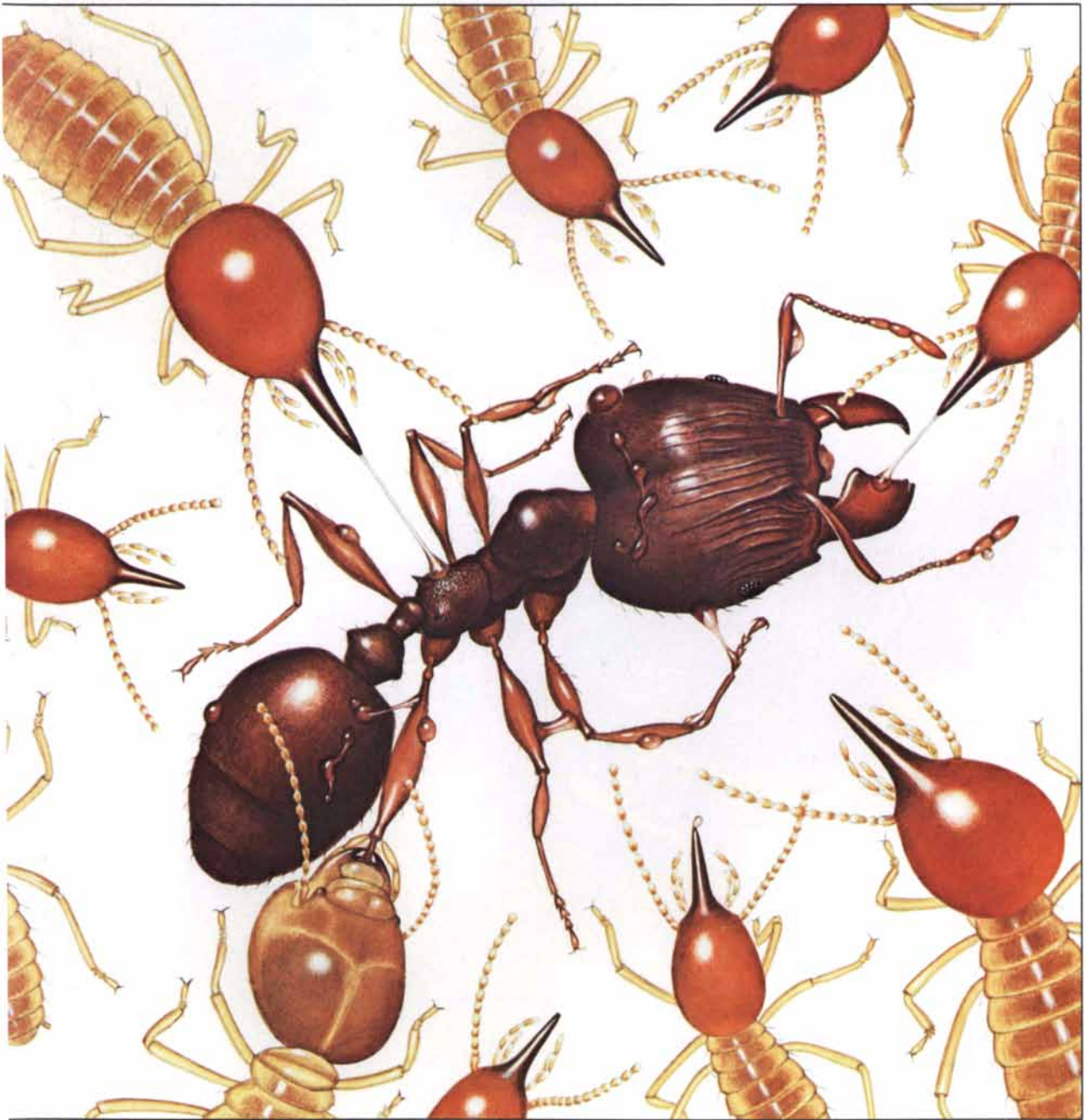


SCIENTIFIC AMERICAN



TERMITE CHEMICAL WEAPONS

\$2.50

August 1983



Why so many Mercedes-Benz automobiles have achieved classic resale value – before they were old enough to be classics.

Each of the Mercedes-Benz automobiles in the picture at left has been shown to be actually worth more money today than the day it was new.

What makes this fact astonishing is that each is a production model and far from rare. And each was built in 1971—little more than a decade ago.

Astonishing consistency

The astonishingly consistent Mercedes-Benz legend of retained value is thus enhanced. A legend composed not just of a few exotic models, so rare and so old that their value could be expected to rise higher as the years pass by, but models you can see on the streets and highways of America every day—such as those in the picture at left.

True, the most money ever paid for a production automobile was paid for a Mercedes-Benz—a 1936 500K Roadster, auctioned in 1979 for four hundred thousand dollars. And the experts can cite numerous other Mercedes-Benz models now worth double, triple, quadruple their original selling prices.

But these same experts can attest to the remarkable overall record compiled by *all* Mercedes-Benz automobiles—sedans as well as coupes and roadsters, diesels alongside

their gasoline-powered counterparts; from the recent past as well as the distant past. A record that is perhaps best expressed in one simple statement: the Mercedes-Benz name is so coveted by American buyers today that after the first three years, the *entire Mercedes-Benz line*—not just a few isolated models—has been shown to retain an average of 84 percent of original value.

Some individual models from some makers might possibly match this figure. In so doing, they only underscore the point: the Mercedes-Benz legend is not based on *some* cars, specially handpicked; it is based on the total resale performance of all models in the line.

Minimal depreciation

The net result has been a series of automobiles so desirable to so many people that their value has refused to tumble—refused, indeed, to more than minimally depreciate as the miles and years have gone by.

Perhaps this is because there have never been quite enough Mercedes-Benz automobiles to satisfy America's demand.

Perhaps it is because their value has never been cheapened

by annual model changes, or face-lifts, or marketing artifice of any kind.

Perhaps—as the engineers would claim—it is simply because they are built to uncommon standards, to serve their owners uncommonly well.

A crucial measure

Not even Mercedes-Benz knows the reason to an absolute certainty. And not even Mercedes-Benz can predict the future course of resale value in this uncertain, unpredictable world of ours.

But that resale value stands as a crucial measure of automotive worth is beyond question. And amid the welter of claims and counterclaims about value retention in the marketplace today, the lessons of Mercedes-Benz resale history—*recent* resale history—cannot be discounted by any automobile buyer. And they should not be ignored.



Engineered like no other car in the world



◀ Clockwise from top: 300 SEL 6.3 Sedan, 280 SL Roadster, 280 SE 3.5 Coupe, 280 SE 3.5 Convertible. 250 CE Coupe, 280 SE Sedan.

When it comes to superior performance, we study our lines very carefully.

Superior printer performance is not a fluke. It evolves from analyzing printed line after printed line. Taking the time to test and retest. After 30 years of manufacturing precision parts, we know that there are no shortcuts.

And so we took the Gemini-10X and methodically put it through its 120 cps pace. We achieved a print head life of over 100 million characters with an extremely precise dot alignment creating each crisp character.

So far so good.

Next, sophisticated performance demanded versatility. A wide choice of character sets, a buffer expandable to 8K, and the ability to interface with all popular personal computers. We added macro

instruction, giving Gemini-10X the capability to perform up to 16 operations with one command. We included as standard a paper feed system that has a friction and fully adjustable tractor feed. Then we even built in the dexterity to print graphics and text on the same line.

Done.

And, of course, staying the best means constant reviewing and fine-tuning. Keeping the Gemini easy to find, easy to afford and so reliable it can be warranted for up to twice as long as its major competitors.

Only the most careful engineering has built the new hard-working Gemini-10X. You'll applaud its performance.

stairTM
MICRONICS • INC

THE POWER BEHIND THE PRINTED WORD.

Computer Peripherals Division
2803 N.W. 12th Street, Dallas/Ft. Worth Airport, TX 75261



ARTICLES

- 28** **TRAUMA**, by **Donald D. Trunkey**
Accidental and intentional injuries in the U.S. take a heavier toll than cancer and heart disease.
- 36** **THE PURIFICATION AND MANUFACTURE OF HUMAN INTERFERONS**, by
Sidney Pestka Recombinant-DNA bacteria are now producing them in large quantities.
- 44** **MAGNETIC FIELDS IN THE COSMOS**, by **E. N. Parker**
The principle of the dynamo accounts well for the magnetic fields of planets, stars and galaxies.
- 66** **INTERSTELLAR MATTER IN METEORITES**, by **Roy S. Lewis and Edward Anders**
The meteorites known as carbonaceous chondrites contain material from outside the solar system.
- 78** **THE CHEMICAL DEFENSES OF TERMITES**, by **Glenn D. Prestwich**
Termite soldiers harass intruders with sophisticated irritants, toxins, anticoagulants and glues.
- 88** **RATIONAL COLLECTIVE CHOICE**, by **Douglas H. Blair and Robert A. Pollak**
Axiomatic analysis shows that no voting system is perfect, but it can help to determine the best.
- 96** **THE STAVE CHURCHES OF NORWAY**, by **Petter Aune, Ronald L. Sack and Arne
Selberg** Going back to the 10th century, they suggest that wood buildings can be permanent.
- 106** **DIGITAL TYPOGRAPHY**, by **Charles Bigelow and Donald Day**
Most type, including this type, is currently generated on the cathode-ray screen by computer.

DEPARTMENTS

- 6** LETTERS
- 8** 50 AND 100 YEARS AGO
- 9** THE AUTHORS
- 12** MATHEMATICAL GAMES
- 22** BOOKS
- 58** SCIENCE AND THE CITIZEN
- 120** THE AMATEUR SCIENTIST
- 128** BIBLIOGRAPHY

BOARD OF EDITORS	Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky
ART DEPARTMENT	Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell
PRODUCTION DEPARTMENT	Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Julio E. Xavier
COPY DEPARTMENT	Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson
GENERAL MANAGER	George S. Conn
ADVERTISING DIRECTOR	C. John Kirby
CIRCULATION MANAGER	William H. Yokel
SECRETARY	Arlene Wright

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of \$2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036-8733/83. \$2.00 + 0.00.

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

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

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

Name _____

New Address _____

Street _____

City _____

State and ZIP _____

Old Address _____

Street _____

City _____

State and ZIP _____



THE COVER

The painting on the cover shows a group of termites attempting to immobilize an intruding ant of the large-headed species *Pheidole megacephala*. One worker termite has grasped the intruder by its right rear leg, but the main defenders of the termite nest are members of the termite soldier caste. Soldiers of this termite species, *Trinervitermes bettonianus*, are of two sizes and defend the nest by squirting an intruder with an irritating glue-like secretion (see "The Chemical Defenses of Termites," by Glenn D. Prestwich, page 78). The intruder has already been hit by several gluey squirts. Most of the small soldiers are out of glue but the soldiers' defense has entangled two of the ant's legs.

THE ILLUSTRATIONS

Cover painting by Tom Prentiss

Page	Source	Page	Source
12-21	Ilil Arbel	87	Barbara L. Thorne, Museum of Comparative Zoology, Harvard University (<i>top</i>); Glenn D. Prestwich, State University of New York at Stony Brook (<i>bottom</i>)
29	Donald D. Trunkey, San Francisco General Hospital	88-94	Ilil Arbel
30-34	Albert E. Miller	97	Ronald L. Sack, University of Idaho
35	Allgemeine Deutsche Automobil Club (<i>left</i>), Albert E. Miller (<i>right</i>)	98	Mittet Foto A/S
36-37	Hoffmann-La Roche, Inc.	99-101	J. Dyck Fledderus
38-42	Bunji Tagawa	102	Ronald L. Sack, University of Idaho
45	High Altitude Observatory of the University of Colorado at Boulder	103	Ronald L. Sack, University of Idaho (<i>top</i>); J. Dyck Fledderus (<i>bottom</i>)
46-54	George V. Kelvin	104	J. Dyck Fledderus
67	Glenn MacPherson, University of Chicago	107	Autologic, Inc.
68-73	Walken Graphics	108-109	Kris Holmes and Edward Bell
74	Mitsuo Ohtsuki, University of Chicago	110	Allen Beechel (<i>top</i>); Charles Bigelow, Stanford University (<i>bottom</i>)
75-76	Walken Graphics	112-116	Allen Beechel
79	Barbara L. Thorne, Museum of Comparative Zoology, Harvard University	117	John E. Warnock, Xerox Corporation (<i>top</i>); Allen Beechel (<i>bottom</i>)
80	Glenn D. Prestwich, State University of New York at Stony Brook	118	Allen Beechel
81-85	Tom Prentiss	119	Peter Karow, URW Unternehmensberatung, Hamburg, and Donald E. Knuth, Stanford University, courtesy of <i>Visible Language</i>
86	Barbara L. Thorne, Museum of Comparative Zoology, Harvard University (<i>top</i>); Manfred Kaib, University of Bayreuth (<i>bottom</i>)	123-126	Michael Goodman

SCIENCE/SCOPE

An easily processed version of a heat-resistant plastic should find new high-temperature industrial and commercial applications, as well as promote more use of advanced composites in such aerospace products as aircraft, engines, and supersonic missiles. The new Hughes Aircraft Company polyimide, which withstands temperatures of 600°F for long periods and much higher temperatures for short periods, can be processed in existing equipment. It uses a simple one-step curing process very similar to state-of-the-art epoxies. By comparison, plastics with equivalent strength and heat resistance require complicated and expensive curing procedures. The new material will be produced and marketed under the trade name Thermid® by National Starch and Chemical Corp. of Bridgewater, N.J.

Two weather satellites are being readied to monitor the Western Pacific through the end of the decade. Under contract to Nippon Electric Company of Japan, Hughes will refurbish one Geostationary Meteorological Satellite (GMS) and build another. The GMS-2 protoflight spacecraft, in storage in Japan since 1981, will be updated and renamed GMS-3a. It is scheduled for launch in August 1984, and will replace GMS-2. The new spacecraft, GMS-3b, will serve as a back-up. The satellites will provide pictures every 30 minutes, and gather other data.

NASA's space shuttle gets off the ground with support from Hughes. Astronauts train for missions on a simulator that uses a Hughes liquid-crystal projector to show what they will see outside cockpit windows. The pictures are brighter and sharper than home projection TV because the projector contains an exclusive device that draws on some of the technology used in liquid-crystal watches. Other Hughes support includes: technicians who adjust and repair flight instruments, test equipment, and ground support equipment; an instrumentation amplifier carried by chase planes; and a unique radar and communications unit that soon will serve as the space shuttle's eyes, ears, and voice.

In what may be the world's biggest aerospace cost reduction program, Hughes and its customers, including the U.S. government, have saved \$1.8 billion during the past 25 years through the ideas and ingenuity of company employees. The savings were documented by the Hughes Cost Improvement Program, in which employees are encouraged to submit cost-reduction or cost-avoidance ideas on prepared forms. Last year 6,931 employees submitted ideas that saved over \$250 million. One novel suggestion was to replace old vacuum pumps, used to hold semiconductor wafers in place during testing, with inexpensive fish tank pumps modified to reverse their air flow. The annual savings was \$100 per pump.

Hughes Research Laboratories needs scientists for a whole spectrum of long-term sophisticated programs. Major areas of investigation include: microwave devices, submicron microelectronics, ion propulsion, lasers and electro-optical components, fiber and integrated optics, and new electronic materials. For immediate consideration, please send your resume to Professional Staffing, Dept. SE, Hughes Research Laboratories, 3011 Malibu Canyon Road, Malibu, CA 90265. Equal opportunity employer.

Creating a new world with electronics



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

Life is an open book. If only he'd open the book.

Reading is thinking. And learning. And growing. But a kid won't read if he doesn't want to. So we're giving kids the incentive. And the books.

We're RIF (Reading is Fundamental), a national program with hundreds of local, community projects that help kids help themselves to books. Books they can choose for themselves. And keep for their own.

RIF works. But RIF only works if people work, too.

That's why we need you—or your organization—to help start a RIF program in your community. We'll help you to start going and start growing. Write RIF, Box 23444, Washington, D.C. 20024.



When it comes to getting kids to read, RIF wrote the book.



This is a public service message on behalf of Reading Is Fundamental and this magazine.

LETTERS

Sirs:

Wilson G. Pond's article "Modern Pork Production" [SCIENTIFIC AMERICAN, May] ignored a nest of questions that should be of considerable interest to producers and consumers alike, that is, to the ethical issues raised by modern intensive methods of production. I shall comment on only one of these questions: our obligation not to treat the pig as a robot. As Pond notes, pigs no longer live in outdoor pens, where they root in the ground and wallow in the mud. They spend their short lives totally confined. While this may have led to some benefits for producers and consumers, namely to increased production efficiency in certain respects, and to some benefits for the pigs, namely to improvements in health during their short lives, it has also had consequences for pigs that may not be beneficial. Confinement production units do not allow behaviors such as rooting and wallowing. The impulse of a pig to engage in such behavior, however, may be strongly genetically determined. Denied the opportunity to engage in such behavior, an intelligent animal such as a pig may suffer boredom or other unpleasant psychological states. . . .

The ignoring of ethical issues in such an article is particularly distressing. As Pond's article indicates, research to increase the efficiency of the pig in regard to feed conversion continues. If the quality of life of the pig continues to be ignored, however, we can expect that in the more efficient production units of the future the pig's life will be even bleaker than it may be today. Hopefully, future technological improvements will be made that will at least allow the pig to live a satisfactory life until it is slaughtered. This means research should be done to develop production systems that will contribute to satisfying the pig's psychological needs as well as to improving the pig's health and increasing the pig's efficiency as a converter of feed. Attention must also be given so that pigs can be transported and slaughtered in humane ways. At present it is not uncommon to break pigs' snouts for the final trip to the slaughterhouse to prevent fighting, pigs are not always rendered unconscious during slaughter, etc. Such practices cause unnecessary pain and so should be eliminated.

HUGH LEHMAN

Department of Philosophy
University of Guelph
Guelph, Ontario

Sirs:

The concerns about animal care ex-

pressed by Professor Lehman are shared by pork producers and animal scientists. The changes that have occurred in the technology of pork production, although they are partly motivated by economics, have also been motivated by concern for the welfare of the pig. The treacheries of outdoor farrowing in bad weather, of excessive losses of newborn pigs because of crushing or smothering by a 300-pound mother lying down to nurse three-pound young and of wallowing in filth and mud were overcome by the move indoors.

The pork producer of today generally is an animal protectionist and an animal welfarist in the best practical sense. Insinuations that the quality of life of the pig is being ignored reflect ignorance or misinformation. There is intense concern among producers, scientists and engineers for identifying and controlling environmental factors that may impede animal well-being. Current research in several university and government laboratories is aimed at characterizing and measuring animal stress and at continuing to develop production facilities and systems that will maximize animal comfort.

The breaking of pigs' snouts for the final trip to the slaughterhouse to prevent fighting, referred to by Professor Lehman, is, like any inhumane treatment, abhorrent to pork producers as it is to all reasonable people. Aberrant and aggressive behavior exists in all segments of society and is condoned no more by those associated with food-animal production than it is by society at large.

The human appetite for animal foods, including pork, appears to have persisted since *Homo sapiens* first appeared on the scene. Pork producers play a vital role in meeting this demand. Some have termed the efforts of the swine industry a vainglorious attempt to help feed the hungry world. Sincere, yes, vainglorious, no. Current trends point toward continued increases in pork production around the world. Pork producers are aware of their obligations as well as their opportunities. Their concerns about proper animal care will motivate them to continue to adapt new technology developed from penetrating and vigorous research currently under way in environmental physiology and animal care. Such adaptations should yield valuable benefits to consumers and the animals serving them.

WILSON G. POND

Research Leader, Nutrition
Roman L. Hruska U.S. Meat Animal
Research Center
Agricultural Research Service
United States Department
of Agriculture
Clay Center, Neb.

The Optimum Shape

The background of the entire page is a complex, abstract geometric pattern. It consists of a dense network of overlapping triangles of various sizes and orientations. The triangles are primarily white, but some are colored in shades of blue and orange. The overall effect is a textured, crystalline surface that appears to be illuminated from the top left, creating a gradient of light from white to dark blue and black.

The Optimum Shape

Researchers at the General Motors Research Laboratories have developed the first integrated system for computer design of mechanical parts with minimum mass.

Optimal Shape Generation automatically optimizes the component shape in a single computer run.

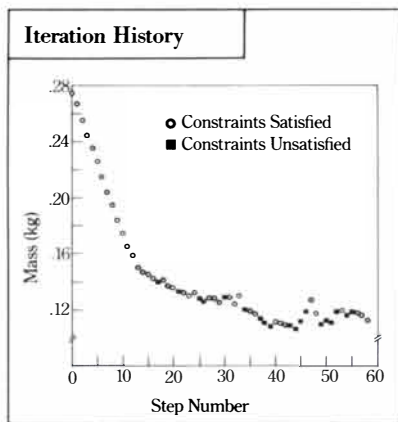


Figure 1: Decreasing mass plotted as a function of design iterations for the component shown in Figure 2.

Figure 2: Shapes as they appear on the CRT screen in the design of a minimum mass automotive component capable of performing under the structural loads. Color changes indicate (blue→yellow→green→red) increasing stress levels within the design limits.

COMPUTER-AIDED design systems automate the processes of generating geometric data and engineering drawings of parts, but they do not determine whether these parts meet structural performance requirements. In an ongoing research project at the General Motors Research Laboratories, a system has been developed that automatically ensures that the design meets structural performance constraints. More important, Optimal Shape Generation provides the component shape with the minimum mass capable of satisfying structural demands in a single computer run, without requiring human interaction with the machine.

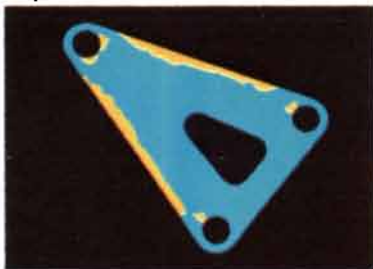
In the last two decades, extensive research has been done in the area of computer design of

structural components. Most of this work has focused on individual aspects of the process. Drs. Jim Bennett and Mark Botkin have succeeded in integrating the process from description of the model through convergence to the optimum solution.

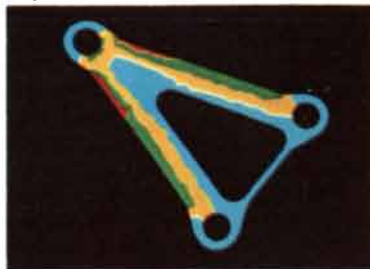
Conventional systems continue distinctions characteristic of age-old "build and test" methods by separating the tasks of design generation and design analysis. Typically, a "designer" uses one computer system to produce engineering drawings of a given part. The task then shifts to an "evaluator" who creates a mathematical model with which to test the design on another computer system. The evaluator determines only whether or not the design meets the requirements. A lengthy interaction between the designer and the evaluator is required to optimize the design. Optimal Shape Generation integrates the process from design generation through design optimization. The system can generate the mathematical model from the design data as the shape changes without requiring additional input, thereby turning the process from a multi-person, multimachine operation into a one-person, one-machine operation.

Since there is no interaction beyond the initial input, a flexible description of the problem is crucial to effective use of the system. The researchers responded to this challenge by developing a geometric format based on a parametric description of the boundary. Defining the problem with geometric data is desirable because it describes the shape of the part in a

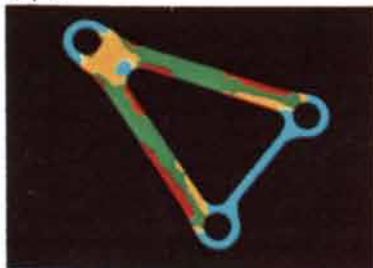
Step 0



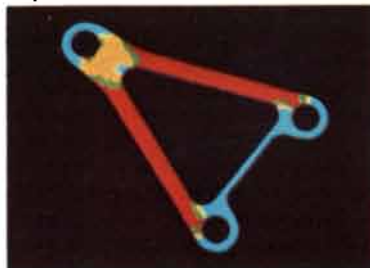
Step 10



Step 33



Step 58



form directly suitable for conceptual visualization.

Because the boundary geometric description must be transformed into an analysis model not once but several times, some type of automatic finite element mesh generation is required. The researchers adapted a mesh generation technique which divides a closed region into triangular elements based on a discrete description of the boundary. The sizes of the elements of the mesh are determined by a characteristic length selected for each problem and are related to the need for accurately describing the geometry. Automatic triangulation is used to create a set of connectivities for the discrete points placed uniformly throughout the part's interior with approximately the same density as the boundary points. The combination of boundary data description and automatic mesh generation permits the system to accommodate major changes in shape from the initial design.

ADEQUACY of the triangular meshes to calculate accurate stress levels was next addressed by the development of an adaptive mesh refinement scheme. By evaluating the solution for the uniform mesh created by the choice of characteristic length and identifying areas where the strain energy density changes rapidly, the system selects the areas of the mesh that require mesh refinement. These refinements can take the form of either adding elements in the area to be refined or increasing the order of the finite element

polynomial interpolation. The former approach has been taken, because it can be implemented automatically and does not require the formulation of new finite elements.

The culmination of the process introduces an optimization routine which directs the design toward a minimum mass configuration. A mathematical optimization technique is used to change the design to that shape giving minimum mass within the structural constraints. This optimization technique is based upon a sequential first-order Taylor series approximation of the constraints and a feasible directions solution of the problem. Periodic mesh refinements are performed throughout the optimization, since the design is continually changing, and the system must predict the stresses and the behavior of the constraints as the design changes.

"By taking an integrated approach," says Dr. Bennett, "we're able to combine the objectives of reducing the mass of the material and meeting structural performance requirements in a single automatic system."

"We expect," adds Dr. Botkin, "that in the future this technique will become the standard way of designing structural components."

General Motors



THE MEN BEHIND THE WORK



Drs. Jim Bennett and Mark Botkin are members of the Engineering Mechanics Department at the General Motors Research Laboratories.

Dr. Bennett holds the title of Assistant Department Head. He attended the University of Michigan as an undergraduate and received his graduate degrees from the same institution in the field of aerospace engineering. His Ph.D. thesis concerned non-linear vibrations. Before coming to General Motors in 1973, he taught aeronautical and astronautical engineering at the University of Illinois.

Dr. Botkin is a Staff Research Engineer. He received his undergraduate and graduate degrees from the University of Missouri at Rolla. His graduate work was in the field of civil engineering, and his doctoral thesis concerned structural optimization. Prior to joining General Motors in 1978, he worked for four years as a consultant to computer applications engineers.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1933: "Out of years of sad experience with two kinds of destructive natural phenomena—earthquakes and hurricanes—and years of study of their forces and effects by scientists and builders, one elementary fact is beginning to emerge. This is that man has little actual need to be in terror of these angrier moods of nature if only he will provide good, substantial buildings. Good buildings will ride safely through earthquakes of major force and will outlast a heavy hurricane. It is only bad buildings that go down before these forces. We need not abandon California and Florida, and we need not even fear to move into them—provided we move into well-built structures when we get there."

"Why should sunspots be cool? A typical sunspot maintains a temperature fully 1,000 degrees centigrade below that of its surroundings, over an area thousands of miles across and for days and weeks on end. Something is obviously happening to keep it cool, and there is convincing evidence that the process of expansion in the vast ascending vortex of gases that constitutes the spot is the cooling agency. New material is continually being fed into the bottom of the vortex, deep below the surface, and as the gases rise (reversing the familiar motion of water running out of a washbasin) the forced expansion cools the newly supplied material faster than the influx of heat from the sides can warm it up. Only as the spent gases flow outward close to the surface does the heating process overtake the cooling and so set an outer boundary to the spot."

"For 88 years *Scientific American* has portrayed the scientific progress of the world. In the first decades of its existence the magazine was the sole semi-popular disseminator of such news. Latterly other magazines have sprung up, some touching only on the spectacular fringe of science. In this enlarged group *Scientific American* stands alone, the only magazine of its class that aims to present in readable style a graphic and reliable account of man's conquest of the secrets of the universe. It does not 'write down' to any group but rather aims at a high standard that will appeal to the scientist, the engineer, the pro-

fessional man and the thinking layman. We cannot hope to please everyone. To those who desire information principally on mechanics or invention we recommend the other magazines that admirably report such news. Getting the proper perspective on scientific progress is *Scientific American's* purpose. While following legitimate scientific advances rather than predicting imaginative future developments, it still will lead, as always, in accuracy and reliability."



AUGUST, 1883: "The introduction within the past two years of the improved gelatine process, by which the time of taking photographs with dry plates has been reduced a thousand times, renders it an easy matter now to obtain with certainty excellent pictures of moving objects and opens up a vast field of experiment for the scientific student. We have lately received some excellent specimens of instantaneous work by Mr. G. G. Rockwood, illustrating the principal proceedings at the opening of the great Brooklyn Bridge. Pictures of frigates covered with flying flags, sailors manning the yard arms and cannons firing from the same ships are among the photographs, and they convey to the mind an idea of the extreme brevity of the time in which the impression must have been made on the sensitive plate."

"The long voyage made in the interest of science a few years since by the *Challenger*, a ship of the British service, awakened a widespread desire upon the part of intelligent people in all portions of the world to learn something further concerning that wide and mysterious domain, the bottom of the sea. A report made by the U.S. steamship *G. S. Blake*, upon her return in February from a somewhat similar mission, has renewed this feeling. The *Blake* is a floating laboratory. She is schooner-rigged and would pass, were it not for certain accessories, for a roomy private yacht. These additions are mainly a heavy boom rigged forward and pivoting upon the foremast, and a high framework over the port bow. The former is used in handling the trawl and the latter is the complex and delicate reel and its belongings by which the miles of wire are paid out or wound in during soundings. Below decks a large proportion of the space is given over to the draughting room, laboratory and storage rooms. Every minute particle brought up during a haul is preserved and labeled. To the ordinary investigator much of this painstaking labor seems wasted, but a little reflection reveals the great utility of all the care. An example is afforded by the results of the first two years of Lt. Comdr. Charles

D. Sigsbee's work upon the *Blake*, which were devoted to the Gulf of Mexico. The execution of the work embraced observations of depths, serial water temperatures and densities, and of currents when possible, together with the collection of specimens of the bottom soil and deposit, and of surface, bottom and intermedial water specimens, all of which served to add vastly to the sum of human knowledge regarding that vast and capricious basin."

