedral space, after Herbert Seifert and C. Weber, now at the University of Geneva, who discovered the manifold in 1932 [*see illustration on page 109*]. Like the three-torus, both manifolds would give rise to a universe with a finite volume but no boundary or edge.

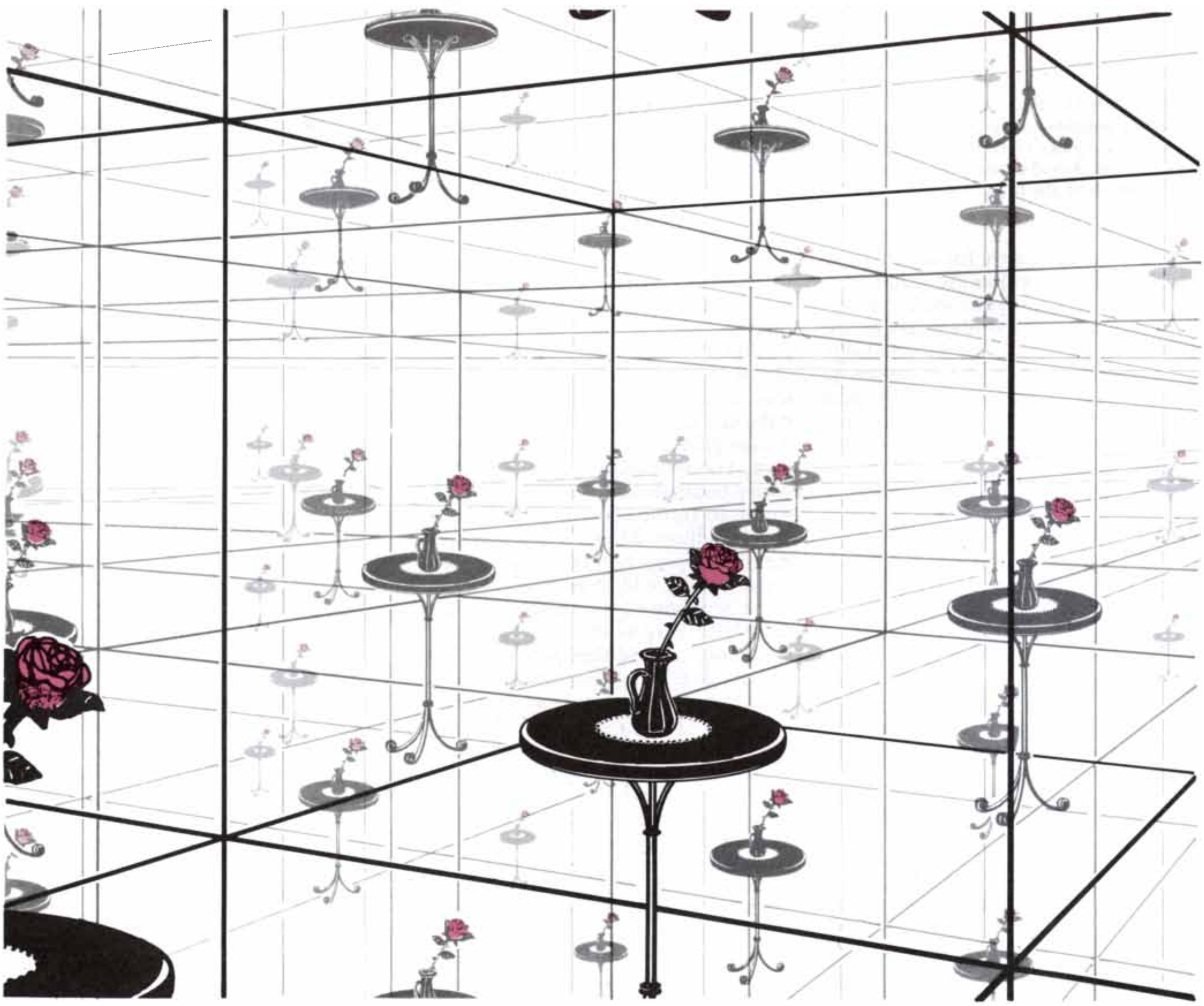
Vast numbers of additional models for the large-scale structure of space can be constructed in a similar way. Because most polyhedrons are irregular, most three-manifolds arise from abstract gluing of the faces of irregular polyhedrons. The description of the gluing can become quite complicated when the number of faces is large.

It may appear there is a certain unreality to the exercise. The nonspecialist would probably grant that the true topology of space is a worthy object of speculation but would likely wonder how far such speculation should be carried. In particular, the nonspecialist might well question the utility of studying "spaces" in the plural. To the topologist such objections miss the mark because they focus only on the metaphorical content of topology. The study of topology can certainly be motivated by problems that arise in other contexts, but topology itself is a theory of pure form, not a theory of the real world. If the structure of space were somehow settled tomorrow, no topologist would give up the study of abstract spaces.

This assertion need not imply that topology is irrelevant to the real world. On the contrary, like other branches of mathematics, topology has many strong and substantive connections with the world, but the connections are indirect. If a particular metaphor, such as the spatial one, starts to wear thin, it is best to abandon the metaphor rather than to abandon the study of the form to which the metaphor gave rise. Experience has shown repeatedly that a mathematical theory with a rich internal structure generally turns out to have significant implications for the understanding of the real world, often in ways no one could have envisioned before the theory was developed. A theory would never reach the mature stage of development in which such applications are recognized if it were constantly burdened with overworked metaphors.

In order to illustrate the scope of topological analysis, it is useful to abandon the cosmological metaphor for the time being in favor of a more terrestrial one. Consider a mechanical system of bars and linkages, such as the one that connects a key in a manual typewriter to the type element. We shall discuss only planar linkages, or in other words assemblies of rigid bars pinned to one another in such a way that all the bars move in only one plane. There must also be at least one anchor point, or fixed base, to which the bars are pinned.

The aim of a theory of mechanical linkages is to analyze the possible motions of a linkage. There are many real mechanical devices to which the analysis applies, and they need not have much physical resemblance to a collection of bars joined together. The study of linkages was much in vogue in the second half of the 19th century, when there was interest in the problem of finding a linkage in which at least one point moves in a straight line. It seemed that a solution to the problem would lead to many practical applications, such as the design of a power train for a steam locomotive. Although a number of elegant theoretical solutions to the problem

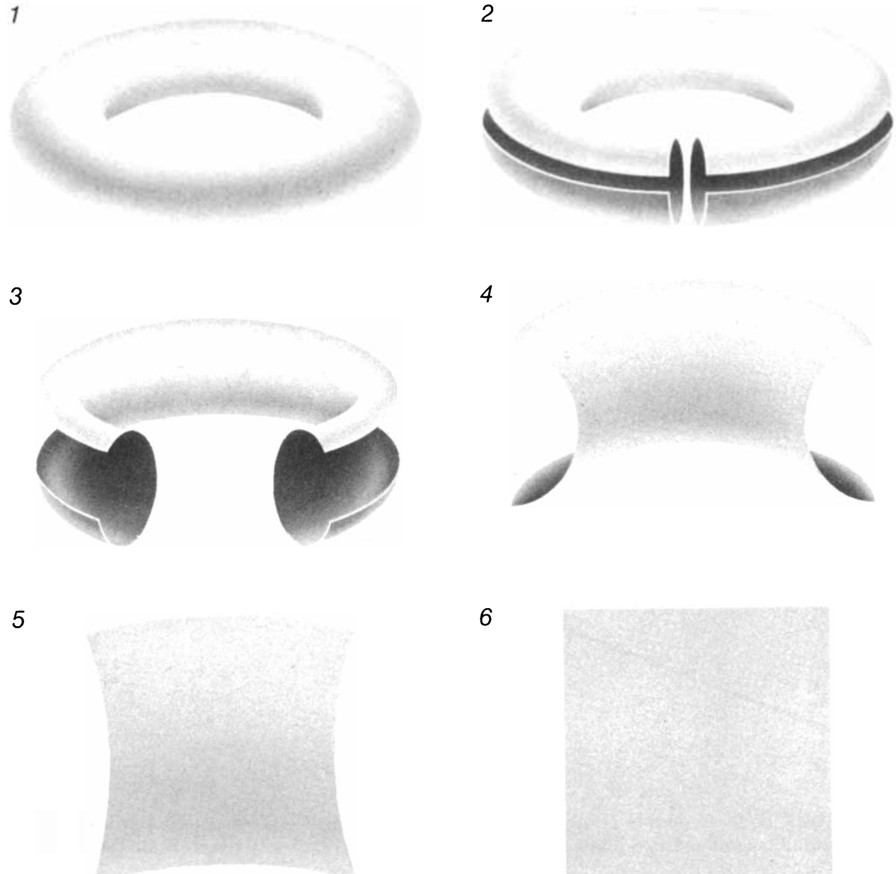


OBSERVER'S VIEW inside a three-torus is similar to the view inside a room whose walls, floor and ceiling are covered with mirrors; there is, however, no mirror reversal of the images. The line of sight passes into, say, the right wall and emerges from the left wall; looking right, therefore, the observer sees the room as it would appear from the left wall. Similarly, looking forward the observer sees the room as it would appear from the back wall, and looking up the observer sees the room as it would appear from the floor. Since the line of sight continues indefinitely across the three-torus, the room appears to be an infinite rectangular lattice extending in all directions. The three-torus is not infinite, however, because the images in the infinite rectangular array are really all images of the same thing.

were found, none of them proved to be a mechanically practical design.

A mechanical linkage can be represented mathematically by a set of line segments in the plane; at some intersections of the lines there may be pivot points, or pins, for the linkage. In the mathematical theory one assumes the lines and pivot points can pass freely through one another. The problem of constructing a physical model whose bars and pins replicate the motions of the idealized linkage is not a trivial one, but it is secondary to the mathematical analysis. It turns out that for any mathematical version of a linkage there is a physical linkage that executes the same motion, although the physical linkage may be much more complicated than its theoretical counterpart and may look quite different.

The set of all possible positions for a mechanical linkage is called the configuration space of the linkage; in most cases it is a topological manifold. Consider the simplest possible linkage, made up of a single bar pinned to an anchor point at one end but otherwise free to move in a plane. The moving end of the bar traces a circle in space, and each point on the circle corresponds to only one position of the linkage. The configuration space is a circle, which can also be viewed as a straight line segment whose ends are abstractly glued together. The circle is a one-dimensional manifold analogous to the two-torus, and every point in the manifold is identified with one position of the linkage.



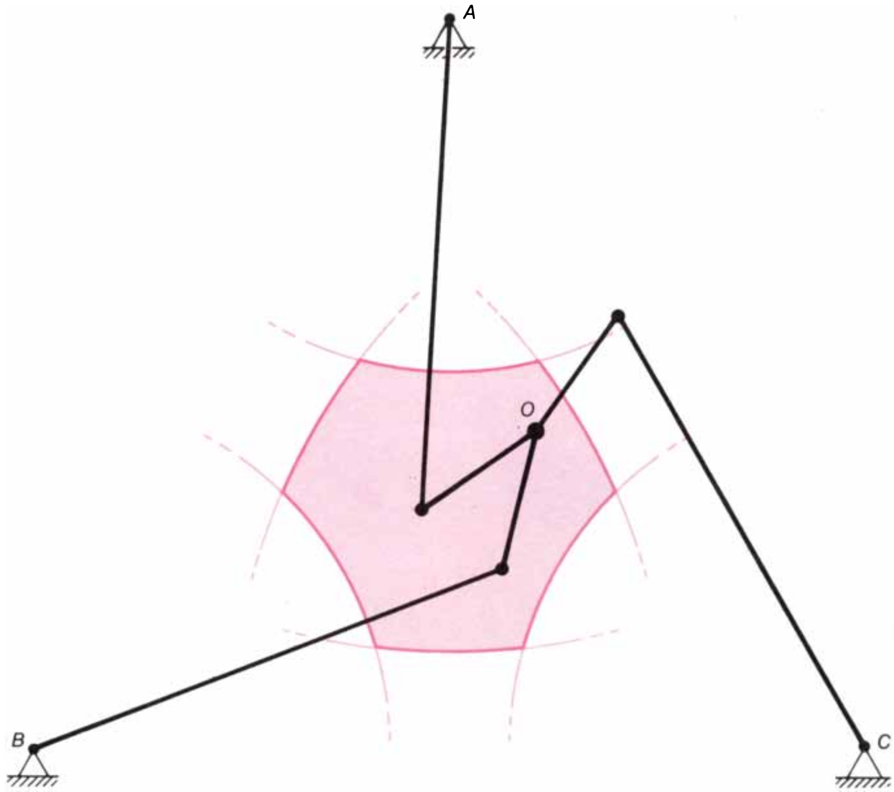
DOUGHNUT WITH ONE HOLE can be slit open and stretched into a square. If the opposite edges of the square are abstractly glued, the resulting surface is topologically equivalent to the doughnut. Since the square is flat like the plane, its geometry is Euclidean; hence from a topological point of view the one-hole doughnut is said to admit a Euclidean geometry.

By pinning another bar to the end of the first bar one obtains a double crank, a mechanical linkage with two degrees of freedom. If the second bar in the linkage is shorter than the first, the free end of the second bar can reach any point in a ring centered on the anchor point. The ring is bounded on the outside by a circle whose radius is the sum of the lengths of the two bars and bounded on the inside by a circle whose radius is the difference between the two lengths. If the second bar is the same length as the first, the free end of the second bar can reach any point in a circle whose radius is equal to the sum of the lengths of the two bars. If the second bar is longer than the first, the trace of the free end is also a ring whose inside radius is equal to the difference between the two lengths. One must not mistake these sets of points, however, for the configuration space of the linkage. The reason is that knowledge about the position of the end point of the second bar does not uniquely determine the configuration of the linkage. For every point reached by the free end of the second bar, the elbow of the double crank can bend in either of two ways.

In order to analyze the configuration space correctly it is easier to consider the possible configurations of the double crank without regard for the position of the free end. Every configuration can be described by two angles, namely the angle between each bar and some fixed direction (say to the right), measured in a counterclockwise sense. The two angles range freely and independently from zero to 360 degrees, but for each bar the angle zero degrees is identified with the angle 360 degrees. If the two angles are plotted on mutually perpendicular coordinate axes in the plane, every point within the square bounded by the lines labeled zero and 360 degrees for each angle corresponds to a different configuration of the double crank. Furthermore, every configuration of the double crank is represented by a point within the square. Since zero and 360 degrees are identified, the top edge of the square is identified with the bottom edge and the left edge is identified with the right edge. The configuration space is the two-torus.

If a third bar is added to the free end of the double crank, any position of the resulting triple crank can be described by giving the three angles of the bars. The angles are again measured counterclockwise from a fixed direction, and again they range freely and independently from zero to 360 degrees. The angles zero and 360 degrees are identified as before. When the three angles are plotted on three mutually perpendicular coordinate axes, every possible position of the triple crank is represented by a unique point in a cube whose opposite pairs of faces are abstractly glued together. The configuration space of the triple crank is therefore equivalent to the three-torus.

All the configuration spaces we have described so far lead to polygons or polyhedrons whose edges or faces can be glued together without distortion. There is no topological rule, however, that forbids abstract gluing when the edges or surfaces fail to be geometrically congruent; in fact, the examples we have given are quite special in that the gluings do not require distortion of the parts to be glued. Consider a mechanical linkage made up of three double cranks, each one fixed to a vertex of an equilateral triangle and pinned at its moving end to the moving ends of the other two double cranks [*see illustration on next page*]. In order to understand the configuration space of the linkage, first plot the set of possible positions for the central pin. Each double crank keeps the



LINKAGE of three double cranks joined by a central pin also gives rise to a two-manifold. The motion is confined to a plane, and the central pin can reach any point inside the hexagon.

pin within a ring centered on the anchor point of that double crank. Hence the central pin can reach any point that lies in the intersection of the three rings corresponding to the three double cranks. The intersecting region is a curvilinear hexagon in the plane.

There is more to the configuration space, however, than one planar hexagon. Remember that for one double crank there are two configurations of the linkage for each point reached by the end of the second crank. Similarly, for each point in the interior of the planar hexagon, the elbow of each double crank can be bent in one of two ways. With three double cranks the total number of configurations for each point in the interior of the planar hexagon is 2^3, or eight configurations. The configuration space for the three double cranks can therefore be assembled by abstractly gluing the edges of eight abstract curvilinear hexagons.

How are the eight abstract hexagons to be glued together in the configuration space? When the central pin of the three double cranks lies on an edge of the hexagon in the plane, one of the double cranks is forced into straight alignment. The alignment can be straight in one of two ways: the bars of the double crank can point in the same direction or in opposite directions. In both cases, however, the configuration of the entire linkage is specified as soon as one specifies the position of the central pin and the sense of the bend of the two unstraightened double cranks.

For each of the two edges of the planar hexagon that can be traced when the first double crank is straight, there correspond only 2^2, or four, distinct edges in the configuration space instead of eight. Along any such edge in the configuration space two hexagons are glued together. The two hexagons represent the two ways in which the first double crank can bend. The opposite edge of the first hexagon in the configuration space is also glued to the opposite edge of the second hexagon, because the first double crank also becomes straight there. The bending of the other two double cranks does not change from the first to the second hexagon. The analysis is identical for all the other edges of the configuration space.

Similarly, when the central pin of the three double cranks lies on a vertex of the hexagon in the plane, two of the double cranks are forced into straight lines. In the configuration space, therefore, there are only two points instead of eight that correspond to each vertex of the planar hexagon, one point for each way in which the third double crank can bend. The four abstract hexagons that correspond to the four ways in which the two straight double cranks can bend when the central pin moves inside the planar hexagon must share a vertex in the configuration space.

The configuration space is a surface that, unlike the planar hexagon, has no corners and no boundary. The surface can be tiled, or covered, with eight hexagons. There are 6×2^2, or 24, edges between the tiles and 6×2, or 12, vertexes where four tiles meet.

The description of the configuration space for the three double cranks that we have given so far is logically complete, because all the abstract gluings have been specified. Nevertheless, it is much more satisfying to carry out the gluings and exhibit the manifold as a closed surface in space. It turns out that such a construction is always possible when the gluing description gives rise to a manifold satisfying a technical condition called orientability. The manifold we have described is orientable and so the gluing can be done, but it is not straightforward.

It was proved in the mid-19th century that every orientable two-manifold is topologically equivalent to the surface of a doughnut with some number of holes. The number is called the genus of the surface. For example, the sphere is a surface of genus 0. The genus of the surface of a one-handle coffee cup, like the genus of the surface of a one-hole doughnut, is 1. The genus of the surface of a pretzel depends on the brand.

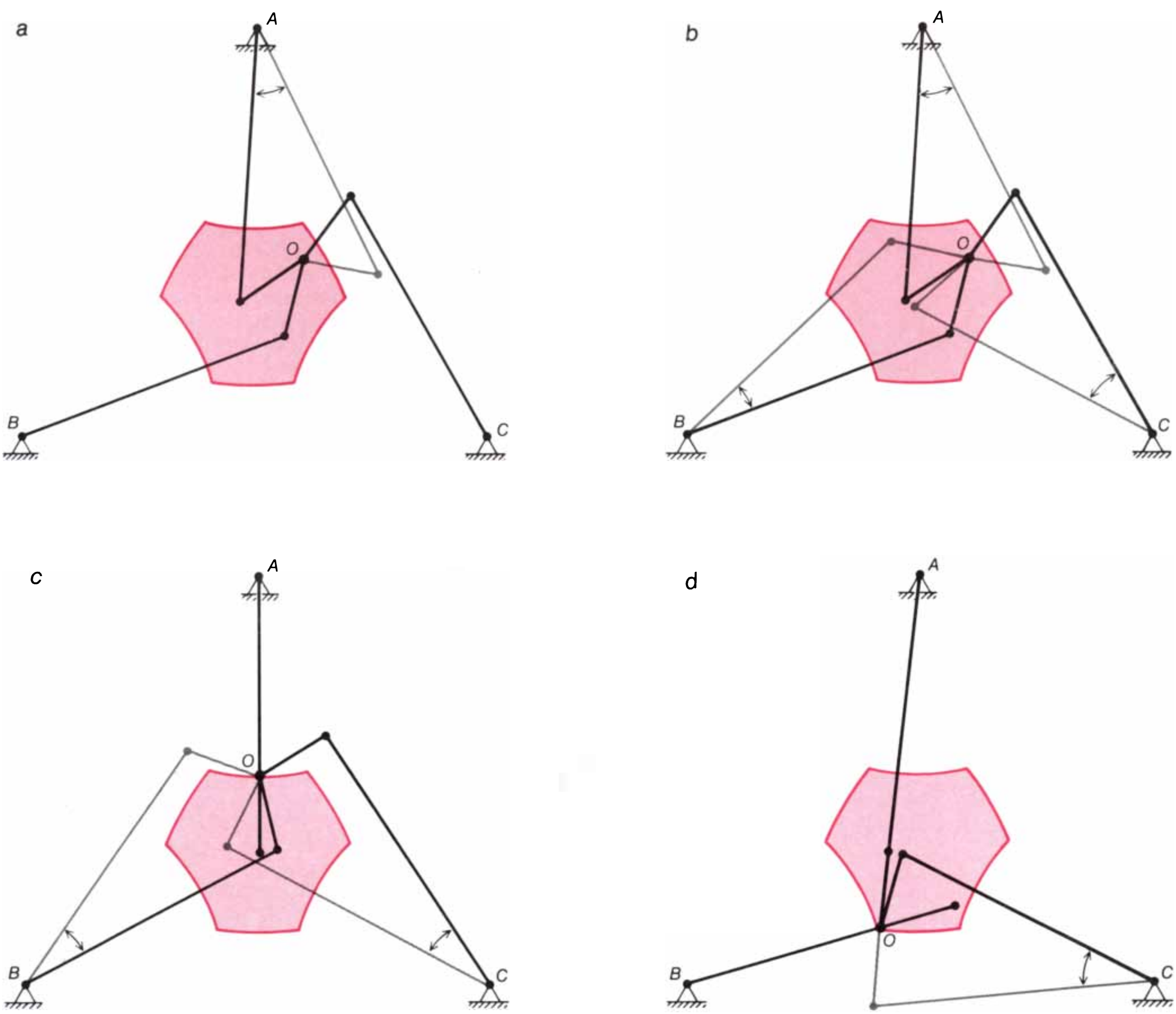
For any surface divided into polygonal cells of arbitrary shape the number of polygonal faces minus the number of edges plus the number of vertexes is a numerical constant that depends only on the surface. Remarkably, the number is independent of the way in which the surface is divided into polygonal cells. The constant is called the Euler number, after the Swiss mathematician Leonhard Euler. The Euler number of a surface of genus n is equal to $2 - 2n$. Because the surface can be curved in space, the polygons need not be planar and their edges may curve almost arbitrarily. For example, a sphere can be divided into eight triangles by connecting the north and south poles to four points along the equator. Because there are no holes on the sphere, its genus is 0, and its Euler number should be $2 - 2 \times 0$, or 2. One can easily verify that there are six vertexes and 12 edges on the surface, and so the Euler number of the eight triangular regions on the sphere is indeed 2. It is worth noting that the number of faces, vertexes and edges in the example is also characteristic of the regular octahedron, which is topologically equivalent to the sphere.

Since the configuration space of the three double cranks can be divided into eight hexagonal faces, with 24 edges and 12 vertexes, the Euler number of the configuration space must be $8 - 24 + 12$, or -4. The genus n of the manifold can be calculated by setting the Euler number -4 equal to $2 - 2n$. It follows that n is equal to 3. The eight hexagons in the configuration space for the three double cranks can be depicted in their proper relation to one another on a three-hole doughnut [*see illustration on next page*].

The visual representation of a manifold such as the three-hole doughnut is satisfying because it is concrete, but it also has a number of disadvantages. For example, many of the symmetries that are present in the abstract description of the manifold must be given up in order to picture the manifold in ordinary space. In the abstract description we initially gave for the configuration space of the three double cranks every hexagon is congruent to every other hexagon. Moreover, a rotation of any one hexagon by 120 or 240 degrees leaves its shape unchanged. In the visual representation of the configuration space, however, most of the abstract symmetry has been lost. The hexagons on the three-hole doughnut are neither congruent to one another nor rotationally symmetrical: a rotation of one of them by 120 or 240 degrees cannot be done without changing its shape.

Another problem with the three-hole doughnut is that the geometric properties of its surface vary from point to point: properties of the surface around the outer rim are different from the properties of the surface near one of the holes. It must be emphasized that the geometric properties we refer to are intrinsic to the surface. Intrinsic geometry can be determined by measurements made on the surface itself, without reference to the surrounding space where the surface is found. It is to be distinguished from the extrinsic geometry of the surface, which describes how the surface is bent in space. For example, if a flat sheet of paper is bent without distortion to form a cylinder or a cone, both the cylinder and the cone have the same intrinsic geometry as the flat sheet, although their extrinsic geometries are quite different.

The fact that a surface appears to bend in a nonuniform way when it is viewed from above is therefore not a reliable indicator of its intrinsic geometry. What is the intrinsic difference between the inner and outer regions of the surface of the doughnut? Imagine that a small piece is cut out of the convex, outside portion of the doughnut and flat-



CONFIGURATION SPACE of the three double cranks is a two-manifold on which distinct points represent distinct configurations, or possible arrangements, of the linkage. Every point inside the curvilinear hexagon traced by the central pin of the linkage can be reached when any one of the double cranks is bent in either of two ways (*a*). The bending of each double crank is independent of the bending of the other cranks, and so every point inside the hexagon gives rise to 2^3, or eight, configurations of the linkage (*b*). Whenever the central pin reaches an edge of the hexagon, only two double cranks can be bent; the linkage can assume only four configurations (*c*). When the central pin reaches a vertex of the hexagon, only one of the double cranks can bend; the linkage can assume only two configurations (*d*).

tened on a table. The piece rips open as it is flattened, much like the peel of an orange. The reverse of the process is often exploited by tailors to form a part of a garment intended to fit a convex shape, such as the bust of a dress. A pointed section called a dart is removed from the fabric and the two sides of the gap are sewn together.

On the other hand, when a small piece is cut out of the surface of the doughnut near the hole, the piece wrinkles and overlaps itself when it is flattened on a table. The tailor can reverse the process by slitting the fabric and sewing a godet, or pointed patch, into the slit. The device is often used to make a skirt that is tight below the knees but flares at the bottom. Whether a finished piece of fabric splits, overlaps or conforms to a flat surface when it is spread out on the surface is an important property of its intrinsic geometry.

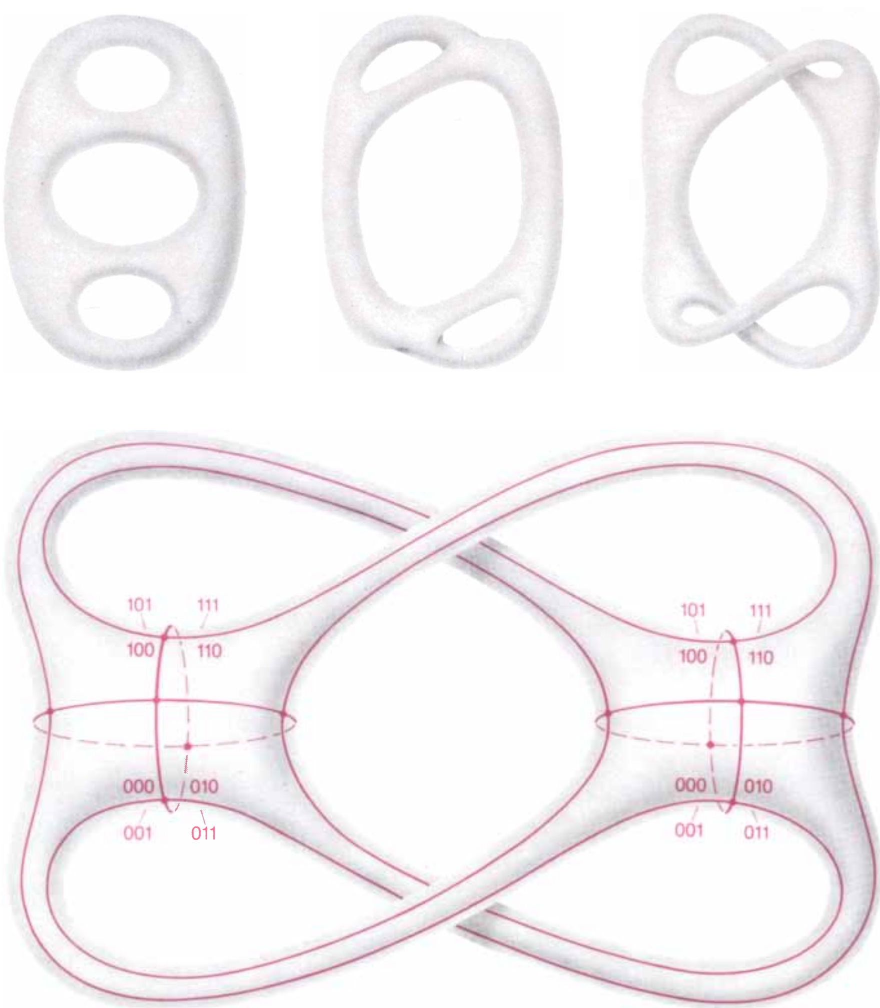
The intrinsic geometry of the surface of a one-hole doughnut varies in a way quite similar to that of the three-hole doughnut. As we have emphasized, the one-hole doughnut and the square whose opposite edges are identified have the same topology. On the square, however, the intrinsic geometry is much simpler than it is on the one-hole doughnut: the intrinsic geometry in a small region around every point in the square is the same as the geometry in a small region of the plane. The property holds even for points on the edges or the vertexes of the square. In other words, the intrinsic geometry in any small region of the square whose opposite edges are identified is the same as that in any other small region on the square. When the intrinsic geometry of a manifold has uniformity of this kind, the geometry is said to be locally homogeneous.

The introduction of the concept of local homogeneity was an important advance in the understanding of two-manifolds. About 100 years ago it was proved that any surface—not only the one-hole doughnut—can be generated in such a way that its geometry is locally homogeneous. Moreover, no manifold can be given more than one kind of locally homogeneous geometry.

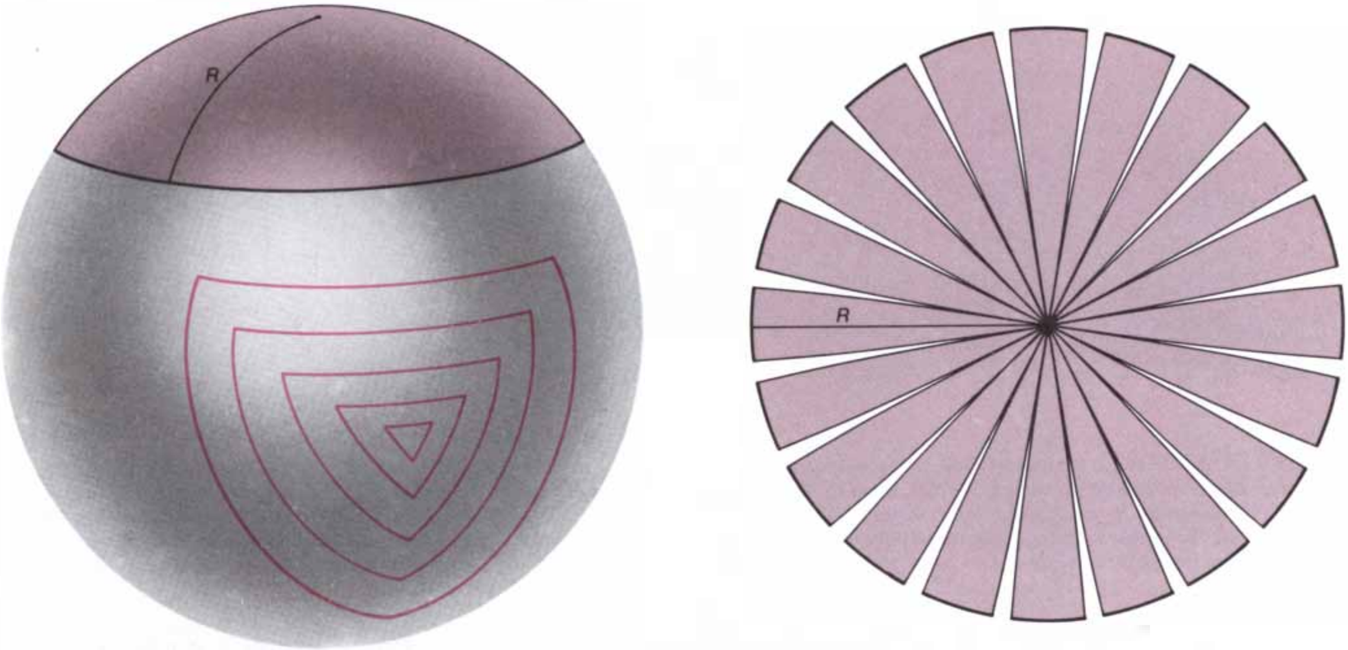
For a surface there are only three kinds of intrinsic geometry that are locally homogeneous. The first kind is the simple Euclidean geometry of the plane. On the plane the circumference of a circle is equal to pi times its diameter, and the sum of the interior angles of a triangle is 180 degrees. The plane is said to have zero Gaussian curvature, which is a measure of the intrinsic shape of a surface first developed by Carl Friedrich Gauss.

The second locally homogeneous geometry is the geometry on the surface of a sphere. A circular cap cut from the surface of a sphere rips open when it is flattened on a plane much like the piece from the convex region of the three-hole doughnut. Hence the circumference of a circle on the sphere is less than the circumference of a circle having the same radius on the plane. The missing circumference suggests the standard name for the locally homogeneous geometry of the sphere: elliptic geometry, from the Greek word for falling short. The interior angles of a triangle constructed on the sphere add up to more than 180 degrees, and the greater the ratio of the area of the triangle to the area of the surface of the sphere, the greater the sum of the angles [*see top illustration on opposite page*]. The sphere has constant positive Gaussian curvature.

As one might expect, a circle cut from a surface having the third kind of locally homogeneous geometry overlaps when it is flattened like the piece from the region near a hole of the three-hole doughnut. The circumference of such a circle is greater than the circumference of the corresponding circle on the plane. The geometry is therefore called hyperbolic geometry, from the Greek word for excess. It is impossible to define a complete hyperbolic surface with any analytic formula, but one can make approximate models of large pieces of such a



EIGHT ABSTRACT HEXAGONS make up the configuration space of the three double cranks; each hexagon corresponds to one of the eight ways the three double cranks can bend to reach one of the points in the interior of the curvilinear hexagon. If the elbow of a double crank is bent clockwise, its configuration is labeled 0; if it is bent counterclockwise, its configuration is labeled 1. The abstract hexagons are labeled with three binary digits. In order to visualize the eight hexagons and the relations imposed on them in the configuration space the hexagons can be placed on the surface of a three-hole doughnut and distorted as if they were made of rubber. The three-hole doughnut has been deformed to the topologically equivalent manifold at the bottom of the illustration in order to show the eight hexagons more symmetrically. The binary digits assigned to the abstract hexagons reflect the pattern of gluings. If the digits for two hexagons match at two of the three positions, the hexagons are glued along two opposite edges; the two edges correspond to the edges of the curvilinear hexagon where the double crank associated with the nonmatching binary digit is straight. Any four hexagons with one matching binary digit meet at a vertex in the configuration space. The diametrically opposite vertexes of the four hexagons meet at a second point in the space. The two points represent straight configurations of the two double cranks associated with the two nonmatching digit positions.

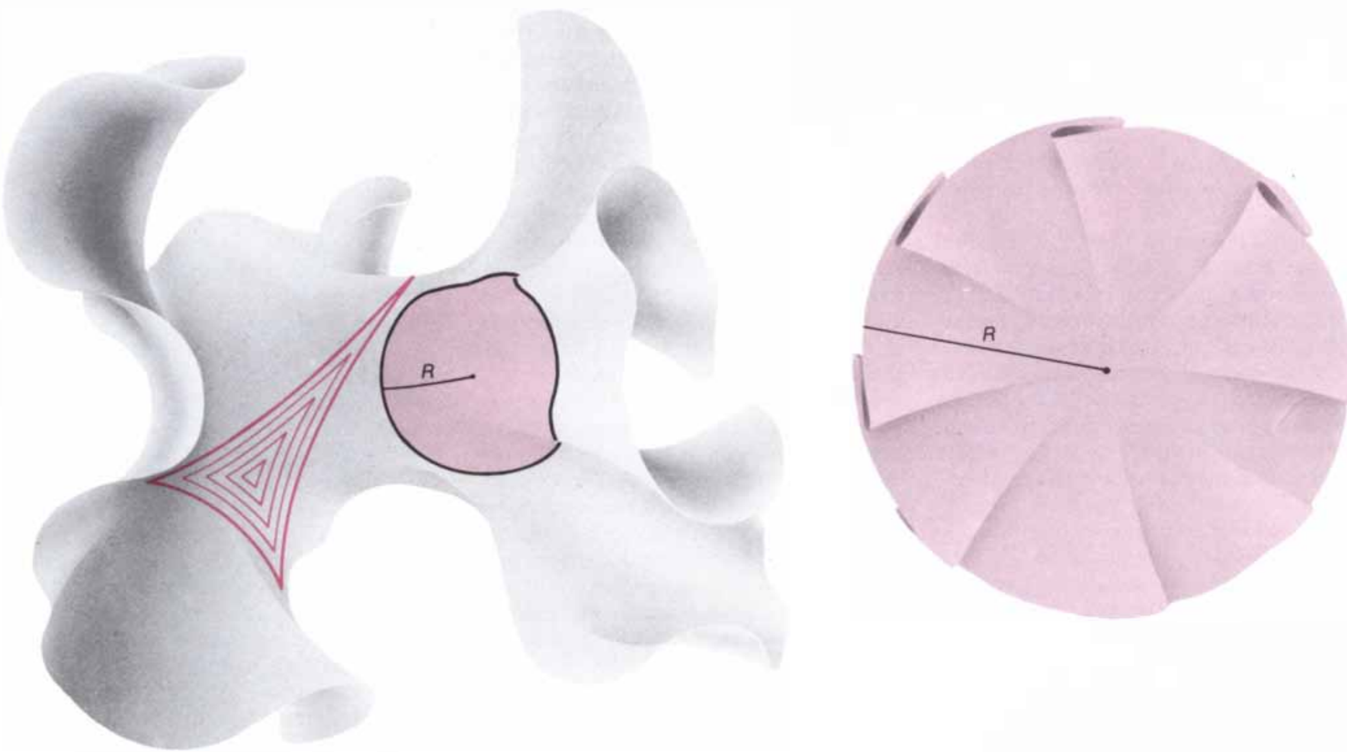


GEOMETRY ON A SPHERE, which is called elliptic geometry, differs from the ordinary Euclidean geometry on the plane. The interior angles of a triangle do not add up to 180 degrees on the sphere as they do on the plane; instead the sum increases with the area of the spherical triangle (*left*). A circular cap cut out of a sphere and flattened on the plane would crack and split as shown at the right. The area of a circle on the sphere is less than the area of a circle that has the same radius on the plane. The sphere has a constant positive curvature.

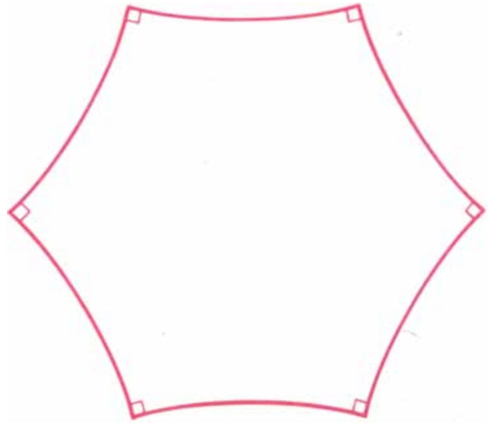
surface [*see illustration below*]. The interior angles of a triangle constructed on the surface add up to less than 180 degrees, and the greater the area of the triangle, the smaller the sum of the angles. The hyperbolic surface has constant negative Gaussian curvature.

It is instructive to show how the three-hole doughnut can be given a locally homogeneous geometry. Remember that each of the eight hexagons from which the manifold was originally constructed can be bent and deformed in any way, as long as no hexagon is cut or torn. The method is then to deform each hexagon in such a way that it has a locally homogeneous geometry and still fits together with the other hexagonal pieces of the manifold as it does on the three-hole doughnut.

All eight hexagons on the three-hole



GEOMETRY ON A SURFACE of constant negative curvature (*left*) is called hyperbolic geometry. The sum of the interior angles of a triangle is less than 180 degrees, and the sum decreases as the triangle grows. A circle cut out of a hyperbolic surface would wrinkle and overlap as shown at the right; its area is greater than the area of a circle that has the same radius on the plane. A paper model of a hyperbolic surface can be made by gluing many equilateral triangles along their edges in such a way that seven triangles meet at each vertex.