"The Examiner of Interferences at the Patent Office, Mr. J. B. Church, has lately rendered a decision in the long-contested telephone case, in which the parties interested were Bell, Gray, Edison, McDonough, Dolbear, Boelker, Blake, Irwin and Richmond. Priority of invention is awarded to Bell for the art by which oral conversation or sounds of any description can be telegraphically transmitted; also for the improvement in the art of transmitting vocal sounds or spoken words telegraphically; also for the acoustic telegraph, including sound producers as well as reproducers on armature plate, the electro-magnet for the same, and a closed circuit passing from the helix of such electro-magnet to the source of undulatory electric energy; also for the telephonic transmitters and the combination in one circuit of two or more disks or diaphragms; also for the combination for rendering audible acoustic vibrations; also for the combination in an acoustic telegraph of an electro-magnet and a polarized armature, and the combination in an acoustic telegraph of an armature plate polarized by induction, a resonant tube and an electro-magnet and circuit connections. Priority of invention is awarded to Edison, although he did not claim it, for the transmitter, consisting of the combination in an electric circuit of a diaphragm and a liquid or equivalent substance of high resistance, whereby the vibrations of the diaphragm cause variations in the resistance of the electric current; also for the combination in a telegraph instrument operated by sound of two or more electrodes placed in an electrolytic liquid, and operating to increase and decrease the resistance of the electric circuit by the movement derived from the diaphragm; also for a spring forming or carrying one electrode and constantly pressing against the other electrode and the diaphragm to maintain the required initial pressure between the electrodes and yield to the movements of the diaphragm. Priority of invention for a telephone receiver, consisting of the combination in an electric circuit of a magnet and a diaphragm supported and arranged in close proximity thereto, whereby sounds thrown upon the line may be reproduced accurately as to pitch and quality, is awarded to McDonough."



Allen Paulson remembers what took him to the top.

“Begin at the bottom and learn everything you can every step of the way.”

That’s the credo of Allen Paulson, owner and operator of the world’s largest privately owned aircraft manufacturing company: Gulfstream Aerospace Corporation.

Starting at 18 as a mechanic for TWA, he’s worked as a flight engineer, a test pilot, a commercial pilot and even today, he’s still flying.

“Behind the controls,

40,000 feet in the air, you can get back to basics,” he says.

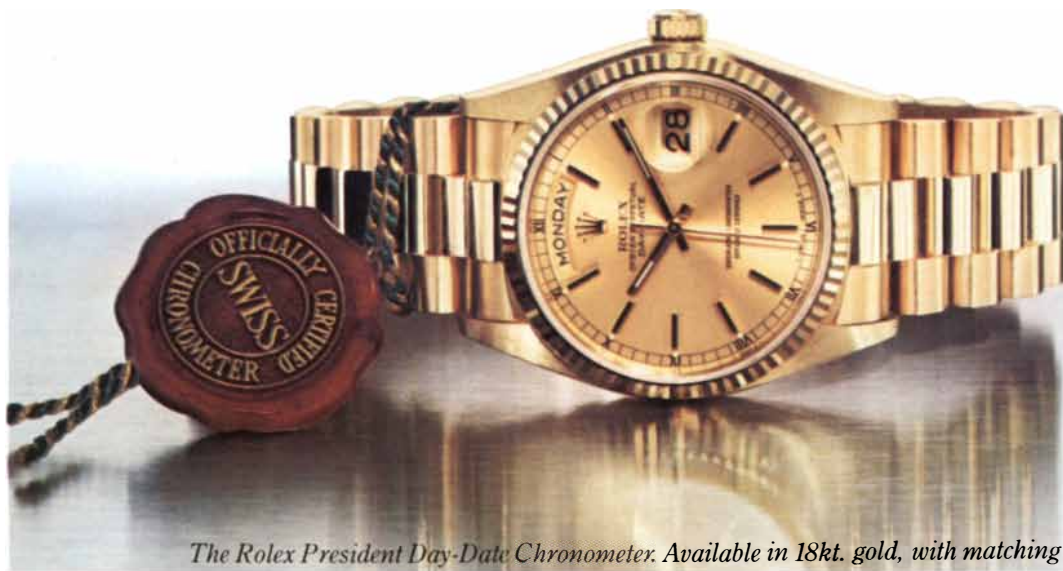
Remembering the basics is probably what’s behind Gulfstream’s continual success. Innovation is never made at the expense of time-tested ideas. Evolution, not revolution, is the Paulson game plan. And Allen Paulson is one of the world’s great planners. You can’t build a plane like the Gulfstream III any other way.

As with any well-built machine, the process is long, the timing precise and the result—a masterpiece.

We, at Rolex, have no trouble understanding this concept. It underlies the construction of every timepiece that bears our name. Like the President® Day-Date Chronometer in 18kt. yellow gold. Planned to give precise information to a demanding president. Like Allen Paulson.



ROLEX



The Rolex President Day-Date Chronometer. Available in 18kt. gold, with matching bracelet.

Write for brochure. Rolex Watch, U.S.A., Inc., Dept. 380, Rolex Building, 665 Fifth Avenue, New York, N.Y. 10022. World headquarters in Geneva. Other offices in Canada and major countries around the world.

Over 300 manufacturers sell their microcomputers with Microsoft software. There's a reason.

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

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers.

That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

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

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard™

products that materially increase the capabilities of Apple® and IBM PC computers. Plus, business and management software such as Multiplan™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

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

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004



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

THE AUTHORS

DONALD D. TRUNKEY ("Trauma") is professor of surgery at the University of California at San Francisco School of Medicine and chief of surgery at the San Francisco General Hospital Medical Center. He attended Washington State University as an undergraduate, obtaining a bachelor's degree there before going on to earn his M.D. at the University of Washington School of Medicine. In 1966 he entered the Army. After two years of service in Germany he returned to the U.S. to complete a residency in surgery at the San Francisco General Hospital. Following an additional year of training at the University of Texas Southwestern Medical School in Dallas he moved to San Francisco General in 1972. Trunkey's main scientific interest is the subject of his article; he also works on shock.

SIDNEY PESTKA ("The Purification and Manufacture of Human Interferons") is head of the laboratory of molecular genetics at the Roche Institute of Molecular Biology. Born in Poland, he came to the U.S. to receive his education. His B.A. is from Princeton University and his M.D. is from the University of Pennsylvania School of Medicine. After serving an internship in pediatrics and medicine in the Baltimore City Hospitals he joined the staff of the National Institutes of Health. He worked at the National Heart Institute and the National Cancer Institute before going to the Roche Institute in 1969. Pestka is adjunct professor of pathology at the Columbia University College of Physicians and Surgeons. He did much of the fundamental work in purifying the interferons and formulating techniques for making them in large quantities. He is an enthusiastic amateur photographer who personally photographed the first interferon crystals.

E. N. PARKER ("Magnetic Fields in the Cosmos") is Distinguished Service Professor of Physics and of Astronomy and Astrophysics at the University of Chicago. He got his B.S. in 1948 at Michigan State University. His Ph.D. in physics was awarded in 1951 by the California Institute of Technology. From 1951 to 1955 he was at the University of Utah, first as a member of the department of mathematics and then as a member of the department of physics. In 1955 he moved to Chicago. He has served as chairman of both departments of which he is currently a member. He writes: "Astrophysical observations provide a number of puzzles in classical physics with the clues to discover the previously unknown physical effects that cause them. A few of us with a skept-

tical disposition make our living recognizing hollow explanations [of such phenomena] and moving in to see what can be done in the way of constructing a real explanation. And this leads to the discovery of many, many hitherto unknown little effects that accumulate over the years to explain some major puzzles. The solar wind is one example; the basic structure of the galactic magnetic field is another."

ROY S. LEWIS and EDWARD ANDERS ("Interstellar Matter in Meteorites") are members of the faculty at the University of Chicago who have a common interest in the chemical aspects of astronomy. Lewis is a senior research associate. He earned his B.A. at Berkeley in 1967. He went to Chicago in 1972, shortly before his Ph.D. in atmospheric and space science was granted by the University of California at Berkeley. He writes: "My main interest is the early history of the solar system and more generally the nucleosynthetic history of the universe. My own contribution is primarily to measure the isotopic composition and abundance of the noble gases trapped in primitive meteorites." Anders is Horace B. Horton Professor of Chemistry; he has been at Chicago since 1955. A native of Latvia, he attended the University of Munich from 1946 to 1949 as an undergraduate before coming to the U.S. His M.A. (1951) and his Ph.D. in radiochemistry (1954) are from Columbia University.

GLENN D. PRESTWICH ("The Chemical Defenses of Termites") is associate professor of chemistry at the State University of New York at Stony Brook. He was graduated from the California Institute of Technology with a B.S. in 1970. His Ph.D. in chemistry, given in 1974, is from Stanford University. He served for three years as a research scientist at Cornell University and the International Centre of Insect Physiology and Ecology in Kenya before moving to Stony Brook. He is the recipient of an Alfred P. Sloan Foundation Fellowship for 1981-83 for his work on the chemistry of substances synthesized by insects. Among Prestwich' interests in relation to insects are steroid and fat metabolism and chemical ecology.

DOUGLAS H. BLAIR and ROBERT A. POLLAK ("Rational Collective Choice") are economists whose interests intersect in the subject of their article. Blair is associate professor of economics at Rutgers University. He got his B.A. at Swarthmore College (1970) before going on to obtain his

INVEST YOURSELF



A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps.

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

VITA Putting Resources to Work for People

1815 North Lynn Street, Arlington, Virginia 22209-2079, USA



THE FOOD AND AGRICULTURAL ORGANIZATION OF THE UNITED NATIONS, Rome, Italy, requires AGRICULTURAL OFFICERS

The FAO/World Bank Cooperative Programme invites applications for anticipated Agricultural Officer appointments of three years initial duration, stationed in Rome, but involving frequent travel to developing countries.

The main duties involve participating in or leading missions concerned with identification and preparation of agricultural investment projects in developing countries, and being responsible for all agricultural aspects of such projects.

Essential qualifications: University degree in agriculture followed by at least 7 years of either field or managerial professional experience in agriculture under diverse conditions in more than one developing country. Practical experience of a broad range of rainfed and irrigated tropical crops. A high standard of report writing is required.

Languages: Full working knowledge of English and limited knowledge of French, or vice-versa.

Salary: Depending upon level of appointment, from US \$32,000 to US \$46,000 per annum net tax-free including cost of living and other allowances.

Please send detailed Curriculum Vitae quoting this advertisement to Mr. T.R. Short, Personnel Officer, DDC, FAO, Via delle Terme di Caracalla, 00100 Rome, Italy.

M.A. (1972) and his Ph.D. in economics (1976) from Yale University. In 1976 he joined the faculty at the University of Pennsylvania. He left Pennsylvania in 1981 to take up his job at Rutgers. Blair's current work concerns the objectives of labor unions and the theory of bargaining. Pollak is Charles and William Day Professor of Economics and Social Sciences at Pennsylvania. He attended Amherst College as an undergraduate, getting his B.A. in 1960. His Ph.D. in economics was granted by the Massachusetts Institute of Technology in 1964. He joined the faculty at Pennsylvania the same year and has remained there since with the exception of a stint as an economist at the Bureau of Labor Statistics.

PETTER AUNE, RONALD L. SACK and ARNE SELBERG ("The Stave Churches of Norway") are structural engineers who are interested in old wood structures. Aune is senior lecturer in timber-structure design at the Norwegian Institute of Technology (N.T.H.) in Trondheim. A native of Norway, he received his training in civil engineering at N.T.H. Sack is professor of civil engineering at the University of Idaho. His degrees in civil engineering are all from the University of Minnesota; they include a B.S. (1957), an M.S. (1958) and a Ph.D. (1964). His study of the stave churches began in 1976 and 1977, which he spent on sabbatical leave as a fellow of the Royal Norwegian Council for Scientific and Industrial Research at N.T.H. Until his retirement in 1979 Selberg was head of the division of steel structures at N.T.H. He is a native of Norway who earned his degrees in civil engineering from N.T.H. His work has mainly focused on suspension bridges. He became interested in stave churches during the renovation of the church at Urnes in 1975.

CHARLES BIGELOW and DONALD DAY ("Digital Typography") combine an interest in letterforms with an interest in computers. Bigelow is assistant professor of computer science and art at Stanford University. He received his B.A. from Reed College in 1967. In 1982 he was given a MacArthur Foundation Prize Fellowship for contributions to the study of typography. With his partner, Kris Holmes, he designs typefaces for computer systems. Day is lecturer in the video department of the California College of Arts and Crafts. Like Bigelow, he attended Reed College. Since 1979 he has been collaborating with a group headed by Daniel Sandin of the University of Illinois at Chicago Circle in the design of a digital image processor. Day's interest in letterforms was aroused partly by a course in calligraphy he took at Reed; Bigelow was in the same class.



"In Just A Few Days, I'll Show You How To Do

REAL MATH

On Your Calculator!"

$$\int_a^b f \quad \sum_{n=1}^{\infty} a_n \quad \frac{df}{dx} \quad \lim_{n \rightarrow \infty}$$

- Quick. •Guaranteed.
- Easy. •Fun, Too!

INTRIGUED BY CALCULATORS? Then you can step up your math skills fast! Use my new method in guidebook form. It's called **CALCULATOR CALCULUS**. This space-travel spinoff is sure-fire, so it has a **simple guarantee** — just return it for an immediate refund if you are not astounded at the problems you're solving with it!

But the point is - you won't want to send it back. For this is the **easiest, fastest shortcut** ever! The day you receive your copy in the mail you'll want to put it to work. It's that exciting and helpful.

My name is Dr. George McCarty. I teach math at the University of California. I wrote this guidebook to cut through the confusion. I guide you with examples you follow step-by-step on your calculator — you do simple exercises — then you solve practical problems with real precision!

POWER METHODS. Need to evaluate functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series! It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more (even differential equations).

Important numerical techniques? Those algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

LOOK AT WHAT USERS SAY: Samuel C. McCluney, Jr., of Philadelphia writes:

"**CALCULATOR CALCULUS IS GREAT!** For ten years I have been trying to get the theory of calculus through my head, using home-study courses. It was not until I had your book that it became clear what the calculus was all about. Now I can go through the other books and see what they are trying to do. With your book and a calculator the whole idea becomes clear in a moment, and is a MOST REFRESHING EXPERIENCE. I program some of the iterative prob-

lems you suggest and it always GIVES ME A THRILL to see it start out with a wild guess and then approach the limit and stop."

Professor John A. Ball of Harvard College (author of the book *Algorithms for RPN Calculators*) writes: "I wish I had had as good a calculus course."

Professor H. I. Freedman of the U. of Alberta, writing in *Soc. Ind. Appl. Math Review*, states: "There can be no question as to the usefulness of this book...lots of exercises...very clearly written and makes for easy reading."

Tektronix Engineer Bill Templeton says "**CALCULATOR CALCULUS** is the best, most clearly written book I have seen for improving your math skills."

I WANT YOU TO DO THIS. Get my complete kit, with a TI-35 calculator, plus its 200 p. Student Math Book, **AND** the guidebook, **ALL** for \$44.95 (for shipping to USA add \$2, or \$5 by AIR; Foreign \$5, or \$10 AIR; in Calif. add \$2.70 tax).

If you already have a scientific calculator, you can invest in the guidebook, "**CALCULATOR CALCULUS**", for only U.S. \$19.95 (to USA or foreign: add \$1 for shipping, or \$4 by AIR; in Calif. add \$1.20 tax).

As pennywise Ben Franklin said, "*An investment in knowledge pays the best dividends.*" **GET STARTED NOW** — Tax deductible for professionals.

MONEY-BACK GUARANTEE! Send for it today. Be sure to give me your complete mailing address with your check or money order. If you want to charge it (Visa or MC), tell me your card no. and exp. date. Prompt shipment guaranteed.

George M. Cart

Thank you! **EduCALC Publications - Dept. A7**
27963 Cabot Road, South Laguna, CA 92677
For fast service, phone MC or VISA orders to (714) 831-2637

ATTEND TELECOM 83 GENEVA

Switzerland—October 26-November 1, 1983 Fourth World Telecommunication Exhibition

TELECOM 83, the fourth in a series of quadrennial World Telecommunication Exhibitions, is organized under the auspices of the International Telecommunication Union (ITU) and its 158 member countries.

Highlighting the general theme "Telecommunications for Everyone," TELECOM 83 offers a unique opportunity for individuals or groups to join in the exchange of ideas, information, and technology in the fields of telecommunications and electronics. Additionally, it will provide an excellent setting for international contacts.

An integral part of TELECOM 83 will be the Fourth World Telecommunication Forum, organized by the ITU and 50 national and international engineering societies from all five continents. Telecom and the World Telecommunication Forums are recognized as the universal and most authoritative meeting of telecommunication engineers and economists.

FORUM 83: The World Telecommunications' Summit

FORUM 83—Parts 1, 2 and 3—will assemble a "brains trust" of several thousand top executives who will present and discuss the planning, financing, management and implementation of the world telecommunication network, and the convergence of computing and communications technologies.

Over 200 presentations will enable participants to hear, meet and question industry and government leaders on current and future developments of the world telecommunication network. Plenary sessions and selected parallel sessions will have simultaneous interpretation.

Nations Working Hand in Hand

The three-day Part One of FORUM 83 will offer presentations by prominent figures in the economic sector as well as the telecommunication field. Speakers will include government leaders, senior corporate managers, chief scientists from industry, and representatives from international and financial organizations. They will discuss the technological requirements of industrialized and developing nations. They will also confront the future need for financing national, regional and international telecommunication development plans.

Participation for each of these parts of FORUM 83 is limited to the first 1500 registrants.

In keeping with the ever-increasing transnational exchange of data, information and broadcasting, FORUM 83 Part Three, a Legal Symposium, will examine the legal aspects of international telecommunication. This concluding session of TELECOM 83 will scrutinize the international regulations relating to transnational information transport.

TELECOM 83 Exhibition

The Telecom 83 Telecommunication Exhibition encompasses all of the exhibition, conference and outdoor floorspace of Geneva's New Exhibition and Conference Centre. This quadrennial Telecommunication Exhibition will be a kaleidoscope showcasing the world's telecommunications and electronics industries, their capabilities and their networks.

The BOOK FAIR at TELECOM 83 will provide an opportunity for delegates to reference current works on telecommunications, electronics and allied fields.

Registration Fees (Swiss Francs)


	Individual	Group*	Officials**
Part 1	600	550	500
Part 2	300	275	250
Part 3	250	250	250
Parts 1 and 2	750	750	750
Parts 1 and 3	750	750	750
Parts 2 and 3	500	500	500
Parts 1, 2 and 3	850	850	850

*two or more from the same company or organization and members of institutions co-sponsoring the Technical Symposium.

**official representatives from administrations of countries/members of the ITU.

Towards a World Network

FORUM 83 Part Two has adopted the theme: "Integration of the World Telecommunication Network." This three-day event will be a far-reaching technical and scientific symposium with wide-ranging appeal. It represents the combined efforts of the ITU's 158 member countries to introduce to those attending, the advances made since FORUM 79 to solidify the world telecommunication network. Some 60 presentations will be made covering the latest innovations and technological trends.



Registration Form

TO: FORUM 83
International Telecommunication Union
Place des Nations
CH-1211 Geneve 20, Switzerland

Please register me for the following Forum '83 session(s).

Part 1 at rate for individuals groups officials

Part 2 at rate for individuals groups officials

Part 3 at rate for individuals groups officials

Name _____

Organization/Job Title _____

Address _____

I enclose a cheque for SwFr of _____ in full payment of my registration.

I have initiated a bank transfer of SwFr _____ to Credit Suisse Geneva a/c Wagons-lits Tourisme Cooks TW 673.050.71. Specify TELECOM 83 Forum.

Date _____ Signature _____

MATHEMATICAL GAMES

Tasks you cannot help finishing no matter how hard you try to block finishing them

by Martin Gardner

EDITOR'S NOTE: Martin Gardner's column "Mathematical Games" will be appearing in this space for two issues.

With useless endeavor,
Forever, forever,
Is Sisyphus rolling
His stone up the mountain!

—HENRY WADSWORTH LONGFELLOW,
The Masque of Pandora

Suppose you have a basket containing 100 eggs and also a supply of egg cartons. Your task is to put all the eggs into the cartons. A step (or move) consists of putting one egg into a carton or taking one egg from a carton and returning it to the basket. Your procedure is this: After each two successive packings of an egg you move an egg from a carton back to the basket. Although this is clearly an inefficient way to pack the eggs, it is obvious that eventually all of them will get packed.

Now assume the basket can hold any finite number of eggs. The task is unbounded if you are allowed to start with as many eggs as you like. Once the initial number of eggs is specified, however, a

finite upper bound is set on the number of steps needed to complete the job.

If the rules allow transferring any number of eggs back to the basket any time you like, the situation changes radically. There is no longer an upper bound on the steps needed to finish the job even if the basket initially holds as few as two eggs. Depending on the rules, the task of packing a finite number of eggs can be one that must end, one that cannot end or one that you can choose to make either finite or infinite in duration.

We now consider several entertaining mathematical tasks with the following characteristic. It seems intuitively true that you should be able to delay completing the task forever, when actually there is no way to avoid finishing it in a finite number of moves.

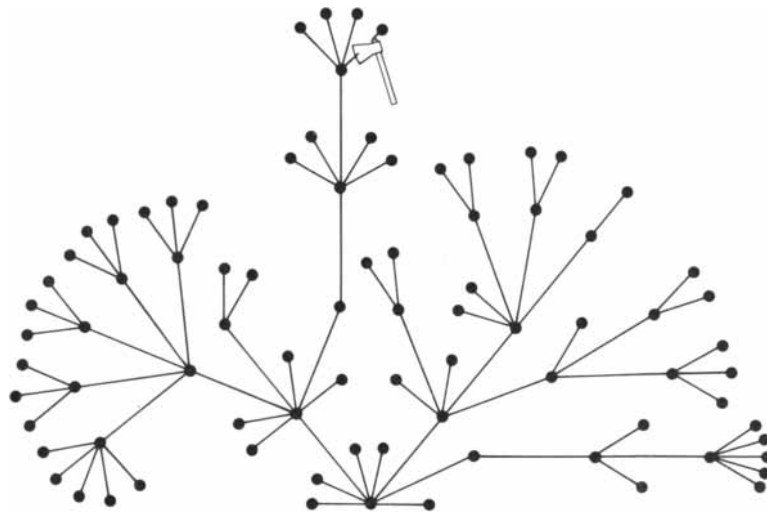
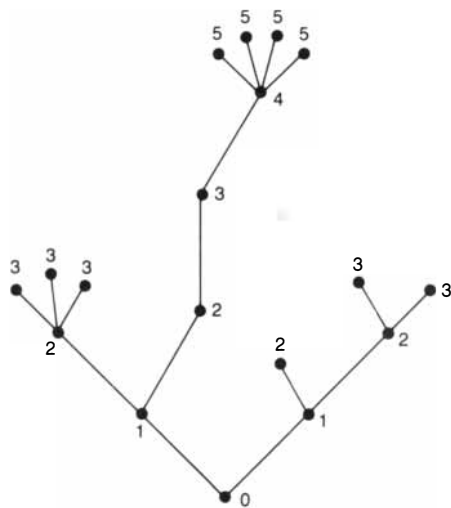
Our first example is from a paper by the philosopher-writer-logician Raymond M. Smullyan. Imagine you have an infinite supply of pool balls, each bearing a positive integer, and for every integer there is an infinite number of balls. You also have a box that contains a finite quantity of numbered balls. Your goal is to empty the box. Each step

consists of removing a ball and replacing it with any finite number of balls of lower rank. The 1 balls are the only exceptions. Since no ball has a rank lower than 1, there are no replacements for a 1 ball.

It is easy to empty the box in a finite number of steps. Simply replace each ball higher than 1 with a 1 ball until only 1 balls remain, then take out the 1 balls one at a time. The rules allow you, however, to replace a ball with a rank above 1 with any finite number of balls of lower rank. For instance, you may remove a ball of rank 1,000 and replace it with a billion balls of rank 999, with 10 billion of rank 998, with a billion billion of rank 987 and so on. In this way the number of balls in the box may increase beyond imagining at each step. Can you not prolong the emptying of the box forever? Incredible as it may seem at first, there is no way to avoid completing the task.

Note that the number of steps needed to empty the box is unbounded in a much stronger way than it is in the egg game. Not only is there no bound on the number of eggs you begin with but also each time you remove a ball with a rank above 1 there is no bound to the number of balls you may use to replace it. To borrow a phrase from John Horton Conway, the procedure is "unboundedly unbounded." At every stage of the game, as long as the box contains a single ball other than a 1 ball, it is impossible to predict how many steps it will take to empty the box of all but 1 balls. (If all the balls are of rank 1, the box will of course empty in as many steps as there are 1 balls.) Nevertheless, no matter how clever you are in replacing balls, the box eventually must empty after a finite number of moves. Of course, we have to assume that although you need not be immortal, you will live long enough to finish the task.

Smullyan presents this surprising re-



A tree-trimming task

Radio Shack Presents The Micro Executive Workstation

- Powerful Built-In Software
- Retains Memory Data When "Off"
- Self-Contained Telephone Modem
- 8K RAM—Expandable to 32K

\$799

8K RAM
Cat. No. 26-3801

\$999

24K RAM
Cat. No. 26-3802



Introducing the TRS-80® Model 100 Portable Computer —User-Friendly Software Makes it Truly Revolutionary

Imagine a computer on your desk so small, it can fit in your in-basket. The second you turn it on, imagine seeing a menu of built-in executive management programs and your own files, ready for immediate use. All revealed on an eight-line by 40-character LCD display positioned just above a full-size keyboard. And when you leave the office, imagine a three-pound computer you can take along, because it works on AC or batteries.

Stop imagining. The new TRS-80 Model 100 is the computer you've been waiting for. As a desk organizer, it's a phone directory, address book, appointment calendar and telephone auto-dialer. It's a personal word processor, as well. There's even a built-in modem to access other computers by phone.

Come see the most revolutionary computer since the TRS-80 Model I at over 6500 Radio Shack stores and participating dealers, including over 400 Radio Shack Computer Centers nationwide.



Send me a free TRS-80 Model 100 brochure today!

Mail To: Radio Shack, Dept. 84-A-221
300 One Tandy Center, Fort Worth, Texas 76102

NAME _____
 COMPANY _____
 ADDRESS _____
 CITY _____ STATE _____ ZIP _____
 TELEPHONE _____

Retail prices may vary at individual stores and dealers.

Radio Shack®
 The biggest name in little computers®
 A DIVISION OF TANDY CORPORATION

sult in a paper, "Trees and Ball Games," in *Annals of the New York Academy of Sciences* (Vol. 321, pages 86-90; 1979). Several proofs are given, including a simple argument by induction. I cannot improve on Smullyan's phrasing:

"If all balls in the box are of rank 1, then we obviously have a losing game. Suppose the highest rank of any ball in the box is 2. Then we have at the outset a finite number of 2's and a finite number of 1's. We can't keep throwing away 1's forever, hence we must sooner or later throw out one of our 2's. Then we have one less 2 in the box (but possibly many more 1's than we started with). Again, we can't keep throwing out 1's forever, and so we must sooner or later throw out another 2. We see that after a finite number of steps we must throw away our last 2, and then we are back to the situation in which we have only 1's. We already know this to be a losing situation. This proves that the process must terminate if the highest rank present is 2. Now, what if the highest rank is 3? We can't keep throwing away just balls of rank ≤ 2 forever (we just proved that!); hence we must sooner or later throw out a 3. Then again we must sooner or later throw out another 3, and so we must eventually throw out our last 3. This then reduces the problem to the preceding case when the highest rank present is 2, which we have already solved."

Smullyan also proves that the game ends by modeling it with a tree graph. By "tree" is meant a set of line segments each of which joins two points, and in such a way that every point is connected by a unique path of segments leading to

a point called the tree's root. The first step of a ball game, filling the box, is modeled by representing each ball as a point, numbered like the ball and joined by a line to the tree's root. When a ball is replaced by other balls of lower rank, its number is erased and the new balls indicated by a higher level of numbered points are joined to the spot where the ball was removed. In this way the tree grows steadily upward, its "endpoints" (points that are not the root and are attached to just one segment) always representing the balls in the box at that stage of the game.

Smullyan proves that if this tree ever becomes infinite (has an infinity of points), it must have at least one infinite branch stretching upward forever. This, however, is clearly impossible because the numbers along any branch steadily decrease and therefore must eventually terminate in 1. Since the tree is finite, the game it models must end. As in the ball version, there is no way to predict how many steps are needed to complete the tree. At that stage, when the game becomes bounded, all the endpoints are labeled 1. The number of these 1 points may, of course, exceed the number of electrons in the universe, or any larger number. Nevertheless, the game is not Sisyphean. It is certain to end after a finite number of moves.

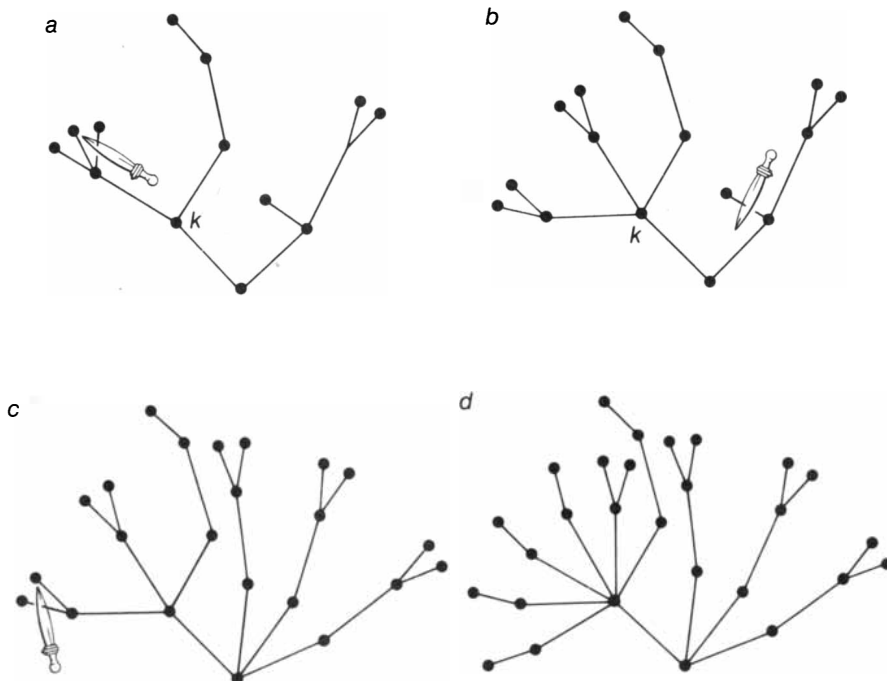
Smullyan's basic theorem, which he was the first to model as a ball game, derives from theorems involving the ordering of sets that go back to Georg Cantor's work on the transfinite ordinal numbers. It is closely related to a deep theorem about infinite sets of finite trees

that was first proved by Joseph B. Kruskal and later proved in a simpler way by C. St. J. A. Nash-Williams. More recently Nachum Dershowitz and Zohar Manna have used similar arguments to show that certain computer programs, which involve "unboundedly unbounded" operations, must eventually come to a halt.

A special case of Smullyan's ball game is modeled by numbering a finite tree upward from the root as is shown in the figure at the left in the illustration on page 12. We are allowed to chop off any endpoint, along with its attached segment, then add to the tree as many new branches as we like, and wherever we like, provided all the new points are of lower rank than the one removed. For example, the figure at the right in the illustration shows a possible new growth after a 4 point has been chopped off. In spite of the fact that after each chop the tree may grow billions on billions of new branches, after a finite number of chops the tree will be chopped down. Unlike the more general ball game, we cannot remove any point we like, only the endpoints, but because each removed point is replaced by points of lower rank, Smullyan's ball theorem applies. The tree may grow inconceivably bushier after each chop, but there is a sense in which it always gets closer to the ground until eventually it vanishes.

A more complicated way of chopping down a tree was proposed by Laurie Kirby and Jeff Paris in *The Bulletin of the London Mathematical Society* (Vol. 14, Part 4, No. 49, pages 285-293; July, 1982). They call their tree graph a hydra. Its endpoints are the hydra's heads, and Hercules wants to destroy the monster by total decapitation. When a head is severed, its attached segment goes with it. Unfortunately after the first chop the hydra acquires one or more new heads by growing a new branch from a point (call it k) that is one step below the lost segment. This new branch is an exact replica of the part of the hydra that extends up from k . The figure at the top right in the illustration at the left shows the hydra after Hercules has chopped off the head indicated by the sword in the figure at the top left.

The situation for Hercules becomes increasingly desperate because when he makes his second chop, *two* replicas grow just below the severed segment [figure at bottom left]. And *three* replicas grow after the third chop [figure at bottom right], and so on. In general, n replicas sprout at each n th chop. There is no way of labeling the hydra's points to make this growth correspond to Smullyan's ball game; nevertheless, Kirby and Paris are able to show, utilizing an argument based on a remarkable number theorem found by the British logician R. L. Goodstein, that no matter what se-



Growth of the hydra

Thunderbird Turbo Coupe. A World Class Touring Car.



First impressions. As you enter Turbo Coupe you notice how the doors curve into the roof to smooth and quiet the air-flow. You sit in an articulated seat. It's firm. Purposeful. Under and side thigh supports adjust to fit your body. The back angle and seat position adjust to align you with the controls. The pneumatic lumbar support adjusts to the curve of your back. Your feet feel how the pedals are positioned for heel-and-toe-operation. You test the five-speed shifter. The throws are short. Precise.

Start the engine. As it warms, consider the specifications. Four cylinders with aluminum pistons. 2.3 liters. Turbocharged, fuel-injected and monitored by Ford's EEC-IV Onboard Engine Control Computer. An engine so efficient it's rated at 145 horsepower (at 4,600 RPM per SAE standard J-1349) and is also rated at 33 estimated highway, 22 EPA estimated miles per gallon.* A triumph of technology over brute force.

Engage first gear... At this point, you will feel how the Goodyear Eagle performance tires, modified struts, four shock rear suspension and stabilizer bars work on the road. But nothing will better describe that feeling than a test drive. For more information, call 1-800-772-2100.

*For comparison. Your mileage may differ depending on speed, distance and weather. Actual highway mileage lower.

Get it together —
Buckle up.

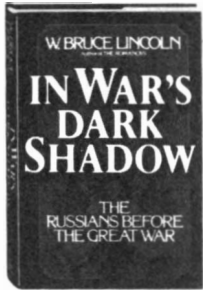
**Have you driven a Ford...
lately?**



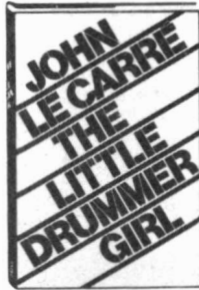
Either.

Choose any 4 books below for two dollars as your introduction to membership in the Book-of-the-Month Club.

You simply agree to buy 4 books within the next two years.



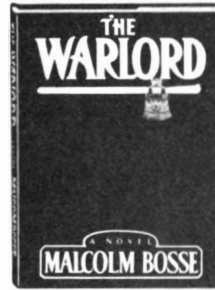
172 Pub price \$19.95



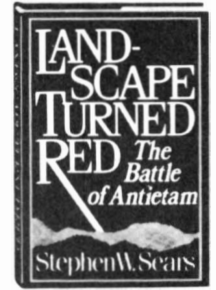
307 Pub price \$15.95



620 Pub price \$25



067 Pub price \$17.95



213 Pub price \$17.95



829 Pub price \$14.95



159 Pub price \$16.95



241 Pub price \$16.95



244 Pub price \$17.95



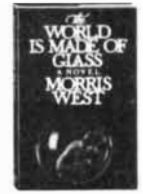
238 Pub price \$19.95



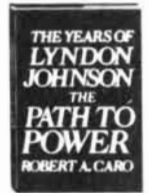
232 Pub price \$17.95



860 Pub price \$19.95



208 Pub price \$15.95



691 Pub price \$19.95



099 Pub price \$22.95



593 Pub price \$15.95



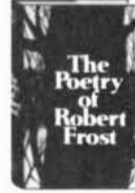
168 Pub price \$19.95



526 Pub price \$22.95



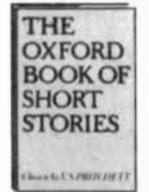
660 Pub price \$14.95



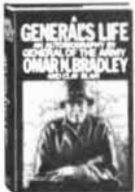
127 Pub price \$17.50



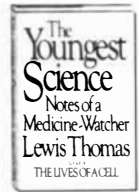
092 Pub price \$14.95



242 Pub price \$19.95



110 Pub price \$19.95



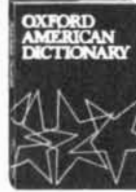
854 Pub price \$14.75



779 Pub price \$14.95



055 Pub price \$19.95



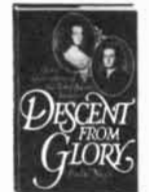
624 Pub price \$14.95



234 Pub price \$14.95



068 Pub price \$17.95



144 Pub price \$25



786 Pub price \$25



431 Pub price \$24.95



722 Pub price \$24.95



201 Pub price \$16.60



740 Pub price \$34.95

Benefits of Membership

Membership in the Book-of-the-Month Club begins with your choice of either 4 of today's best books for \$2 or one of the extraordinary works offered here. Because our prices are generally lower than the publishers' prices, you will save on works such as these throughout your membership. In fact, the longer you remain a member, the greater your savings can be. Our Book-Dividend® plan, for which you become eligible after a brief trial enrollment, offers savings from 50% to 75% off the publishers' prices on art books, reference works, classics, books on cooking and crafts, literary sets and other contemporary works of enduring value. All Book-of-the-Month Club books are equal in quality to the publishers' originals; they are not condensed versions or cheaply made reprints.

• As a member you will receive the *Book-of-the-Month Club News*® 15 times a year

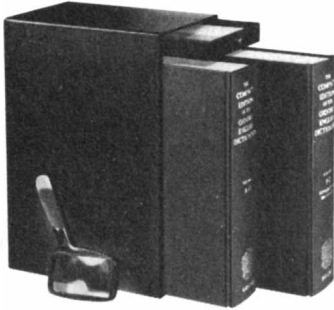
(about every 3½ weeks). Every issue reviews a Selection and 150 other books that we call Alternates, which are carefully chosen by our editors.

- If you want the Selection, do nothing. It will be shipped to you automatically. If you want one or more Alternates—or no book at all—indicate your decision on the Reply Form and return it by the specified date.
- *Return Privilege:* If the *News* is delayed and you receive the Selection without having had 10 days to notify us, you may return it for credit at our expense.
- *Cancellations:* Membership may be discontinued, either by you or by the Club, at any time after you have bought 4 additional books.
- Join today. With savings and choices like these, no wonder Book-of-the-Month Club is America's Bookstore.

Or.

Choose one of the 5 valuable sets below as your introduction to membership in the Book-of-the-Month Club.

You simply agree to buy 4 books within the next two years.



The Compact Edition of The Oxford English Dictionary
for **\$24.95** (Pub price \$150)

The "most complete, most scholarly dictionary of the English language" — *The Christian Science Monitor*. Through photoreduction, the original 13-volume set has been reproduced in this two-volume *Compact Edition*. Magnifying glass included.



The Story of Civilization by Will and Ariel Durant
for **\$29.95** (Pub prices total \$335.45)

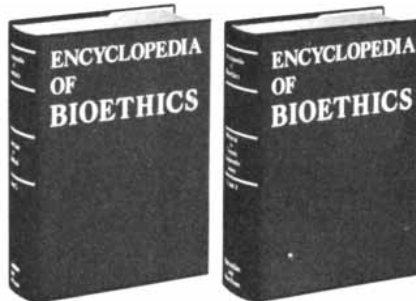
For almost half a century Will and Ariel Durant traced the continuity of world history to show the foundations of society today. The Durants' 11-volume illustrated masterwork is history come alive in all its dimensions. The enormous scope of the work covers ancient and recent civ-

ilization, including Oriental as well as Western history. A Book-of-the-Month Club exclusive for almost 50 years, it is both a reference work of undisputed authority and a lively portrait of the rulers and leaders who have shaped the course of human events.



The Encyclopedia of Philosophy
for **\$24.95** (Pub price \$175)

The most comprehensive encyclopedia of philosophy ever published, this set traces ancient, medieval, modern, Eastern and Western thought. An essential and rewarding reference source for home libraries.



The Encyclopedia of Bioethics
for **\$19.95** (Pub price \$125)

The new two-volume edition of the first comprehensive reference work in the increasingly important field of bioethics. It includes the views of 285 contributors and has 315 articles on the scientific, sociological, legal and ethical aspects of the major medical and scientific issues of today.



The Decline and Fall of the Roman Empire by Edward Gibbon Edited by J. B. Bury
for **\$24.95** (Pub price \$300)

The definitive Bury edition of the most acclaimed history of all. Illustrated 7-volume set, newly available on long-lasting acid-free paper, quarter-bound in leather.

Book-of-the-Month Club, Inc., Camp Hill, Pennsylvania 17012

A170-8

Please enroll me as a member of Book-of-the-Month Club and send me *either* the four books I've listed in the "Either" boxes at right, billing me \$2, plus shipping and handling charges, *or* the one set I've checked in the "Or" column, billing me for the appropriate amount, plus shipping and handling charges. In either case, I agree to buy four additional books from the Club over the next two years. A shipping and handling charge is added to each shipment.

Name _____ (Please print plainly)

Address _____ Apt. _____

City _____

State _____ Zip _____

Prices shown are U.S. prices. Outside the U.S., prices are generally higher.

BOOK-OF-THE-MONTH CLUB®

America's Bookstore® since 1926.

Either.

Indicate by number the four books you want

3-04

Or.

Check one box only

912. Compact OED **\$24.95**

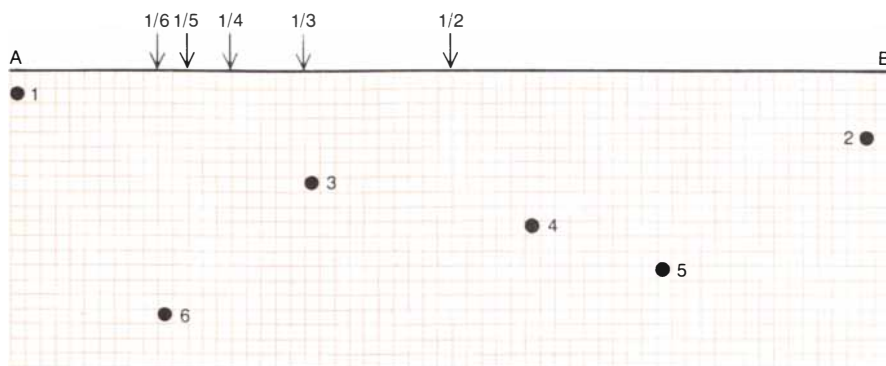
913. The Story of Civ. **\$29.95**

917. Ency. of Philosophy **\$24.95**

951. Decline and Fall of the Roman Empire **\$24.95**

953. Ency. of Bioethics **\$19.95**

3-64



One way to place six spots on the line A—B

quence Hercules follows in cutting off heads, the hydra is eventually reduced to a set of heads (there may be millions of them even if the starting form of the beast is simple) that are all joined directly to the root. They are then eliminated one by one until the hydra expires from lack of heads.

A useful way to approach the hydra game is to think of the tree as modeling a set of nested boxes. Each box contains all the boxes reached by moving upward on the tree, and it is labeled with the maximum number of levels of nesting that it contains. Thus in the first figure of the hydra the root is a box of rank 4. Immediately above it on the left is a 3 box and on the right is a 2 box, and so on. All endpoints are empty boxes of rank 0. Each time a 0 box (hydra head) is removed the box immediately below gets duplicated (along with all its contents), but each of the duplicates as well as the original box now contains one less empty box. Eventually you are forced to start reducing ranks of boxes, like the ranks of balls in the ball game. An inductive argument similar to Smullyan's

will show that ultimately all boxes become empty, after which they are removed one at a time.

I owe this approach to Dershowitz, who pointed out that it is not even necessary for the hydra to limit its growth to a consecutively increasing number of new branches. After each chop as many finite duplicates as you like may be allowed to sprout. It may take Hercules much longer to slay the monster, but there is no way he can permanently avoid doing so if he keeps hacking away. Note that the hydra never gets taller as it widens. Some of the more complicated growth programs considered by Dershowitz and Manna graph as trees that can grow taller as well as wider, and such trees are even harder to prove terminating.

Our next example of a task that looks as if it could go on forever when it really cannot is known as the 18-point problem. You begin with a line segment. Place a point anywhere you like on it. Now place a second point so that each of the two points is within a different half of the line segment. (The halves are taken to be "closed intervals," which means that the endpoints are not considered "inside" the interval.) Place a third point so that each of the three is in a different third of the line. At this stage it becomes clear that the first two points cannot be just anywhere. They cannot, for example, be close together in the middle of the line or close together at one end. They must be carefully placed so that when the third point is added, each will be in a different third of the line. You proceed in this way, placing every n th point so that the first n points always occupy different $1/n$ th parts of the line. If you choose locations carefully, how many spots can you put on the line?

Intuitively it seems as if the number should be endless. A line segment obviously can be divided into as many equal parts as you like and each may contain a point. The catch is that the points must be serially numbered to meet the task's conditions. It turns out, astonishingly, that you cannot get beyond 17 points!

Regardless of how clever you are at placing 17 points, the 18th will violate the rules and the game ends. In fact, it is not even easy to place 10 points. The illustration at the left shows one way of placing six.

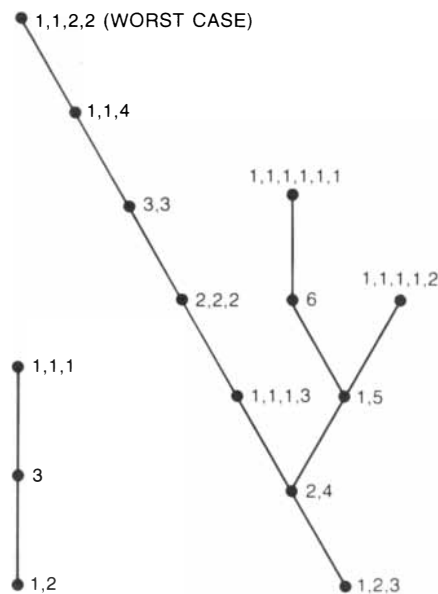
This unusual problem first appeared in *One Hundred Problems in Elementary Mathematics* (problems 6 and 7) by the Polish mathematician Hugo Steinhaus. (Basic Books published a translation in 1964, and there is now a Dover soft-cover reprint.) Steinhaus gives a 14-point solution, and he states in a footnote that M. Warmus has proved 17 is the limit. The first published proof, by Elwyn R. Berlekamp and Ronald L. Graham, is in their paper "Irregularities in the Distributions of Finite Sequences," *Journal of Number Theory* (Vol. 2, No. 2, pages 152-161; May, 1970).

Warmus, a Warsaw mathematician, did not publish his shorter proof until six years later in the same journal (Vol. 8, No. 3, pages 260-263; August, 1976). He gives a 17-point solution, and he adds that there are 768 patterns for such a solution, or 1,536 if you count their reversals as being different.

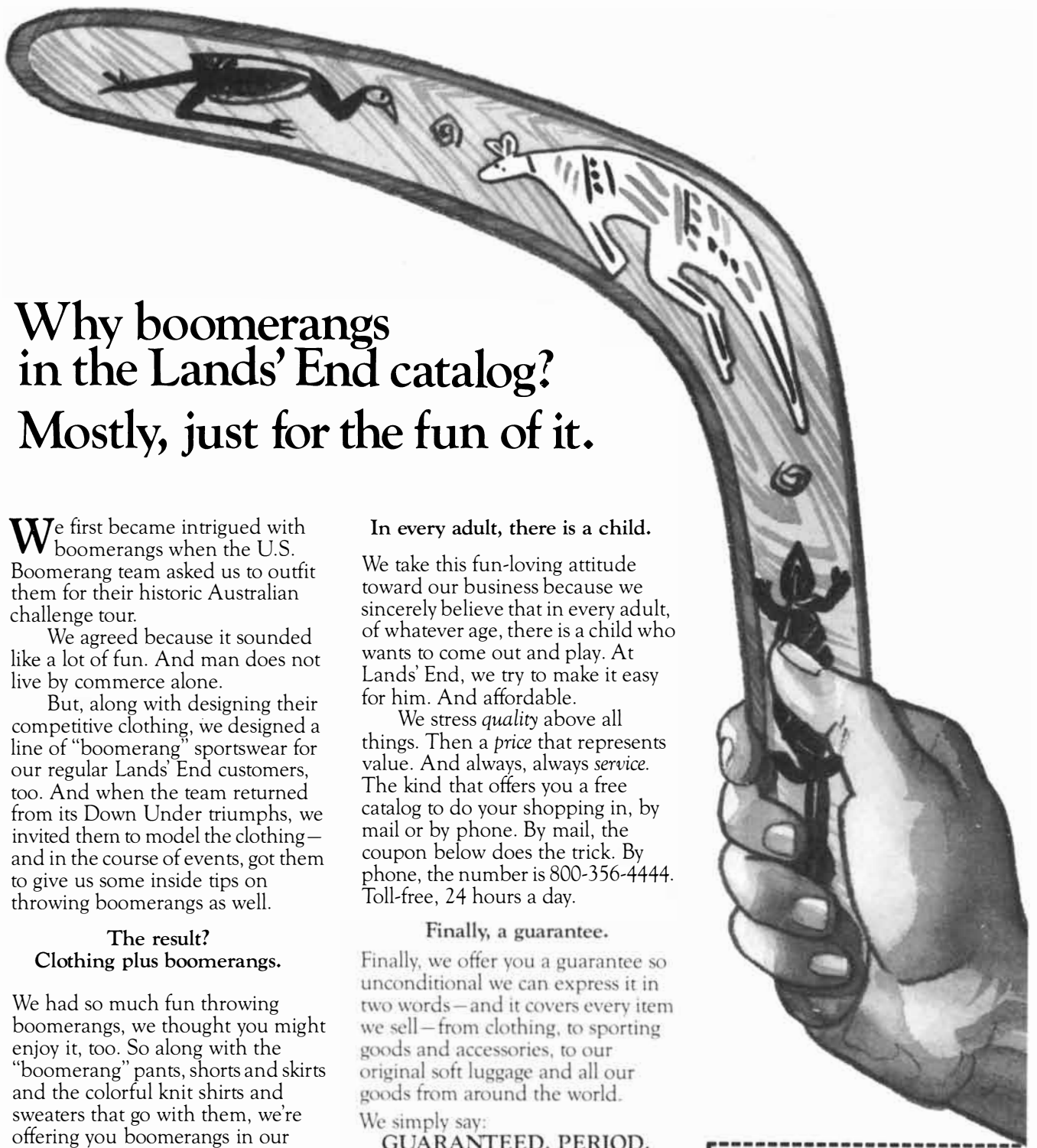
Our last example of a task that ends suddenly in a counterintuitive way is one you will enjoy modeling with a deck of playing cards. Its origin is unknown, but Graham, who told me about it, says that European mathematicians call it Bulgarian solitaire for reasons he has not been able to discover. Partial sums of the series $1 + 2 + 3 + \dots$ are known as triangular numbers because they correspond to triangular arrays such as the 10 bowling pins or the 15 pool balls. The task involves any triangular number of playing cards. The largest number you can get from a standard deck is 45, the sum of the first nine counting numbers.

Form a pile of 45 cards, then divide it into as many piles as you like, with an arbitrary number of cards in each pile. You may leave it as a single pile of 45, or cut it into two, three or more piles, cutting anywhere you want, including 44 cuts to make 45 piles of one card each. Now keep repeating the following procedure. Take one card from each pile and place all the removed cards on the table to make a new pile. The piles need not be in a row. Just put them anywhere. Repeat the procedure to form another pile, and keep doing it.

As the structure of the piles keeps changing in irregular ways it seems unlikely you will reach a state where there will be just one pile with one card, one pile with two cards, one with three and so on to one with nine cards. If you should reach this improbable state, without getting trapped in loops that keep returning the game to a previous state, the game must end, because now the state cannot change. Repeating the procedure leaves the cards in exactly



Trees for solitaire with three and six cards



Why boomerangs in the Lands' End catalog? Mostly, just for the fun of it.

We first became intrigued with boomerangs when the U.S. Boomerang team asked us to outfit them for their historic Australian challenge tour.

We agreed because it sounded like a lot of fun. And man does not live by commerce alone.

But, along with designing their competitive clothing, we designed a line of "boomerang" sportswear for our regular Lands' End customers, too. And when the team returned from its Down Under triumphs, we invited them to model the clothing—and in the course of events, got them to give us some inside tips on throwing boomerangs as well.

The result?

Clothing plus boomerangs.

We had so much fun throwing boomerangs, we thought you might enjoy it, too. So along with the "boomerang" pants, shorts and skirts and the colorful knit shirts and sweaters that go with them, we're offering you boomerangs in our catalog, too. Authentic Australian boomerangs—in keeping with our obsession with quality in everything you get from Lands' End.

And because fun, and the clothing to wear while having it, is at the heart of the business we're in as Direct Merchants, you won't be surprised to see a colorful spread of beautiful kites in fanciful designs, along with instructions on how to fly them. And another set of pages devoted to skate sailing and Lands' End clothing suitable on the water.

In every adult, there is a child.

We take this fun-loving attitude toward our business because we sincerely believe that in every adult, of whatever age, there is a child who wants to come out and play. At Lands' End, we try to make it easy for him. And affordable.

We stress *quality* above all things. Then a *price* that represents value. And always, always *service*. The kind that offers you a free catalog to do your shopping in, by mail or by phone. By mail, the coupon below does the trick. By phone, the number is 800-356-4444. Toll-free, 24 hours a day.

Finally, a guarantee.

Finally, we offer you a guarantee so unconditional we can express it in two words—and it covers every item we sell—from clothing, to sporting goods and accessories, to our original soft luggage and all our goods from around the world.

We simply say:

GUARANTEED. PERIOD.
And we mean it.

LANDS' END
DIRECT MERCHANTS

of fine wool and cotton sweaters, Oxford button-down shirts, traditional dress clothing, snow wear, deck wear, original Lands' End soft luggage and a multitude of other quality goods from around the world.

Please send free catalog.

Lands' End Dept. Q-21
Dodgeville, WI53533

Name _____

Address _____

City _____

State _____ Zip _____



Or call Toll-free:
800-356-4444

(Except Alaska and Hawaii call 608-935-2788)



Conjugate pairs 1,1,2,3,3 and 2,3,5

erations on any partition other than the consecutive one would return a diagram to its original state, you would have proved that all Bulgarian games graph as trees and therefore must end when their root is reached.

If the game is played with 55 cards ($k = 10$), there are 451,276 ways to partition them, so that drawing a tree would be difficult. Even the 15-card tree, with 176 points, calls for computer aid. How are these numbers calculated? Well, it is a long and fascinating story. Let us say partitions are *ordered*, so that 3, for example, would have four ordered partitions (usually called "compositions"): $1 + 2$, $2 + 1$, $1 + 1 + 1$ and 3. It turns out that the formula for the total number of compositions is simply $2^n - 1$. But when the partitions are unordered, as they are in the solitaire card game, the situation is unbelievably disheveled. Although there are many recursive procedures for counting unordered partitions, using at each step the number of known partitions for all smaller numbers, an exact asymptotic formula was not obtained until recent times. The big breakthrough was made by the British mathematician G. H. Hardy, working with his Indian friend Srinivasa Ramanujan. Their not quite exact formula was perfected by Hans A. Rademacher in 1937. The Hardy-Ramanujan-Rademacher formula is a horribly shaggy infinite series that involves (among other things) pi, square roots, complex roots and derivatives of hyperbolic functions! George E. Andrews, in his standard textbook on partition theory, calls it an "unbelievable identity" and "one of the crowning achievements" in the history of his subject.

The sequence of partitions for $n = 1$, $n = 2$, $n = 3$, $n = 4$, $n = 5$ and $n = 6$ is 1, 2, 3, 5, 7, 11, and so you might expect the next partition to be the next prime, 13. Alas, it is 15. Maybe all partitions are odd. No, the next partition is 22. One of the deep unsolved problems in partition theory is whether, as n increases, the even and odd partitions approach equality in number.

If you think partition theory is little more than a mathematical pastime, let me close by saying that a way of diagramming sets of partitions, using number arrays known as the Young tableaux, has become enormously useful in particle physics. But that's another ball game.

The Crown Jewel of England.



100% Grain Neutral Spirits.

© 1983 SCIENTIFIC AMERICAN, INC

BOOKS

Offshore oil adventures, biblical plants, the geometry of behavior, soccer madness

by Philip Morrison

OFFSHORE ADVENTURE: A PICTORIAL HISTORY OF THE NORWEGIAN PETROLEUM INDUSTRY, by Thorvald Buch Hansen, Odd Jan Lange, Håkon Lavik and Willy Håkon Olsen; color photographs by Leif Berge. Universitetsforlaget, distributed in the U.S. by Columbia University Press (\$44). NEW TECHNOLOGIES IN EXPLORATION GEOPHYSICS, by H. Roice Nelson, Jr. Gulf Publishing Company, Houston, Tex. (\$29.95). The long Norwegian coastline, deeply cut by the breathtaking fjords, stretches from the Skagerrak 1,000 miles north well across the Arctic Circle to Nordkapp. Offshore all the way lies a shallow continental shelf, the largest in Europe, three times the land area of Norway under a couple of hundred feet of seawater, stormy, cold but not ice-covered. Under the summer sun the sea glows blue all the long northern day; in winter gloom the storm waves can crest 30 feet high. As you read these lines there are about 3,000 people hard at work out there, heavy industrial work, around the clock.

What they do, of course, is win oil and gas. They work mostly at the southern end of the long coast, close to the median line across the North Sea where Norway's claim to the shelf abuts Britain's. Twelve hours on and twelve off they work, on great steel decks held high above the waves, roughnecks straining at cables, tool pushers at the complex consoles of the central control room. It is Norwegians who now carry out, two weeks at sea for three ashore, the oil-field tasks first developed amidst the bayous and along the Texas and Louisiana coasts. Helicopters by entire squadrons carry the crews and the mail into and out of Stavanger; the powerful workboats fetch pipe and cement and drilling mud and fresh water. Out flows the oil to the big tankers moored at the loading buoy far out of sight of any land; the gas is piped across the shallow sea floor not to Norway but to foreign landfalls at Teesside and Emden. Closer to the coast the shoreline is everywhere flanked by the deep Norwegian Trench, and a pipeline crossing that 1,000-foot depth was beyond the technology of the

1970's, although one is now being laid from new fields.

In *Offshore Adventure* clear text and strikingly beautiful photographs document the Norwegian oil industry in human, technical and institutional terms since its beginnings in 1969, when a Phillips Petroleum mobile drilling rig struck oil and gas in the Ekofisk field 250 kilometers out of Stavanger; that notable flare is shown lighting the sea. Production began on the Ekofisk in 1971. Statoil, the national oil corporation and a major operator, is now able to mark its 10th birthday, in part by sponsoring this book. The scale of the Norwegian industry is not large by world standards; some 40 percent of Norwegian oil comes to the U.S., where it slakes only a 30th part of the American thirst for imported oil. Still, that earns a tidy couple of billion dollars a year, about the value of what the Norwegians extract for themselves.