ANGLES OF A REGULAR HEXAGON can be reduced in size if the hexagon is allowed to grow on a hyperbolic surface. In the illustration the hexagon grows until each interior angle is equal to a right angle. The eight hexagons in the configuration space of the three double cranks must meet four hexagons at a vertex, and so all six interior angles of each hexagon must be right angles. If the eight hexagons are topologically deformed into right-angled hyperbolic hexagons and abstractly glued as before, the resulting two-manifold has a constant curvature; its geometry is said to be locally homogeneous. The geometry of the configuration space is thereby greatly simplified.

doughnut meet four at a vertex. If the hexagons were Euclidean, the angles at any vertex would add up to 480 degrees, which is impossible. If the hexagons were spherical, the sum of the four angles at a vertex would be even greater than 480 degrees, which is also impossible. In the hyperbolic plane, however, the larger the polygon, the smaller the interior angles. A sufficiently large hexagon in the hyperbolic plane must have interior angles of 90 degrees; four such hexagons would fit snugly together at a vertex. Hence if the eight hexagons are placed on the surface of a hyperbolic plane and inflated on that plane until each interior angle shrinks to 90 degrees, the manifold constructed by gluing the eight hexagons will have a locally homogeneous, hyperbolic geometry. The manifold cannot be directly visualized in its new form, but the geometric properties of the manifold are much simpler.

The reader may enjoy verifying that the surface of any doughnut with two or more holes can be cut into hexagons that meet four at a vertex. The resulting manifold can be given a hyperbolic geometry by constructing it from right-angled, hyperbolic hexagons. A more traditional procedure is to cut the surface open into a polygon whose vertexes all meet on the surface at one point. The three-hole doughnut, for example, can be split open into a dodecagon, or 12-sided polygon, as well as into eight hexagons. If the surface is sufficiently complex, the polygon derived from the split must have at least six vertexes. If all six vertexes are to fit together properly, the interior angles of the polygon must be reduced. The reduction is done by allowing the polygon to grow in the hyperbolic plane. When the edges of the polygon are glued in pairs, the new surface is topologically identical with the original surface, but it has the locally homogeneous geometry of the hyperbolic plane.

There are only four finite surfaces for which the locally homogeneous geometry is not hyperbolic, because the polygons that arise from the cuts in the surfaces have fewer than six sides. The one-hole doughnut gives rise to a square, and all four corners of the square can be abstractly glued without changing the interior angles of the square. Since no further deformation is needed, the locally homogeneous geometry given to the two-torus is Euclidean. Similarly, the sphere and a nonorientable two-manifold called the projective plane are given an elliptic geometry, and a nonorientable two-manifold called the Klein bottle is given a Euclidean geometry.

A three-manifold can be curved in much the same way as a surface can: every two-dimensional slice of a positively curved three-manifold would split open if it were placed in ordinary Euclidean space, and every two-dimensional slice of a negatively curved three-manifold would wrinkle and overlap. Elliptic geometry, Euclidean geometry and hyperbolic geometry all have their three-dimensional counterparts.

In 1976 one of us (Thurston) began to suggest that locally homogeneous, hyperbolic geometry is the key to the understanding of almost all three-manifolds. The development has come as a surprise to many topologists, because three-manifolds are so much more complicated than two-manifolds. Whereas any orientable two-manifold can be specified and listed according to its genus, every three-manifold, like a tangled loop of string, seems to have its own distinct properties and resists fitting into any larger pattern. On closer scrutiny, however, larger patterns have begun to emerge. The patterns depend on the fact that many three-manifolds can be given a locally homogeneous geometry.

How can such a simple geometric structure be imposed? A procedure quite similar to the one we have described for two-manifolds works for a large number of cases. The manifold is cut apart into a polyhedron, and one must determine how many vertexes of the polyhedron are to be fit in place when it is abstractly glued back together. For example, in the Seifert-Weber space all 20 corners of the dodecahedron that generates the space are abstractly glued together. The solid angle formed at each vertex of a Euclidean dodecahedron is much too large for 20 such vertexes to fit together at a point. If the dodecahedron is placed in three-dimensional hyperbolic space, however, it can be expanded until the solid angle at each vertex is small enough to pack 20 equal vertexes around a point [*see illustration on the cover of this issue*]. When opposite faces of the hyperbolic dodecahedron are abstractly glued together after a rotation of three-tenths of a turn, the resulting manifold is a Seifert-Weber space with a locally homogeneous, hyperbolic geometry.

The Poincaré dodecahedral space is also derived by gluing the faces of a dodecahedron, but the vertexes are glued together in five groups of four. The solid angle at the vertex of an ordinary dodecahedron is slightly too small to pack tightly around a point in groups of four, but a suitably large dodecahedron in a positively curved space has corners that are just the right size. The enlarged dodecahedron makes it possible to construct a Poincaré dodecahedral space with a locally homogeneous, elliptic geometry [*see illustration at right on opposite page*]. In this context it is worth mentioning how a locally homogeneous geometry is given to the three-torus. In the construction of the manifold the eight vertexes of a cube are abstractly glued. Because the eight corners can fit together at a point without distortion, the locally homogeneous geometry of the three-torus is Euclidean.

Lest the reader be misled, we must point out that the preceding examples are not really typical because they are highly symmetrical. When a three-manifold is defined by gluing the faces of an irregular polyhedron, more care must be taken to give the polyhedron a shape that leads to a locally homogeneous geometry for the three-manifold. The shapes of the polyhedral faces that are glued together must match, and the angles between the faces that surround any edge must add up to 360 degrees.

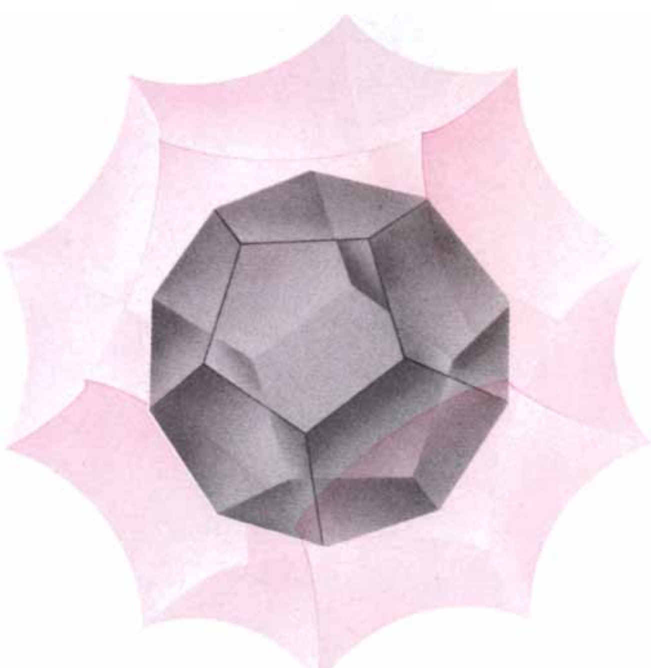
There are at least two major differences between the geometry of two-manifolds and the geometry of three-manifolds. First, there are five more kinds of locally homogeneous geometry that can be given to three-manifolds, in addition to the three we have mentioned. The additional geometries come about because in three or more dimensions an intrinsic curvature is defined for each two-dimensional slice that passes through a point. A locally homogeneous geometry need not have the same curvature in all the two-dimensional slices. Nevertheless, an understanding of the intrinsic curvature of all eight geometries can be based on the simpler geometries of two-manifolds.

The second difference between the two- and three-manifolds might seem to present insurmountable complications. It is possible to combine three-manifolds in such a way as to yield new three-manifolds that cannot be given a locally homogeneous geometry. Fortunately, topologists know how to split a three-manifold into primitive pieces by purely topological methods.

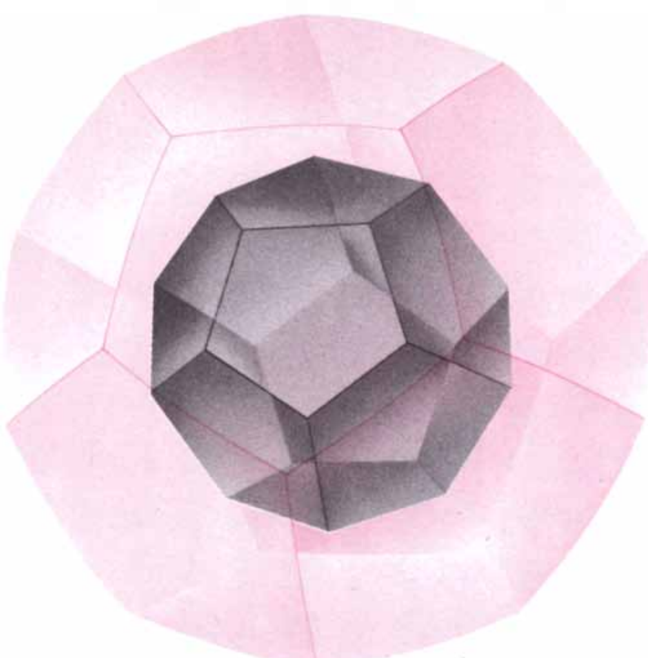
One of us (Thurston) has proposed that after a three-manifold has been cut up into its primitive pieces, each of the resulting pieces does in fact admit a locally homogeneous geometry of one of the eight possible types. The conjecture has been proved for wide classes of manifolds, and it has been empirically tested for many other examples either by hand or with the aid of a computer, and it has never been found wanting. It now seems unlikely a counterexample will be found.

The empirical studies also suggest that for most three-manifolds the complications of three-manifold geometry do not come into play. Indeed, it has been proved that, in a certain technical sense of the word, "most" three-manifolds can be given a locally hyperbolic geometry. The finding is fortunate, because hyperbolic three-manifolds have many beautiful properties. For example, in 1971 G. D. Mostow of Yale University proved that if a three-manifold can be given a locally hyperbolic geometry, the geometry is completely determined by the topology. A consequence of Mostow's theorem is that all manifolds having a locally homogeneous geometry can in principle be classified. Moreover, for hyperbolic three-manifolds the theorem gives a rough-and-ready test of identity. When a manifold is deformed into its geometrically tractable form, its volume can be measured, and the theorem guarantees that the volume depends only on the topological type of the manifold. It can often be quite difficult to distinguish two manifolds in their arbitrary topological form, and so the volume turns out to be a handy signature for each manifold.

With these results in mind it is worth returning to our initial speculations about the topological structure of the universe. Observational evidence suggests the universe is homogeneous everywhere and has either elliptic, hyperbolic or Euclidean geometry. There is also strong support for the theory that the universe is currently in a stage of expansion that has continued since the beginning of the big bang. It is interesting to speculate on what the distant future holds in store, but there are essentially only two possibilities. One is that the mutual gravitational attraction of the matter in the universe will finally



SEIFERT-WEBER SPACE can be given a locally hyperbolic geometry if the dodecahedron that generates the space is allowed to grow in hyperbolic space. The growth of a polyhedron in hyperbolic space is similar to the growth of a polygon on a hyperbolic surface. When the dodecahedron grows, the solid angle at each vertex shrinks, and so each vertex becomes progressively sharper. The abstract gluings that lead to the Seifert-Weber space specify that all 20 vertexes of the dodecahedron must meet at a point. The solid angle at each vertex must therefore be shrunk in the hyperbolic space until all 20 of the solid angles are small enough to fit together around a single point.

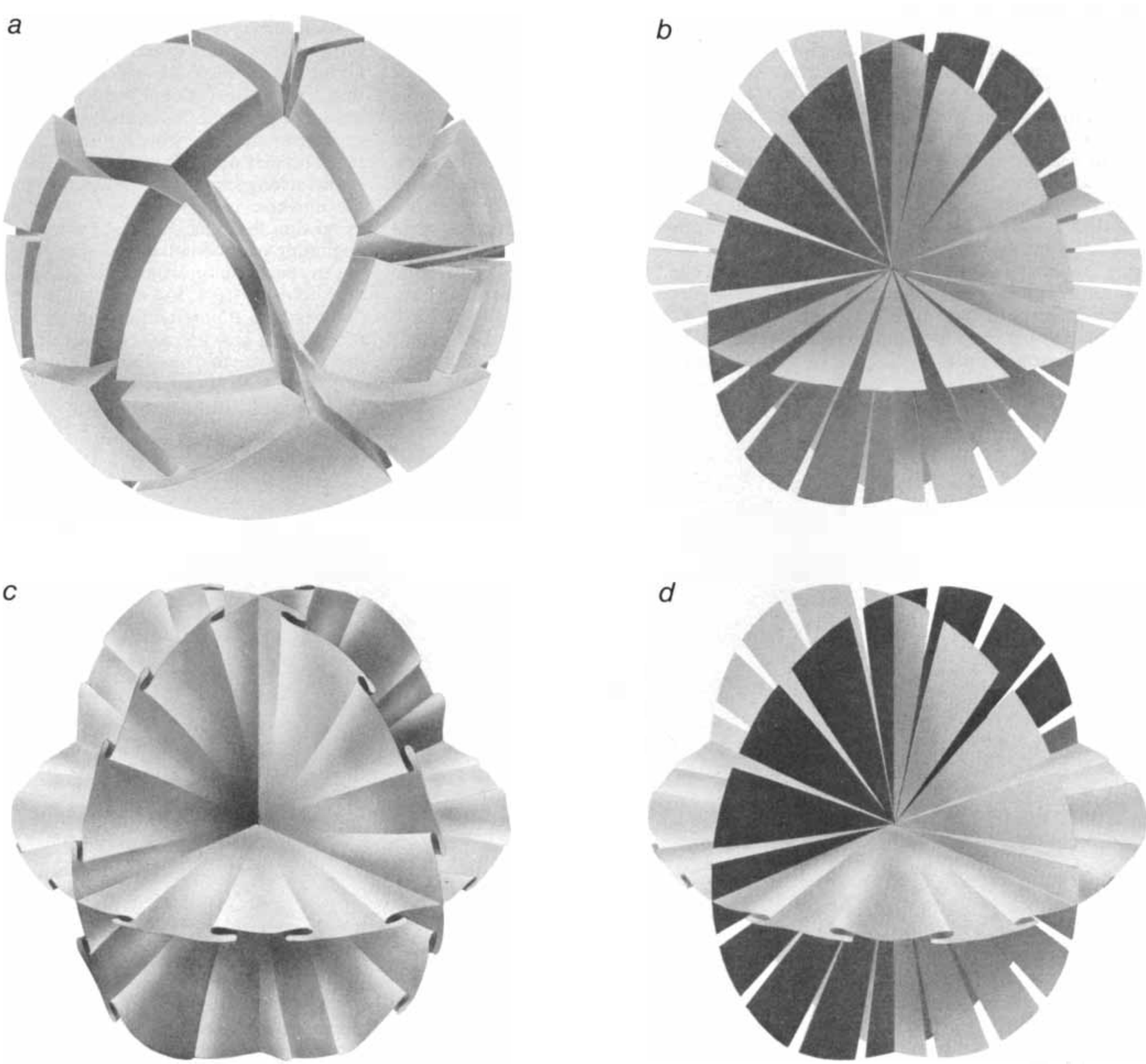


POINCARÉ DODECAHEDRAL SPACE is also generated by gluing pairs of opposite faces of a dodecahedron, but one member of each pair is matched to its counterpart after a rotation of one-tenth of a turn instead of the three-tenths of a turn required for the Seifert-Weber space. The abstract gluings lead to the identification of four vertexes of the dodecahedron at every vertex in the manifold. The solid angle at each vertex of an ordinary dodecahedron is slightly too small for four dodecahedrons to fit snugly around a point, but the solid angles can be increased if the dodecahedron is inflated in elliptic space. The effect is the reverse of the inflation in hyperbolic space.

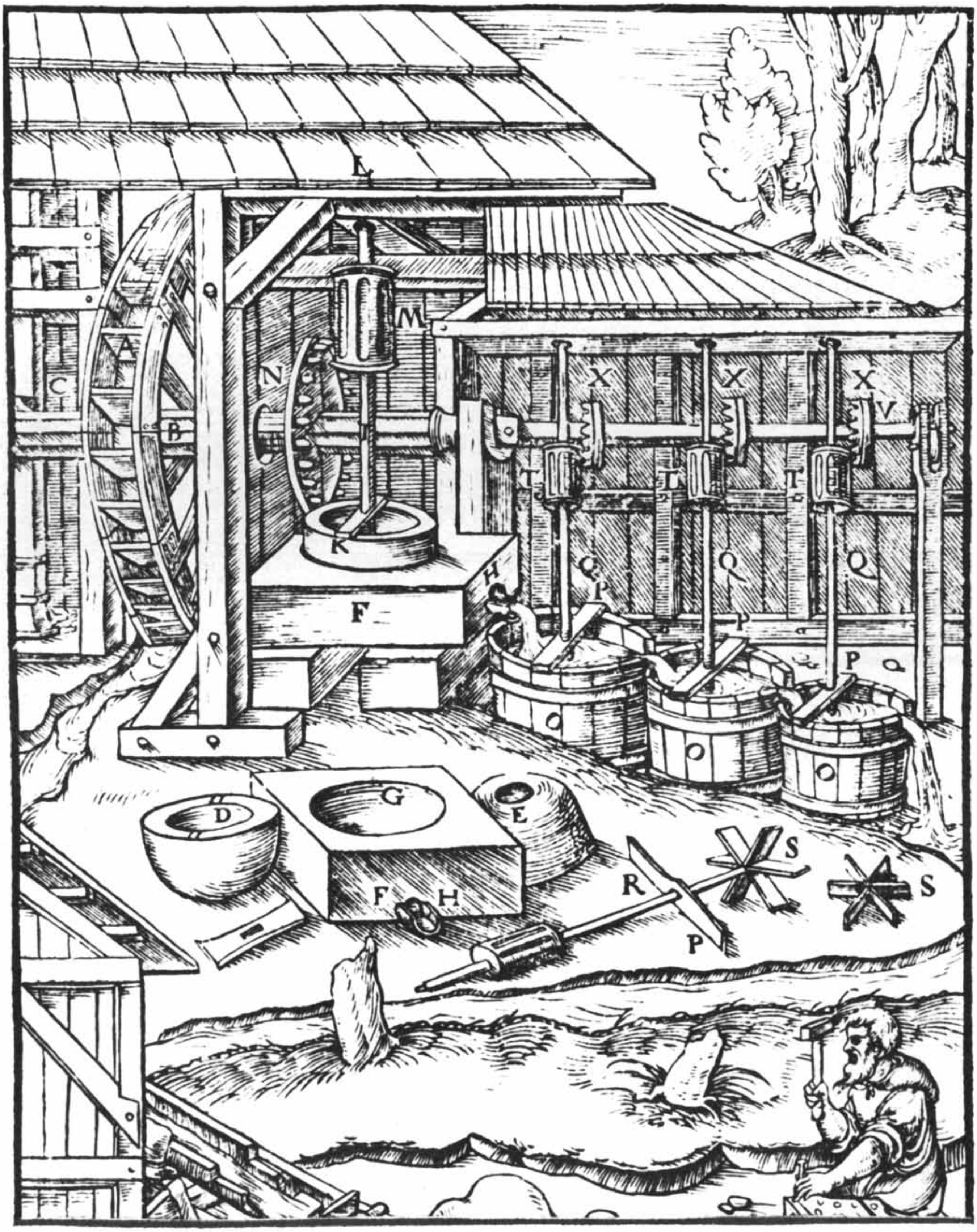
halt the expansion and cause the universe to recollapse in a "big crunch." A second possibility is that the gravitational attraction is not strong enough to halt the expansion, and the universe will expand forever.

One consequence of the general theory of relativity is that the ultimate fate of the universe depends on its geometry. If the universe has an elliptic geometry, it will eventually recollapse. If it has a hyperbolic geometry, it will expand forever. If it has a Euclidean geometry, it will also expand forever, but the rate of the expansion will approach zero. In principle it would be possible to determine the geometry of the universe by laying out a huge triangle and accurately measuring its interior angles. If the sum of the angles were greater than 180 degrees, the geometry of space would be elliptic; if the sum of the angles were equal to 180 degrees, the geometry would be Euclidean and if the sum were less than 180 degrees, the geometry would be hyperbolic. In practice cosmologists try to estimate the density of the matter in the universe and the rate of the expansion, because the geometry of the universe can be correlated with the two measurements. If the density is high enough for a given rate of expansion, the universe will recollapse.

There is a widespread misconception, however, that the curvature of the universe determines whether the universe is finite or infinite in extent. It is often asserted that if the universe is finite, its geometry must be elliptic and, conversely, that if the geometry of the universe is hyperbolic, the universe must be infinite. The Seifert-Weber space, which is a finite three-manifold with a locally hyperbolic geometry, shows that neither of these beliefs is true. Indeed, most finite topological models of space are three-manifolds like the Seifert-Weber space with a locally hyperbolic geometry. Such manifolds yield models of a finite universe that expands forever.