Close up, this is an industry on a dramatic scale. One photograph shows a rigger handling the block and hook of the biggest construction crane in Stavanger. The smooth form of forged steel weighs 109 tons; its great flukes engage a dozen or so loops of eight-inch steel hawsers, able to lift a hotel module of eight floors of accommodations, ready to place on the platform deck. The Statfjord B platform, the biggest to date, reaches almost 900 feet from sea bottom to crane tip. The hexagonal array of concrete storage and ballast tanks that forms its heavy base is first fitted with four towering hollow concrete columns. The entire assembly is then sunk in a fjord; it is pumped out with care to rise vertically under a 40,000-ton steel deck structure set afloat on barges. The columns lift the platform high above the level of the water for permanent service. One grand picture records the day *Queen Elizabeth II* sailed into Stavanger harbor. Beside her floated the Statfjord B steel deck, with much of the liner's look, all sheer white walls and regular openings, and across the harbor the four columns awaiting the deck towered far above the city skyline. By now those shipyards have poured seabound concrete by the megaton.

Another photograph evokes the inhospitality of the North Sea: a helicopter looks down at a layer of undifferentiated whiteness, out of which eerily feather three crane ends, a tower and a gas flame 30 feet long. Then there are the specialized workboats, powerful and wonderfully maneuverable, shipping a crosswise-mounted reversible propeller as well as the usual propulsion screws aft, able to turn on a silver krone, to judge by the wakes shown from overhead. That the local shipyards of western Norway were able to build drilling platforms competitive around the world and a line of supply vessels salable to the Russians and the Japanese is part of the Viking story: there is no drawing-board surrogate for experience with the powers of the ocean. Statoil itself has a new consulting office in Beijing, where it offers expertise to the Chinese state oil company on offshore exploration.

There have been troubles, both the tragic and the merely grave. The open sea is unforgiving even to platforms. The Bravo blowout on the Ekofisk in 1977 drew Red Adair and Boots Hansen to stem the wild flow, but not before half a tankerful of oil had stained the sea; the safety management had been inadequate. The safeguards mount. In 1980 the living platform *Alexander L. Kielland*, a converted drilling rig, had one of five steel legs torn off in a near-hurricane. A hundred and twenty-three people were lost, many while watching the evening movie; the verdict was a bad weld. The safeguards are never enough. The book gives more detail about unions, safety, finance and company structure than it does about technology, but the pictures display the engineering well. Only the diving technology, reportedly now mounting a very spirited response to the sea's cold challenge, is slighted. One also misses an index.

The third-largest seismic exploration company in the world is GECO, Geophysical Company of Norway. At this point the popular account of Norway's seaward oil makes contact with *New Technologies in Exploration Geophysics*, an insider's sharp appraisal of the turbulent state of the art of seeking invisible oil underground. Published first as a set of lively and opinionated journal articles, the book is ephemeral, but it gives the reader a quick look at an ebullient segment of contemporary high technology.

Oil is worth finding. Finding it depends on many remarkable techniques of geophysics. Yet 95 percent of all expenditures in this wealthy industry on geophysics are made in the area of reflection seismology. The sources of the seismic signal are classically small explosions, now augmented with air bubbles underwater and with vibrating plates on land, four heavy trucks shaking in synchronized FM. A hundred geo-

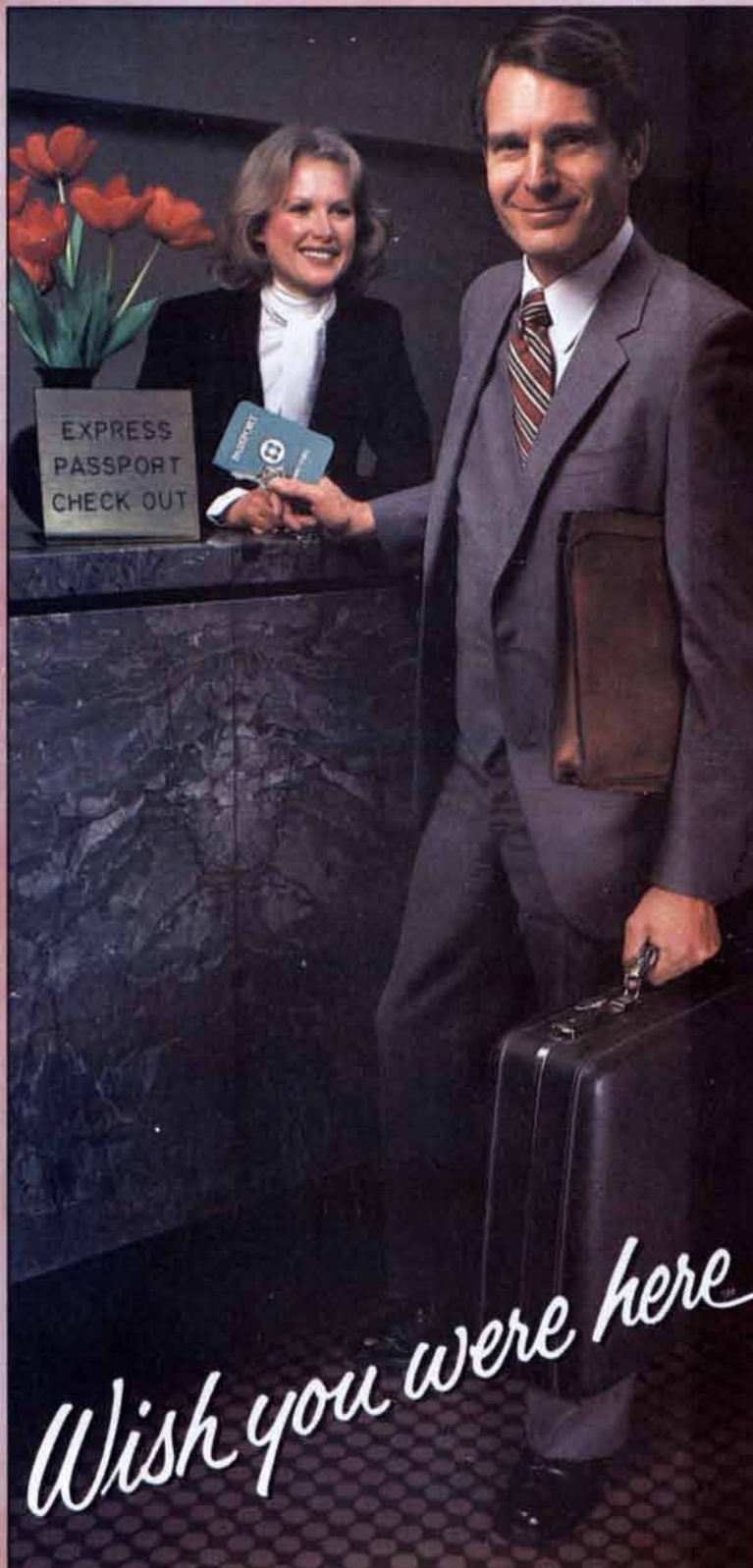
Why we make checking out of a Hyatt just as pleasant as staying at one.

The way we figure it, we give you so many extra touches at Hyatt, it would be a pity to have a long check-out line spoil them for you.

That's why when you check out of a Hyatt you simply hand in the passport you received when you checked in. You're checked out. Your bill is mailed to you within 24 hours, so there's absolutely no waiting in line. Absolutely no way you can forget how pleasant your stay was.

Any hotel can give you the essentials, but it's those extra touches... like express check-out, that make the trip.

For reservations at 109 hotels worldwide, call your travel planner or 800 228 9000.



Wish you were here

Chicago

In Illinois Center overlooking Michigan Avenue and Lake Michigan.

Dallas Soar above the lively downtown Reunion area at Hyatt.

Miami Hyatt is next to the Miami Convention Center with its sophisticated communications capabilities.

New York Grand Hyatt on fashionable Park Avenue at Grand Central.

San Francisco Hyatt on Union Square is an elegant hotel, located in the heart of the city's finest shopping.

HYATT  HOTELS

phones are set out to record every costly pulse. A crew will have on the frugal end a dozen or two people at work, and 100 or more where labor costs are low and many receivers are worthwhile. Where once all was cables from the dynamic mikes, nowadays the long dispersive trains of waves are collected by radio telemetry. Meanwhile channel multiplexing grows apace, and the bit stream becomes a digitized flood. The processing of three-dimensional seismic volumes, often in a hurry to learn where the likely blocks lie before someone else does, may need to handle a gigabyte per day per complex crew. Naturally the computer becomes a bottleneck. These firms are good customers for the vector-array processors feeding the supercom-

puters of Control Data and Cray. The aim is an interpretation of all those endless wiggles to map a physical volume with recognizable strata, all their folds in proper place.

One of the many photographs here shows the multilayer model of loaded silicone rubbers built to scale for the Staffjord production area. Once finished, the variable-density sculpture in the tank yields admirable scans to sonic probes under computer control. The interpreters learn fast. The next step is interactive interpretation at the terminal; new ingenuities of 3-D presentation are coming on stream. Right now there is a scheme of building a virtual volume of images by a vibrating mirror synchronized with the high-resolution

video output from the computer. The color pictures reproduced here are impressive, even though they lack the 3-D impact.

All these data have to be stored somewhere, and the big roll stuffed on the laboratory shelf with those old surveys from the Falklands, say, seems a poor match to the FM shakers, the telemetry and the Cray. Still the reader is not quite prepared for the complaint that the footprint of the INMARSAT data-transmission satellite system misses Oklahoma City, Denver and Calgary, "one of the problems that have kept satellite data transmission out of the oil patch." One gathers that the satellite link is on the way. It would be hard to suggest any other book that so clearly paints the

RENAULT TURBO



GIVE YOURSELF A WHOLE NEW SLANT ON PERFORMANCE—A TURBO SPORTS COUPE

image of an economic technology in such a full flood. Knowledge is bits, and enough bits is new oil. The last chapter suggests that the university supply of the trained people to examine all those costly, swift 3-D images calls for close attention and generous money inputs, not high corporate bids to lure away the best faculty. Is there a plan?

PLANTS OF THE BIBLE: A COMPLETE HANDBOOK TO ALL THE PLANTS, by Michael Zohary. Cambridge University Press (\$16.95). Three little golden piles lie side by side in the photograph. They are the ripe grains of barley, emmer wheat and einkorn wheat, still grown in some fields in the Holy Land as they have been for 80 centuries. "Unlike the

modern cereals, these grains still bear their shaft after threshing." Here too are ears of the wild wheat, frustrating to the would-be domesticator because the heads tend to shatter and disarticulate into single spikelets that fall to the ground before the harvest. The epic search for a mutant form that would not shatter became the domestication of wheat; those tough little bits of shaft were a small price to pay for the wholeness of the ear.

The ethnobotany in Scripture is plainly matchless for its literary quality and influence. It offers a rich view of one agricultural world, at a time about 3,000 years back and a place near the very center of origin of the crops and practices that still nourish our lives. Profes-

sor Zohary is a distinguished lifelong student of the flora of the Middle East; his own photographs add beauty and conviction to his careful plant-by-plant account of 128 species from the biblical green world. The Bible—Old Testament and New Testament together—names only 110 species; the others arise from the recognition that entire groups of plants, for instance the thorny and thistly ones, are often implied by a single term.

Plants enter the texts in everyday ways, in accounts of ritual and often unforgettably as figures of speech. Figs are still not gathered from thistles; plainly not every Hebrew or Greek usage in the long texts can be unambiguously matched to a plant species, even call-

FUEGO

Renault Fuego Turbo Fuel-injected intercooled turbo with 11.6 psi maximum boost 5-speed overdrive transmission Front wheel drive Power-assisted front discs Michelin TRX radials on cast aluminum wheels Air conditioning AM-FM stereo radio with quad speakers Tach-to-turbo gauge telemetry Leather-wrapped steering wheel, and more \$11,095*
*List price Tax, license, destination charges, optional or regionally required equipment extra

American Motors

FROM THE PEOPLE WHO TURBOCHARGED FORMULA ONE RACING.

RENAULT
THE ONE TO WATCH

ing on the ancient translations into the Greek, the Latin, the Aramaic and the Syriac, works much older than the European renderings and less subject to the natural error of naming biblical plants by European analogues. Botanical expertise is the major tool of the task, and to its aid the learned author, a botanist long steeped in the voluminous philological record and its living counterparts, brings the cognates, particularly in Arabic, still in use on the land.

There is no balm in Gilead, and there never was. The pungent balsam resin comes from the small fruits of a thorny desert shrub. The spade has yielded up the equipment for commercial extraction of the balm within the ruins of En Gedi on the Dead Sea. Judea was famed in antiquity for its balsam. We can forgive the scholars of King James's time, who ought to have written the name of a larger gum tree, the storax. The storax too is not found now in Gilead (the central region of the high plateau across the Jordan), but it seems probable that it grew there once, as it grows and is slashed for a sticky exudate in highland Anatolia today.

The tares sown among the wheat "while men were sleeping" was not mere parable; the grass darnel is shown in a clear photograph. Its grains are so much like those of wheat in size and shape that they cannot be removed by sieve or winnowing fan, and they may embitter the ground meal. Like rye and oats, Syrian scabious, a second form of tare, has become a crop plant on its own, although an unwanted one. It belongs to the teasel family, but long centuries of unintended selection have formed it into a plant whose seeds are indistinguishable from those of wheat. Not a stealthy enemy but the husbandman himself has unwittingly sown, reaped, threshed and sieved the tares; sometimes now he can take only a bitter harvest, the unwanted crop overwhelming the desired one.

This delightful book is handsome, inexpensive and compact. It is an ornament among the ponderous old volumes on its much-gleaned topic. In addition to the main plant list, with a page and a photograph or two for each species, it provides an admirable apparatus of indexes, maps and an introductory textual history, geography and ecology. Not every question is answered; one would like more than one can have. The forbidden fruit of the tree that was in the midst of the garden is simply not identified in the text. It may have been an apple, as widespread tradition tells us, since apples have been grown there, "eastward in Eden," toward Turkey and Syria, since about 4000 B.C., but the right word does not stand on the page. Manna is still a problem too. A sweet white exudate of the hammada shrub, a plant widespread in the southern Sinai, is shown here. A

decade ago the Bedouins were still collecting those hardened droplets to use in cakes. But quantitative calculation fails; our author concludes that all the possible sources of exudate together could not provide "much more than a tidbit for the hungry people." It is no wonder that the word *man-ha* means "What is that?"

DYNAMICS—THE GEOMETRY OF BEHAVIOR: PART 1, PERIODIC BEHAVIOR, by Ralph H. Abraham and Christopher D. Shaw. Aerial Press, Inc. P.O. Box 1360, Santa Cruz, Calif. 95061 (\$29). This book is an aspiring experiment in pedagogy, a systematic and extended introduction to an important branch of mathematics, one that promises a wide range of applications in addition to its powerful results of principle. It is called dynamics, since it grew from the problems and the methods of classical mechanics, turned much more general, undaunted by the nonlinearity of the real world. Its aim is to model a wide variety of systems by geometric means, seeking not the transient details of change but the overall outcome: "prediction forever" for systems from biological clocks to interacting species, parts of the world not to be grasped by the simpler formalism of the past.

The experiment is the presentation of this rich discipline at an ambitious if beginning level, almost entirely visually. The brief verbal explanations are in the form of captions to a long and steadily growing series of uniformly presented and color-coded line diagrams. The abstraction gains strength from the generality and growth of the sequence as a whole; it is revealed by pages drawn in a way recalling the pseudo realism of the thought experiments dear to Niels Bohr. The oscillators are boxes replete with cables and meters, the pendulum swings by a sturdy bearing, and so on. Capsule biographies with images of the great pioneers from Galileo to Lord Rayleigh and Balthasar van der Pol open the book, and an occasional rock group or electrocardiogram subject appears in an enlivening caricature. Nearly all formulas are suppressed in the text, to appear safely caged in a tight appendix.

About a third of the way into the volume we see a familiar system, the simple pendulum. The representation has opened up the to-and-fro and the encircling swings into a flow mapped in the phase plane, rotation rate against angle. Friction is no difficulty for these geometric methods; the central point is revealed as an attractor, toward which every nearby trajectory in phase will sooner or later come. By two-thirds of the way the phase plane has grown to three dimensions, representing a pendulum and the phase of its forcing oscillator, and even to four, although now a second

reduction occurs, and a two-dimensional phase diagram for coupled oscillators is open to study on the page. The discourse is rich enough to contain the phenomena of entrainment, with its braided orbits, repellers caught between attractors as they wind. A fuller account draws out the entire invariant torus in the three-dimensional phase space, we see basins and their separatrices and we arrive at the edge of classical dynamics, as it stood in about 1950. Ahead are two more volumes of this Visual Mathematics Library, one on stable and chaotic systems and the other on the bifurcation theory and its classifications of dramatic change.

The experiment is not to be evaluated yet; the authors are persuaded of its importance in our fragmented culture, and eager for the success of their nonmathematical readers. They do, however, have very high standards for the visual grasp of complicated meaning. Ralph Abraham is professor of mathematics at the University of California at Santa Cruz, a chief contributor to the growth of modern dynamics, and Christopher Shaw is an artist and illustrator of diverse training and interest in science. Their work is a challenge above all to the small industry of textbook authors and instructors in beginning college physics and mathematics everywhere. In a digitally powerful world, in a world that needs and builds models for systems far from that lamp swinging gently in Pisa, ought we not to begin to enrich the exact solutions of simple systems, all those sines and cosines and linearized frictionless models, with the masterful if approximate geometries of our own day? This important book sets out one path to that future.

SOCCER MADNESS, by Janet Lever. University of Chicago Press (\$17.50). The enormous ring is the Maracanã Stadium in Rio de Janeiro, largest in the world; it can hold 220,000 fanatic enthusiasts, separated from the playing field—and the referee—by a 10-foot moat. It is the flower of Brazil's garland of 4,000 stadiums, with a seating capacity overall about a third more than all the "civic and private stadia, bowls and ballparks" of the U.S., even though Brazil has about half our population. Brazil is first in the world for stadiums and for soccer.

Dr. Lever is a reflective sociologist whose functional study of this phenomenon arose by chance but is now solidly set in a decade of inquiry in the field. She has interviewed everyone from the peerless Pelé to the earnest fans, including a sample of 200 working-class men in Rio. Though her sample was not a properly random one, she capably took pains to see that it mirrored the city census for age, marital status, religion, race and income

within that class. Virtually every male followed professional soccer. Only one of the men in the sample had not watched the television display of the legendary final World Cup match Brazil won in 1970, a victory followed by a welcome back to Rio that became a national holiday, brought a crowd of two million into the streets and set off a two-day carnival ecstasy that saw 44 dead and 1,800 injured. A fourth of the men in the sample were strong fans who listen to every soccer game they can, go to at least a couple of games a month and score nearly 100 percent on the test items about players set in the interview. Half of the men in the sample make soccer a routine part of their life, see a game a month and miss out on some of the trickier questions.

How did it happen? The Fédération Internationale de Football Association (FIFA) has ruled the global professional sport since 1904; it binds 147 national members, second only to the United Nations. Soccer—most of the world calls it football—is clearly a handcrafted British product of high quality, like parliaments and the Rolls-Royce, that became dispersed over the world in the halcyon days of imperial Britain. The port cities of Latin America, then hosts to large British colonies, acquired the game early. Everywhere it moved from being a sport of the gentry to being a professional game appealing to the working masses; it is still the case that the volunteer managers of the teams are amateurs, in contrast to the big business of the gate receipts, the national football lottery and the television coverage. Among the top officers of FIFA only a full-time secretary general in Zurich draws a salary. The U.S. offers “an anthropological oddity”: our mass team sports are mostly, if not yet entirely, businesses run for profit. In Brazil and most other countries the football club is really a club, usually a money-losing affair with a shifting set of volunteer patrons and officials and a devoted dues-paying cadre of supporters, who may benefit from athletic facilities and others open to members. A team could hardly move to another city: civil war!

It was the play of history that seems to have made football so preeminent in Europe and Latin America. The game arrived suddenly, to find nearly everywhere no indigenous team sport on the scene. Its simple rules, its open, visible, varied and uninterrupted flow and its freedom from equipment allowed it to catch on. In the U.S. and in commonwealth countries such as Australia and India, however, professional baseball or cricket were already in place. In most countries soccer was the first professional team sport and remains the only one. It is the first team sport for amateur participation as well; indeed, there is no bet-

ter predictor for a Rio man's fandom than his early personal engagement in the sport. It is not only the players who are fostered on the sandlots but also the lifelong fans.

Dr. Lever sees soccer as a profound and benign paradox of social cohesion in Brazil. It ritualizes conflict, both that on the field and that in objective social reality. The strongest rivalries in Brazilian soccer are intracity ones. There are a dozen Rio soccer clubs. Flamengo is the team of the masses, its followers millions of the urban poor, even of the steep favelas. Fluminense is thought of as a club of the elite; its applicants are screened. Criminals and the handicapped are ineligible for the Fluminense club, except “those maimed while fighting for their country or while in the service of the [club].” Its own players, “worshiped on the field,” are barred from most of its social activities. This dramatic and genuine conflict is played out, however, within rules. Picture a game between Flu and Fla, 100,000 fans of each filling Maracanã, giant flags, firecrackers, confetti, team shirts, intensity—and one referee.

Soccer is no mere opiate. It integrates the nation, bonding as it divides, serving both pluralism and unity. Intercity U.S. sport emphasizes integration; Brazilian intracity soccer dramatizes cultural pluralism, in a playful sublimation of sharp conflict. Such rule-bound drama does not survive times of intense social unrest; games in Northern Ireland see the warring communities kept apart in the stadium by barbed wire and riot preparations, and the 1969 football playoff between Honduras and El Salvador, at a period of deep animosity over jobs and immigration, catalyzed a month of armed combat with thousands of casualties. The sport promotes nation building, and nationalism is the most passionate and broadest loyalty of our time. Indeed, governments can and do exploit sport; the author does not blink that fact, but she judges that the game is worth the candle, that the glowing reality of voluntary organization and civil unity outweighs the shoddy propaganda ploys of a transient regime.

There is a yawning gap in the map of unity. In Brazil football is not for women. They belong in church, men only in the stadium. The sample confirms this in detail, as do the stated views of Brazilian fans at every level. There are signs of change; Pelé has spoken up for it. Sandlot games with girls taking part will mark the real change, whenever it comes. Lever has given the reader a small book as well written as it is thoughtful; the role of sport in human society is deserving of more study, and this account is a happy example painted in the bright colors and sharp contrasts of Brazilian life.

Introducing

PABSoft

MATH MASTER SERIES

Matrix Master \$24.95
+, -, *, /, trn., inv., IO, etc.

Poly Master \$24.95
+, -, *, /, int., diff., etc.

- * Callable from BASIC
- * Fast Machine Code
- * Loads Routines as needed
- * TRS-80 MODS I, III, IV, Apple II+, IBM PC
- * Complete Math Library
- * When ordering, please describe your system, (RAM, DOS, etc.)
- * Send for Scientific—Engineering Catalog

PAB Software, Inc.
P.O. Box 15397
Fort Wayne, IN. 46885
(219) 485-6980

Master-Visa accepted

Trademark Credits
TRS-80: Tandy Corp.
Apple: Apple Computer Corp.
IBM: IBM Corp.

BINOCULARS



Precision Optics From
CELESTRON®

Our Giant Series Binoculars (80mm diameter!) allow you to explore the Moon and Outerspace as well as your own backyard. Celestron also offers extremely compact roof prism and classical binoculars, and superb quality Telescopes, Spotting Scopes and Telephotos. FREE product, price and ordering information by returning the coupon or send \$3.00 for 52-page color catalog.

CALL TOLL FREE 1-800- 421-1526

CELESTRON INTERNATIONAL
P.O. Box 3578-SA, 2835 Columbia St.
Torrance, CA 90503

Name _____
Address _____
City _____ State _____ Zip _____

Trauma

Accidental and intentional injuries account for more years of life lost in the U.S. than cancer and heart disease. Among the prescribed remedies are improved preventive efforts, speedier surgery and further research

by Donald D. Trunkey

Trauma is the medical term for a personal injury or wound. Including both accidental and intentional injuries, physical trauma is the principal cause of death among Americans between the ages of one and 38. In 1982 there were about 165,000 deaths from trauma in the U.S., and for each death there were at least two cases of permanent disability. Statistics compiled by the Department of Health and Human Services indicate that for Americans between the ages of 15 and 24 the combined death rate from motor-vehicle accidents, homicides and suicides has risen by 50 percent since 1976. Among young whites motor-vehicle accidents are the leading cause of death, accounting for about 40 percent, whereas among young blacks homicide is the leading cause of death, accounting for approximately the same percentage. In large cities black males have a one-in-20 chance of being murdered before the age of 30. Increased urban violence has been a major contributor to the rise in the national homicide rate: from 8,464 in 1960 to more than 26,000 in 1982. Overall the death rate for American teenagers and young adults is 50 percent higher than it is for their contemporaries in other industrialized societies.

Because trauma primarily affects people at or near the beginning of their most productive work years, its cost measured in lost productivity from both death and disability is high: more than \$63 million per day in lost wages from accidental trauma alone, according to a recent estimate by the National Safety Council. The total annual cost of accidental trauma, including lost wages, medical expenses and indirect work losses, comes to about \$50 billion.

Trauma patients currently take up a

total of about 19 million hospital days per year in the U.S., more than the number needed by all heart-disease patients and four times the number needed by all cancer patients. In the past decade the death rate from heart disease and stroke has fallen by 22 and 32 percent respectively. In contrast the death rate from accidents has risen by about 1 percent per year since 1977. Trauma is clearly a major medical and social problem in the U.S. To a large extent, however, it is being neglected by physicians, hospital administrators, government officials and the general public.

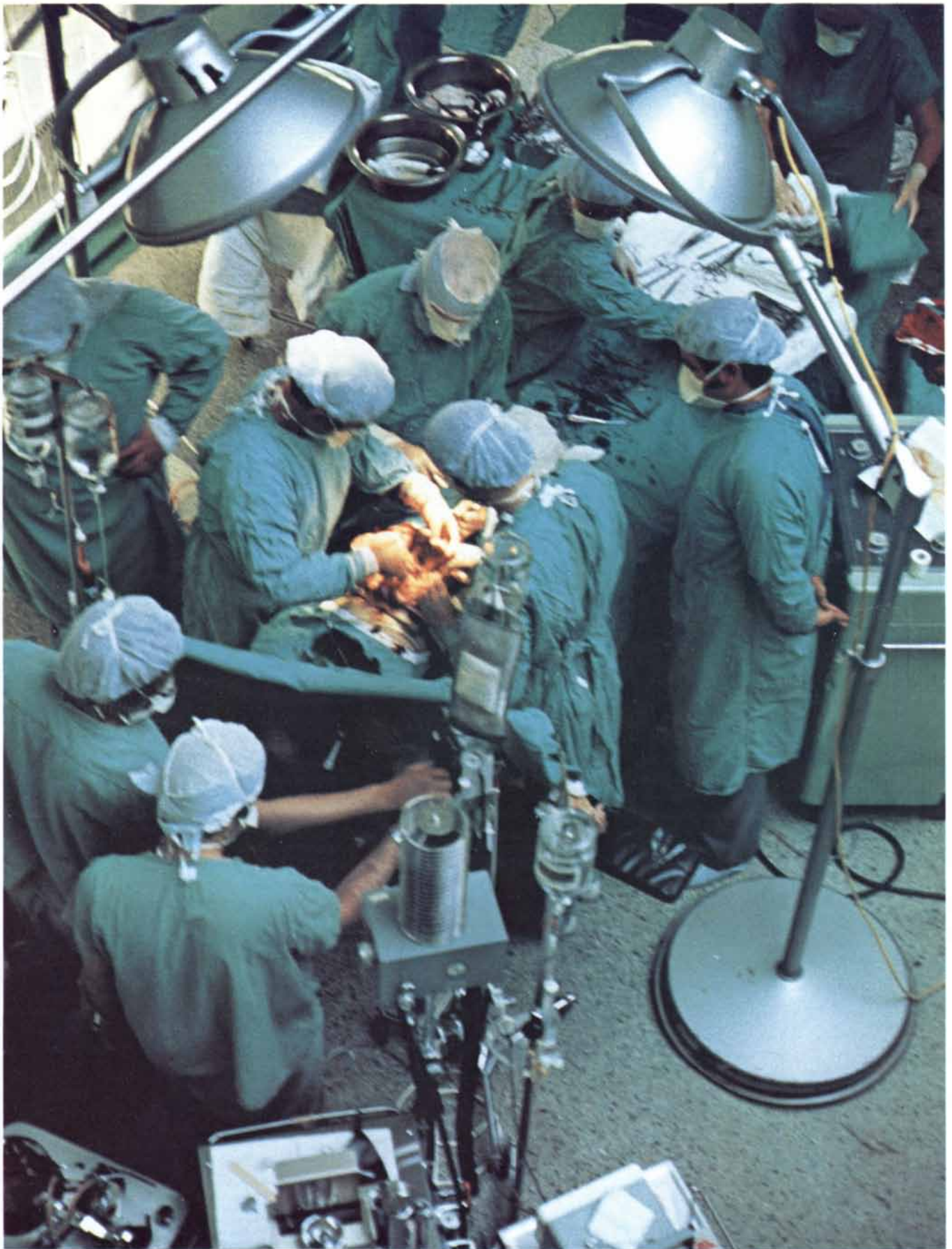
Data from several parts of the country show that death from trauma has a trimodal distribution: when the death rate is plotted as a function of time after injury, three peaks appear in the resulting graph [see illustration on page 31]. The first peak, characterized as "immediate deaths," represents people who die very soon after an injury. Invariably these deaths are caused by lacerations of the brain, the brain stem, the spinal cord, the heart or one of the major blood vessels. Only a fraction of the patients in this category could in principle be saved, even under the most favorable medical conditions.

The second peak, characterized as "early deaths," represents people who die within the first few hours after an injury. These deaths are usually caused by major internal hemorrhages of the head, the respiratory system or the abdominal organs, or by multiple lesser injuries resulting in severe blood loss. Almost all injuries of this type are considered treatable by currently available medical procedures. The interval between injury and definitive treatment, however, is critical to the probability of recovery.