SPHERICAL "SLICE" of a curved three-manifold, analogous to the circular pieces of the two-manifolds shown in the illustrations on page 117, cannot be fitted into ordinary Euclidean space without deformation. A positively curved three-manifold would split open everywhere (*a*); every two-dimensional slice of the manifold has the curvature of an ordinary sphere (*b*). Similarly, every two-dimensional slice of a negatively curved three-manifold would wrinkle and overlap as if it were cut out of a hyperbolic surface (*c*). A three-manifold whose curvature varies with direction can still have a locally homogeneous geometry, as long as the pattern is the same at every point. For example, one slice of the manifold could wrinkle and overlap in ordinary space, whereas two other slices could split apart (*d*).



MULTIPURPOSE ORE MILL powered by a single water wheel at the upper left was illustrated in Georgius Agricola's *De re metallica,* published in 1556. Gold ore was processed in the following steps. First the ore was crushed by a cam-lifted stamp (*c*), just visible to the left of the water wheel. Next the crushed ore was ground to a powder in a pair of millstones to the right of the wheel. Two spare dome-shaped upper millstones (*d, e*) are on the ground on each side of a spare lower millstone; one of the upper millstones is turned upside down to show the hole that admitted the crushed ore. The outlet for the powdered ore (*h*) in the lower millstone deposited the powder into the first of three settling tubs (*o*). The slurry of powdered ore in the tubs was agitated by paddles driven by cogs (*x*) attached to the axle of the wheel. The agitation separated the heavier gold from the lighter dross, which eventually spilled from the last of the settling tubs.

Medieval Roots of the Industrial Revolution

The revolution is generally dated to the arrival of steam power in the 18th and 19th centuries. It is now clear that long before then a significant role was played by water-powered machines

by Terry S. Reynolds

The origins of modern industry are often dated only to the late 18th and early 19th centuries, when manual labor was displaced by steam-powered machines, first in the cotton-textile industry and later in other industries. This period is commonly called the Industrial Revolution, a term strongly suggesting that there had been a sharp break from developments in the preceding centuries.

The history of water power in medieval and early modern Europe presents a different picture. Powered machinery had begun to displace manual labor long before the 18th century, and in some areas of Europe it had done so on a substantial scale and in many industries. In other words, the rise of European industry should more properly be regarded as an evolutionary process going back at least to the eighth or ninth century, when European engineers began to aggressively apply water power to industrial processes.

Although water power can be harnessed by a variety of devices, the commonest device is of course a wheel fitted with blades or buckets. Such a wheel can be mounted either horizontally or vertically. Until the introduction of the water turbine in the 1830's horizontally mounted water wheels were the simpler of the two types. In the early horizontal water mill the lower end of a vertical shaft carried a small horizontal wheel consisting only of blades. The upper end of the axle was linked directly, without gearing, to a rotating millstone. This kind of mill was cheap to build but usually delivered no more power than a donkey or a horse (less than one horsepower). Even that power it generated wastefully, operating at an efficiency of only 5 to 15 percent. The horizontal mill was also not as easily adapted as the vertical one to tasks other than milling grain. Hence it did not play an important role in the early evolution of water-powered industry in Europe.

Vertical water wheels, on the other hand, did. There were two main subtypes of vertical wheel: undershot and overshot. The undershot was the simpler. It had flat radial blades attached to its circumference, and it was energized by the impact of water flowing under the wheel and pushing against the blades. The undershot wheel could work in almost any stream as long as there was enough water flowing at a modest speed, but it worked most effectively in a confined channel. Its typical output was three to five times as great as that of a horizontal wheel (namely about two to three horsepower) and its efficiency was 20 to 30 percent.

With the overshot wheel water was fed over the wheel into "buckets" built into the wheel's circumference. There the weight of the water, rather than its impact, turned the wheel, with each bucket discharging its water at the bottom of the wheel and returning empty to the top to begin the cycle anew. Overshot wheels were usually more expensive to build than undershot or horizontal wheels, since they called for a dam and an elevated water channel, and they could not handle large volumes of water. With a small volume and a head of water ranging from three to 12 meters, however, they were capable of operating at an efficiency of from 50 to 70 percent and delivering anywhere from two to 40 horsepower. (The average was from five to seven horsepower.)

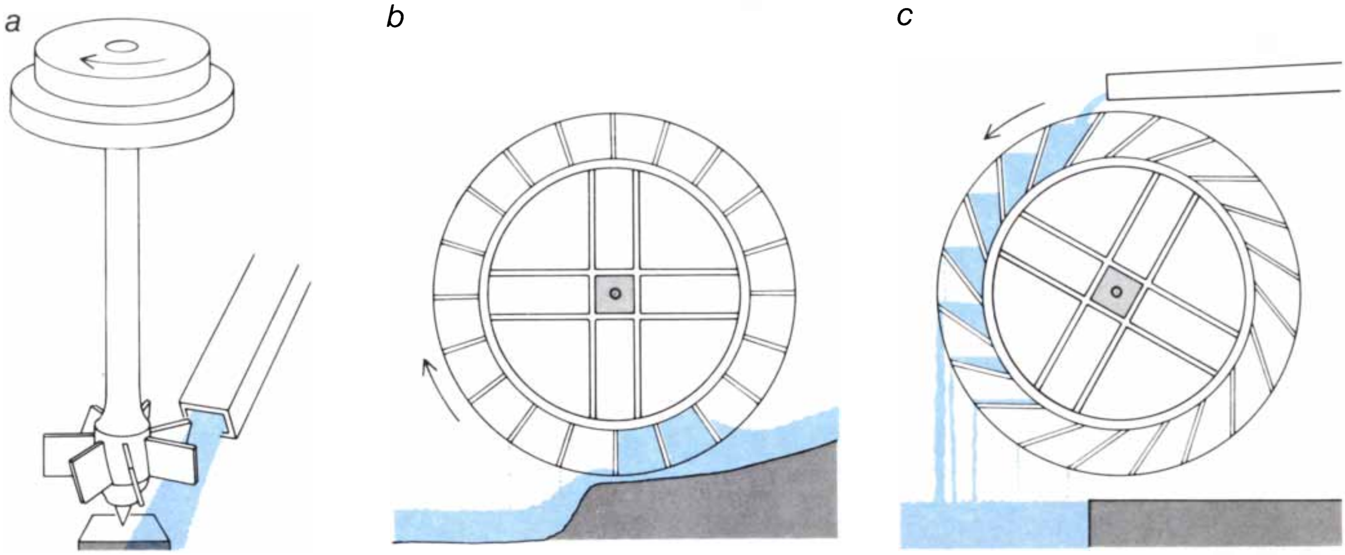
Although ancient engineers devised both undershot and overshot wheels, neither were widely built. For example, in the first century B.C. the Roman engineer Vitruvius described the undershot wheel in a section of his *De architectura* dealing with machines rarely employed. Indeed, there are fewer than a dozen known literary references to the use of water power in all antiquity.

In medieval Europe social and economic conditions increased the need for such power and initiated the trend toward replacing manual labor with powered machines. One of the most critical elements in the changing technological climate of western Europe was the monastic system, based on the rules laid down early in the sixth century by St. Benedict. These rules had two features that encouraged the introduction of water power. First, monks were to devote certain rigidly regulated periods to manual labor, to reading and study and to spiritual duties such as meditation and prayer. Second, the monastery was to be self-sufficient and to isolate itself from worldly influences.

These rules provided an incentive for the development of water power because only by harnessing power for time-consuming manual tasks such as grinding grain could the monastery ever become self-sufficient or give the monks time for study and prayer. The monastic order that was perhaps the most aggressive in developing water power was the Cistercian. By 1300 there were more than 500 Cistercian monasteries. Virtually all of them had a water mill, and many had five or more mills.

One other social class contributed to the diffusion of water power in the medieval West. It was the feudal nobility. The nobles saw in the introduction of water power an additional means of getting revenue from the peasantry. In a number of areas of Europe lords imposed on their serfs the obligation to bring their grain for grinding only to the lord's mill. That milling monopoly, at first limited to grain mills, was sometimes extended to other water-powered processes, for example fulling: the finishing of wool cloth. Thus in western Europe two major groups—the clergy and the landed nobility—developed an interest in expanding water power. They were later joined by a third group: the merchant class, which saw in milling a way to make a profit.

Other economic pressures in medie-



MEDIEVAL WATER WHEELS rotated either horizontally on a vertical axle (*a*) or vertically on a horizontal axle (*b, c*). The horizontal wheel, an ancestor of the turbine, was called a Norse mill; it was inefficient and was not widely used for anything but flour milling. The earliest vertical wheel (*b*) is known as an undershot wheel because the water passes under it; its chief virtue is its low cost and simplicity of installation. The overshot wheel (*c*) usually requires either a substantial fall of water (three meters or more) or a dam to provide such a fall.

val Europe contributed to the extension of water power. By the seventh century the labor surplus that had plagued the Roman Empire at its height and may have discouraged the adoption of water power had completely disappeared. The ensuing labor shortage encouraged the adoption of laborsaving devices such as the water wheel. Europe's geography may also have contributed to the development. The heart of medieval European civilization lay in the drainage basins of rivers that flowed into the Bay of Biscay, the English Channel and the North Sea. In this region were hundreds of small to middle-size streams with a fairly regular flow, which was convenient for the development of water power. The heart of Classical civilization, on the other hand, was in the Mediterranean basin, where because of the dry climate stream flow tended to be erratic and seasonal.

As a result of these social, economic and geographic factors the use of water power in Europe steadily grew, particularly from the ninth century on. By 1500 water wheels were in operation throughout Europe. At some sites the concentration of powered machines was quite comparable to that in the factories of the 18th- and 19th-century Industrial Revolution. Let us look at three elements of medieval water-wheel technology: growth in numbers, growth in applications and concentrations of power.

The best source of data for the number of water mills in any part of medieval Europe is William the Conqueror's late-11th-century census of his newly acquired English domain. The areas of England under Norman rule in the late 11th century had 5,624 water mills at more than 3,000 locations. This amounted to one mill for approximately every 50 households. Moreover, in some areas the mills were close together: as many as 30 mills might be found along 16 kilometers of the same stream. Whether these mills had horizontal, undershot or overshot wheels and what work they did was not recorded. Most of them probably ground grain, a tedious task that if done by hand would have consumed from two to three hours of each housewife's day.

Even figures as incomplete as these do not exist for other European countries of the same period. Presumably some of those areas were technologically ahead of England. Later evidence, however, indicates the substitution of water power for manual labor must have been growing at a rate at least comparable to that in England. For example, in 1694 the Marquis de Vauban, a French military engineer, estimated that France had 80,000 flour mills, 15,000 industrial mills and 500 iron mills and metallurgical works. This is a total of more than 95,000 mills, although some of them, particularly the flour mills, were powered not by water but by wind.

Even in industrially backward regions of Europe, such as Russia and Poland, water wheels were common before the introduction of steam power. A 1666 survey of the northern tributaries of the Dnieper River in Russia, from the Sula River to the Vorskla River, lists 50 dams and 300 water wheels. One of these tributaries alone, the Udai River, had 72 water mills. By late in the 18th century the part of Poland under Austrian occupation had more than 5,000 water mills.

The steady numerical growth of water wheels was accompanied by a steady geographic diffusion. By the 13th century water wheels were turning throughout Europe: from the Black Sea to the Baltic, from Britain to the Balkans, from Spain to Sweden.

Ancient engineers had applied the rotary motion of the water wheel in only two ways: in the flour mill and in what was called a noria, a wheel to raise water. In the noria the motion was harnessed directly; it was not transformed by any kind of gearing. In the flour mill, however, right-angle gearing at one end of the wheel's axle changed the speed of rotation and transformed the wheel's motion in the vertical plane into motion in the horizontal plane in order to turn a millstone. Beyond this ancient engineers evidently did not go, although one poem of the fourth century Roman poet Ausonius suggests the possibility of wheel-driven saws.

Beginning in the ninth century millwrights started to extend the developments of antiquity. For example, medieval European millwrights applied the vertical water wheel to several processes that, like water-raising with the noria, called for rotary motion in the same plane as the wheel itself. One was the grinding and polishing of metals in cutlery mills. These mills are first mentioned in documents dating from early in the 13th century. In such mills gears were installed not to alter the plane of rotation but to step up the rotational speed of the water-wheel axle and in some instances to shift the direction of the plane of rotation to grindstones mounted on shafts set at right angles to the water-wheel axle.

Other examples of new applications of water power that harnessed the rotary motion of vertical water wheels in the same plane were lathes (the earliest evidence of the use of water power for this purpose was in the 14th century), pipe borers (in the 15th century), rollers for producing metal sheets and rotary cutters for cutting the sheets (also in the 15th century), fans for mine ventilation, mine hoists and ball-and-chain mine pumps (all in the 16th century).

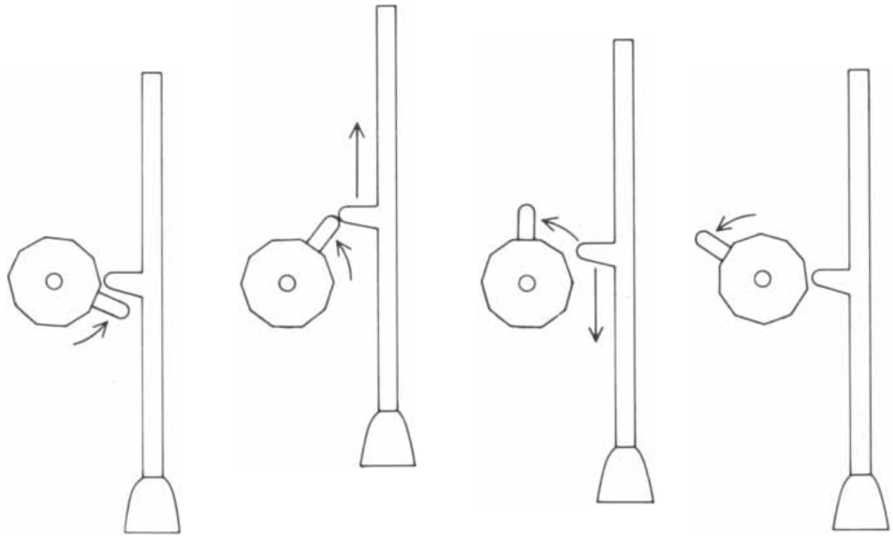
In a similar way medieval engineers extended the combination of the water wheel and vertical-to-horizontal plane gearing. As early as the ninth century in France traditional water-powered flour mills were modified to turn millstones not for grinding wheat into flour (the only Roman use of the combination of gearing and the water wheel) but for grinding malt for beer mash. Later vertical-to-horizontal plane gearing was applied in order to replace manual labor in such activities as grinding metal ores into powder.

By the 11th century vertical-to-horizontal gearing arrangements had been further developed to turn millstones set on edge, which crushed rather than ground. Water mills with such edge-rolling stones may have served to crush oil out of olives as early as the 11th century and were definitely serving this purpose and others by the 12th century. Early in the latter century the mills were adopted by the tanning industry: they reduced oak bark to powder preparatory to the leaching process that extracted the tannin. Water-powered edge rollers may also have been used to crush sugarcane in Sicily as early as the 12th century. Later such mills served to crush mustard seed (the earliest evidence of the application of water power for this purpose is in the 13th century), poppy seed (also in the 13th century) and dyeing substances (in the 14th century).

Although the rotary motion of water wheels—speeded up, slowed down or translated into another plane of rotation—could be applied to many tasks, other tasks called not for rotary motion but for linear motion. For example, many industries employed pummeling or hammering actions. They included washing wool cloth, crushing ore for a smelter, forging iron and separating the fiber from flax plants. Medieval technicians between the 10th and the 15th centuries devised two solutions to the problem of transforming rotary motion into the linear motion needed to actuate hammers: the cam and the crank.

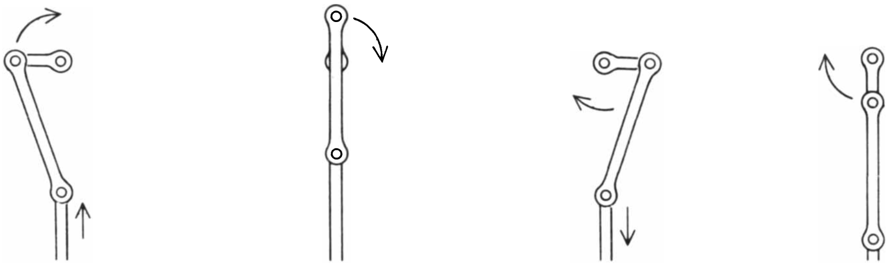
The cam was the earlier of the two devices and was long the more widely applied. It was a simple device, basically a small projection fixed on an axle. Not a medieval invention, it had appeared in small-scale mechanisms in antiquity but had never been applied to large-scale production devices. It came of age in medieval Europe, and in conjunction with water wheels it served mostly to actuate hammers. By 1500 European engineers had developed two forms of water-actuated hammer: the vertical stamp and the recumbent stamp. In the vertical stamp a cam on a horizontal shaft rotated against a similar projection on a vertical rod with a hammerhead at its lower end. As the cam rotated it lifted the rod until contact was lost; the rod then dropped, delivering the hammer blow. With the recumbent stamp the cam was rotated against the hammer end of a horizontal rod pivoted at the other end, first lifting the hammer and then, as the rotation continued, dropping it.

Water-actuated trip hammers could

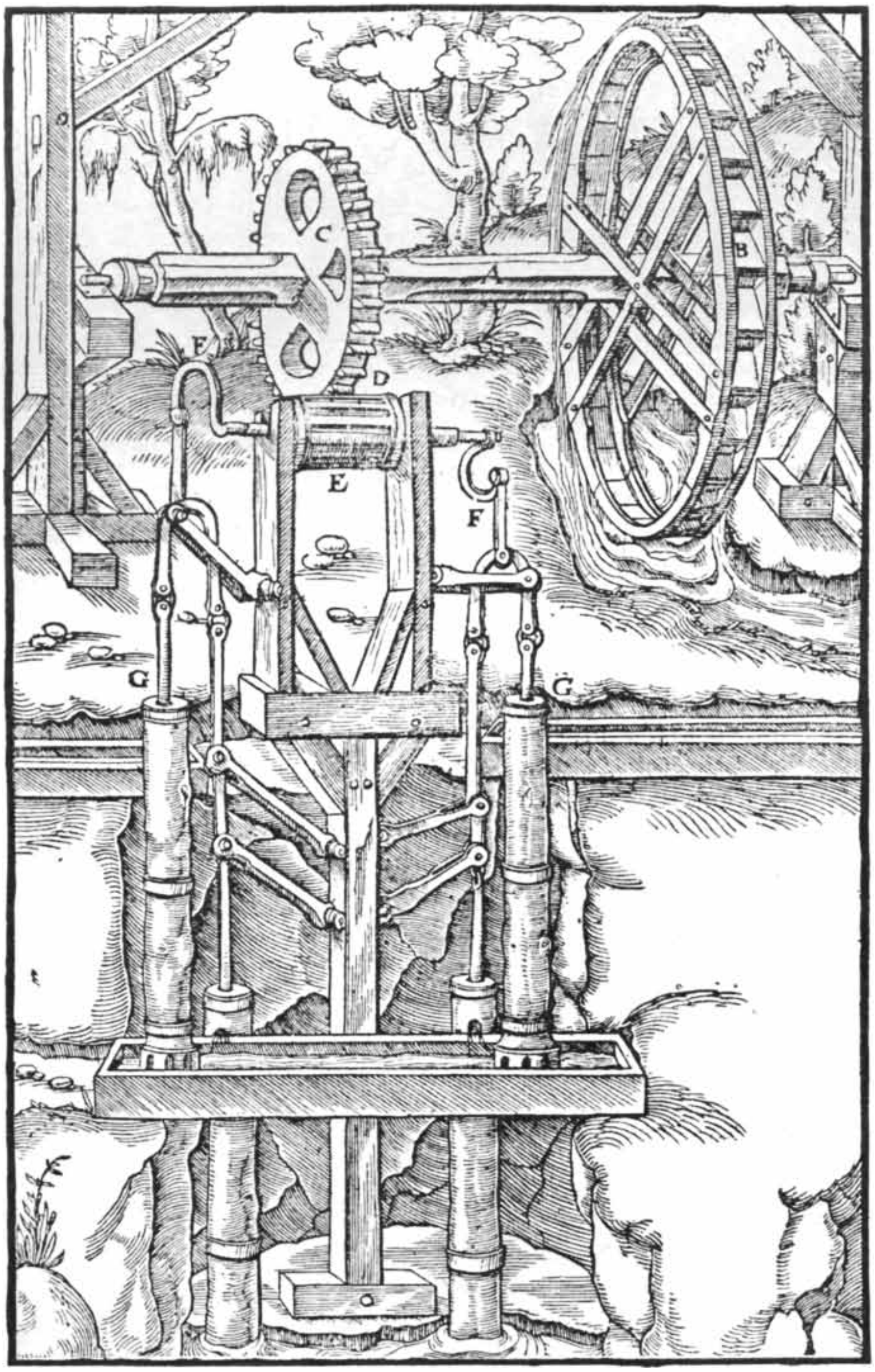


TRANSFORMATION OF ROTARY MOTION into linear motion can be achieved by having a cam on the axle of the wheel. The rotating cam engages a matching cam on a stamp shaft; as the shaft turns, the stamp is first raised and then dropped to deliver a powerful impact.