The third peak, characterized as "late deaths," represents people who die days or weeks after an injury. In almost 80 percent of these cases the cause of death is either infection or multiple organ failure. Here time is less of a factor than the quality of medical care and the extent of medical knowledge. In what follows I shall discuss the pathology of each peak in somewhat greater detail, with particular reference to the prospects for reducing the rate of mortality and disability resulting from the associated set of medical conditions.

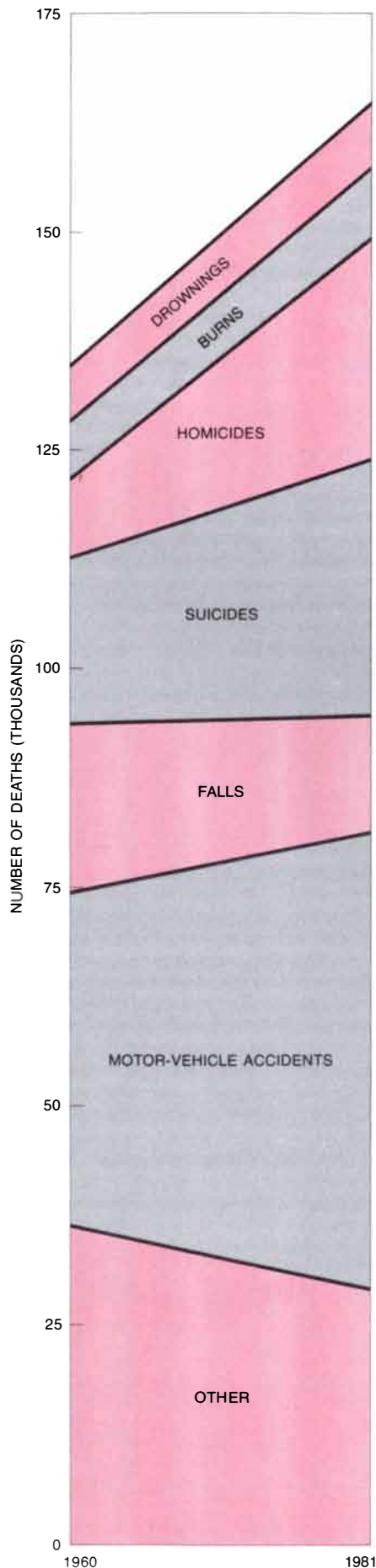
More than half of all trauma deaths are classified as immediate. The small number of patients in this category who could be saved are those in the few large cities where rapid transportation is available and special facilities called trauma centers are in operation. A trauma center is a hospital where the medical staff has made a commitment to provide 24-hour "in house" coverage by surgeons, anesthesiologists and supporting staff to care for trauma patients. Recent medical records from two of these centers, one in Seattle and the other in San Francisco, indicate that if there are signs of life at the scene of an accident or on the way to the hospital, 20 percent of the patients who are classified as "dead on arrival" can be resuscitated in the emergency room and will eventually leave the hospital without permanent neurological damage.

This remarkable rate of recovery will probably never be achieved in most suburban and rural settings, because of the longer time it usually takes there between injury and definitive treatment. The only way to reduce the number of immediate deaths in these circumstances is through prevention. Perhaps



OPERATING ROOM at San Francisco General Hospital is the scene of efforts by surgeons, anesthesiologists and other specialists to treat a critically wounded patient. Medical records from trauma

centers established in such major metropolitan hospitals indicate that the elapsed time between an injury and definitive surgical care is a critical factor in determining the survival rate of trauma victims.



as many as 40 percent of all deaths from trauma could be averted by the introduction of various prevention programs. Most of these programs involve controversial social issues, however, and so their chances of success are unpredictable. I shall cite here just a few of the more important trauma-prevention programs that have been proposed.

According to the Insurance Institute for Highway Safety, between 50 and 60 percent of the fatal motor-vehicle accidents in the U.S. are caused by drunk drivers. Efforts to reduce drunk driving by increasing the penalties for infractions have generally failed in the U.S., and similar programs in Europe have had only mixed results. For example, reports from a number of Scandinavian countries indicate that after mandatory jail sentences for drunk driving were imposed a significant reduction in fatal accidents was observed. In time, however, there was usually a reversion to the same mortality rate that had prevailed before the stronger measures were introduced. Rehabilitation programs for drunk drivers, introduced in several parts of the U.S., have also been found to be ineffective.

In spite of this generally negative record there is some evidence that the suspension or revocation of a driver's license after a drunk-driving conviction can have a significant effect on the subsequent rate of drunk-driving arrests in the affected population. Recently a "grass roots" group called Mothers against Drunk Drivers (MADD) was organized in California to promote such stronger measures to reduce the carnage caused by drunk drivers. The impact of this campaign, which is now spreading to other parts of the country, remains to be seen.

Another vexing social issue with a bearing on the current rate of trauma focuses on the mandatory use of safety devices such as automobile seat belts and motorcycle helmets. Legislation requiring the use of seat belts has been introduced in at least 20 countries. The results of these measures vary, depending on the degree of enforcement and compliance. So far the best record has been achieved in Australia, where after a law was passed requiring the use of seat belts there was a 27 percent decrease in the death rate from motor-

vehicle accidents. Mandatory seat-belt legislation has not been popular in the U.S., however, in part because of the active resistance of groups opposed to such forms of Federal regulation.

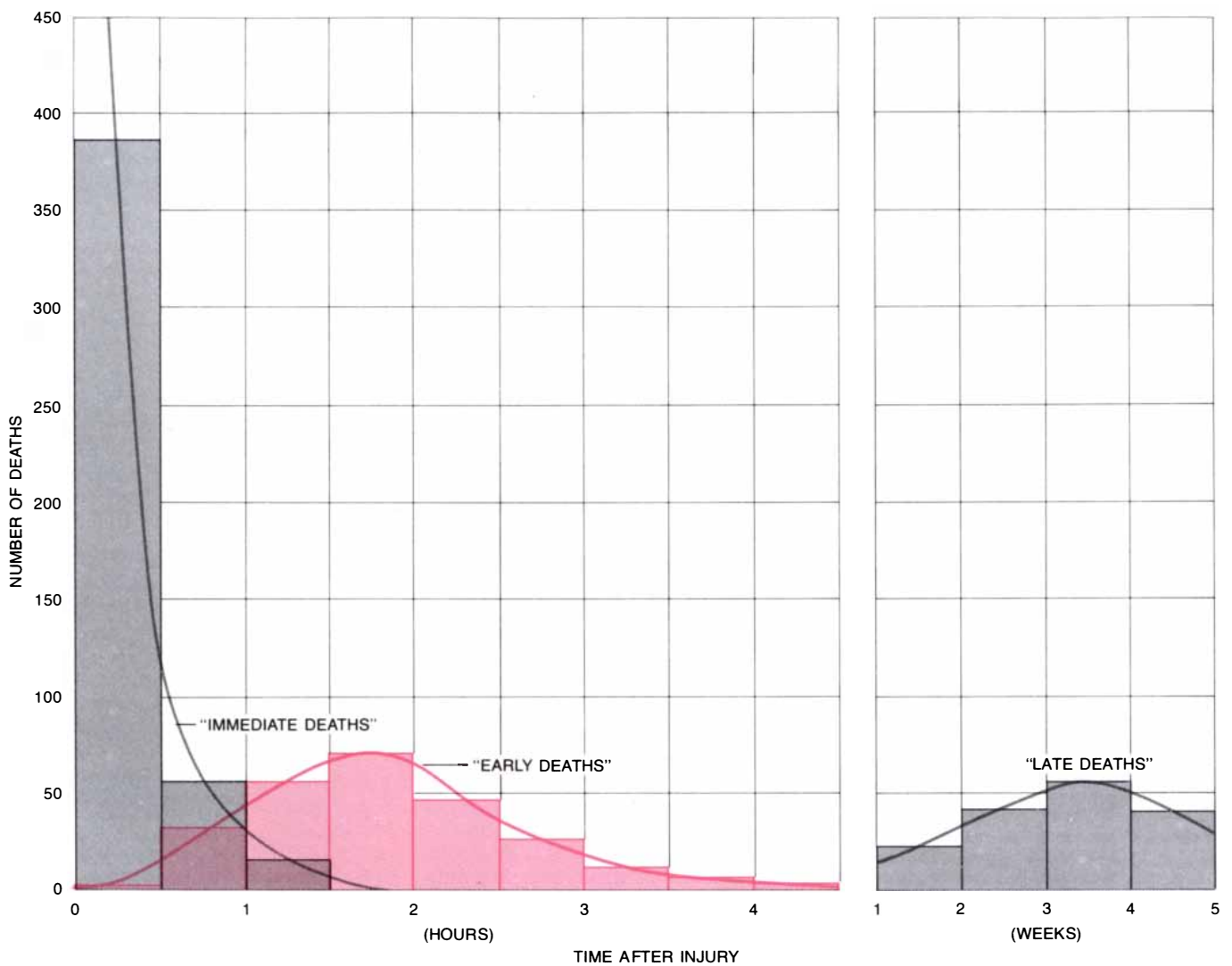
The situation is equally unsettled in the case of laws requiring the use of motorcycle helmets. Beginning in 1967 a Federal highway-safety standard required that all states enact and enforce motorcycle-helmet laws. In the next decade fatalities from motorcycle accidents decreased by 50 percent nationally. Then in 1976 Congress revoked the sanctions against states not in compliance with the Federal standard. Over the next three years 27 states repealed or weakened their motorcycle-helmet laws. The result so far has been a 40 percent increase in the death rate from motorcycle accidents in those states. A recent study sponsored by the National Highway Traffic Safety Administration concluded that "the use of a safety helmet is the single most critical factor in the prevention or reduction of head injury" from motorcycle accidents.

The burden placed on society by unhelmeted motorcyclists can be demonstrated. In one study of 71 motorcyclists admitted to Denver General Hospital it was found that only 38 percent were covered by commercial medical insurance or workman's compensation; most of the unpaid bills were borne by the taxpayers. Similarly, in a survey done at the Maryland Institute for Emergency Medical Services it was found that 40 percent of the 65 motorcyclists hospitalized did not have insurance coverage.

Perhaps the most controversial trauma-prevention issue in the U.S. is that of handgun control. According to advocates of greater restrictions on the availability of handguns, more than 11,000 of the 26,000 murders recorded in the U.S. in 1982 were committed with such weapons. In addition, suicides and accidents involving handguns accounted for at least 10,000 deaths. Both figures are extraordinarily high, particularly in comparison with other industrialized societies, where handguns are controlled. There are at present some 60 million handguns in the U.S., and it can be argued that it would be difficult, if not impossible, to eliminate them entirely, even if the political will to do so were to prevail in Congress. In the meantime other approaches to handgun control, such as the institution of mandatory jail sentences for the criminal use of a handgun, might help to reduce the death rate from gunshot wounds. Laws of this kind have been passed in several states, and the results, particularly in Michigan, seem to be quite positive.

Another controversial issue is that of the decriminalization of narcotic drugs. In the 15 years that I have worked

TRENDS IN MORTALITY from trauma in the U.S. are plotted according to the cause of death in this graph, representing data gathered by the National Center for Health Statistics. The figures for 1960 are either actual totals for that year or averages for the period from 1952 through 1963; figures for 1981 are based on a 10 percent sample of that year's deaths.



TRIMODAL DISTRIBUTION of trauma deaths is observed when the death rate for a large enough sample of such deaths is plotted as a function of time after injury. The first peak (*“Immediate deaths”*) corresponds to people who die very soon after an injury; the deaths in this category are typically caused by lacerations of the brain, the brain stem, the upper spinal cord, the heart or one of the major blood vessels. The second peak (*“Early deaths”*) corresponds to people who

die within the first few hours after an injury; most of these deaths are attributable to major internal hemorrhages or to multiple lesser injuries resulting in severe blood loss. The third peak (*“Late deaths”*) corresponds to people who die days or weeks after an injury; these deaths are usually due to infection or multiple organ failure. The graph is based on a sample of 862 trauma deaths recorded over a two-year period by the author’s group at San Francisco General Hospital.

as a surgeon at the University of California at San Francisco General Hospital Medical Center the number of victims of penetrating trauma (primarily gunshot and stab wounds) has increased to approximately 40 percent of the total trauma caseload. Almost all of these injuries are related to drug trafficking. Most drug addicts must pay for their habit by illegal means, and violent crime is a common recourse. The decriminalization of drugs could help to solve at least this part of the drug problem.

Supporters of decriminalization argue further that the prevalence of drug abuse is not significantly dependent on the legal status of the drug in question. The experience of this country in the 1920’s suggests that the consumption of alcohol was not reduced by prohibition; indeed, it may even have been increased.

There is no reason to believe drug abuse would be greatly affected one way or the other by decriminalization. What is certain is that many of the negative social effects accompanying drug abuse would be alleviated.

Finally, there is the problem of burn injuries, which in many respects is representative of the larger trauma-prevention problem. More than two million Americans per year suffer from burns of one kind or another, and of them some 70,000 are admitted to a hospital. Of the latter group 8,000 or so eventually die of their burn injuries. More than a third of these deaths are attributable to cigarette smoking. The average American cigarette contains additives in both the paper and the tobacco that cause the cigarette to burn for approximately 28 minutes. If these additives were omitted, the

cigarette would burn out in less than four minutes. As it happens, most furniture, upholstery and mattresses made in the U.S. need more than four minutes’ exposure to a burning cigarette for ignition. The problem and the solution are obvious. Omitting the incendiary additives from cigarettes would not change the taste of the cigarette smoke, but it would make smoking safer by reducing fire-related deaths, disabilities and property losses.

Of course, the cigarette manufacturers are not about to remove these additives voluntarily. That change undoubtedly calls for Federal legislation. Just as in the case of motor-vehicle accidents caused by drunk driving, fires caused by cigarette smoking bring death and disability to innocent people as well as to the individuals responsible for the acci-

dents. Beyond the question of controlling cigarette additives, other burn-prevention proposals call for measures to promote such practices as the manufacture of flame-resistant clothing, the installation of smoke alarms, the shortening of cords on appliances and the reduction of water-heater temperatures to prevent scalds of small children.

Although the preceding discussion is not a complete catalogue of trauma-prevention issues, it does indicate some of the problems such proposals face in the U.S. Clearly these issues are complex and impinge on long-established social customs. Nevertheless, prevention remains the only feasible way to reduce the toll of immediate deaths from trauma. It is not only the most effective way to save lives but also the cheapest: crisis intervention after the fact is always expensive. Ultimately prevention could also help to reduce the other two death peaks from trauma, topics I shall now address.

Roughly 30 percent of the deaths from trauma fall into the category of early deaths. This category can in turn be subdivided into two major pathological conditions: neurological injuries and various kinds of hemorrhage. According to a recent nationwide survey of head and spinal-cord injuries, head injuries account for about .2 percent of all hospital admissions in the U.S. On this basis one can calculate that roughly 34,000 cases of traumatic intracranial bleeding are treated annually in the U.S.

The results of another recent study, done by a group at the Health Sciences Division of Virginia Commonwealth University, point to the need for prompt management of such head injuries. The

Virginia group found that if surgical intervention for intracranial bleeding was delayed for more than four hours after an injury, the most probable outcome was death or permanent disability. If definitive surgical care was provided within four hours after an injury, however, the likelihood of a favorable outcome was significantly enhanced.

The need for prompt, definitive surgical intervention is also critical in the treatment of patients with injuries resulting in hemorrhage. For the sake of discussion hemorrhage can be divided into three grades: severe, moderate and minor. In cases of severe hemorrhage the rate of blood loss exceeds 150 milliliters per minute. In the first 10 minutes of severe hemorrhage the patient will lose at least 1,500 milliliters of blood, or roughly a third of his blood volume. If this rate continues unchecked, the patient will lose more than half of his blood volume within 20 minutes of the injury. In such cases little can be done in the prehospital setting to control the hemorrhage. Prompt, definitive surgical care offers this patient his only chance of survival.

In cases of moderate hemorrhage the bleeding rate is between 30 and 150 milliliters per minute, and there will be a life-threatening blood loss within an hour of the injury. Rapid transport of the patient to a place where prompt surgical intervention is available is also the preferred treatment. Patients with minor hemorrhage (bleeding rates of less than 30 milliliters per minute) may have the "luxury" of an hour or more before surgical intervention is necessary. In addition intravenous lines started in the prehospital setting may keep up with the bleeding. In any case the main point re-

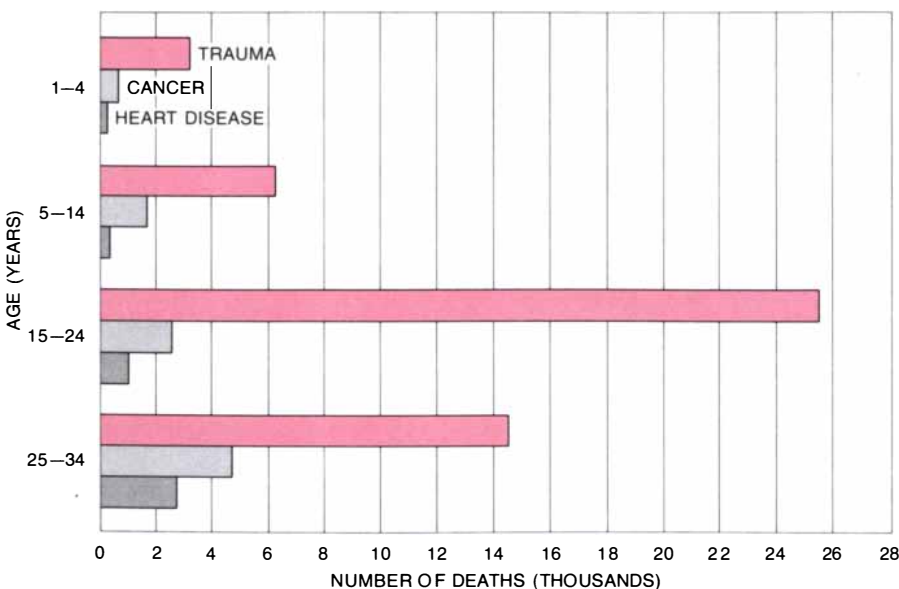
mains that many early deaths from trauma could be prevented by rapid transport of the injured patient to a trauma center.

Are trauma patients in the U.S. receiving the kind of timely medical care these studies indicate they need? Several sources show that with the exception of a few communities having modern trauma centers the answer is no. For example, in 1960 investigators from the Surgeon General's office in Texas examined the deaths of 606 soldiers injured in accidents and treated in community hospitals. They found that 96 of the patients would have survived if appropriate treatment had been administered in time, and an additional 103 patients might have been saved if they had been treated appropriately. Another study, done by workers at the University of Michigan in 1969, showed that 28 of 159 patients who died as a result of trauma were inadequately treated. Still another study, reported by a group at Johns Hopkins University in 1972, showed that a third of the deaths resulting from motor-vehicle accidents involving abdominal injuries in the Baltimore area could have been prevented by prompt surgical intervention.

I have been personally involved in several studies of this kind in the San Francisco Bay area. The first study, reported in 1974, compared the death records from the trauma center at San Francisco General Hospital with those from several community hospitals in the surrounding area. The results showed that patients with injuries from motor-vehicle accidents treated in a hospital without a trauma center had a significantly greater chance of dying than those treated in the one with a trauma center. A subsequent study compared deaths caused by motor-vehicle accidents in one part of California where there was no trauma center (Orange County) with those in a part of the state that had a single designated trauma center (San Francisco). Again the outcome was significantly worse in the region without a trauma center.

The latter finding led to a follow-up study, initiated by physicians in the Orange County area. The data were re-evaluated by an independent group of general surgeons, neurosurgeons and emergency-room physicians. Their report showed that non-neurological trauma care was inadequate in the hospitals without a designated trauma center. As a consequence five trauma centers were established in the Orange County area in 1980.

Another pertinent study was recently completed by the same group. It showed that the preventable-death category in Orange County dropped from 73 percent to 4 percent when patients were tak-



THREE LEADING CAUSES OF DEATH among young Americans are compared. The mortality figures, compiled by workers at the National Center for Health Statistics, are for 1977.

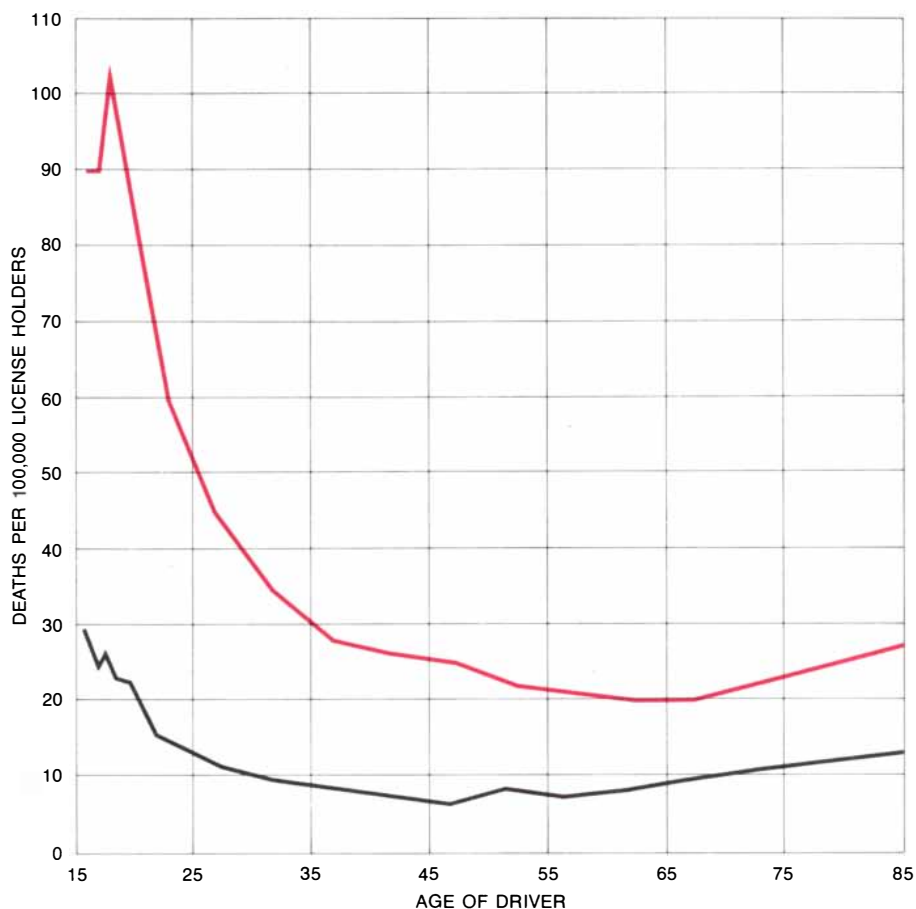
en to a trauma center rather than to a conventional hospital. Furthermore, the group found that none of the patients in the study died as a result of bypassing a conventional hospital in order to get to a trauma center. This finding emphasizes the importance of regional trauma care. Numerous other studies in various parts of the country lend further support to the conclusion that injured patients taken to hospitals without a trauma center are at a marked disadvantage. On the average the incidence of preventable deaths resulting from inadequate trauma care was found in these studies to vary between 30 and 40 percent.

Perhaps the most comprehensive study of this kind was done in 1980 by Daniel K. Lowe and his associates at the Oregon Health Sciences University. The study included 23 hospitals in a six-county region around Portland. The region has a total extent of 5,724 square miles and a population of 1,257,450, distributed over urban, suburban and rural areas. Originally 763 trauma patients were enrolled in the study; of these 104 had minor injuries and were excluded from further consideration. Of the 659 remaining patients there were 105 cases of inappropriate care, as determined by an independent trauma panel composed of general surgeons, neurosurgeons and emergency-room physicians. Of the 278 deaths registered 135 occurred in the hospital, and of these 34 were judged preventable by the panel. The latter group included 15 patients with brain injuries and 19 with various kinds of hemorrhage.

Of particular interest in the Portland results was the finding concerning the response time of the surgical consultants. In general the surgical consultants (there were 304 in all) took an average of 1.26 hours to get to the hospital after being called into the case. Neurological consultants responded somewhat more promptly: the average was .98 hour. The independent panel considered delayed response to be a significant factor in some of the cases of inadequate care. This finding draws attention to another problem: the popular misconception that any physician can treat a trauma patient adequately in a hospital emergency room. The emergency-room physician can start resuscitation, but a surgeon is almost always needed to provide definitive care. The sooner this care is provided, the better the outcome will be.

These findings all lead to one conclusion: There is a major shortfall in the delivery of trauma care in the U.S. The number of preventable deaths resulting from the existing system (or nonsystem) of trauma care is clearly unacceptable. What can be done to organize a better system of trauma care in the U.S.?

The concept of organized trauma care



YOUNG MALE DRIVERS account for a disproportionately high percentage of deaths from motor-vehicle accidents, as is shown in this graph of the death rate for male drivers (color) and female drivers (black). Curves, for 1978, are from the Insurance Institute for Highway Safety.

is not a new one; indeed, it can be traced back in military history to antiquity. The earliest mention of organized battlefield care is in the *Iliad*. According to Homer, Greek soldiers wounded in the fighting for Troy were carried off the battlefield and cared for in barracks (called *klisiai*) or on nearby ships. The *Iliad* contains references to 147 different wounds and implies a mortality rate of 77 percent.

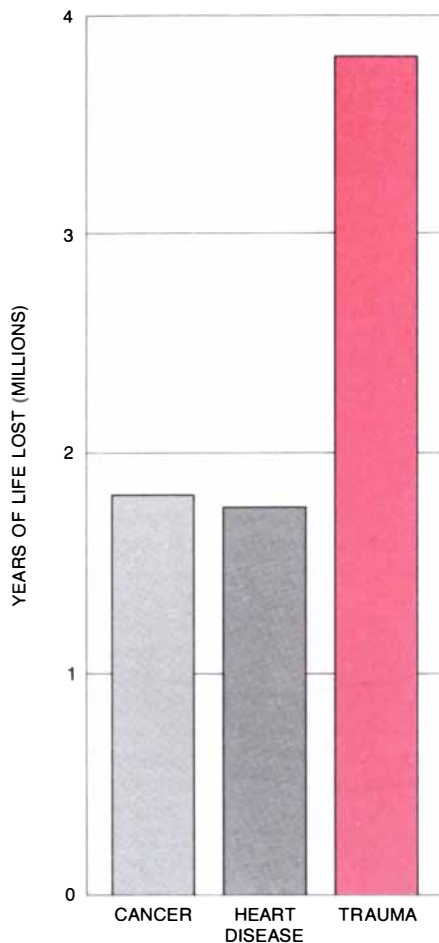
The Romans also had considerable experience with emergency care for the injured; as early as 480 B.C. wounded soldiers were reportedly assigned to the care of patrician families. By the first century A.D. special hospitals (called *valetudinaria*) had been established along the borders of the Roman Empire to care for wounded legionaries. Archaeologists have identified at least 25 of these structures, which were quite sophisticated in design.

In the early 19th century Baron Dominique Jean Larrey, the chief surgeon in the army of Napoleon, made two improvements in the care of wounded soldiers that have persisted to modern times. The first was the *ambulance volante* ("flying hospital"), an innovation that sharply reduced the time it took to provide definitive care to the wounded.

Before the advent of Larrey's horse-drawn ambulances injured soldiers often lay on the battlefield for periods of a day or more. Larrey's second innovation was to concentrate the casualties in one area and to operate on them as close to the front lines as possible.

During World War I the time lag between injury and surgery was still between 12 and 18 hours, and the overall mortality rate was 8.5 percent. The time lag was reduced in World War II to between six and 12 hours, and the mortality rate fell to 5.8 percent. Perhaps the most dramatic reduction in the time lag from injury to definitive surgical treatment came during the Korean conflict. A decision was made in the U.S. Army Medical Corps to bypass the battalion aid station and to take injured soldiers directly from the battlefield to a mobile army surgical hospital (M.A.S.H.). The average time lag between injury and definitive care during the Korean conflict was between two and four hours, and the mortality rate was 2.4 percent.

This tactic was further improved on during the U.S. involvement in Vietnam, where casualties were taken directly from the battlefield to the corps surgical hospital. According to one study, the average time lag between injury and de-



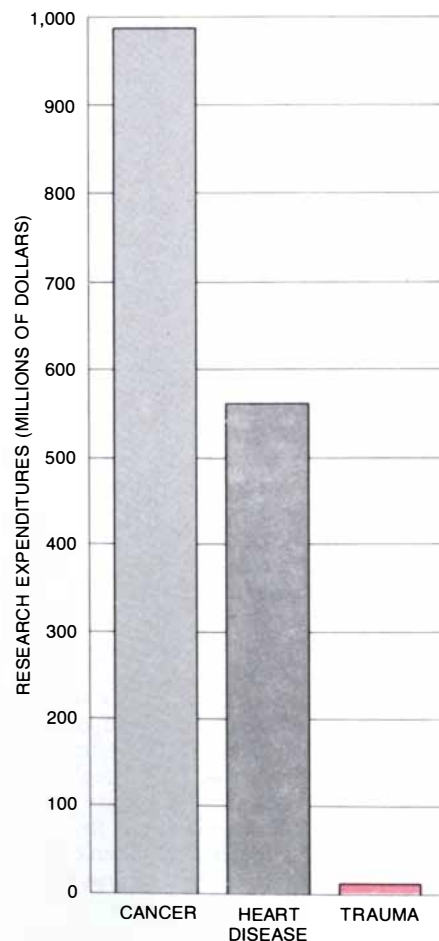
finite surgical care was reduced to 65 minutes, and the mortality rate fell to 1.7 percent. This military experience, one might think, should have served as an incentive and a model for the improvement of civilian trauma-care systems. With the exception of a few isolated instances, however, that has not happened.

One example of excellent regional trauma care can be found in West Germany. During the late 1960's West German health officials observed the U.S. methods of providing battlefield care in Vietnam. In 1970 the decision was made to apply these procedures throughout most of West Germany, establishing trauma centers along the main autobahns. Integral to the trauma-center concept in West Germany is rapid prehospital transport, which primarily entails the use of helicopters but also includes ground vehicles. There are now 32 air-rescue stations in the country with a standard mission radius of 50 kilometers. It is estimated that 90 percent of the population are within 15 minutes of a trauma center.