CAM PRINCIPLE WAS APPLIED in a rock-crushing mill illustrated in *De re metallica*. Cams on the axle engage and then release a tappet (*g*) attached to each of the iron-shod stamps.



ROTARY MOTION WAS ALSO TRANSFORMED into linear motion by the crank. This one requires a flexible joint between the shaft being raised and lowered and the driven shaft.



PAIR OF CRANKS driven by an overshot wheel rotating a two-gear train are converting rotary motion into linear motion in this illustration from *De re metallica*. The linear motion is being conveyed to pistons of two pairs of mine pumps. The lower pump of each pair lifts water from the mine shaft to a trough in the foreground for further raising by the upper pumps.

have been used instead of modified millstones in the beer mills of the ninth century, but the first industries to definitely take up hydraulic hammers were the fulling and hemp industries of the 10th and 11th centuries. After wool has been woven into cloth it must be pounded or pummeled in a cleansing solution. This action serves three purposes. First, it washes the cloth and removes much of the remaining sheep's grease from it. Second, it shrinks the wool so that it can be safely sewn to size afterward. Third, it felts the wool fibers, strengthening the weave.

In antiquity and on into early medieval times fulling had been done by hand. The water wheel and the cam-actuated trip hammer mechanized the process beginning in the 11th century. By the 13th century mechanized fulling was done over much of western Europe. In England, for example, the earliest water-powered fulling mill recorded dates from 1185. By 1327 there were 130 such mills, and before the end of the century the English woolens industry had shifted almost entirely to sites where water power was available.

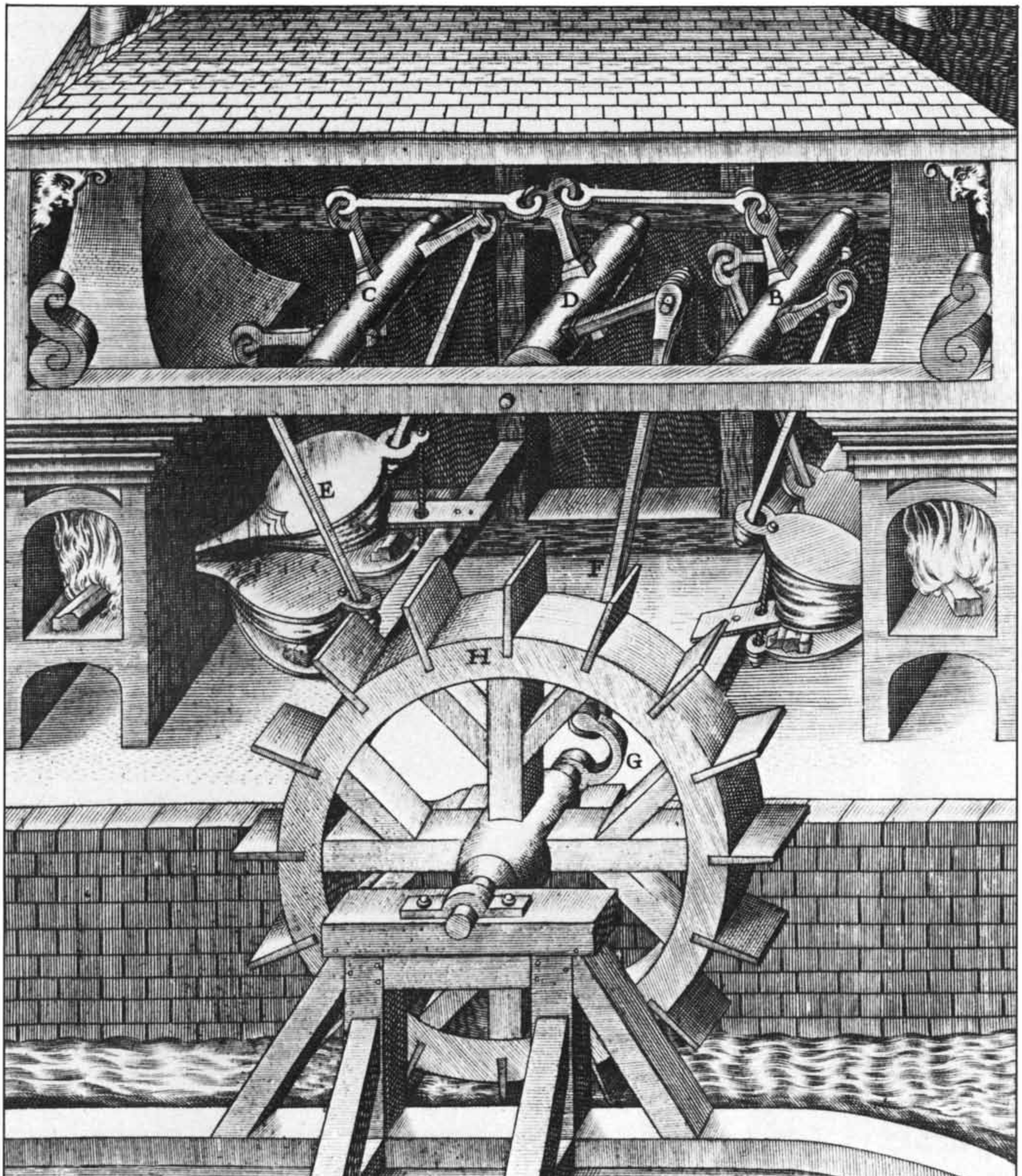
The hemp industry was also among the first industries to adopt the mechanized trip hammer. Traditionally hemp fibers, which are formed into rope and cord, had been separated from the woody plant tissue by being pounded and picked by hand. Water-powered hammers were substituted for this manual labor in Alpine France late in the 10th century and early in the 11th. By the 12th century there were water-powered hemp mills throughout France.

As time passed many other tasks were taken over by water-powered hammers. Ever since paper had been invented in Asia the pulp for making it had been produced by manually pounding rags in water. Western Europeans learned how to make paper only at the beginning of the 12th century. By the late 13th century European papermakers had taken a step that Chinese and Arabic ones never had: they substituted the water-powered trip hammer for the manual hammering. By the early 17th century England alone had 38 water-powered paper mills. By 1710 the number was 200 and by 1763 it was 350.

One of the most important European industries to be partially mechanized by the combination of the vertical water wheel and the cam was the iron industry. Early in the Middle Ages, European ironmasters smelted ore in a small furnace with air supplied to the burning mixture of charcoal and ore by a hand- or foot-powered bellows. The process did not yield temperatures high enough to liquefy iron. Thus almost daily the ironmaster had to shut his furnace down and take it apart to remove the sponge-like mass known as bloom, consisting of a porous mixture of metallic iron and slag. In order to get a usable form of iron the ironworkers had to heat and hammer the bloom repeatedly, with each cycle further consolidating the iron and eliminating the slag. The bloom, like the ore, was heated in a furnace with a draft that was supplied by a manually operated bellows.

The introduction of water power significantly affected both processes. Water-powered hammers in iron forges may have appeared as early as the 11th century; they were certainly in service by the 13th century, and they were commonplace in the 14th. Cam-actuated bellows, powered by water wheels, were operating in forges by early in the 13th century and were common in the 14th.

By late in the 14th century the combi-

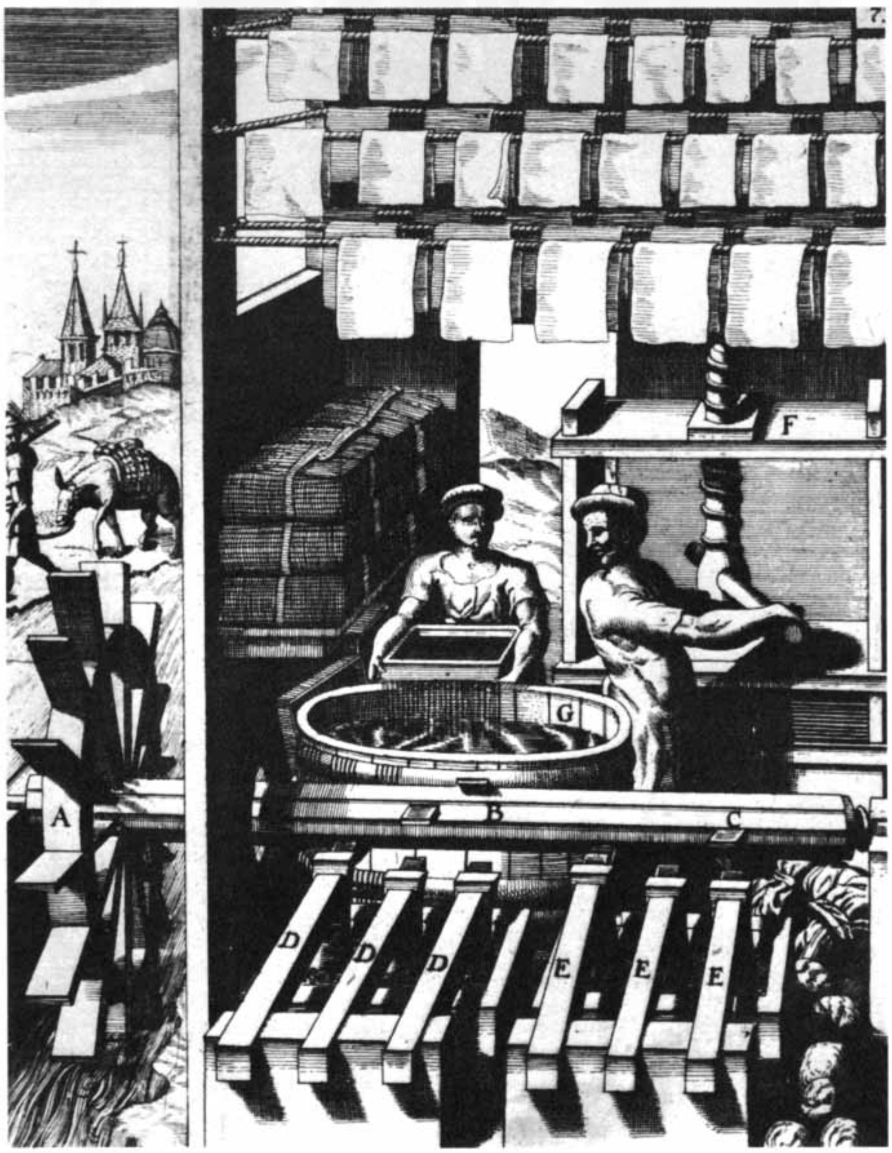


SIXTEENTH-CENTURY SMITHY had bellows driven by an undershot wheel to achieve high furnace temperature. Again a crank converted rotation of the shaft into reciprocating motion. The plate is in Agostino Ramelli's *Le diverse et artificiose machine,* dated 1588.

nation of a vertical water wheel and a cam-actuated bellows had moved from the forge to the smelting furnace and iron production underwent an even more radical change. Larger and more powerful bellows, made possible by water power, enabled ironmasters to get higher temperatures in their smelters, so that the iron was liquefied. The bottom of the smelting furnace could be tapped, and the liquid metal would run out to solidify in "pigs." Shutting down and taking apart the furnace to get the bloom came to an end, and iron production was transformed from a batch process into at least a semicontinuous one, with significant reductions in labor requirements. This new application of water power spread rapidly. For example, in the Siegen area of Germany all 38 pig-iron works and steelmaking forges were relying on water power by 1492.

Still other industries were transformed by the combination of the water wheel and the cam. In water-powered sawmills the cam served to pull down a ripsaw; the saw was pulled back up by a spring pole. Mills of this kind are first mentioned in documents of the early 13th century. Apparently the practice spread rapidly. In 1304 deforestation in the area of Vizille in southeastern France was being blamed in part on the proliferation of water-powered sawmills. Two centuries later the water wheel and the cam were being employed to lift hammers for the crushing of ore and to operate piston pumps for mine drainage.

The alternative to the cam for converting rotary motion into linear motion was the crank (or crankshaft). Known in China by the second century A.D., the crank appeared in Europe somewhat later. Cranks may have been used to rotate millstones by hand late in Classical times; they appear in mature form in Europe only in the ninth century, when the Utrecht Psalter depicts a crank attached to a hand-powered grindstone. Late in the medieval period the crank was combined with the water wheel and began to replace the cam for certain tasks. The double action provided by the crank was more advantageous than the single action of the cam in water-powered pumps, sawmills and bellows. The crank was also combined with manually operated grippers and drawplates in wire-drawing mills, which emerged in the 14th or 15th century.



PAPER MILL relied on an undershot wheel with cams (*c*) to lift and drop hammers (*d, e*) that reduced rags to pulp. Thereafter the process was manual: the pulp was transferred to a vat (*g*), dipped out with a seive and formed into a sheet in a press (*f*). The sheets were hung on racks to dry. The plate is from Georg Andreas Böckler's *Theatrum machinarum novum* of 1662.

By the 16th century at least 40 different industrial processes in Europe had come to depend on water power. The trend continued in the ensuing centuries. As one example, water power was first applied to spinning silk sometime between 1300 and 1600. In the silk mills water-powered spindles twisted individual silk fibers into thread. By 1700 there were 100 silk mills in northeastern Italy alone. The large silk mill erected early in the 1700's by Thomas Lombe at Derby in England, powered by the River Derwent, was a multistoried structure with a work force of 300.

Between 1550 and 1750 water power was applied to other processes as well. It was harnessed to bore the barrels of cannons and muskets, to thresh grain (with rotary flails), to agitate mixtures of ore and water and to pulverize the raw materials of glassmaking. Edge rollers were applied to such new activities as the preparation of snuff, cement, potter's clay and gunpowder. By 1692 there were 22 gunpowder mills in France, some on a scale matching the British textile mills of the late 18th and early 19th centuries. By the middle of the 18th century water-actuated hammers had been applied to the crushing of bone for fertilizer and of chalk for whitewash, and complex water-driven transmission systems, including cranks, had mechanized glass polishing.

Just as the water-powered trip hammer had mechanized the hemp industry by the 12th century, so it had penetrated the linen industry before the middle of the 18th century. The making of linen had been completely dependent on manual labor. After the flax plants for the linen had been harvested the stems were put in water to rot. The stems were then beaten to separate their fibers. When the fibers had been spun into thread and the thread had been woven into cloth, the cloth was washed and then beaten with light wood hammers to toughen the weave and give the cloth a sheen.

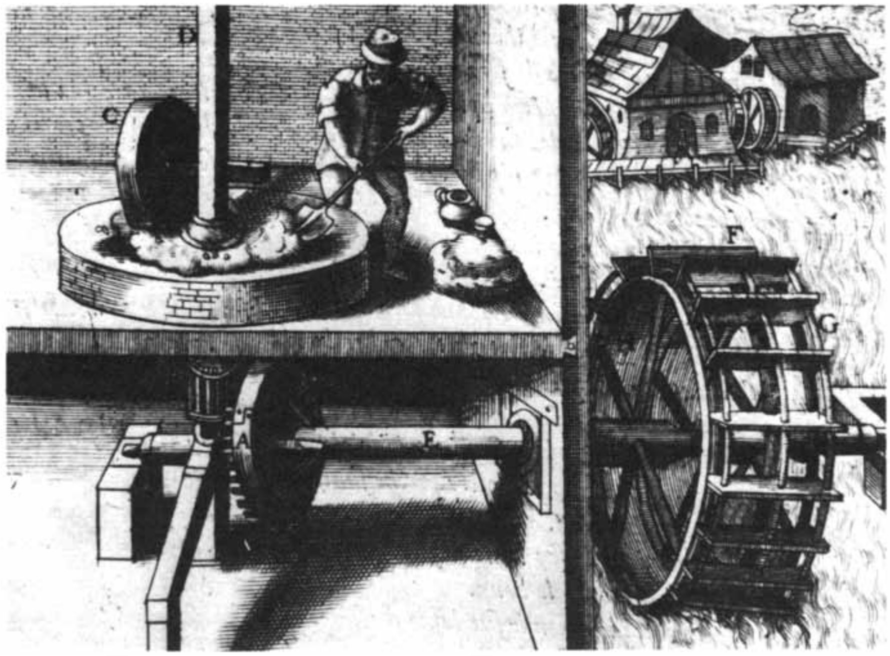
Late in the 17th century and early in the 18th European technicians mechanized several of these steps. Water-powered scutching mills relied on wood blades, rather like fan blades, to separate the fibers from the rotted flax stems. Linen-washing mills harnessed the combination of a water wheel and a cam or a crank to drive scrubbing boards of corrugated wood, through which the wet linen was drawn. Beetling mills then used water-actuated wood hammers to toughen and give a sheen to the linen cloth as it was drawn over large wood rollers. In Ulster alone more than 200 water-powered linen plants were established between 1700 and 1760.

The traditional beginnings of the Industrial Revolution in the late 18th century are dated to the early English cotton-textile mills. It is true that no cotton-textile production had been mechanized with water power before the 1770's. Nevertheless, water power not only had mechanized many nontextile industries but also had affected several processes in the production of textiles other than those of cotton. As we have seen, the fulling of wool, the spinning of silk and several stages of linen production had been taken over by water-powered machinery well before 1770.

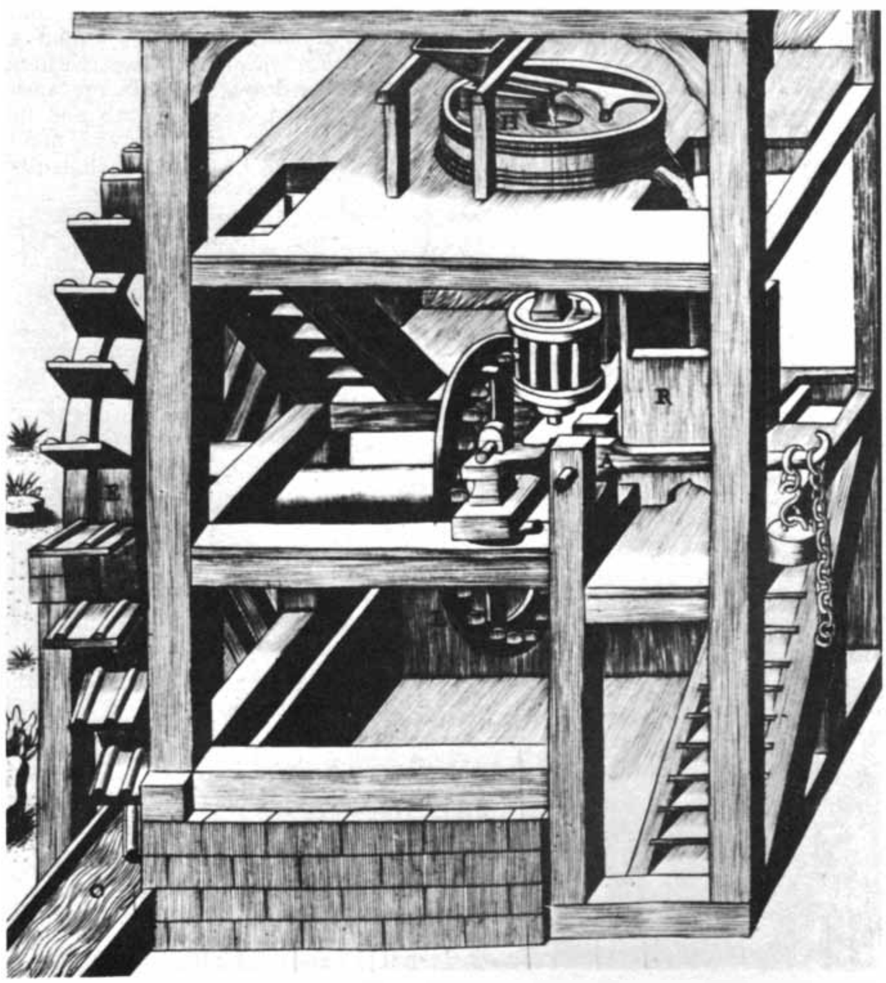
By the same token, the cotton-textile mills of the late 18th century were not unique in replacing manual labor with powered machinery. Neither were they unique in the amount of power they concentrated in a single location. Most of the early mechanized cotton mills had available perhaps 10 to 20 horsepower. The replacement in these mills of water power by steam power, beginning in the 1790's, made little difference at first because until well into the 19th century the average power of the steam engines was less than 20 horsepower. Even as late as 1835 the average mechanized cotton mill had available no more than about 35 horsepower. Concentrations of that much energy were not at all unknown in water-powered mills between the ninth century and the middle of the 18th.