More important than the prehospital system is the system of integrated trauma care within the hospitals. West German hospitals have been classified according to their ability to provide trauma care. Furthermore, there is an in-house team of surgeons in every designated trauma center on a 24-hour basis. The teams include not only surgical residents but also a chief surgeon. The other important members of the trauma team are a neurosurgeon and an anesthesiologist. The chief trauma surgeon also cares for the patient in the postoperative period, including the time spent in the intensive-care unit. Overall the quality of surgical care is excellent.

The West German system also has a strong rehabilitation program, the primary goal of which is to get the accident victim back to gainful employment as soon as possible. I do not mean to imply that the system there is perfect. Some of

MISMATCH between the cost of trauma, in terms of the number of years of life lost, and the national effort to solve the trauma problem, in terms of dollars spent on research, is particularly striking in the context of a comparison with the corresponding figures for cancer and heart disease. The bars in the top chart are based on an estimate published in the Surgeon General's report for 1975. The bars in the bottom chart are based on 1982 figures from the National Institutes of Health; they refer only to research funds spent under the auspices of the National Cancer Institute, the National Heart, Lung, and Blood Institute and (in the case of trauma research) the National Institute of General Medical Sciences.



the trauma centers are not as strong as they should be, but in general the system is an excellent model for the U.S.

As a consequence of this regionalized system the mortality rate from motor-vehicle accidents in West Germany has dropped from 16,000 per year in 1970 to 12,000 per year at present, a reduction of 25 percent. It is probably more than a coincidence that the magnitude of this reduction is remarkably close to the preventable-death estimate made in most American studies (between 30 and 40 percent). By applying simple arithmetic and assuming that the 4,000 additional German patients who now survive each year return to work, a rough estimate of the financial benefit to that society can be made. If one assumes that each survivor over the past 10 years now earns the equivalent of \$10,000 per year and pays \$2,500 in taxes, the gross national product of West Germany would be increased by \$220 million per year and tax revenues would rise by \$55 million. The value of a trauma center, therefore, lies not only in a reduction in deaths and disabilities but also in a positive financial contribution to society. If the U.S. were to introduce a similar system and could achieve the same reduction in mortality, then over the first 10 years this country's G.N.P. could be increased by more than \$2 billion and the additional taxes paid would amount to more than \$550 million.

The final category in this discussion, late deaths, accounts for approximately a fifth of all trauma deaths. Of these deaths 80 percent are attributable to infection and multiple organ failure. The two conditions seem to be causally related. The common risk factors that have been identified so far include shock, head injury, peritoneal contamination and malnutrition, all of which can lead to infections late in the course of a patient's injury. This development may in turn be related to the failure of the patient's immune system, but the exact causes have not been elucidated. Once infection is obvious the patient often develops progressive organ failure. The resulting mortality rate is high and is directly related to the number of organ systems involved.

The answer to the question of why the trauma patient is at risk for infection and multiple organ failure can only come from further research. Even this solution, however, is not without difficulties. At present the U.S. spends very little of its research funds on trauma. National priorities are clearly directed to cancer and heart disease, even though trauma accounts for more years of life lost than cancer and heart disease combined. One solution would be to establish a National Institute of Trauma, on the model of the National Cancer Insti-

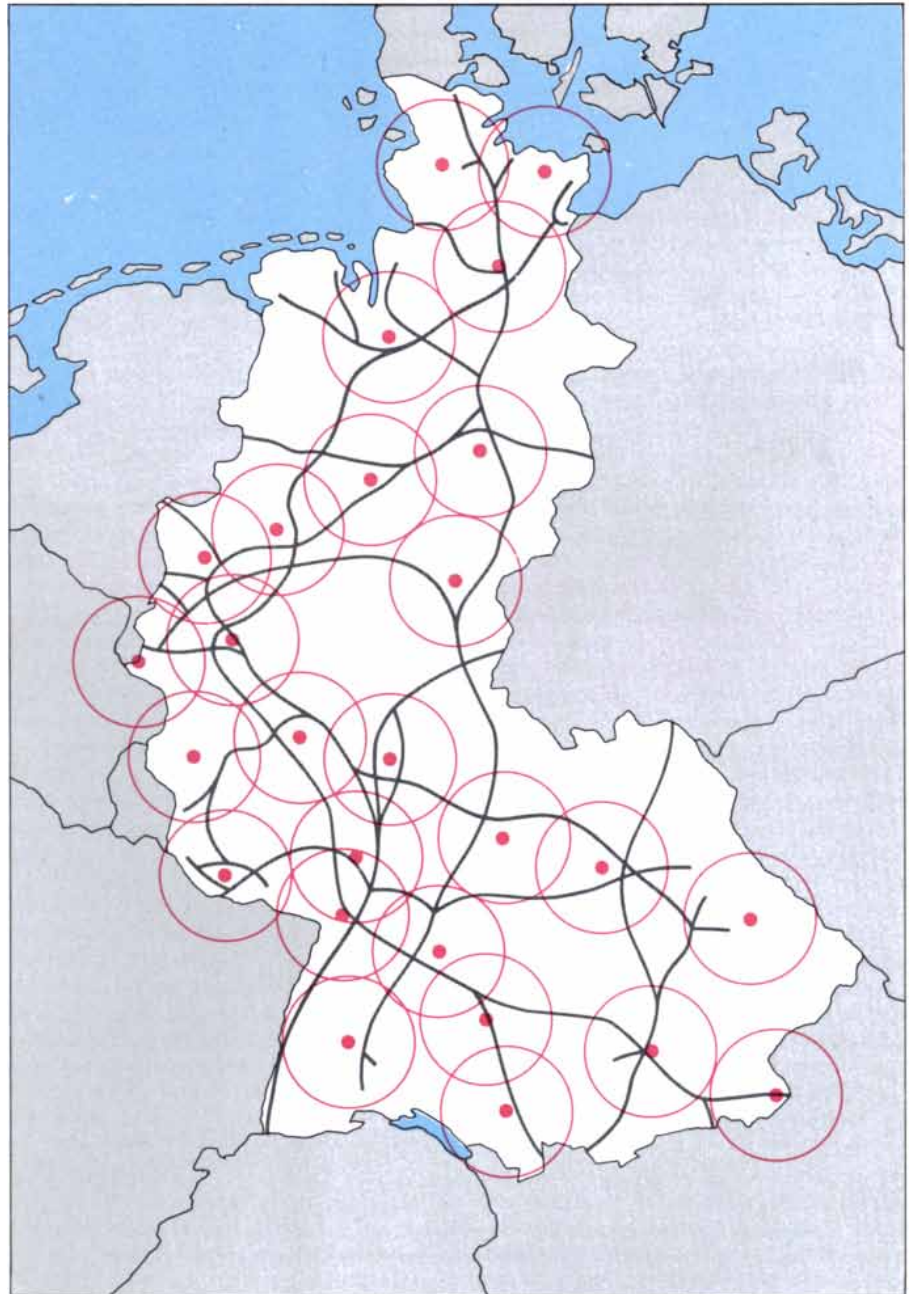
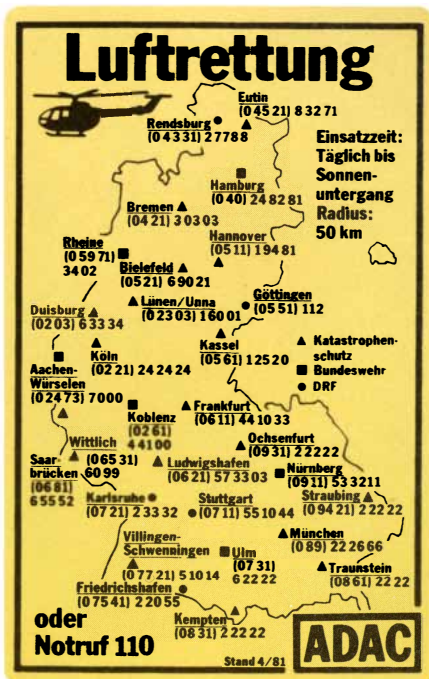
tute or the National Heart, Lung, and Blood Institute. Such an institute could serve many purposes. It could focus on trauma both as a medical issue and as a social one. It could approach trauma in such a way as to place equal emphasis on prevention, health-care delivery and research aimed toward the solution of the late-deaths problem. It could also serve as a focus for innovative ideas in research funding. For example, since drunk driving is a major contributor to the national trauma rate, perhaps it would be feasible to devote part of the tax on alcoholic beverages to help support trauma research.

Finally, there is the problem of rehabilitation. One of the most pronounced deficiencies in trauma care in the U.S. is the lack of an integrated rehabilitation system. Most disabling injuries are caused by neurological and orthopedic injuries. With the exception of some excellent rehabilitation centers for spinal-cord injuries the U.S. has not placed enough emphasis on returning the injured patient to work. This approach should involve not only physical rehabilitation but also job retraining and treatment of the emotional trauma that often accompanies physical trauma. In order to have an effective rehabilitation system the nation must also address

some of the existing worker's compensation laws and disability-reimbursement programs. For example, an employed person or union member currently takes five times as much disability time as a person who is self-employed. Any proposed rehabilitation system must address the disincentives that affect the decision to return to work.

In summary, trauma is a seriously neglected public-health problem in the U.S. Each of the three peaks in the trimodal distribution of trauma deaths has

its own set of associated problems. If the U.S. is to achieve a significant reduction in the rate of mortality and disability from trauma, each of these issues must be addressed vigorously. The solutions will not be easy, and they will inevitably engage some controversial social issues. It is my contention, however, that the U.S. can no longer afford the present rate of preventable death and disability resulting from trauma. The search for solutions to the trauma problem must become a national priority.



MODEL SYSTEM for the delivery of trauma care has been established in West Germany. In the past decade or so specialized trauma centers have been set up at hospitals along the main autobahns. Most patients are transported to the designated trauma centers by helicopter. The red dots on the map at the right designate the system's principal air-rescue stations; the red circles mark the standard

50-kilometer operating radius associated with each station. According to West German health officials, 90 percent of the population are now within 15 minutes of a trauma center. The smaller version of the map reproduced at the left is on a sticker distributed by the German automobile club ADAC. Telephone numbers accompany names of the stations; an alternative emergency number (110) is also given.



INTERFERON IS MANUFACTURED by bacteria in this battery of 400-liter fermentation tanks at Hoffmann-La Roche, Inc., in Nutley, N.J. *Escherichia coli* cells into which recombinant DNA carrying a gene for human alpha interferon has been inserted are grown in

a culture medium. The bacteria proliferate, synthesizing the protein interferon along with thousands of their own proteins. When the bacterial cells reach maximum concentration, they are killed, discharged from the fermentation tanks and concentrated by centrifugation.

The Purification and Manufacture of Human Interferons

Their promise is still not fulfilled, but now their genes have been isolated and cloned in E. coli. The bacteria are making interferon in quantity and the bacterial product is undergoing clinical trials

by Sidney Pestka

In the 1930's several investigators described the phenomenon of viral interference, whereby the infection of an animal by a virus seemed somehow to protect it against subsequent infection by another virus. In 1957 Alick Isaacs and Jean Lindenmann of the National Institute for Medical Research in London found an agent of viral interference: a protein, released by cells exposed to a virus, that enables other cells to resist viral infection. They called it interferon.

The great promise of interferon as an antiviral agent was evident from the moment of its discovery, which was supported by an independent report of similar findings by Y. Nagano and Y. Kojima of the Institute for Infectious Diseases in Tokyo. The promise was implicit in the fact that the protein is not directed against any one virus but rather protects cells against a wide range of viruses. It does more than that, affecting a number of different cellular activities in ways that suggest therapeutic possibilities. It is a potent substance: a very little of it goes a long way. As a natural cell product it seemed likely (given the correct dose level) to be safer than most new experimental drugs.

Yet the original promise has still to be fulfilled. For some 20 years after the discovery of interferon new complexities and difficulties seemed to turn up as often as promising leads. For one thing, interferon soon was found to be not a single protein but a large family of proteins, varying from species to species and present in multiple forms even within a species. Its mode of action proved to be indirect and is still not clearly understood. Above all, it is secreted by cells in minute amounts and was extremely difficult to purify. For many years it was hard to accumulate enough interferon for effective clinical trials. What was available was a crude preparation containing some interferons and a large amount of other protein. In the absence

of purification it was impossible to pin down the specific activities of individual interferons, to determine their structure and so differentiate clearly among them, and to assess their safety and efficacy.

In the past few years the situation has changed. First a number of human interferons were purified, making it possible to begin to understand their structure and to categorize their activities. Then, very quickly, the availability of recombinant-DNA technology led to the isolation of human-interferon genes, their cloning in bacteria, the production in quantity of recombinant human interferon by fermentation and its purification by means of monoclonal antibodies. The first trials to test dose levels and the side effects of the purified bacterial product in human beings began in 1981. Now extensive tests of interferon's efficacy against viral diseases and cancers are under way. There has been a remarkably rapid transfer of new biological technology from the laboratory to pharmaceutical trials.

Isaacs and Lindenmann discovered interferon made by chick cells exposed to influenza virus and found it protected other chick cells, but not the cells of other animals, against infection by viruses. Whereas interferon is not virus-specific (in keeping with the concept of viral interference), it seemed to be species-specific, with each animal species manufacturing its own interferon. Then things got more complicated: it was found that a species makes several kinds of interferon, each of which has a particular spectrum of activity in other species.

Here I shall discuss only the human interferons, of which three classes have been well defined. For historical reasons they have long been designated as leukocyte interferon, fibroblast interferon and immune interferon. Now it is known that a particular class is not made by a single cell type, and so a new nomenclature has been proposed: alpha (for leu-

kocyte), beta (for fibroblast) and gamma (for immune) interferon. Even this is an oversimplification. Many different members of the alpha class have already been isolated and characterized. There are reports that the beta class may have several members. So far only one member of the gamma class has been isolated and purified.

The reason interferon's protective effect is not limited to a single virus is that (unlike an antibody) it does not interact with the attacking virus; interferon interacts with the cell it protects. The manufacture of interferon is induced when a virus introduces its genetic material (DNA or RNA) into the cell. The presence of the foreign material (probably in the form of a double strand of RNA) causes the cell to synthesize and secrete molecules of interferon. The secreted interferon binds to a specific re-



PASTE OF E. COLI is removed from the centrifugation vessel. The packed cells are then resuspended and broken open; cellular debris is separated by centrifugation. Nucleic acids and other viscous materials are removed and proteins in extract are passed through monoclonal-antibody column to purify interferon.

ceptor on the surface of other cells and in doing so apparently triggers several different mechanisms within these cells, initiating the synthesis of a number of proteins (some of which have been identified). The proteins somehow cause the cells to resist the usual effects of viral infection: multiplication of the virus, lysis (bursting) of the cell and liberation of progeny viruses. Much remains to be learned about interferon's mode of action, but the experiments of Thomas C. Merigan, Jr., of the Stanford University School of Medicine and others have established that even crude preparations have some clinical effect against certain viral infections.

As I mentioned above, interferon has multiple biological activities other than

those contributing to the antiviral effect. Two are of particular interest. The protein appears in many cases to inhibit the proliferation of cells. It also somehow stimulates the activity of certain cells of the mammalian immune system: the lymphocytes called natural killer cells, which have a role in the destruction of foreign cells and perhaps of cancer cells. Either or both of these effects could explain evidence reported over the years, notably by Hans Strander of the Karolinska Institute, that interferon may promote the regression of some tumors.

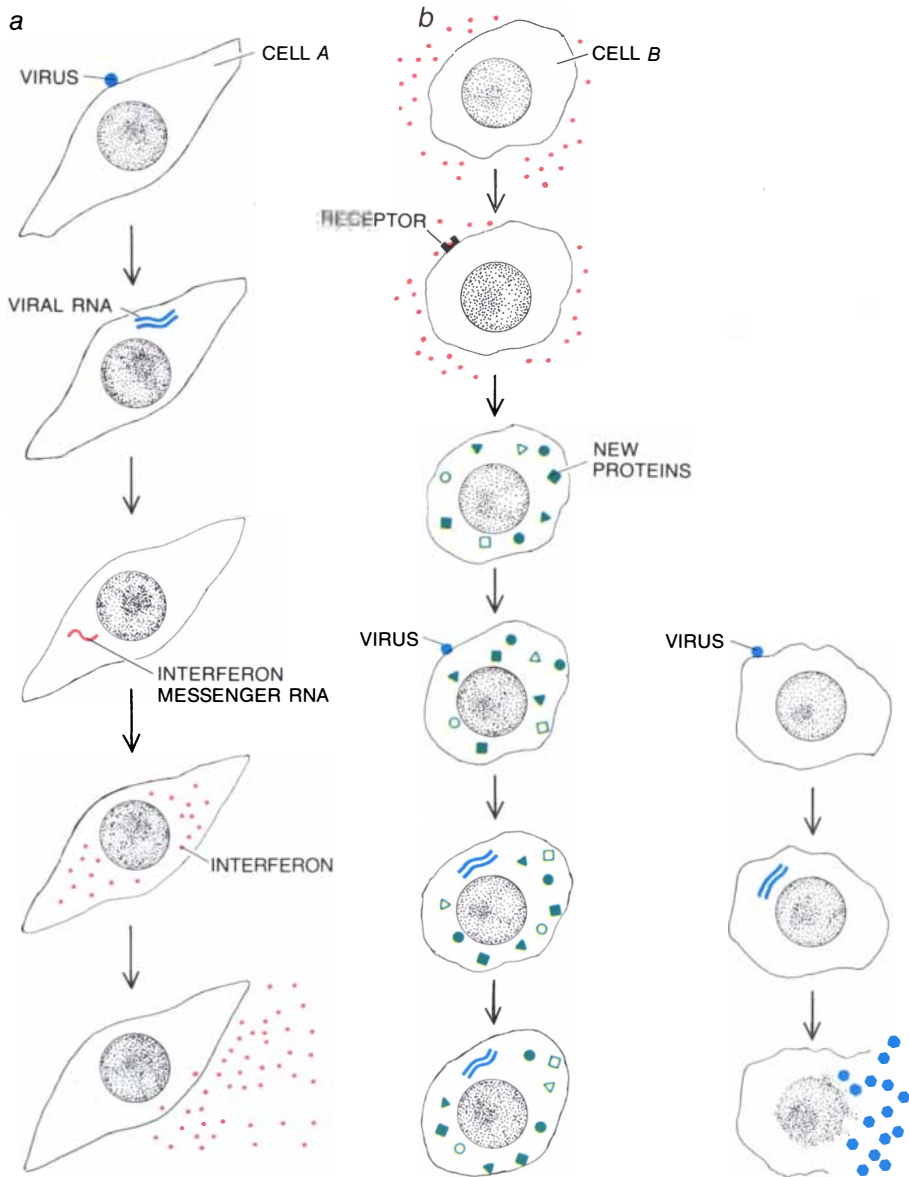
Both the sporadic cancer studies and those attempting to assess interferon's efficacy against viral diseases suffered from the serious shortage and high cost of interferon and even more from the

fact that the major "interferon" available was really a mixture of various proteins of which less than 1 percent by weight was interferon itself. Aside from antiviral activity no observed effect of the mixture could dependably be attributed to interferon. Purification of the protein was therefore a matter of high priority.

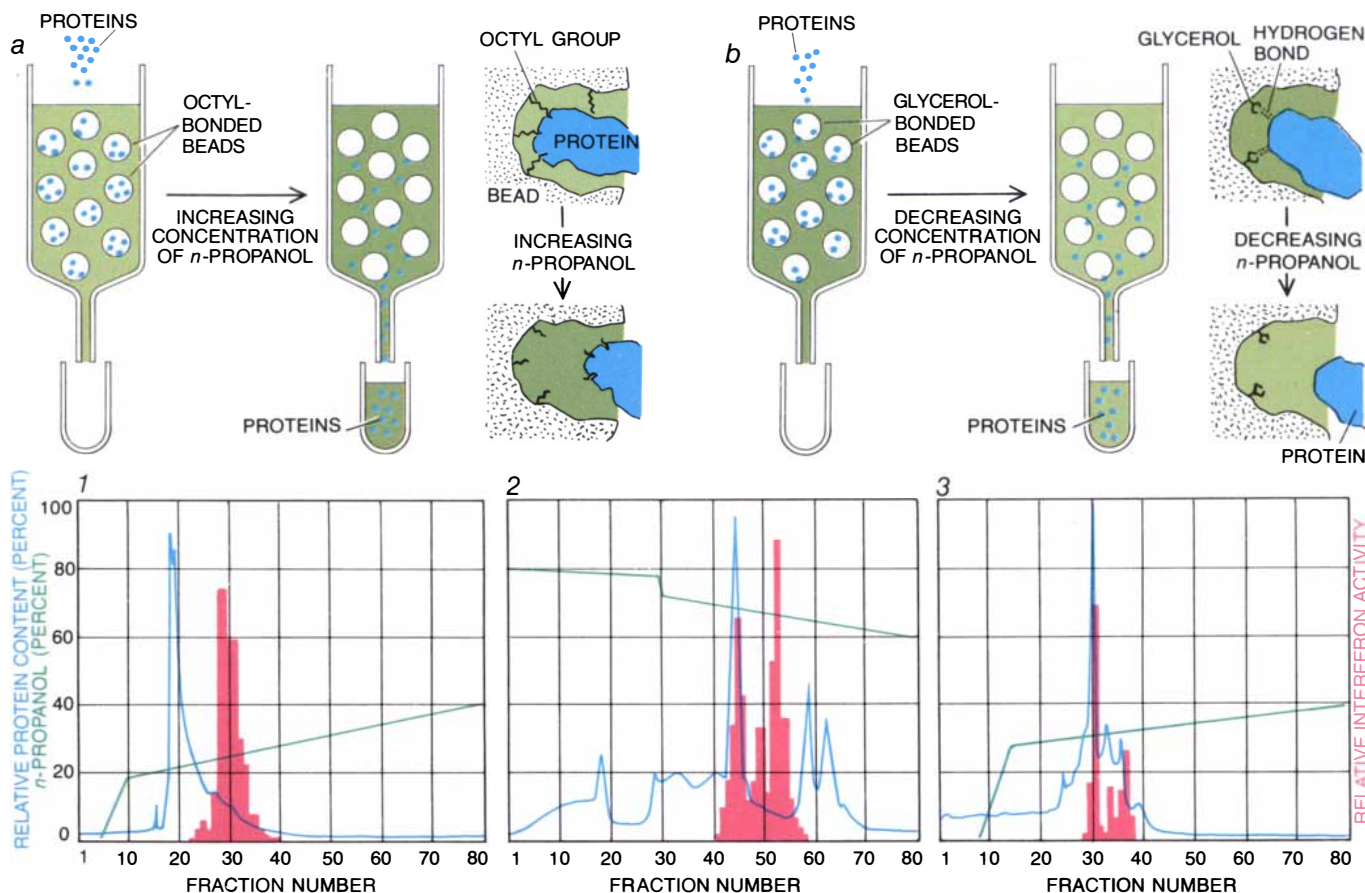
In 1977 we took on the task of purifying human alpha interferon in my laboratory at the Roche Institute of Molecular Biology. The first requirement was a large supply of crude leukocyte interferon. Our method of preparing it was essentially the one developed several years before by Kari Cantell of the Finnish Red Cross Blood Center. Human leukocytes (white blood cells) are incubated with an inducing virus, either Sendai virus or Newcastle-disease virus. Where Cantell had used human or bovine blood serum as a component of the culture medium, we substituted the milk protein casein; as a single protein it proved to be easier to remove in the initial concentration steps than the multiple proteins of serum were. We found the yield could be improved by substituting leukocytes from patients with chronic myelogenous leukemia for normal leukocytes, and we were able to obtain large supplies of leukemic cells (which are removed in therapy of the disease) from the M. D. Anderson Hospital and Tumor Institute in Houston.

After incubation overnight the cells and viruses are removed by centrifugation. What remains is a crude interferon preparation: it contains some induced interferon along with any other proteins whose secretion may have been induced by the virus and also all the normal cell secretions. The objective is to remove all the proteins other than interferon in a series of purification steps. At each step one must assay the concentrate for interferon activity in terms of standard units of "specific activity." The usual assay, measuring the extent to which a sample inhibits the destruction of cells by a virus, took three days. Philip C. Familletti and Sara Rubinstein found a way to cut the time to less than 16 hours and thereby speeded up the purification process considerably.

Knowing how much difficulty had attended efforts to purify interferon by conventional techniques, we decided to try what is called high-performance liquid chromatography. Chromatography methods in general involve adsorbing a crude mixture to some solid support and eluting different fractions with a solvent. In high-performance liquid chromatography the starting mixture is adsorbed to very fine beadlike particles packed in a column and the solvent is pumped through the column. Sidney Udenfriend, Stanley Stein and Peter Böhlen of Roche had managed to



SYNTHESIS AND ANTIVIRAL ACTIVITY of interferon proceed in stages. The protein is made and secreted (a) by a cell infected by a virus. It is apparently the presence in the cell of double-strand viral RNA that gives rise to the synthesis of interferon messenger RNA, which is thereupon translated into interferon. The secreted interferon binds to specific receptor molecules on the surface of another cell (b). The binding of the interferon triggers a number of changes in cellular activity, including the synthesis of proteins that render the cell resistant to infection by a virus, which in an unprotected cell (right) would replicate and burst the cell.



INTERFERON WAS PURIFIED by means of high-performance liquid chromatography (top). In the reverse-phase method (a) silica beads to which octyl groups have been bonded are packed into a chromatography column; proteins pumped into the column bind tightly to the octyl groups. Increasing concentrations of the solvent *n*-propanol in an aqueous solution are passed through the column; the *n*-propanol releases different proteins from the column in the order of their increasing affinity for octyl groups. In the normal-phase method (b) glycerol is bonded to the beads. In the presence of a high *n*-pro-

panol concentration, proteins pumped into the column form hydrogen bonds with the glycerol. As the *n*-propanol is diluted proteins are released in the order of their increasing ability to form hydrogen bonds with the glycerol. Purification data (bottom) show how reverse-phase chromatography (1, 3) was alternated with normal-phase chromatography (2). The protein content of successive fractions was monitored and each fraction was assayed for interferon activity (red bars). Interferon-rich fractions were pooled for further purification. Eventually a pure interferon fraction was isolated (3).

separate peptides (short chains of amino acids) by means of reverse-phase liquid chromatography, in which the solid phase (the beads) is coated with an organic material that is hydrophobic, or water-repellent; the mobile phase (the solvent) is more polar, or water-attracting. Stein, Menachem Rubinstein and I undertook to apply the method to purify the alpha interferon.

Ethyl alcohol, the accepted solvent, did not work; the protein remained adsorbed to the beads. We decided to try a somewhat less polar solvent, *n*-propanol, even though proteins are not very soluble in it and the interferon might precipitate as it was being eluted. As it turned out, a gradient of *n*-propanol worked. As we pumped increasing concentrations of the solvent through the column, different protein fractions (having different affinities for the solid support) were eluted successively and collected in tubes. Each fraction was assayed for interferon activity. Fractions that were relatively rich in interferon were applied to another column for further purification. By alternating the re-

verse-phase process with normal-phase chromatography (in which the beads are coated with hydrophilic groups and the solvent is less polar) Rubinstein purified human alpha interferon, in just a few steps, about 80,000-fold. The specific activity of the purified interferon was from 200 to 400 million units per milligram. Subjected to gel electrophoresis, which separated the proteins on the basis of molecular size, the interferon yielded a single band at a molecular weight of 17,500, and the protein in that single band was active. In other words, human alpha interferon had been purified to homogeneity.

During the purification process we had been surprised to note that interferon activity was purified in several different fractions, indicating that we were isolating a number of different species of the protein. Chemical characterization of the purified products showed this was indeed the case. When we treated the proteins with trypsin, an enzyme that cleaves the long chain of amino acid subunits at specific sites, different

purified products yielded somewhat different sets of peptides. When we had enough of each of several purified interferons, we proceeded to analyze their amino acid composition (the number of each of the 20 different amino acids present) with a sensitive amino acid analyzer Stein had built. Again there was evidence for the presence of multiple species.