Precise data on the power output of water wheels do not exist before 1700. In the period between 1700 and 1800, however, before the traditional wood wheel was replaced by iron wheels, water turbines and the steam engine, widely scattered sources provide enough information to allow an approximate calculation of the power output of some of the traditional wheels. I have collected data on 40 wheels, of both the undershot and the overshot type, from scattered millwright manuals, engineering texts, encyclopedias and other sources. They indicate that the average power output at the shaft was between five and seven horsepower. Thus the concentration of three or four wheels of average size at a single site would represent a power concentration roughly equal to that of the mechanized English cotton mills.

Concentrations of this kind, although



EDGE-ROLLER MILL differed from the mills that ground flour in that the top stone rolled over the bottom one instead of rubbing against it. Introduced in the 11th or 12th century, such mills were applied to crushing olives for the extraction of oil or pressing sugarcane for the extraction of sugar. Plate is from Vittorio Zonca's *Novo teatro di machine,* published in 1607.



LARGE UNDERSHOT WHEEL powers two gears that transform vertical rotation into horizontal rotation in this illustration of a 16th-century flour mill from Ramelli. The millstones are installed in the top story of the building; the flour falls into a bin below, beside the gears.

not commonplace, certainly existed. Monasteries provide some examples. For instance, as early as the ninth century the abbey of Corbie, near Amiens, had water mills with as many as six wheels. The monastery of Royaumont, near Paris, had a tunnel two and a half meters in diameter and 32 meters long in which there were mounted separate water wheels for grinding grain, tanning, fulling and working iron. In 1136 the abbey of Clairvaux, near Troyes, had wheels that ground grain, fulled cloth and tanned leather.

Similar concentrations of power existed elsewhere. In the 14th century the millers of Paris operated 13 water mills under the Grand Pont. Even earlier, late in the 12th century, the millers of Toulouse built three dams across the Garonne River; the largest of them, the Bazacle dam, was 400 meters long. These dams served 43 horizontal water mills. By the 13th century a division of capital and labor characteristic of the early British cotton mills had emerged at Toulouse: the mills were owned by investors and the millers were employees.

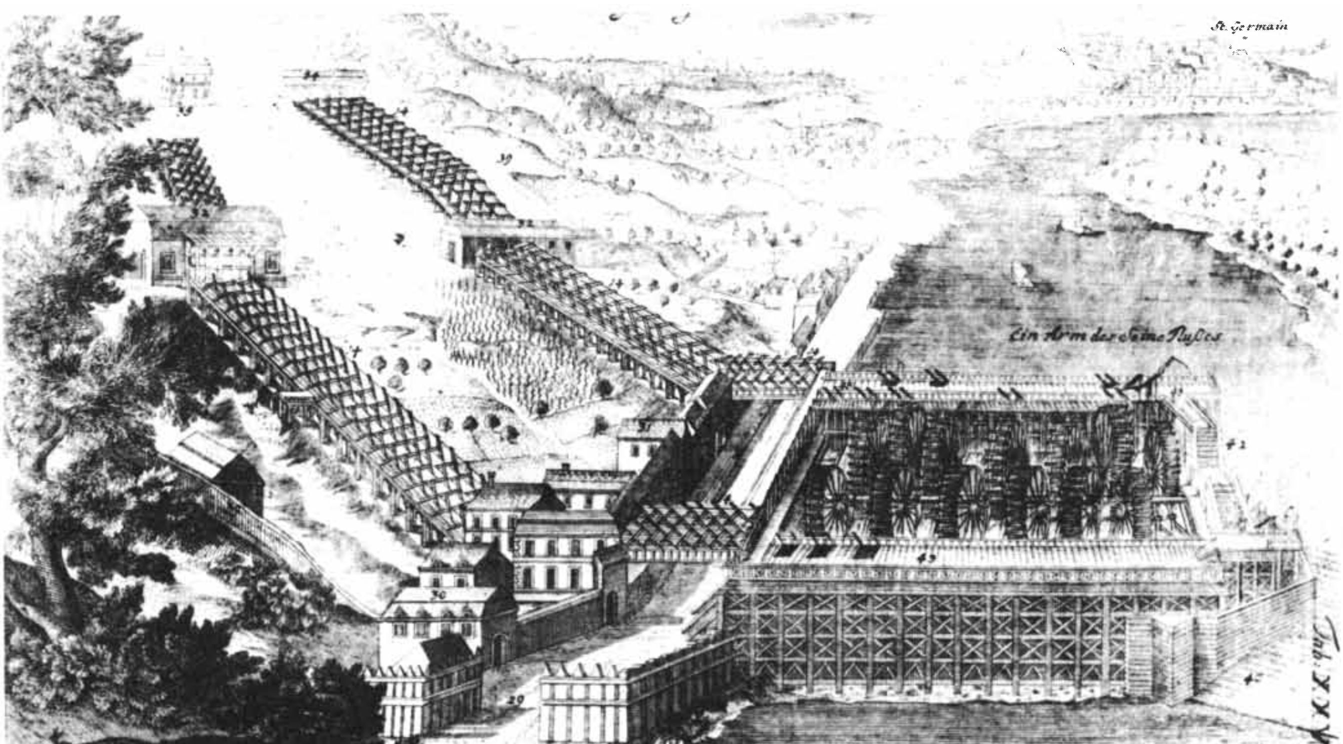
Elaborate water-power installations are also to be found in early modern Europe. For example, in the 1680's a Flemish engineer, Rennequin Sualem, designed and built for Louis XIV an elaborate water-powered complex at Marly-le-Roi on the Seine. There a dam diverted water to a set of 14 undershot wheels, each wheel 11 meters in diameter and 2.3 meters wide. The wheels drove 221 pumps at three levels by means of a complex linked set of cranks, rocking beams and connecting rods and lifted the river water 153 meters to an aqueduct a kilometer away. The transmission system was so cumbersome and inefficient, however, that the 14 wheels yielded only 150 horsepower.

Roughly contemporary with the Marly works was the Grand Rive mill, a paper plant in the Auvergne region with seven water wheels and 38 sets of hammers. In the 1720's Russian engineers built a large dam at Ekaterinburg in the Urals; an industrial complex powered by water from the dam consisted of 50 water wheels that drove 22 hammers, 107 bellows and 10 wire-drawing mills. By 1760 the British Royal Gunpowder Factory at Faversham in Kent included 11 water wheels. At about the same time engineers in Cornwall built what was known as the tower engine; it had 10 overshot wheels, mounted one above the other, linked by connecting rods to two large mine pumps.

Some pre-1800 power concentrations were regional rather than confined to a single site. For example, in the Harz Mountain region of Germany mining engineers began constructing a complex network of dams, reservoirs and canals to turn wheels that powered mine pumps, wire-drawing engines, ore-washing plants, ore-crushing mills and the bellows of furnaces and forges in about 1550. By 1800 this system included 60 dams and reservoirs, all within a four-kilometer radius of Clausthal, the center of the mining district. The largest dam in the system, the masonry Oderteich dam, built between 1714 and 1721, was 145 meters long, 18 meters high and 47 meters thick at the base. The dams eventually fed water to 225 wheels through a network of 190 kilometers of canals. The aggregate power of the system almost certainly exceeded 1,000 horsepower.

Water power also spread to the New World. Near Potosí in the Bolivian Andes, Spanish engineers exploiting rich silver deposits began in 1573 to build a system of dams, reservoirs and canals to bring water to ore-crushing mills. By 1621 the system included 32 dams. A main canal five kilometers long carried water to 132 ore-crushing mills near the city. The system generated more than 600 horsepower.

Hence it is clear that the mechanized cotton mills of England in the late 18th and early 19th centuries represented no radical break with the past, either in the replacement of manual labor with machines or in the concentration of large amounts of power. The substitution of water-powered machinery for manpower and the concentration of water-powered industry were trends well under way before that time. The British textile mills were simply the culmination of an evolutionary process that had its origins in medieval Europe and even in the Classical Mediterranean.



ARRAY OF 14 WHEELS on the Seine at Marly-le-Roi, 14 kilometers west of Paris, was erected in the 1680's. Developing from 300 to 500 horsepower at the shaft but only from 80 to 150 horsepower after losses in the pumps and the mechanical transmission are taken into account, the wheels pumped water to an aqueduct 153 meters above the river. It was carried to several of Louis XIV's palaces.

A MEMORABLE QUALITY.

1984 Mercury Topaz

Get it together—buckle up.

With the capacity to make a quarter of a million operations every second, Ford's EEC-IV can:

1. Save fuel, by making sure that your engine has the precise air-fuel mixture it needs for optimum performance. No more, no less.
2. Help your engine run better, by instantly measuring and adjusting for such external conditions as altitude, temperature, and engine load.
3. Get you off to a better start in the morning, by helping to prevent stall-outs and reduce warm-up time.
4. Help your service technician at maintenance intervals, by keeping an electronic "diary" of your engine's performance history.
5. Perform one final quality control check on itself before the car leaves the factory, to make sure everything is working together properly.

For obvious reasons, other car companies would like to offer an on-board computer with the memory, speed and accuracy of Ford's EEC-IV. But so far, there's no such animal.

And yet, EEC-IV is just a small part of our commitment to build the best cars and trucks we possibly can.

Come find out more about all our 1984 models, and get the details on the Lifetime Service Guarantee.**

We think a few minutes with your Ford or Lincoln-Mercury dealer will convince you of one very important thing: that Ford is, indeed, a car company to remember.

You're going to love the quality. Ford **Quality is Job 1.**

*EEC-IV not available with all engine/transmission combinations.
**Offered by participating dealers.

FORD • LINCOLN • MERCURY • FORD TRUCKS • FORD TRACTORS

THE AMATEUR SCIENTIST

People listening to a bell can perceive sounds the bell does not really make

by Jearl Walker

Some bells ring with musical quality; others only clang. What determines the quality of a bell and the frequencies it emits? Strange to relate, the note that is perceived by a listener may not be produced by the bell. This month I shall examine the acoustics of bells in some experiments designed by G. Theodore Wood, a colleague at Cleveland State University. We concentrated on a ship's bell but also examined several handbells lent to us by Katherine Marshall of Cleveland.

We were aided by three references. In the 19th century Lord Rayleigh examined the eight bells at his parish church in Terling, laying the groundwork for all later study of bells. Recently Stephen David Kelby and Robin Paul Middleton, then undergraduate students at the University of Birmingham, did experimental work on handbells. The source of the illusory note from a bell has been investigated by Arthur H. Benade of Case Western Reserve University.

The key to musical sounds of any kind lies in the study of acoustic resonance. When a bell is struck by its clapper or some other object, it is made to vibrate. What one hears from the bell are the sound waves generated by the vibration. Only certain of the vibrations generate sound waves loud enough to be heard. This special set of vibrations forms what are called resonance modes.

The study of resonance is usually introduced with a simple system in which a string is made taut between parallel supports. When the string is plucked, waves race from the point at which it is plucked to the support points, are reflected and travel back along the string. For the sake of simplicity let us assume that the waves have the sinusoidal shape shown in the illustration on page 134. Each wave has a length that is the distance along the string required for the wave pattern to repeat. It also has a frequency that is the number of times per second (designated in hertz) a crest passes through a point on the string.

Suppose two identical waves are sent along the string in opposite directions. Reaching the support points, they are reflected and pass through each other, maintaining their initial amplitudes. (I shall ignore their loss of energy as they pull on the string and beat against the air.) The shape of the vibrating string at any instant is established by the interference of the waves at that instant. For example, if the waves are in step, the string is pulled into a sinusoidal pattern. Where the waves have crests, the string has crests that are twice as high; where the waves have valleys, the string has valleys that are twice as deep. Such interference is said to be constructive. At other instants the crests of one wave are in the valleys of the other wave. Then the string is flat because any pull upward is exactly countered by a pull downward. Such interference is said to be destructive. At instants that are intermediate the shape of the string is established by some intermediate amount of interference.

I can control the type of waves I send along the string by controlling the frequency generated by the initial vibration of the string. For nearly all choices of frequency the interference of the waves leaves the string with virtually no vibration. For a special set of frequencies—the resonance frequencies—the interference gives rise to a repeating pattern on the string that beats vigorously enough against the air to be heard.

When I pluck the string, a great many waves are generated simultaneously. Most of them have the wrong frequencies, result in no repeated pattern and are not heard. The few waves with a resonance frequency force the string into vigorous, repeated vibrations.

The simplest of the resonance patterns, called the fundamental or the first harmonic, always results from a pluck. The ends of the string do not vibrate because they are tied in place. The center of the string vibrates the most. Points on the string displaying no vibration are called nodes; those displaying the maximum vibration are antinodes. Hence the pattern for the fundamental has an antinode at the center and a node at each end. To create this pattern the waves must have a length that is twice the length of the string. The interference at the center of the string varies smoothly between a large crest, a flat string and a deep valley. The vigorous beating against the air sends out sound waves that have the same frequency as the string waves.

Other patterns of interference yield higher frequencies of sound. The next simple pattern, the second harmonic, forms if the lengths of the string waves are equal to the string's length. This pattern has three nodes (at the end points and at the center) and two antinodes. Although waves are traveling over the string, they always interfere destructively at the center and therefore alternate between destructive and constructive interference at the antinodes. The frequency of the sound generated by the string is the same as that of the waves on the string, which is twice the frequency of the fundamental.

One can find other resonances for the string by considering shorter wavelengths (and higher frequencies) for the waves on the string. The values for the frequencies fall into the neat mathematical sequence known as a harmonic series: each frequency is an integral multiple of the frequency of the fundamental. For example, the third harmonic (three antinodes and four nodes) has a frequency that is three times the fundamental's. The fourth harmonic (four antinodes and five nodes) has a frequency that is four times the fundamental's. When a string is plucked, the fundamental and some of the higher harmonics are generated because the plucking generates string waves with the appropriate wavelengths. It also generates waves with inappropriate wavelengths, which fail to move the string significantly and therefore are not heard.

Usually only the lower harmonics are loud enough to be heard, but which of them are generated depends partly on where the string is plucked. For example, suppose I pluck a string at its midpoint. The fundamental is sure to be generated because it has an antinode there. The second harmonic cannot be generated, however, because it demands a node at the very place I force to vibrate with the pluck. Any other harmonic calling for a node at the center is also absent. The other harmonics will be generated, some (like the fundamental) strongly and others weakly.

My model of string resonance is an ideal one. With a real string the higher resonance frequencies are not exactly integral multiples of the fundamental frequency. The term harmonic is reserved for frequencies with an exact fit to the mathematical sequence. The term partial is employed to describe resonances that fail to fit the sequence. Therefore when a string is plucked, the fundamental and several partials are excited. Part of the richness of a stringed instrument can depend on how much the partials differ from the ideal harmonics.

The striking of a bell is similar to the plucking of a string under tension. Vibrational waves race throughout the bell, interfering with one another. Most of the waves have wavelengths that are inappropriate for creating any strong, repeating pattern of vibration. Those waves do not contribute to the sound of the bell. The few waves with a resonance length interfere to force the bell into repeating vibrational patterns. The bell, beating against the air, generates sound waves having the same frequencies as those vibrational waves.

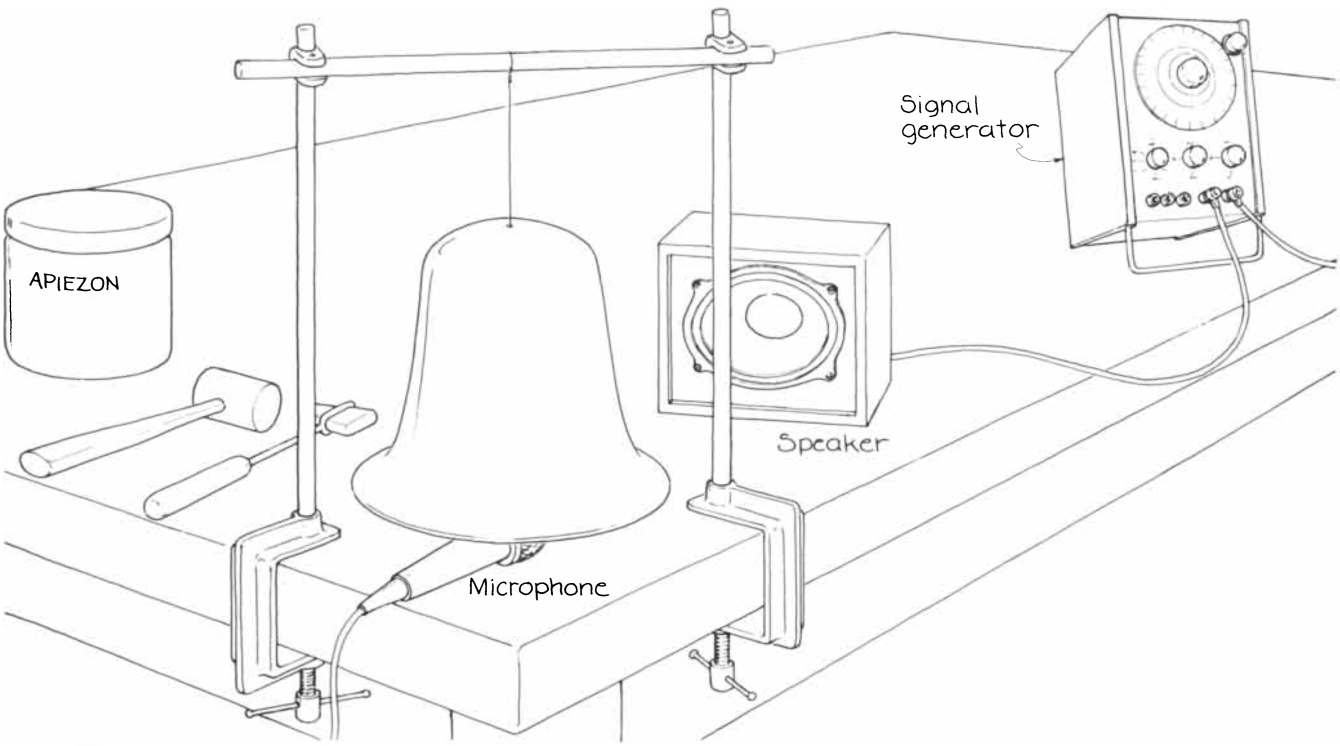
The resonance frequencies do not form a harmonic series and therefore are partials. Still, according to Benade and others, an observer might think the bell has emitted a harmonic series. If a listener is presented with a purely harmonic series from any generator of sound, he perceives only the fundamental frequency. In effect the brain examines the harmonics, picks out the fundamental and brings that frequency to consciousness. The assignment of frequency is easy when a full set of, say, the first five harmonics is present. Suppose, however, one of them is missing. The assignment is still made, even if the missing harmonic is the fundamental itself. Somehow the brain compares the sounds, realizes they are part of a harmonic series and then still perceives the fundamental.

Suppose the sounds presented to a listener do not form any harmonic series. Then the listener might perceive them in several ways. If the brain concludes that they approximately fit a harmonic series, the listener is forced to perceive the fundamental of the series even if that frequency is absent in the sounds. Instead the brain might utilize most of the sounds to achieve a best fit to a harmonic series and leave one or more of them separate. Then the listener might be forced to simultaneously perceive the absent fundamental and the partials left out of the best fit. Judging the frequency of the perceived fundamental introduces a problem. Commonly an error of one octave is made. For example, middle C (C_4) on a piano has a frequency of 261.6 hertz, but a listener might interpret the sound as being an octave higher, that is, C_5, which has twice the frequency of middle C.

Such analysis and interpretation apply to the sounds from a bell. The nonharmonic sounds generated by the bell may be forced by the brain into a best fit for a harmonic series. The perceived fundamental of that series may actually be absent from the sound. Moreover, the assignment of a frequency value to the perceived fundamental may be in error by an octave (or a factor of two). If some of the partials from the bell are not forced into a best fit, the listener might hear them along with the perceived fundamental.

A tuned bell is one designed to generate the frequencies and vibrational patterns displayed in the top illustration on page 135. The lowest frequency (the persistent one) is called the hum note. The next-higher frequency, called either the fundamental or the prime, is twice the hum note. (Since the term fundamental here can be confused with the term applied to string resonance, I shall call this frequency the prime.) The next several partials are referenced to the prime. The third note is a musical minor third up in frequency from the prime, that is, its frequency is 1.2 times the prime one. The next note is up from the prime frequency by a musical fifth (a factor of 1.5). The note called an octave is twice the prime frequency. The note called an upper third is 2.5 times the prime frequency. Also shown in the illustration are two other partials of a tuned bell.