There had been earlier reports that different crude preparations of human alpha interferon were somewhat different in the electric charge of the molecules, but this had not been attributed to variations in the molecules' amino acid sequence, the ultimate determinant of a protein's individuality. Interferons were thought to be glycoproteins (proteins to which sugar chains are attached), and the charge heterogeneity was thought to be due to differences in carbohydrate content. John A. Moschera looked at five purified alpha species and found no detectable carbohydrate in any of them, however. Soon Christian B. Anfinsen's group at the National Institute of Arthritis, Metabolism, and Digestive Dis-

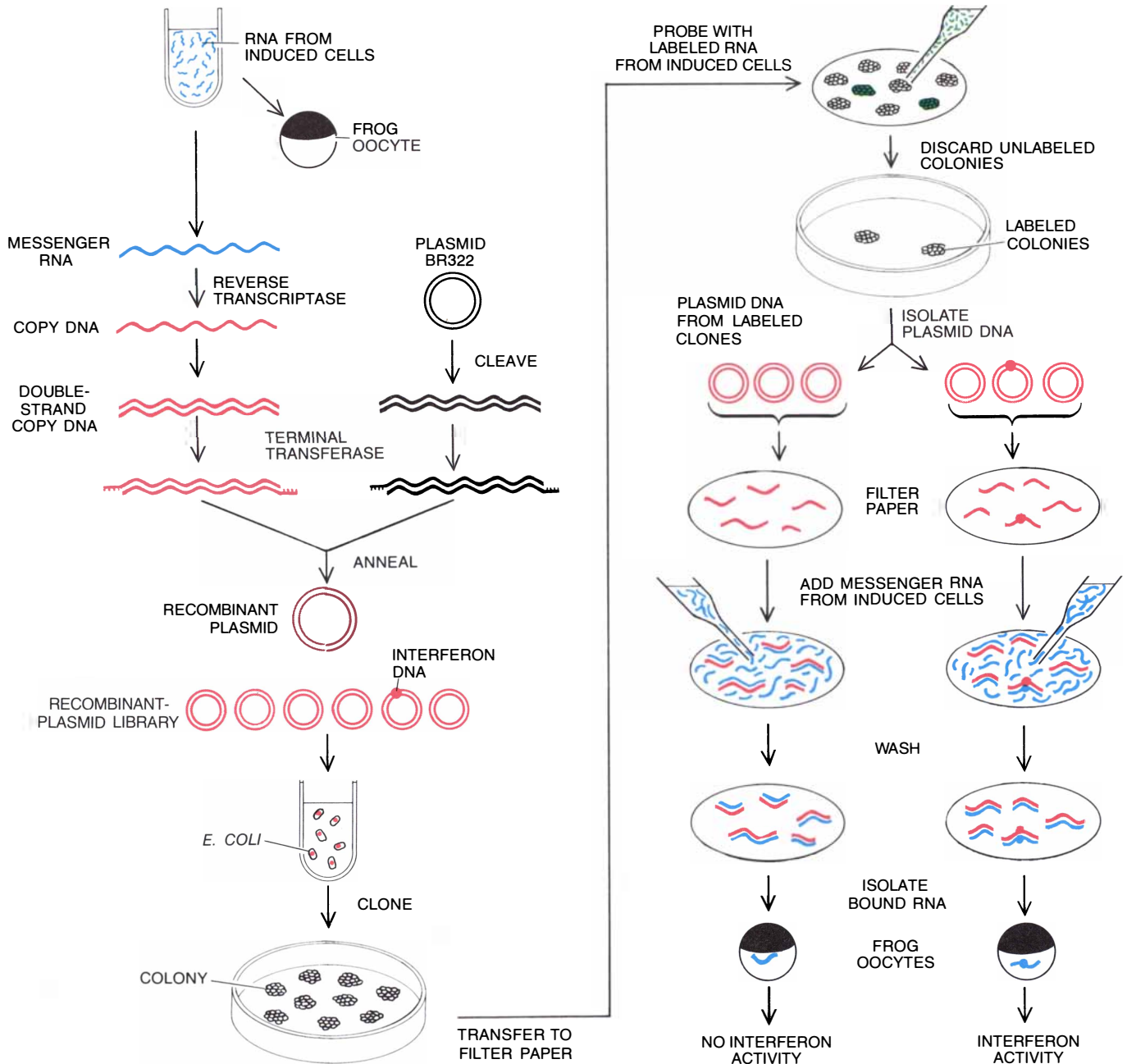
eases and Geoffrey Allen and Karl H. Fantes of the Wellcome Research Laboratories in Britain isolated multiple alpha interferon species; Allen and Fantes looked for carbohydrate and did not find any. Some interferon species may indeed turn out to be glycosylated, but the dogma that all interferons are glycoproteins is invalid.

With the amino acid compositions in hand, several groups went on to tackle the actual amino acid sequences. Kath-

ryn C. Zoon and her colleagues in Anfinsen's laboratory, collaborating with Michael W. Hunkapiller and Leroy E. Hood of the California Institute of Technology, got the first partial sequence, for the amino terminus (the beginning) of their human alpha interferon. A few months later Warren P. Levy of my laboratory and John E. Shively of the City of Hope Medical Center determined the amino-terminal sequence of one of our alpha species. At two sites

Zoon's sequence and ours showed a different amino acid. The particular amino acids involved were not likely to be mistaken for each other, and so we felt both sequences must be correct. Here was more clear evidence that the alpha interferons are a family of closely related proteins, a finding that was confirmed when Levy and Shively, Zoon and her colleagues and Allen and Fantes eventually obtained more sequences.

Although our alpha interferon was



CLONING OF INTERFERON GENE begins with the induction of interferon synthesis in leukocytes. Messenger RNA from the cells is assayed for interferon activity in frog oocytes. If a sample shows interferon activity, the RNA is copied into DNA; a second strand is made with DNA polymerase. The "copy DNA" (cDNA) is treated with terminal transferase to produce "sticky ends." Meanwhile a vector, plasmid BR322, is cut open and complementary sticky ends are synthesized. The cDNA is annealed with the plasmid DNA to make a "library" of recombinant plasmids, which are introduced into *E. coli* cells. Clones of identical cells are grown from individual *E. coli*.

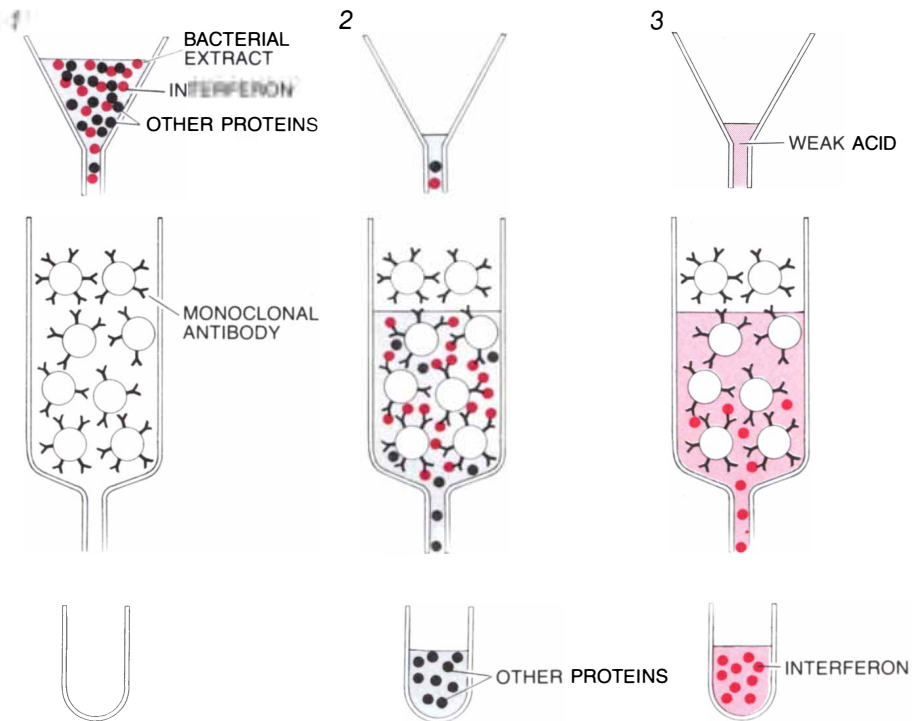
To find a clone harboring plasmids carrying interferon DNA the colonies, replicated on filter paper, are probed with radioactively labeled RNA from induced cells under conditions such that the probe hybridizes with, and thus identifies, DNA encoding induced proteins, including interferon. The plasmid DNA of the labeled colonies is cut into linear fragments, batches of which are fixed to filter paper. RNA from induced cells is added; it hybridizes with DNA encoding induced proteins. The hybridized RNA is separated from the DNA and assayed in oocytes to identify plasmid-DNA batches encoding interferon. Procedure is repeated for individual clones of positive batches.

the first interferon to be purified in such a way that its composition could be determined, Ernest Knight, Jr., of E. I. du Pont de Nemours & Company had purified some human beta (fibroblast) interferon a few years earlier. Now Knight, Yin H. Tan of the University of Calgary and Stein, Heinz-Jürgen Friesen and Moschera in my laboratory worked out the terminal sequences of purified beta interferon. The sequences were all the same; so far only a single beta interferon has been isolated by each group, and the proteins all seem to be identical.

With homogeneous human alpha and beta interferon available it was finally possible to assess the biological activity of pure interferons rather than crude mixtures. The antiviral activity of interferons turns out not to be strictly species-specific. Most of the human alpha (leukocyte) interferons are about as active in protecting bovine cells as they are in protecting human cells; two of them are much more active on bovine cells than on human cells. Human beta (fibroblast) interferon has very low antiviral activity on bovine cells, however. Working with our purified alpha species, Marian Evinger in my laboratory and Nathan O. Kaplan of the University of California at San Diego found that antiproliferative activity too is a property of interferon itself, but that some versions of the protein are much better than others in this respect. As for the stimulation of killer-cell activity, Ronald B. Herberman and John R. Ortaldo of the National Cancer Institute found that almost all the pure interferons do enhance the ability of certain lymphocytes to destroy target cells.

There remained the problem of obtaining enough pure interferon for intensive study of its mode of action and for clinical trials. We might have accumulated enough by large-scale induction of leukocytes and purification of the crude material, but a better method was at hand. In the mid-1970's recombinant-DNA techniques had been developed whereby the DNA containing the gene for a particular protein can be inserted into the bacterium *Escherichia coli* and be cloned, or replicated in many copies, and expressed: translated into protein. We realized that by such methods we could obtain a large supply of particular interferon genes for study and also get the bacteria to manufacture a particular interferon—eventually perhaps on an industrial scale.

Actually we had taken the first step a few years before. Ordinarily one begins the gene-cloning process by isolating, from cells that make a lot of the desired protein, the messenger RNA encoding the protein. (Messenger RNA is the nucleic acid into which the DNA of a gene is transcribed by the cell; it is subsequently translated by the cellular ma-



MONOCLONAL ANTIBODIES are exploited to purify recombinant interferon. A crude interferon mixture extracted from bacteria is poured onto a column packed with beads to which monoclonal antibodies against alpha interferon have been linked (1). The interferon binds to the antibodies and remains attached as all other proteins pass through the column (2). A weak acid dislodges the interferon (3); the solution is neutralized and the interferon is concentrated.

chinery to make a protein.) Interferon messenger RNA is present in leukocytes at a very low level, so that it is hard to identify. In 1974 James L. McInnes of my laboratory, working with Edward A. Havell and Jan Vilček of the New York University School of Medicine, had extracted messenger RNA from cells induced to manufacture interferon. They added the RNA to cell-free extracts containing the cellular translation machinery and succeeded in translating the RNA into active interferon. Later Ralph L. Cavalieri was able to insert the messenger RNA into frog oocytes (precursors of egg cells); the oocytes translated the messenger. By assaying the product of the cell-free system or of the oocytes for interferon activity one could determine whether or not a particular RNA preparation included at least some interferon messenger. M. N. Thang of the Institut de Biologie Physico-chimique and Paula M. Pitha of the Johns Hopkins University School of Medicine developed similar methods for measuring interferon messenger RNA.

Once Familletti had optimized the production of interferon messenger RNA in induced leukocytes, Shuichiro Maeda, Russell McCandliss and Alan Sloma prepared complementary DNA from the RNA and spliced the DNA into plasmids: small circles of bacterial DNA. This yielded a large "library" of recombinant plasmids, each carrying a DNA copy (cDNA) of some messenger

RNA from the induced leukocytes. The plasmids were introduced into *E. coli* cells and individual clones (colonies descended from an individual cell) were isolated. The next problem—the hard part of the procedure—was to find, in the vast library of recombinant plasmids, those with DNA encoding interferon.

If we had been able to begin with a purified interferon messenger RNA, that would have been a straightforward task. DNA and RNA are chains of the subunits called nucleotides; the sequence of nucleotides encodes genetic information. RNA transcribed from a stretch of DNA is complementary to it in sequence and will hybridize with it, the two strands pairing to form a double strand. A pure interferon RNA can therefore be labeled with a radioactive isotope and used as a probe that anneals to the sought-after DNA, whose identity is revealed by autoradiography.

The reader will remember, however, that we did not have pure interferon RNA; we had begun with a mixture of RNA's whose interferon-specific activity could be revealed only after the fact, by translating the RNA and assaying the products for interferon activity. In this situation we had to devise an indirect, two-stage procedure. In the first stage we screened all the bacterial colonies to find those with plasmids carrying cDNA made from the RNA of induced cells and therefore perhaps carrying interfer-

cloned an alpha-interferon DNA that was different from ours. The beta gene was cloned by Tadatsugu Taniguchi of the Japanese Foundation for Cancer Research (who had in fact reported success a few months before the rest of us), by Walter Fiers of the State University of Ghent, by Norman H. Carey and others at G. D. Searle & Co., Ltd., and also by us in collaboration with workers at Genentech, Inc.

Plasmid 104 served as a probe with which to search through a standard library of human DNA and find related interferon genes. So far we have isolated 16 distinct sequences for human alpha interferons; some of the same sequences and a few different ones have been isolated by other workers. None of the interferon DNA's contains intervening sequences, the noncoding stretches of DNA that interrupt protein-coding regions in most mammalian genes. The nucleotide sequences of the DNA's confirm what was indicated by the amino acid sequences of purified interferons: the alpha interferons are a class of related proteins, each one encoded by one of a family of related genes.

Plasmid 104 did not include the entire coding sequence for an alpha interferon; because of the vagaries inherent in the construction of recombinants the beginning of the gene was missing. At Genentech David V. Goeddel and his co-workers exploited plasmid 104 as a probe for searching through a library of cDNA until they found an entire gene. That gene needed to be reconstructed to achieve its expression in *E. coli*. First Goeddel removed from it a segment coding for the leader, a peptide that signals the cell to secrete the protein and is cleaved during secretion. Then Roberto Crea put together a segment of DNA carrying an initiation codon, a three-nucleotide signal to translate messenger RNA into protein; the segment was spliced to the gene just before the beginning of the segment coding for interferon's amino acids. Finally the Genentech workers attached, before the initiation codon, a regulatory sequence containing the bacterium's own signal to begin the transcription of DNA into messenger RNA. The result was a chimeric gene combining a bacterial regulatory region with the human gene's coding region for interferon.

When the reconstructed chimeric gene was inserted into *E. coli*, the bacteria proceeded to synthesize large quantities of human interferon: about as much per liter of bacterial culture medium as could have been produced by the leukocytes from 100 blood donors. Recombinant human leukocyte interferon *A*, as we designated it, behaves like its counterpart made from human cells. In an early animal experiment done by Noel

Stebbing at Genentech it prevented a viral disease, encephalomyocarditis, in squirrel monkeys. Before the bacterial product could be tested in human beings, however, rigorous purification was necessary to remove contaminating bacterial proteins. We turned to monoclonal antibodies.

An antibody is a protein of the immune system that recognizes and binds to a foreign protein, or antigen. Since 1975 it has been possible to prepare large amounts of monoclonal antibodies: antibodies directed against a specific antigen. The availability of our purified interferon had enabled Theophil Staehelin of F. Hoffmann-La Roche & Co., Ltd., to prepare monoclonal antibodies directed against specific interferon molecules. Monoclonal antibodies that were specific for alpha interferon were linked to beads packed into chromatography columns. To purify bacterial interferon Staehelin, Hsiang-fu Kung and Donna S. Hobbs poured onto the column an extract of *E. coli* cells that had synthesized recombinant interferon *A*. The interferon, and only the interferon, bound to the antibodies; the other components, including any bacterial toxins, went right through the column. Then an acid solution was passed through the column to elute virtually pure interferon.

The availability of large amounts of very pure interferon opened the way to clinical trials, which I shall describe below. It also meant we could crystallize interferon, the first step toward analysis of the protein's three-dimensional structure by X-ray crystallography. David L. Miller and Kung have prepared crystals of human recombinant interferon *A*.

Early clinical trials of interferon were hampered, as I have suggested, by the short supply and high cost of interferon synthesized by human cells and by the impurity of the preparations, which clouded assessment of interferon's own effect and also limited dosages (since large amounts of the crude material could not be administered to patients). The purification of recombinant human alpha interferon in quantity removed all these impediments. After appropriate tests for safety in animals, the bacterial product prepared by Hoffmann-La Roche, Inc., was approved for trials in human beings. In January of 1981 Jordan U. Gutterman of the M. D. Anderson Hospital initiated a clinical trial designed to establish the safety, toxicity and side effects of various blood levels of recombinant interferon *A* in cancer patients.

More than 500 patients have been given interferon in trials conducted in various academic institutions and coordinated by Zofia Dzierwanowska of Hoffmann-La Roche and her col-

leagues. The most frequent side effects were those noted earlier with crude interferon: fever, chills, muscle aches, mild gastrointestinal upset, fatigue and anorexia. Patients seemed to develop tolerance to acute "flu-like" side effects, but fatigue and anorexia increased with dosage and duration of treatment. There was also some decrease in the white-blood-cell count, which reversed within a few days, and in a few instances there was some elevation of liver enzymes that seemed to have no deleterious effect.

Some of the studies were designed not only to assess tolerance but also to evaluate antitumor response. Most of the patients had advanced cancer resistant to conventional therapy. Even in this group some tumor regressions were observed in kidney cancer, malignant melanoma, multiple myeloma and a few other malignancies. Some results suggested that recombinant interferon *A* may be effective against Kaposi's sarcoma in patients with acquired immune deficiency syndrome (AIDS). Cancer of the breast, lung and colon appeared to respond only minimally or not at all. These are very preliminary results, to be sure. Additional trials of interferon *A*'s efficacy in cancer and viral diseases are under way.

The detection of cancers may also be improved by interferon. It was known to alter the protein composition of the cell surface. Recently Paul B. Fisher, I. Bernard Weinstein and Soldano Ferrone of the College of Physicians and Surgeons of Columbia University, Jeffrey Schlom and John W. Greiner of the National Cancer Institute and I have found that recombinant interferon *A* can increase the expression of certain tumor-associated proteins on the surface of malignant-melanoma, colon-cancer and breast-cancer cells. Such an effect may make it possible, with monoclonal antibodies to these tumor antigens, to diagnose a developing cancer many months earlier than is now possible.

Interferon *A* is only the first bacterial interferon to be tested. Other alpha interferons are becoming available for trials; beta interferon and gamma interferon will follow. Some classes, species and combinations of interferon may be more effective than others against particular diseases and under particular conditions. Moreover, the physician will not be limited to the set of natural interferons. One can break up interferon genes and splice the pieces to make new genes that are translated into hybrid interferons. We and other workers have already experimented with such hybrid molecules. It may become possible, as more is learned about the mechanisms of these proteins' various activities, to tailor interferon molecules to optimize particular effects.

Magnetic Fields in the Cosmos

The dynamo mechanism for the generation of magnetic fields explains why Venus has no field, why the sun has an oscillating field and why the dominant galactic field is parallel to the plane of the galactic disk

by E. N. Parker

If nuclear and gravitational forces were the only forces at work in the universe, the broad pattern of cosmic evolution would be one of gradual thermal degradation punctuated by occasional explosive events. The cosmos would resemble the serene—and monotonous—heavens of classical conception. There is, however, a cosmic agitator: the magnetic field. Although only a small part of the available energy in the universe is invested in magnetic fields, they are responsible for most of the continual violent activity of the cosmos, from auroral displays in the earth's atmosphere to stellar flares and X-ray emission, and the massing of clouds of interstellar gas in galaxies.

Within the solar system spacecraft have carried magnetometers near every planet from Mercury to Saturn. Mercury, the earth, Jupiter and Saturn have fields of their own, but Venus and Mars do not. The fields range in strength from 3.5×10^{-3} gauss at the poles (Mercury) to eight gauss at the poles (Jupiter). The earth's field, in comparison, is about .6 gauss at the poles. The *Voyager II* spacecraft, which carries a magnetometer, should pass close to Uranus in 1986. Magnetic fields are expected on both Uranus and Neptune because of their similarity to Jupiter and Saturn.

In all but a few cases the magnetic fields of stars cannot be detected directly from the earth, but their existence can be inferred from the presence of activity similar to solar magnetic activity. On such evidence it appears that most stars have magnetic fields at least as strong as the sun's. Some classes of stars do, however, have fields strong enough to be measured directly, for example by measuring the extent to which the two states of polarization of a given spectroscopic line are shifted from the position of the unpolarized line. Magnetic A stars have fields of up to 34,000 gauss; some white-dwarf stars have fields of from 10^7 to 10^8 gauss, and pulsars have fields of 10^{12} gauss. The disk of our galaxy is permeated by a field of between 2×10^{-6} and 3×10^{-6} gauss, and other galax-

ies appear to have fields at least as strong. There is even evidence, although it is controversial, that a weak magnetic field fills the space between galaxies.

Surprisingly, there is a single, generic explanation for the ability of bodies as different as a dense, cold planet and a tenuous, hot galactic disk to generate a magnetic field. The explanation, first worked out for the earth, comes from the discipline of magnetohydrodynamics: the study of the interaction of a moving, electrically conducting fluid and a magnetic field.

The cosmos is filled with fluids capable of carrying electric currents. Most of the fluids are hot ionized gases, but in planets they are internal reservoirs of molten metal. The energy released in the interior of planets and stars and in assemblages of stars by the action of nuclear and gravitational forces keeps these fluids in turbulent motion. The magnetic fields entrained in the fluids are stretched and folded by the fluid motion, gaining energy in the process. In other words, the turbulent fluids function as dynamos, devices that convert mechanical energy into the energy of magnetic fields. The ability of magnetic fields to feed (as living organisms do) on the general energy flow from the interior of planets, stars and galaxies explains why the fields flourish in and around nearly every celestial body.

Mathematical models of dynamos with various physical characteristics convincingly simulate the fields produced by astronomical bodies. The shell model of the sun's convective zone, for example, predicts a waxing and waning field that reverses polarity at regular intervals, as the sun's field does. Given the lack of detailed information about factors such as the motions of the gas within the sun, however, the dynamo models are inevitably less than definitive. In some instances the available information raises questions that cannot yet be answered. One example is the ability of Mercury to generate a field. Moreover, the dynamo mechanism by itself cannot account for the exceptionally strong

fields of some stars. Because of such gaps in information, the rival hypothesis that there are primordial fields, trapped for example in the stable core of stars since their formation, cannot be disproved. The balance of the evidence, however, indicates that the planets, the sun, most stars and the galaxy function as colossal dynamos.

A dynamo converts the energy of motion of a conductor into the energy of an electric current and a magnetic field. A simple laboratory dynamo consists of a metal disk rotating on an axle and over a conducting coil the axis of which is aligned with the axle of the disk. The coil is electrically connected to the disk and the axle by brushes. A current passed through the coil induces a magnetic field aligned with the axis of the coil. The electrons in the disk moving through this field are subjected to a force at right angles to their direction of motion and the direction of the magnetic field, that is, a force directed along the radius of the disk. The direction of the force is given by the familiar right-hand rule. If the disk, viewed from above, is rotating counterclockwise and the magnetic field is directed upward, the induced electromotive force causes a current to flow from the axle to the periphery of the disk. The current then flows from the disk through the brush to the coil, amplifying the magnetic field generated by the coil, which in turn amplifies the current flowing through the disk. Magnetohydrodynamic dynamos operate on the same principle, although the electrons are not confined to coiled wires and can move throughout a body of fluid. Explaining how a moving conductive fluid amplifies a magnetic field becomes complicated, however, if the argument is pursued in terms of induced electromotive forces, the resulting currents and the magnetic fields associated with the currents.

Here some conceptual shortcuts are useful. The magnetic field can be visualized as consisting of lines of force, the closed loops along which a compass

needle would align itself. The strength of the field in any given volume of space can then be represented schematically by the number of lines that traverse the volume or, less conventionally, by the thickness of the lines. Following a point made by Hannes Alfvén, one of the founders of magnetohydrodynamics, one may regard the field lines as being "frozen" into the conducting fluid or "attached" to the particles of which the fluid is composed. The field moves with the fluid, and the field lines are deformed by the fluid motions as if they were rubber bands. If the particles to which a field line is "attached" are moving at right angles to the field line but at different speeds, the field line will be stretched. In accordance with such con-

ventions the stretching corresponds to a gain in the strength of the field. The energy of the motion of the particles is converted into the energy of a magnetic field, and induced electromotive forces drive the current associated with the field just as they do in the disk dynamo.

It should be noted that the operation of either type of dynamo calls for the initial presence of at least a weak magnetic field or a weak current (the current in the coil in the case of the disk dynamo). Therefore the dynamo mechanism does not explain how the magnetic fields of planets and stars may have originated but rather how they are amplified and maintained in spite of the continual sapping of the field by the dissipation of the associated current. Other mechanisms

that will be discussed below explain how the weak fields the dynamo mechanism amplifies could be created.

The ubiquity of magnetic fields suggests that the circumstances under which a magnetohydrodynamic dynamo develops are not particularly stringent. The first requirement is the presence of an electrically conducting fluid capable of supporting the currents associated with the field. Most of the universe is filled with such fluids. Most planets have a molten metallic core; stars are composed of ionized gases, and nearly all space is filled with a gas that has enough free electrons to be a good conductor of electricity. In most instances the gas is not as conductive as a



MAGNETISM OF THE SUN is apparent in a computer-graphics image of the sun's corona. Since the gas of the corona is ionized, the distribution of its electrically charged atoms reflects the form of the sun's magnetic field. This false-color image is a computer analysis of a photograph of the corona made in India during the total solar eclipse of February 16, 1980. The photograph was made by a research group from the High Altitude Observatory of the University of Colorado at Boulder and Southwestern at Memphis College with a special

camera developed by Gordon A. Newkirk, Jr. In the camera a radially graded filter is interposed between the lens and the photographic plate. The filter blocks more light toward the center of the image than toward the periphery, preventing the bright inner corona from washing out the detail of the rest. The colors correspond to the intensity of the light at the wavelength of 6,400 angstrom units from each part of the corona; white areas have the highest intensity and deep violet areas the lowest intensity. Dark lines are contours of equal intensity.

solid metal, but the current-carrying capacity of a body is proportional to its area as well as to its conductivity; hence the broad extent of the gas makes it capable of carrying large currents. The cold, dense, nonconducting atmosphere and the rocky, poorly conducting surface of planets such as the earth are the only effective electrical insulators to be found anywhere in the universe.

The second requirement is a pattern of fluid motion that amplifies the magnetic field. In principle many types of fluid motion would amplify the field, but the naturally occurring combination of nonuniform rotation and cyclonic convection is a particularly effective recipe. All that is necessary for such a combination is a rotating body either containing or composed of a convecting fluid. As a consequence of the convection the rotation of the fluid tends to be nonuniform, and as a consequence of the rotation the convection is cyclonic.

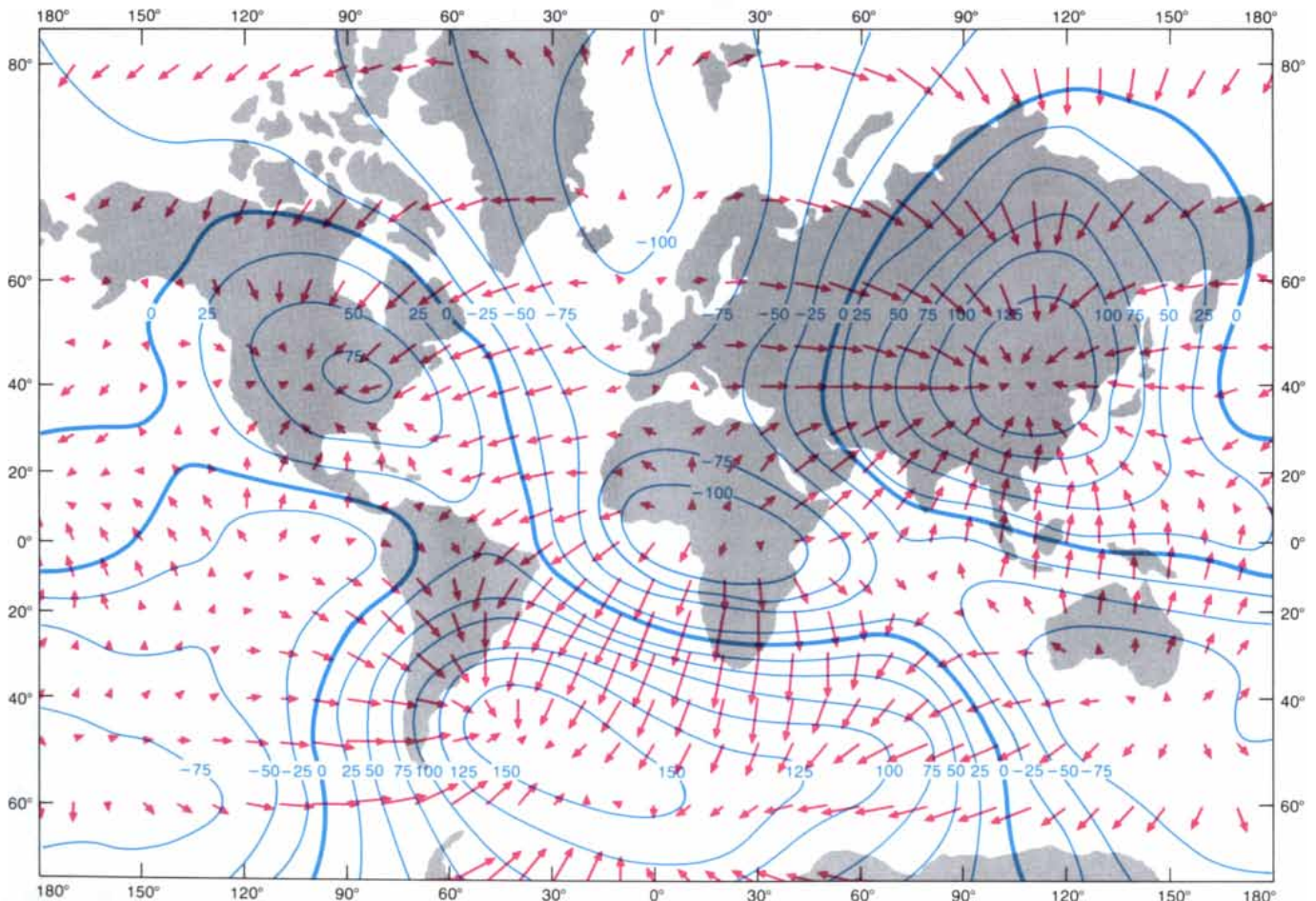
The most familiar example of this pattern of motion is the circulation of the earth's atmosphere. The variation of the direction of the prevailing winds with latitude reflects the nonuniform rotation of the atmosphere. For example,

the tropical trade winds are easterlies because air whose angular momentum is determined at higher latitudes loses rotational speed as it moves farther from the earth's axis of rotation; hence in the equatorial zone the air is moving slower from west to east than the earth's surface. The atmosphere's convective cells (the high- and low-pressure regions) are given a cyclonic twist by the rotation of the earth. As a result rising (low pressure) areas spiral counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere; the reverse is true of sinking (high pressure) areas. The earth's atmosphere does not function as a dynamo simply because there are no free electrons to carry the necessary electric current. Similar motions can be expected, however, in the molten core of the earth, in the convective zone of the sun and on a gigantic scale in the gaseous disk of the galaxy, all of which are sufficiently conductive to function as dynamos.