When a tuned bell is rung, the hum note and several of the other partials are excited. Each partial has associated with it a certain vibrational pattern of nodes and antinodes. I shall label a line of nodal points (no vibration) according to whether it runs up and down the bell (a longitudinal line) or around it (a latitudinal line). The hum note has two longitudinal nodal lines along which the bell does not vibrate. (The pattern is also depicted in the illustration by a circle representing an overhead view of the bell.) Midway between the nodal lines are the antinodes of the vibration. Since striking the bell forces it to vibrate at that point, one of the antinodal lines for the hum note must pass through the same point. The ensuing waves of vibration racing around the bell set up the rest of the pattern. For example, nodal lines ap-



A laboratory bench set up for studying the resonances of a ship's bell pear at 45 degrees longitude on each side of the strike point.

The vibrational pattern for the prime frequency is similar except that it has a latitudinal nodal line. Hence when the prime is excited, there are two nodal lines running up and down the bell and there is also a nodal line running around it. The position of this last line depends on the distribution of mass in the bell. The top illustration on the opposite page shows the vibrational patterns for the other partials of the bell. In each case I have located the latitudinal nodal lines only approximately.

These patterns can be labeled according to the number of longitudinal and latitudinal nodal lines they have. For example, the pattern for the hum note is called a (2,0) mode to represent the two longitudinal nodal lines and the absence of any latitudinal nodal lines. Both the minor third and the fifth are labeled (3,1). Although these patterns are identical in their types of nodal lines, they differ in their latitudinal nodal lines and in the frequencies of the waves responsible for the patterns. A bell cannot vibrate in the (1,0) mode or any mode that lacks longitudinal nodal lines. Such a pattern would require that the circumference of the bell's mouth oscillate in size. The bell is too strong to allow such stretching and contraction of the rim.

The frequency relations I have described are true only for a tuned bell. Such a bell is musically pleasing because of those relations. When a listener hears the sounds from the bell, he might force them into a harmonic series consisting of the prime, the octave and the partials that are at three and four times the frequencies of the prime. Then the listener will be likely to perceive the prime as the fundamental of the series. The hum note and the other frequencies left out of the series might be perceived separately from the prime.

Alternatively the listener may unconsciously choose a harmonic series of the hum note and the partials that are integral multiples of its frequency. Then the bell will seem to ring at the hum note and perhaps also at the minor third (which is left out of the series). If the listener is asked to estimate the frequency of the bell, he might give the frequency of the hum note or make the octave error and give the prime frequency. Although what is heard and what frequency is assigned to the sound are the result of an unconscious process, predicting the outcome is difficult.

When a bell has not been tuned, prediction is even more difficult. Wood and I examined an apparently untuned ship's bell with a mouth 25 centimeters in diameter. It was a heavy brass bell suspended from a horizontal support that was clamped onto table stands. The clapper had been removed. To ring the bell we tapped it with hammers that had a head of hard plastic, hard rubber or steel. The softer blow with the rubber head generated mostly the lower frequencies; the harder blow with the steel head generated the higher frequencies as well. A microphone under the mouth of the bell was connected to a computer Wood had programmed to separate the frequencies in the sound and display them on a monitor along with their relative amplitudes. The computer could also produce a graph of the frequencies. The analysis was limited to frequencies lower than 5,000 hertz.

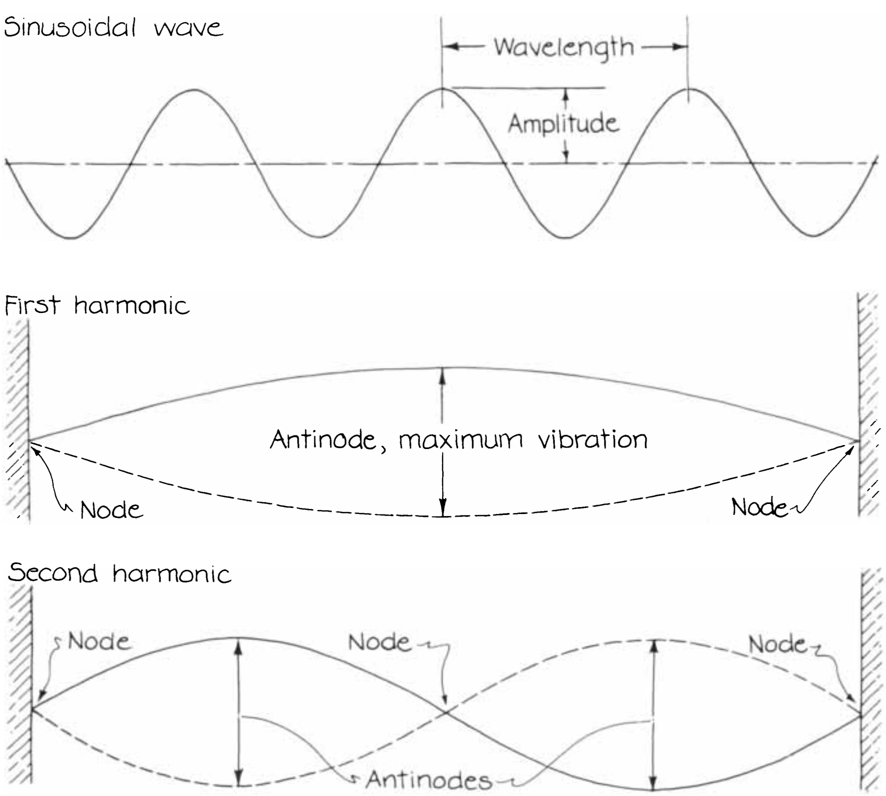
A tap about halfway up the bell produced a hum note with a distinct beat, or warble. The computer revealed that the hum note was split into two frequencies at 811 and 819.9 hertz. We were hearing those frequencies and their interference with each other. Periodically the two sounds arrived at the ear so as to reinforce each other. At intermediate times they arrived so as to cancel (or almost cancel) each other. We heard the true frequencies and also an extra sound (at the average of the two hum notes) that varied in loudness at a rate equal to the difference between the true frequencies. Thus the beat in the hum note was at about nine hertz.

Lord Rayleigh had explained such beats in the output of bells in terms of the distribution of mass around the bell. Suppose the bell is perfectly symmetrical in its mass distribution. Then striking the bell anywhere should excite the hum note. The strike point is automatically an antinode and the rest of the pattern is determined by its position. A different choice of strike point results in the same pattern but shifted so that again the strike point is at an antinode. Regardless of the strike position, the same frequency is generated for the hum note.

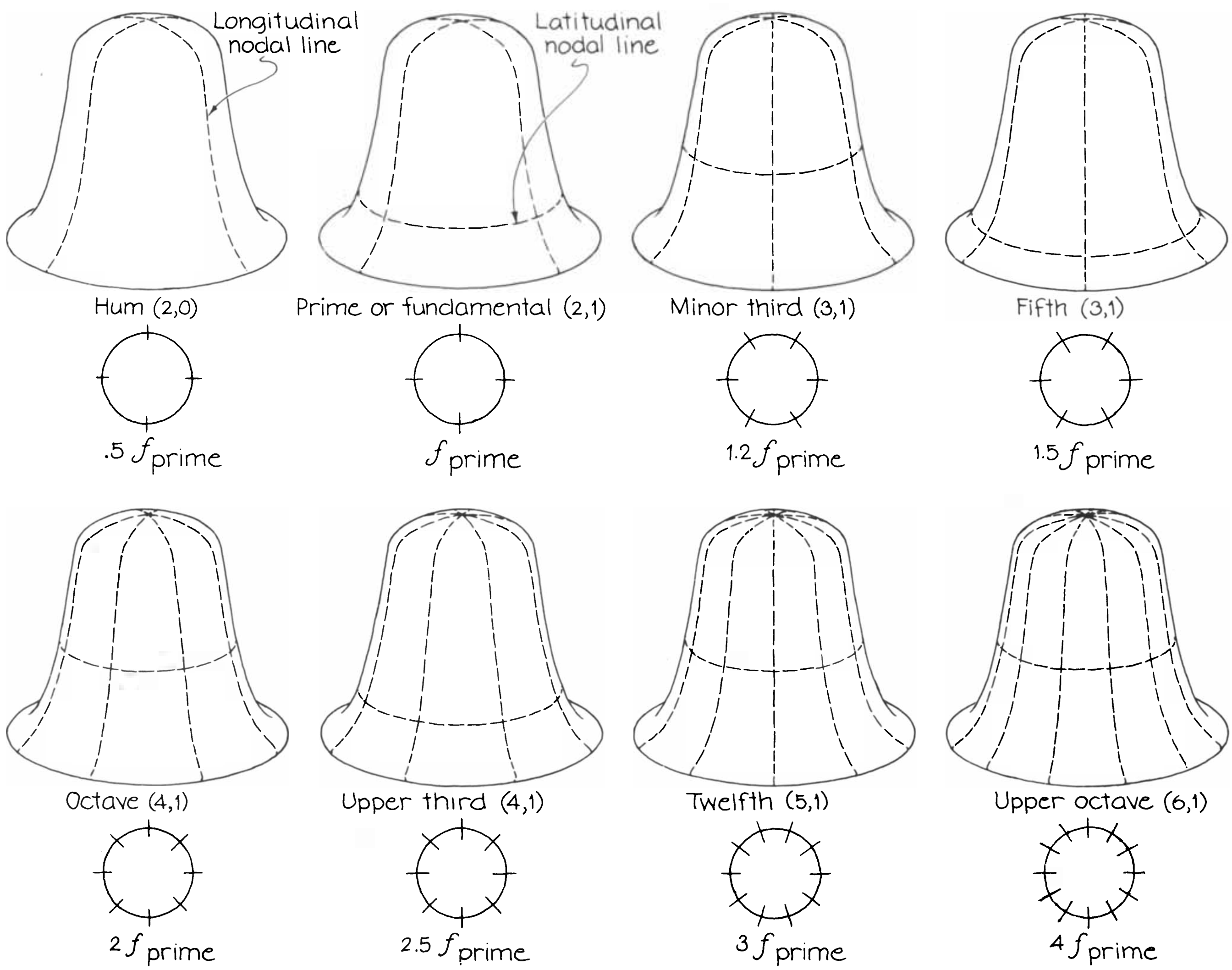
If the bell is asymmetric in its mass distribution, the symmetry of the vibrational pattern for the hum note is broken. There now can be two vibrational patterns for the hum note, one of them shifted in longitude by 45 degrees from the other. Hence the nodal lines for one pattern fall along the antinodal lines for the other. These patterns vibrate with different frequencies. Provided the frequencies are close enough, they can beat against each other when both patterns are excited.

On an asymmetric bell the nodal and antinodal lines are fixed in place. Thus a tap at one strike point on the bell might excite only one of the patterns. A tap at another strike point might excite the other pattern. Taps at still other points might excite both patterns. If only one pattern is excited, only its frequency is emitted. When both patterns are excited equally, both frequencies are emitted with equal loudness and the beats resulting from their interference are most noticeable.

Wood and I set out to find the nodal and antinodal lines for each hum-note pattern. Halfway up the bell we tapped point by point along a latitude line. At one point we heard only a single hum note and no beats. We marked this point as being on the antinodal line for the first



Waves on a string

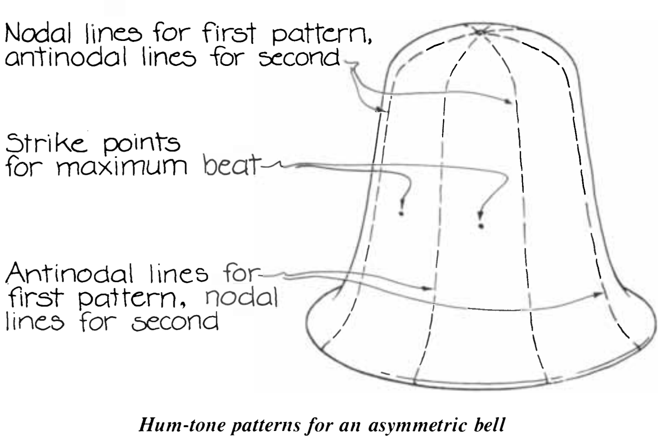


Nodal lines on a bell

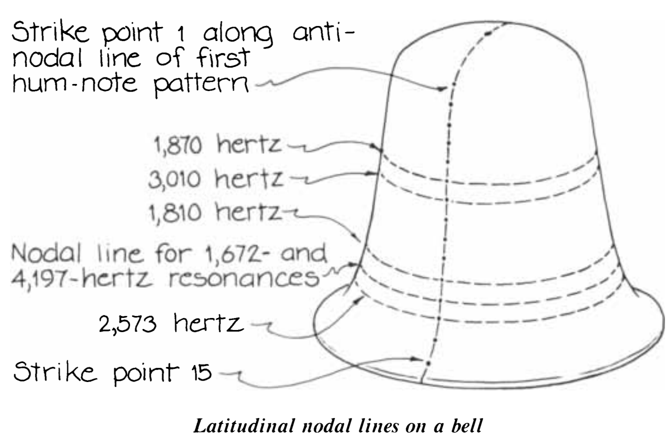
hum-note pattern. Since it was on the nodal line for the second pattern, only the first pattern was being excited. The computer determined the frequency to be 819.9 hertz.

As we explored farther along the latitude line beats developed, indicating we were exciting both patterns. Taps farther along led to less beating and finally a pure tone for the hum note again. This time the frequency was 811 hertz. A tap at this point on the bell excited only the second pattern because it was on an antinodal line for that pattern and on a nodal line for the other pattern. We continued around the bell, marking points of pure tone and maximum beating. We verified that the two patterns for the hum notes were shifted from each other by 45 degrees.

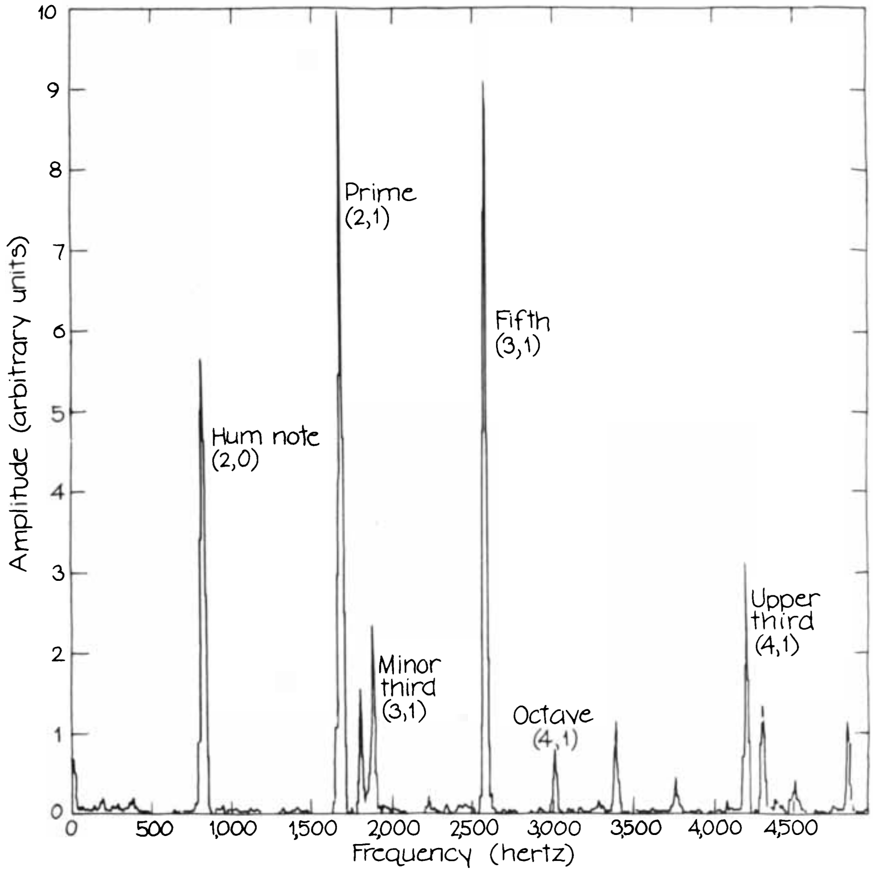
Wood and I reasoned we could alter the frequencies of the hum notes by adding mass to the bell. We stuck Apiezon sealing compound (modeling clay would serve the same purpose) on the rim of the bell at a point on an antinodal line of the second pattern. We figured that when we tapped to excite only the second pattern, the added mass would retard the vibration, decreasing its frequency. The frequency did fall from 811 hertz to 808.4. Since the mass was on a nodal line of the first pattern, it should



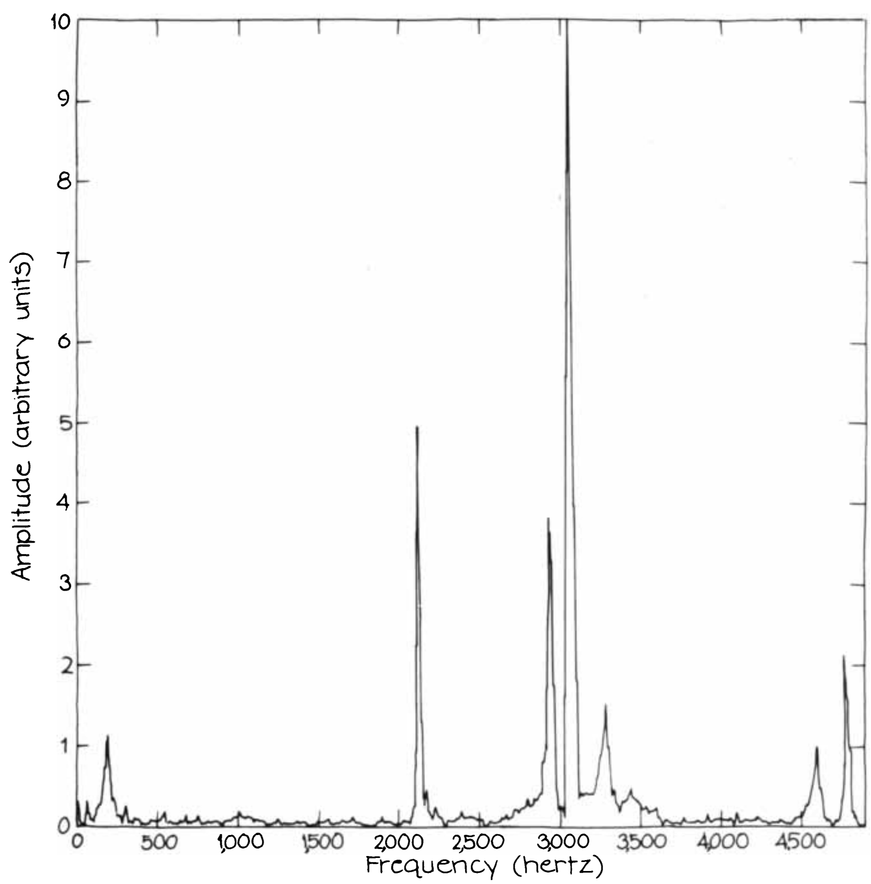
Hum-tone patterns for an asymmetric bell



Latitudinal nodal lines on a bell



Resonance frequencies of strike point 8



Resonance frequencies of a crystal bell

not alter the frequency of that pattern. Indeed, when we tapped at the point that would excite only the first pattern, the frequency was approximately the same.

We next removed the sealing compound and explored down a longitudinal line that was coincident with an antinodal line of the first pattern. Starting at the top, we laid out 15 strike points separated by half an inch. We were searching for latitudinal nodal lines of the higher-frequency partials. In particular we wondered if the sequence of frequencies and vibrational patterns fell into the same order as they would for a bell that had been tuned.

As we tapped our way down the side of the bell, the resonance frequency at 1,672 hertz remained strong except at the 12th strike point; there it apparently had a latitudinal nodal line. The resonance frequency at 1,810 hertz had a broad latitudinal nodal line throughout regions of the seventh through the 12th strike points, entirely disappearing at the 11th point. The resonance at 1,870 hertz displayed a wide latitudinal nodal line centered at the sixth strike point. The resonance at 2,573 hertz was strong at the upper strike points, increased to the loudest resonance at the 10th point and then almost disappeared near the rim of the bell. The 3,010-hertz resonance showed a narrow latitudinal nodal line at the seventh strike point. The 4,197-hertz resonance disappeared entirely at the 12th strike point.

With these clues we attempted to connect the resonance frequencies with the vibrational patterns. The hum notes were already established at 811 and 820 hertz. We figured that the prime frequency was at 1,672 hertz, for two reasons. That frequency is approximately twice the hum-note frequency, which would be true for a tuned bell. Moreover, that frequency has a latitudinal node through the 10th strike point, an appropriate position for the prime.

The other resonance frequencies were harder to identify, and none of them fell into the mathematical sequence of a tuned bell. There are two conditions for identifying such a pattern. First, the resonance frequency must be within a few hundred hertz of the predicted (tuned) values, as referenced to the prime. Second, the pattern must have an appropriate latitudinal node. With these conditions the patterns for the resonances at 1,870, 2,573, 3,010 and 4,197 hertz resemble the patterns called respectively the minor third, the fifth, the octave and the upper third.

Further clues could be found by tapping around a latitudinal line and counting the number of nodes and antinodes for each resonance frequency. The asymmetry in the mass of the bell should split each resonance frequency, but the resolution of the computer analysis was apparently too coarse to find the splits. Nevertheless, the asymmetry fixes in place the nodal patterns for each resonance, so that one can search for the patterns along a latitude line as we did for the hum note. We did not search for these extra clues, being content to find the bell was substantially untuned.

Wood attempted to identify the overall note of the bell in two ways. In one experiment he matched the note of the bell with an electrically generated musical scale. The best match was slightly above G_5. (The piano-key note is at 784 hertz.) Next he matched the bell with the output of a signal generator. Here the best match was at 819 hertz. Apparently what he (and I) heard from the bell was close to its hum note. Perhaps the resonances at 820, 1,672, 2,573, 3,010 and 4,197 hertz fit closely enough to a harmonic series for us to make the identification. If that is the case, the fundamental of the series would be near the hum-note frequency. The higher frequencies might disappear so quickly, however, that one can only hear and recognize the hum note.

Wood and I also examined the frequencies (up to 5,000 hertz) emitted by four of Marshall's handbells. None of the bells was thought to be tuned, although their hum notes were noticeable. One bell, a lead-glass one blown in Portugal, had a pleasing sound but a perplexing number of frequencies. The second bell was a Swiss cowbell that was quite unmusical. The third was a sterling-silver bell with a great many frequencies. The last was a Buddhist prayer bell with high frequencies. Although we could determine the hum-note frequency of this bell, we failed to identify the higher frequencies with any vibrational pattern.

If you would like to examine the resonance of bells, I suggest you use a home computer to analyze the frequency spectrum. Programs for such an analysis are sold for many popular makes of computer. A more difficult approach would be the one taken by the makers of bells. They search for resonance frequencies by listening for beats between the sound of the bell and a tone they produce either by singing or by playing a tuned instrument.

The task is easier if you slowly sweep a signal generator through a range of frequencies. Listen to the sound from a speaker connected to the generator as you are ringing the bell. When the generator's frequency is near a bell frequency, you can hear beats. You can then eliminate some of the resonance frequencies by striking the bell along its latitudinal nodal lines. If the bell is asymmetric in its mass distribution, you can also eliminate some of the frequencies by striking along longitudinal nodal lines. In this way you can unravel what frequencies and what vibrational patterns are characteristic of the bell.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

WIN AT CHECKERS. Millard Hopper. Dover Publications, Inc., 1956.

EXPERIMENTS IN ARTIFICIAL INTELLIGENCE FOR SMALL COMPUTERS. John Krutch. Howard W. Sams & Co., Inc., 1981.

COMPUTER GAMESMANSHIP. David Levy. Century Publishing, 1983.

THE HOMELESSNESS PROBLEM

THE DISAFFILIATED MAN: ESSAYS AND BIBLIOGRAPHY ON SKID ROW, VAGRANCY, AND OUTSIDERS. Howard M. Bahr. University of Toronto Press, 1970.

PRIVATE LIVES/PUBLIC SPACES: HOMELESS ADULTS ON THE STREETS OF NEW YORK CITY. Ellen Baxter and Kim Hopper. Community Service Society, Institute for Social Welfare Research, 1981.

MORE THAN SHELTER: A COMMUNITY RESPONSE TO HOMELESSNESS. A. S. Lauriat and P. McGerigle. United Community Planning Corporation and the Massachusetts Association for Mental Health, 1983.

THE HOMELESS MENTALLY ILL. American Psychiatric Association Task Force Report, H. Richard Lamb, Chairperson. American Psychiatric Association, in press.

UNDERSEA VOLCANOES

THE GALÁPAGOS RIFT AT 86° W: 3. SHEET FLOWS, COLLAPSE PITS AND LAVA LAKES OF THE RIFT VALLEY. Robert D. Ballard, Robin T. Holcomb and Tjeerd H. van Andel in *Journal of Geophysical Research,* Vol. 84, No. B10, pages 5407–5422; September 10, 1979.

BIRTH OF AN OCEAN: THE CREST OF THE EAST PACIFIC RISE. Jean Francheteau, Thierry Juteau, David Needham and Claude Rangin. Centre National pour l'Exploitation des Océans, 1980.

PETROLOGY OF THE OCEAN FLOOR. R. Hékinian. Elsevier Scientific Publishing Company, 1982.

ABUNDANCES, DISTRIBUTION AND SIZE OF VOLCANOES IN THE PACIFIC OCEAN AND IMPLICATIONS FOR THE ORIGIN OF NON-HOTSPOT VOLCANOES. Rodey Batiza in *Earth and Planetary Science Letters,* Vol. 60, No. 2, pages 195–206; September, 1982.

THE DISTRIBUTION OF INTRAPLATE VOLCANISM IN THE PACIFIC OCEAN BASIN: A SPECTRAL APPROACH. N. M. Ribe and A. B. Watts in *Geophysical Journal of the Royal Astronomical Society,* Vol. 71, No. 2, pages 333–362; November, 1982.

IMMUNOTOXINS

ANTIBODY-DIRECTED CYTOTOXIC AGENTS: USE OF MONOCLONAL ANTIBODY TO DIRECT THE ACTION OF TOXIN A CHAINS TO COLORECTAL CARCINOMA CELLS. D. Gary Gilliland, Zenon Steplewski, R. John Collier, Kenneth F. Mitchell, Tony H. Chang and Hilary Koprowski in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 77, No. 8, pages 4539–4543; August, 1980.

ANTIBODY CARRIERS OF DRUGS AND TOXINS IN TUMOR THERAPY. *Immunological Reviews,* Vol. 62; 1982.

CHIMERIC TOXINS. Sjur Olsnes and Alexander Pihl in *Pharmacology and Therapeutics,* Vol. 15, No. 3, pages 355–381; 1982.

NUCLEOTIDE SEQUENCE OF THE STRUCTURAL GENE FOR DIPHTHERIA TOXIN CARRIED BY CORYNEBACTERIOPHAGE β. Lawrence Greenfield, Michael J. Bjorn, Glenn Horn, Darlene Fong, Gregory A. Buck, R. John Collier and Donald A. Kaplan in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 80, No. 22, pages 6853–6857; November, 1983.

FORM AND FUNCTION IN FISH SWIMMING

MATHEMATICAL BIOFLUID DYNAMICS. M. J. Lighthill. Society for Industrial and Applied Mathematics, 1975.

A COMPARATIVE MECHANOPHYSIOLOGICAL STUDY OF FISH LOCOMOTION WITH IMPLICATIONS FOR TUNA-LIKE SWIMMING MODE. T. Y. Wu and G. T. Yates in *The Physiological Ecology of the Tunas,* edited by Gary D. Sharp and Andrew E. Dizon. Academic Press, 1978.

FISH BIOMECHANICS. Paul W. Webb and Daniel Weihs. Praeger Publishers, 1983.

SYMBIOTIC STARS

SYMBIOTIC STARS—A BINARY MODEL WITH SUPER-CRITICAL ACCRETION. G. T. Bath in *Monthly Notices of the Royal Astronomical Society,* Vol. 178, No. 1, pages 203–217; January, 1977.

IUE OBSERVATIONS OF SYMBIOTIC STARS. M. Hack in *Advances in Ultraviolet Astronomy—Four Years of IUE Research,* edited by Y. Kondo, J. M. Mead and R. D. Chapman. NASA Conference Proceedings No. 2238, 1982.

THE NATURE OF SYMBIOTIC STARS: PROCEEDINGS OF THE IAU COLLOQUIUM NO. 70, HELD AT THE OBSERVATOIRE DE HAUTE PROVENCE. Edited by Michael Friedjung and Roberto Viotti. D. Reidel Publishing Co., 1982.

THE PECULIAR VARIABLE STAR R AQUARII AND ITS JET. Minas Kafatos and Andrew G. Michalitsianos in *Nature,* Vol. 298, No. 5874, pages 540–542; August 5, 1982.

MULTILINGUAL WORD PROCESSING

WRITING SYSTEMS OF THE WORLD. Akira Nakanishi. Charles E. Tuttle, 1980.

A HISTORICAL STUDY OF TYPEWRITERS AND TYPING METHODS: FROM THE POSITION OF PLANNING JAPANESE PARALLELS. Hisao Yamada in *Journal of Information Processing,* Vol. 2, No. 4, pages 175–202; February, 1980.

THE MATHEMATICS OF THREE-DIMENSIONAL MANIFOLDS

THREE-DIMENSIONAL MANIFOLDS, KLEINIAN GROUPS AND HYPERBOLIC GEOMETRY. William P. Thurston in *Bulletin (New Series) of the American Mathematical Society,* Vol. 6, No. 3, pages 357–381; May, 1982.

THE GEOMETRIES OF 3-MANIFOLDS. Peter Scott in *The Bulletin of the London Mathematical Society,* Vol. 15, Part 5, No. 56, pages 401–487; September, 1983.

MEDIEVAL ROOTS OF THE INDUSTRIAL REVOLUTION

LE MOULIN A EAU: UNE REVOLUTION TECHNIQUE MEDIEVALE. B. Gille in *Techniques et civilisations,* Vol. 3, pages 1–15; 1954.

MAN AND WATER: A HISTORY OF HYDROTECHNOLOGY. Norman Smith. Peter Davies, Ltd., 1976.

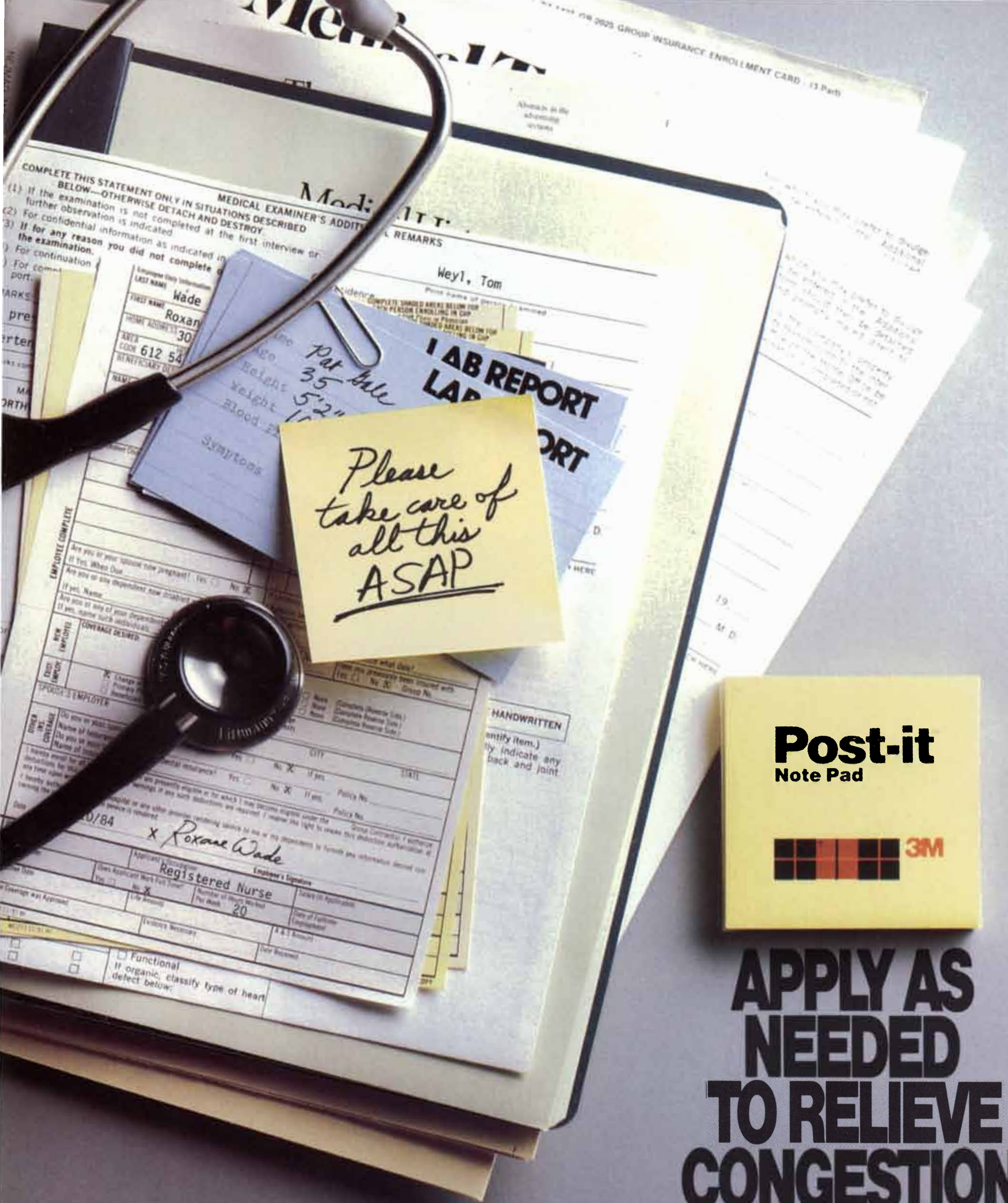
THE MEDIEVAL MACHINE: THE INDUSTRIAL REVOLUTION OF THE MIDDLE AGES. Jean Gimpel. Holt, Rinehart and Winston, 1976.

STRONGER THAN A HUNDRED MEN: A HISTORY OF THE VERTICAL WATER WHEEL. Terry S. Reynolds. The Johns Hopkins University Press, 1983.

THE AMATEUR SCIENTIST

VIBRATIONS OF PLATES. John William Strutt, Baron Rayleigh, in *The Theory of Sound: Vol. 1.* Dover Publications, Inc., 1945.

THE VIBRATIONS OF HAND BELLS. S. D. Kelby and R. P. Middleton in *Physics Education,* Vol. 15, pages 320–323; 1980.



APPLY AS NEEDED TO RELIEVE CONGESTION.

Post-it™ Notes quickly alleviate the symptomatic overcrowding in today's medical offices and hospitals. Our repositionable adhesive means they come off as easily as they go on. And, together with the bright color, makes certain your messages get noticed.

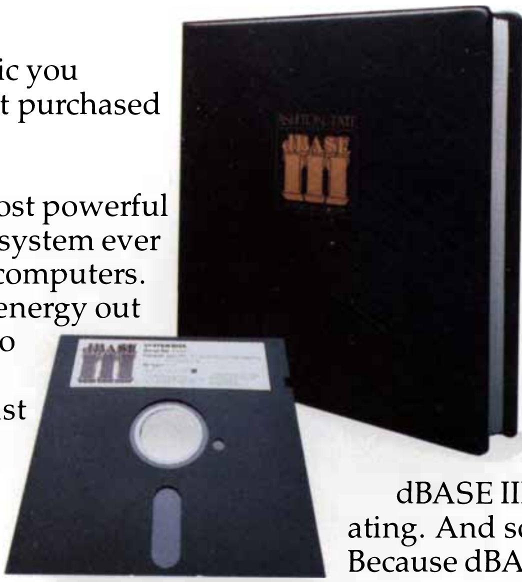
Call 1-800-328-1684 for a free sample. Then get more from your office manager or local stationer. With Post-it Notes, you'll give day to day operations a real shot in the arm.

Commercial Tape Division/3M "Post-it" is a registered trademark of 3M.

3M


dBASE III
TM
dBASE III
SYSTEM DISK
Serial No: 100001
Format: IBM PC, XT and 100% compatibles
PC-DOS 2.0
Version: 1.0
ASHTON·TATE

More power to you.



Remember the magic you expected when you first purchased a PC?

It's here.

dBASE III™ is the most powerful database management system ever created for 16-bit microcomputers. It pulls every ounce of energy out of your PC and puts it to work.

On top of that, it's fast and it's easy.

You've never seen anything like it.

With dBASE III, you'll be able to totally automate your business with real, full-blown applications.

You'll be able to handle accounting, inventory control, payroll, information management and a variety of other sophisticated applications in real estate, medicine, insurance, finances, investments and many other fields.

dBASEIII can handle over a billion records per file, limited only by your computer system. It can do repetitive tasks until you ask it to stop. It can be used in an automated mode while you're doing something else. It can even make choices for you and perform different tasks on different days, without being reminded.

Easier than ever.

dBASE III uses powerful yet simple commands that are the next best thing to speaking English.

If you're unsure of a command, HELP will tell you what to ask for. If you don't know what command comes next, a command assistant does. All you have to know is what you want it to do.

Our new tutorial/manual will have you entering and viewing data in minutes rather than reading for hours.

And to make matters easier, you get a full screen report set-up for simple information access.

Faster than no time at all.

dBASE III isn't just fast. It's ultra-fast. Operating. And sorting. Even faster, is no sorting. Because dBASE III keeps your records in order, so you really don't have to sort anything. Unless you want to. Then watch out!

What about dBASE II®?

It's still the world's best database management system for 8-bit computers. And it's still the industry standard for accounting, educational, scientific, financial, business and personal applications.

Tap into our power.

For the name of your nearest authorized dBASE III dealer, contact Ashton-Tate, 10150 West Jefferson Boulevard, Culver City, CA 90230. (800) 437-4329, ext. 333. In Colorado, (303) 799-4900.

ASHTON·TATE™

INTRODUCING THE 325e. BMW AFFIRMS ITS DOMINANCE IN THE CATEGORY IT CREATED.

Many years ago, into a world forced to choose between the sports car and the passenger sedan, BMW introduced an innovation that proved prophetic.

A way of merging the two into a whole greater than the sum of its parts.

Car and Driver, taking note of the achievement, wrote that "BMWs have reigned as the definitive sports sedans for nearly twenty years now. The world's car companies perennially take them apart to see what makes them tick."

Now there's a new candidate for such dissection—an effort that will prove exhilarating for driving enthusiasts and chastening for a world of late entrants into the sports sedan genre.

It's called the BMW 325e.

HIGH TECHNOLOGY DEDICATED TO HEIGHTENING YOUR PULSE RATE.

The 325e is a $20,970† paragon of high performance. Central to its prodigious performance is BMW's technologically ingenious 2.7-liter, 6-cylinder 'Eta' engine. It develops high torque at low to medium range engine speeds, thus offering exceptional response as you move through the gears.

This response is enhanced by a newly refined version of BMW's Digital Motor Electronics. A system so unerring it also manages the BMW engine that powers the current Formula One Grand Prix championship car.

In the BMW 325e the result is an aggressively smooth engine that delivers soul-stirring performance with a remarkable EPA-estimated [23] mpg, 36 highway.*

But as Road & Track states and BMW engineers concur, "the concept of performance encompasses a good deal more than the various aspects of acceleration."

It encompasses, for example, excellence in deceleration as well. For that reason the 325e is equipped with disc brakes on all four wheels. And they're vented in front to increase fade resistance.

The 325e has also been equipped with a newly engineered sport suspension that's fully independent with anti-roll bars at the front and rear for confidence-inspiring handling and a crisp yet supple ride.

A NEW INTERIOR ARCHITECTURE.

Inside the BMW 325e there are no sacrifices required in the name of high performance. Totally new BMW sport seats can be molded to your needs with a myriad of orthopedic adjustments.

While the instrument panel incorporates such informative BMW innovations as a Service Indicator that determines your individual driving style and recommends when routine services are due. And an onboard computer informs you of range on remaining fuel, average speed and outside temperature as well as other useful data.

The 325e exudes the same attentiveness to detail and quality that characterizes all BMWs and elicited this from Motor Trend: "doors close with a nice solid clunk, gear changes are crisp as cold celery and the steering as precise as a dial indicator."

In sum, it expands the prerequisites for all those cars seeking credibility as high-performance sports sedans. Every nuance of the 325e's performance has been heightened to elevate driving from a mere pastime into passion.

THE ULTIMATE DRIVING MACHINE.

†Manufacturer's suggested retail price. Actual prices established by dealers. Taxes, freight, optional equipment and other dealer charges are extra. Price is subject to change without notice.
*Fuel efficiency figures are for comparison only. Your actual mileage may vary, depending on speed, weather and trip length. Actual highway mileage will most likely be lower. © 1984 BMW of North America, Inc. The BMW trademark and logo are registered. European Delivery can be arranged through your authorized U.S. BMW dealer.