The first cosmic dynamo to be successfully modeled was the terrestrial one. A necessary precondition for the development of the model was an accu-

rate picture of the interior of the earth. This was supplied by the study of the paths and velocities of seismic waves spreading out from the focus of an earthquake. By 1940 it was known that the inner half of the earth's radius (the radius is 6,400 kilometers) is occupied by molten metal with an electrical conductivity not much less than that of ordinary solid iron. At the center of the liquid core there is a small solid core of crystalline metal with a radius about an eighth that of the earth. The small solid core plays no essential role in generating the magnetic field, and so it is ignored in most discussions of the subject. The outer half of the radius of the earth consists of the mantle and the crust. The hot silicate minerals of these layers are relatively poor conductors of electricity and for the present discussion can be treated as insulators. In this context the mantle is largely an impediment that hides the fluid motions and electric currents in the core and masks rapid small-scale fluctuations of the magnetic field itself.

The dipolar (two-pole) magnetic field detected at the surface of the earth is necessarily associated with circular electric currents flowing from east to



LOCAL ANOMALIES in the earth's magnetic field probably correspond to convection cells in the earth's core. These maps show the residual or irregular component of the earth's field, the part that remains when a north-south dipole field is subtracted from the total ob-

served field. The arrows represent the horizontal component of the residual field. The longest arrows correspond to a field strength of about 10,000 gammas. (One gamma equals 10^{-5} gauss.) The contours connect points on the surface of the earth where the vertical

west in the molten metal core. A simple calculation indicates that associated with the field are currents of some 2×10^9 amperes. At first attempts to explain the generation of the earth's field appealed to a variety of atomic effects, such as the thermoelectric effect, as being the agencies potentially capable of producing the separation of charge required to cause the necessary flow of current. In 1945, however, Walter M. Elsasser pointed out that advances in atomic and solid-state physics made it possible to rule out the thermoelectric effect and all other atomic processes as the cause of the earth's field. He emphasized that the only remaining possibility was for the currents supporting the field to be induced by the motion of the conductive liquid across the lines of magnetic force of the field itself. He was proposing the dynamo effect.

The implication of this explanation of the earth's field is that the fluid outer core is in motion, something for which there is no direct evidence because of the interposition of the mantle. The variation of the earth's field with time, however, can be explained only by motions in the core. Records of the direction and

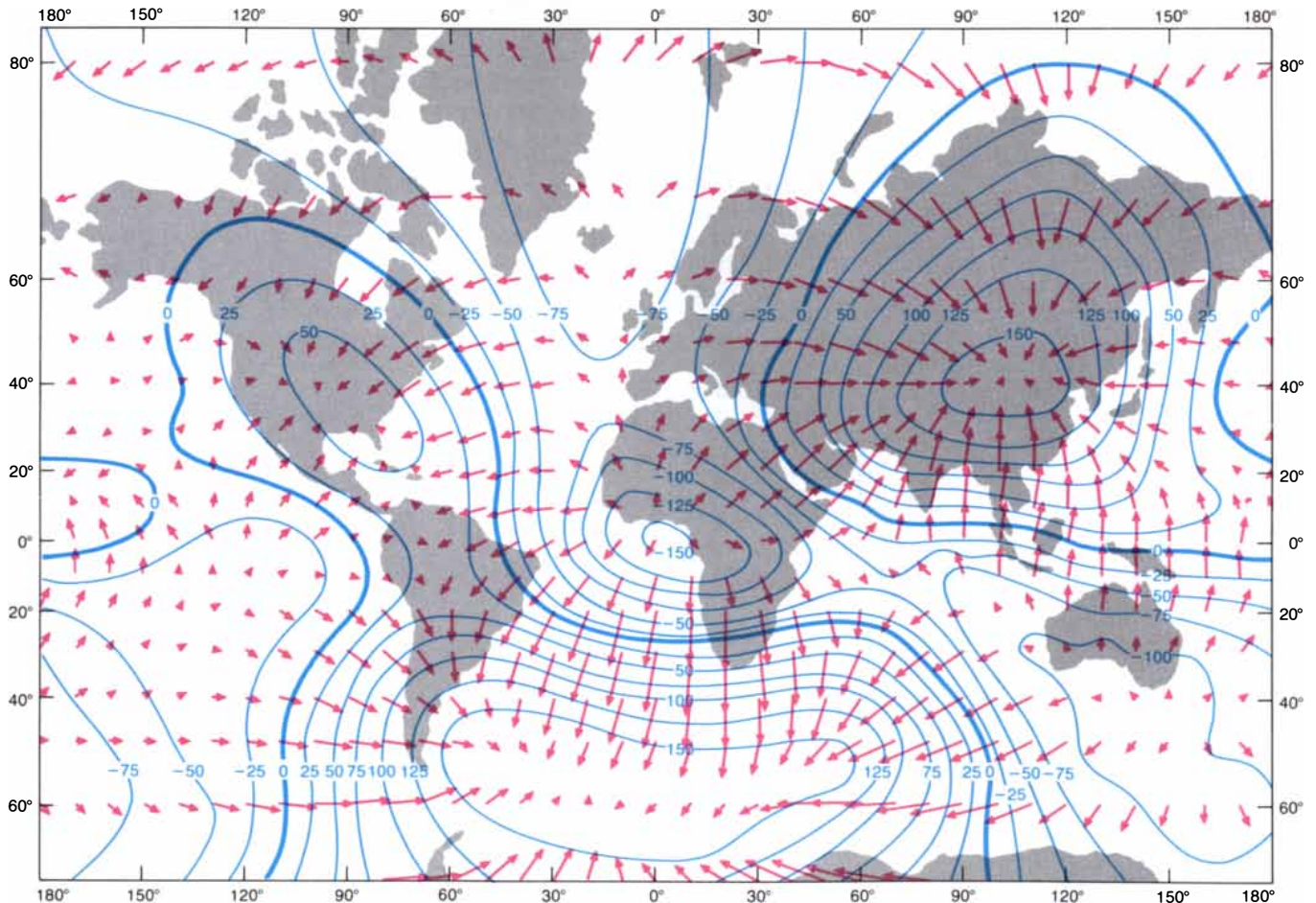
magnitude of the field at the surface have been kept since the time of Carl Friedrich Gauss, the great 19th-century mathematician and physicist for whom the unit of magnetic-flux density is named. They show a dozen or more identifiable local anomalies in the field, anomalies that have dimensions of several thousand kilometers and amplitudes on the order of 10 percent of the dipole field. The anomalies change slowly with time, increasing or decreasing over lifetimes of several centuries. Elsasser pointed out that the changing magnetic pattern corresponds to a broad convective pattern of a dozen identifiable cells in the liquid core.

The known cells lie at low and middle latitudes. They have a definite tendency to drift westward at a rate of about .18 degree per year. The drift suggests that the surface of the core, through which the field emerges, rotates slower than the overlying mantle. The drift rate corresponds to a fluid velocity of .3 millimeter per second (about a meter per hour) at the surface of the core.

Elsasser noted that the most obvious explanation for the slow rotation of the core is the action of the Coriolis force on

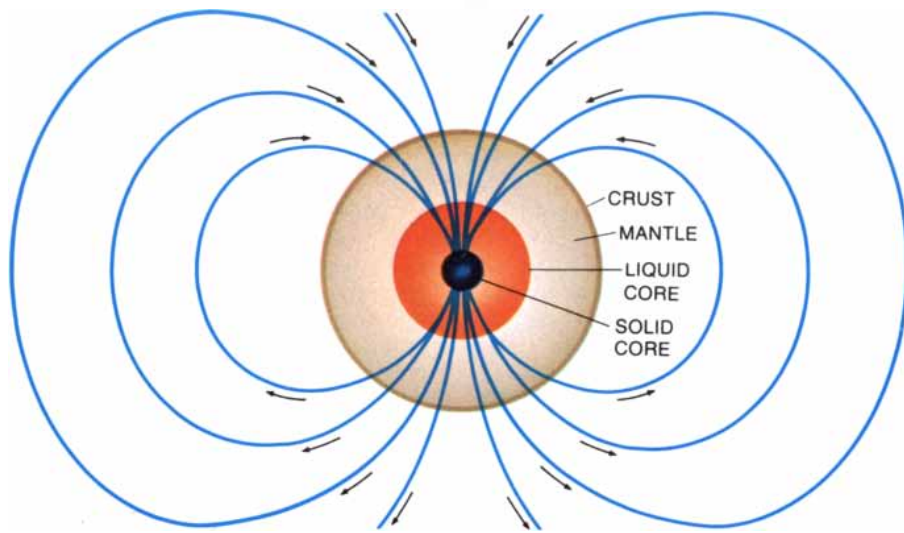
the rising and sinking fluid in the convective cells. The conservation of angular momentum requires that the angular velocity of the rising fluid decrease as it moves farther from the spin axis of the earth. It follows that the core surface rotates faster at high latitudes than it does at low latitudes and that at low latitudes the inner part of the core rotates faster than the surface.

Elsasser set out to calculate the effect of the assumed motion of the electrically conducting fluid in the core on the dipole magnetic field. The first important fact to emerge from his investigation was that the primary magnetic field in the core of the earth is an east-west field, at right angles to the main component of the field at the surface. It is called the azimuthal field; in general that part of a field lying in the planes perpendicular to the axis of rotation of a magnetic body is called an azimuthal field and that part lying in the planes through the axis is called a meridional field. The azimuthal field of the earth's core is created by the stretching of the north-south lines of force of the dipole field as they are carried around in the rotating fluid of the core. The part of a field line lying



component of the residual field is of equal magnitude. The contours are labeled in hundreds of gammas; negative numbers correspond to field directed downward from the horizontal. The magnetic anomalies drift westward at a rate of about .18 degree per year. For exam-

ple, the magnetic anomaly in central Africa in 1907 (map at left) had drifted to the west coast of Africa by 1945 (map at right). The westward drift of the anomalies shows that the surface of the fluid core, from which fields emerge, rotates slower than the overlying mantle.



TERRESTRIAL DYNAMO operates in the fluid outer part of the earth's core. The nature of the core has been deduced from the paths of seismic waves spreading out from the foci of earthquakes. The outer core is known to be fluid (that is, to have a negligible resistance to shearing compared with its high resistance to compression) because it transmits pressure waves but not shear waves. The outer core is mainly composed of iron that segregated from other elements in the interior of the earth soon after its formation. Although iron itself can be a magnet, at temperatures as high as those in the core ferromagnetic materials lose their magnetic properties. The magnetism of the core is due to the electric currents flowing through it rather than to the ferromagnetism of the iron. Liquid iron has a conductivity not much less than that of solid iron, and so the core is able to carry the electric currents associated with the earth's magnetic field.

near the axis is carried around faster than those parts lying farther from the axis. The nonuniform rotation stretches the north-south field lines in an east-west direction. The resulting field is directed toward the east in the Northern Hemisphere and toward the west in the Southern Hemisphere.

As the field lines are carried around, the azimuthal field gets steadily stronger. The amplification continues until it is balanced by an opposing factor such as the tension of the magnetic lines of force or the resistive decay of the associated electric current. Edward C. Bullard of the University of Cambridge, recognizing the importance of Elsasser's work, took up the problem and showed that the azimuthal field in the core may be hundreds of times stronger than the dipole field observed at the surface. The dipole field is about .3 gauss on the surface at the Equator, but the azimuthal field in the core may be 100 gauss or more.

The dipole field at the surface of the earth must then be a secondary effect of the azimuthal field. The idea was that the upwelling of fluid in convective cells in the core deforms the lines of force of the azimuthal field, giving rise to loops of field in the meridional planes of the core that represent a net dipole field. The mechanism seemed promising, but demonstrating that convective motion would generate a dipole field proved to be a formidable task.

The problem lay in devising a mathematical representation of the velocity field and the magnetic field that dealt

efficiently with the endless interactions of the two, and in devising an adequate model of the fluid motions. Bullard developed a way of representing the fields that facilitated the calculations. He then introduced the fluid velocities one might expect from a few steady, approximately circular convective cells. His calculations showed that the convection produced a variety of small-scale magnetic fields with which the convection interacted in turn, producing fields on even smaller scales. He pursued the computations to progressively smaller scales without finding conclusive evidence for the generation of a net, large-scale dipole field.

In the 1930's T. G. Cowling of the University of Leeds, who had been studying the interaction of a conducting fluid and a magnetic field, had proved a theorem stating that fluid motion cannot generate a perfect dipole field, or any field with rotational symmetry about an axis. Being aware of Cowling's theorem, Bullard deliberately placed his convective cells irregularly around the core on the hypothesis that the roughly axisymmetric field of the earth might be the residue of an asymmetric field in the core generated by these motions. Given the negative results of Bullard's calculations, it looked very much as though there might be a "super-Cowling theorem" to the effect that no magnetic field can be generated by *steady* fluid motion, whether the motion is accompanied by asymmetric fields or not. It was a black day and a dark "theorem" for those in-

terested in the origin of the magnetic field of the earth.

I had the good fortune to work with Elsasser for a couple of years around 1954. Bullard and Elsasser corresponded frequently, and on one occasion Bullard came for a visit of several days; I could not help becoming interested in the question of the origin of the earth's field. Bullard's conjecture that there was a "super-Cowling theorem" made a strong impression on me. As a result I began to wonder what effect *nonsteady* fluid motions would have on the magnetic fields. What if the upward and downward motions in the fluid were switched on strongly but briefly and then were switched off for an extended period? For one thing rapid, short-lived fluid motions would simplify the calculations. When the motion was switched off, the electrical resistance of the fluid would cause the field to "relax" and smooth out; the complicated small-scale fields would disappear, leaving only the large-scale residue and the strong azimuthal field.

When I began to consider intermittent convection, it immediately became apparent that the more cyclonic the convection was, the more effective it would be in generating the dipole field. Cyclonic fluid motion raises and rotates the lines of force of the azimuthal field, deforming them into helices. In the Northern Hemisphere, where the field runs from west to east, a rising cell creates a loop of field whose outer part is directed northward and whose inner part is directed southward. In the Southern Hemisphere the azimuthal field has the opposite direction, but the rotation of the cyclonic cells is also reversed, so that the helix has the same direction of rotation. Each loop in a helix has a small meridional component. When the loops diffuse and merge with neighboring loops during the time the fluid motion is switched off, they create a general meridional circulation of field that is recognized as the dipole field. Thus the intermittent cyclonic convection generates a net dipole field.

In 1958 Arvid Herzenberg of the University of Manchester proved that a magnetic field can be generated by steady fluid motions as well as by intermittent motions. Hence Bullard's conjecture proved to be wrong in the end, although it served the purpose of forcing inquiry in a fruitful direction. He had encountered a practical difficulty rather than a fundamental one: the fluid motions he chose had a relatively weak dynamo effect and the corresponding dynamo equations were correspondingly difficult to solve. It has since been shown that essentially all fluid motions have at least some small ability to generate a magnetic field. Ironically the exceptions are the simple motions that are symmetric about a plane or an axis, the

motions one would choose in seeking the simplest mathematical solutions to the dynamo equations. The essential ingredient for the generation of a field is that the motion of the fluid be helical: the fluid must rotate about its direction of motion as it streams along. Thus cyclonic convection even in the absence of nonuniform rotation is sufficient to generate a magnetic field. The naturally occurring combination of cyclonic convection and nonuniform rotation is, however, about the most efficient scheme that can be devised.

The test of the model of the terrestrial dynamo was whether or not it could be shown to amplify the dipole field at a rate high enough to balance the decay of the field. The magnetic field in a current-carrying body decays in a characteristic time (the resistive magnetic relaxation time) that is proportional to the current-carrying capacity of the body. The capacity in turn is proportional to the conductivity multiplied by the

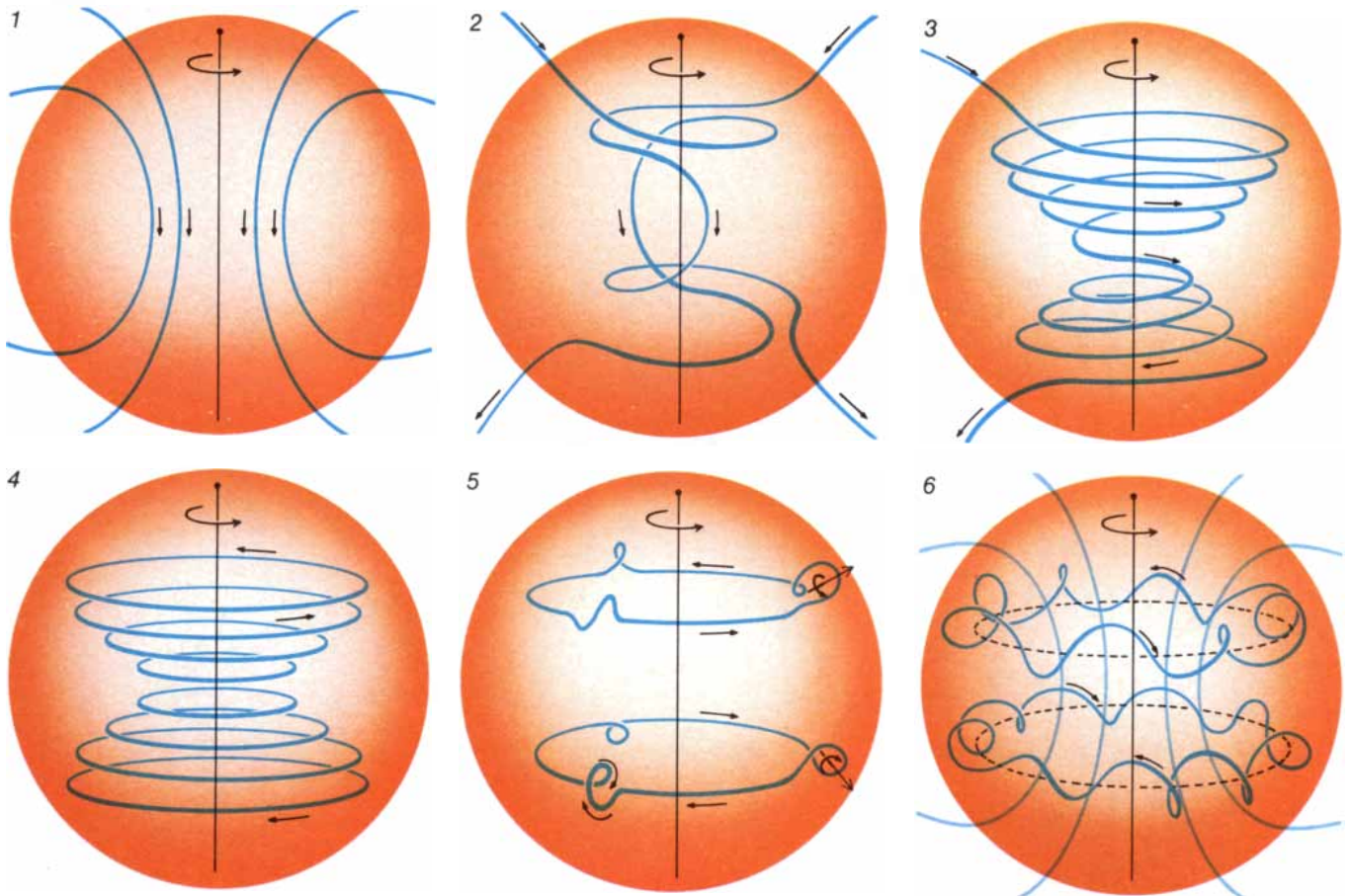
cross-sectional area. Because of its size the earth's core has a long magnetic relaxation time. If the generation of the field ceased today, it would be 30,000 years before the electric current and the magnetic field strengths fell to half their present values.

The strength of the azimuthal field is determined by the number of times the field lines are wrapped around the earth in their 30,000-year lifetime. Therefore an azimuthal field of 100 gauss or more in the core at the Equator may be associated with the dipole field of .3 gauss at the Equator. The rate at which the dipole field is generated from the azimuthal field can then be calculated on the basis of a convective velocity of one meter per hour at the surface of the core. The model of the terrestrial dynamo based on the known constraints (the strength of the dipole field, the resistivity of molten iron and the fluid velocity at the surface of the core) easily regenerates the field at a rate equivalent to replacing the dipole field once every

30,000 years, that is, at a rate equal to the rate of decay.

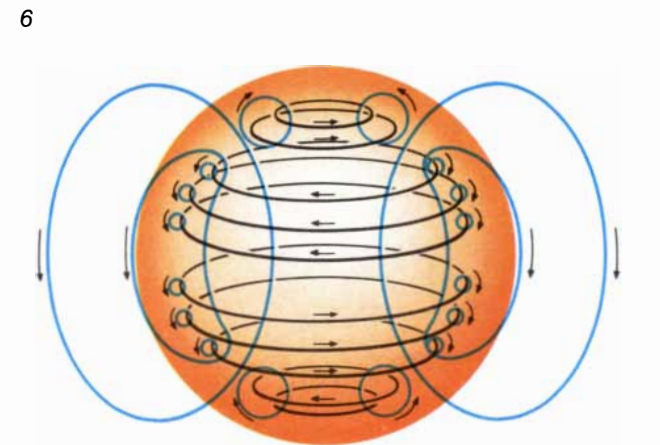
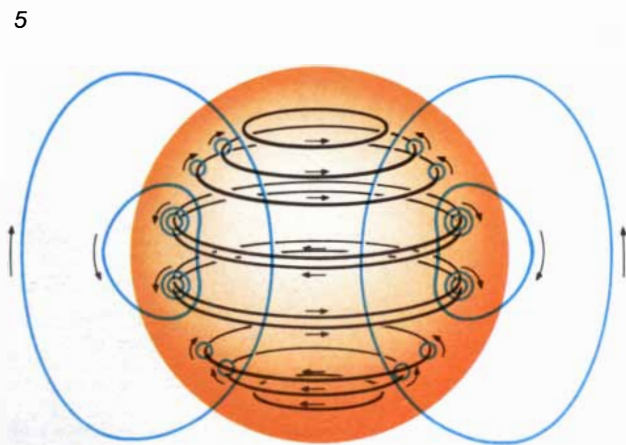
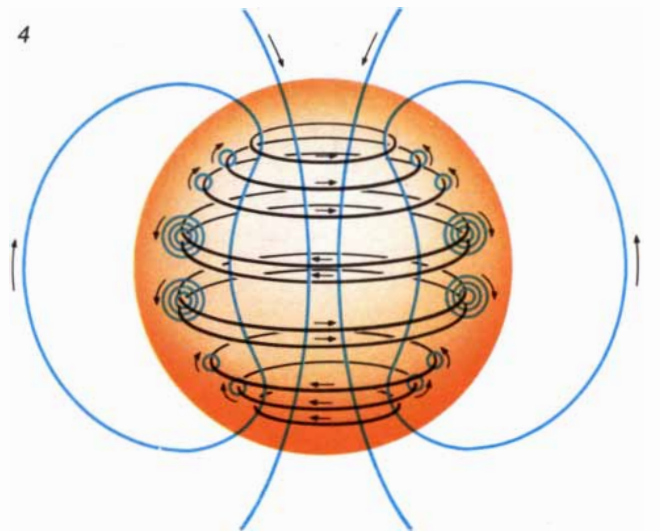
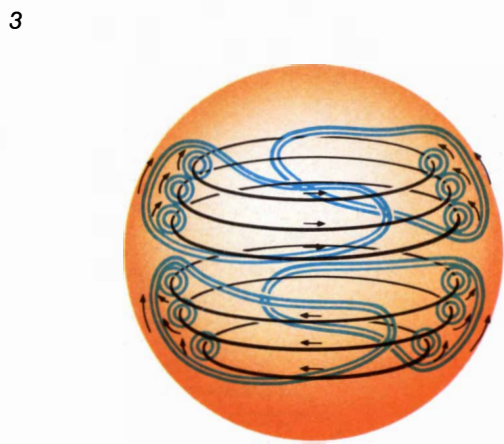
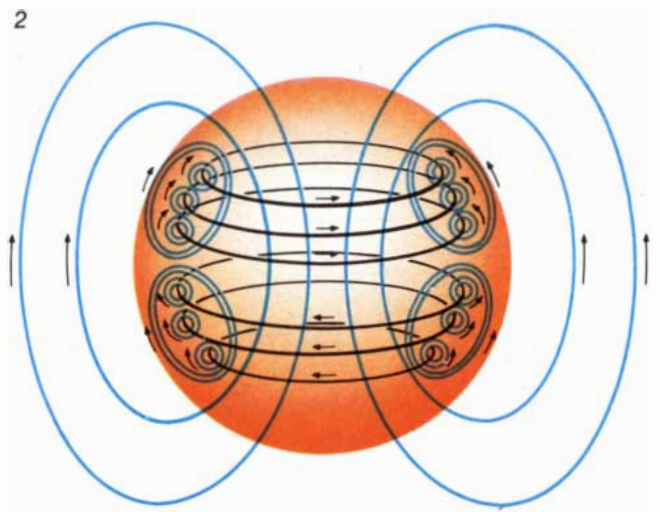
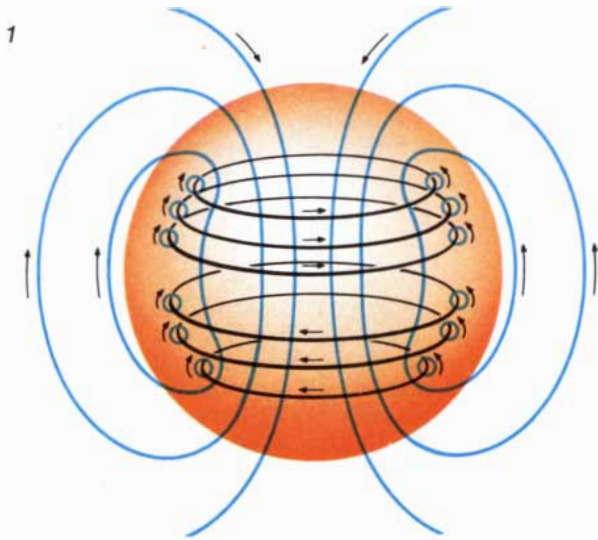
Another property of the earth's field that stands in need of explanation is its ability to reverse polarity. In 1955 S. K. Runcorn of the University of Newcastle upon Tyne pointed out that the alignment of grains of magnetic iron oxide in lavas indicates the direction of the local magnetic field at the time of their crystallization. He went on to show that this paleomagnetic record indicates the earth's field has reversed polarity abruptly (in only about 1,000 years) at random intervals of from 100,000 to 10 million years. It has been shown since then that a sudden change in either the location or the strength of the convective cells can throw the dynamo into a chaotic state in which the local fields run in the wrong direction and generate a large-scale field of the opposite polarity. The required changes are modest (as little as 20 degrees of latitude) and therefore plausible.

In spite of the success of the hydro-



DYNAMO MODEL of the generation of the earth's field holds that the field lines are stretched by fluid motions in the core; as a result the field is amplified at a rate high enough to balance the rate of decay. A much simplified version of the process is shown here. The north-south lines of force of the dipole field (1) are drawn out in the east-west direction (2) by the nonuniformly rotating fluid in the core. The angular velocity of the fluid decreases with distance from the spin axis of the earth, and the part of a field line in the inner core is carried around faster than the part near the core surface. The resulting azimuthal field is directed eastward in the Northern Hemisphere and westward in the Southern Hemisphere. Over their 30,000-year

lifetime the field lines are wrapped around the core many times and the azimuthal field gets strong (3). If the original dipole field is subtracted from the spiraling field created by nonuniform rotation, the result is a purely azimuthal field (4). Upwelling fluid in cyclonic convection cells creates loops in the rings of azimuthal field (5). Rising fluid spirals counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. As a result the outer part of a loop in either hemisphere is directed north and the inner part is directed south. When the loops in the helices into which the azimuthal field lines are deformed diffuse and merge, they create the large-scale meridional field detected at the earth's surface as a dipole field (6).



REVERSAL OF THE EARTH'S MAGNETIC FIELD might be caused by a sudden increase in the velocity of convection in the earth's core. Ordinarily the small meridional loops (blue) into which the azimuthal field (black) is carried by upwelling fluid have time to merge, creating the large-scale dipole field (1). If the velocity of convection increases (2), the local meridional fields do not have time to merge before the nonuniform rotation of the core begins to stretch them in the east-west direction. The new azimuthal field runs in the direction op-

posite to that of the existing field at low latitudes and in the same direction at high latitudes (3). The addition and cancellation of the fields shifts the old azimuthal field to higher latitudes and produces a reversed azimuthal field at low latitudes (4). Convection produces reversed meridional loops of field from the reversed azimuthal field. The general meridional field created by the diffusion of these loops crowds out the original field (5, 6). The scheme shown here is a model; the mechanism by which the field reverses polarity is not known.

magnetic model of the terrestrial dynamo it would be misleading to suggest that the problem of the earth's field has been completely worked out. So far it has been demonstrated that fluid motions of the kind thought to exist in the core of the earth could in principle generate the observed magnetic field. The fluid motions used in the formal mathematical demonstration, however, and even those used in the sophisticated numerical experiments done recently were simplified in one or another way to make the calculations tractable. Ultimately the goal is to start from first principles, such as the rate of rotation and the energy source that drives convection. Instead of postulating a certain pattern of circulation one might deduce a precise quantitative model of the fluid motions and of the local magnetic fields in the core. The various effects of the convective flow on the field and vice versa might then be calculated.

In all probability it will be some time before the problem can be solved completely, largely because of the lack of hard information about the source of the energy driving convection. Both the heat released by the decay of the radioactive isotope potassium 40 and the settling of a denser phase of iron through a lighter liquid phase may contribute to convection. Similarly, although sudden changes in the pattern or rate of convection could cause reversals of polarity, it is not known why there should be sudden changes. Some recent studies of coupled mathematical systems indicate that such systems have repeated erratic episodes, but it has not been shown how these results can be applied to convection in the core.

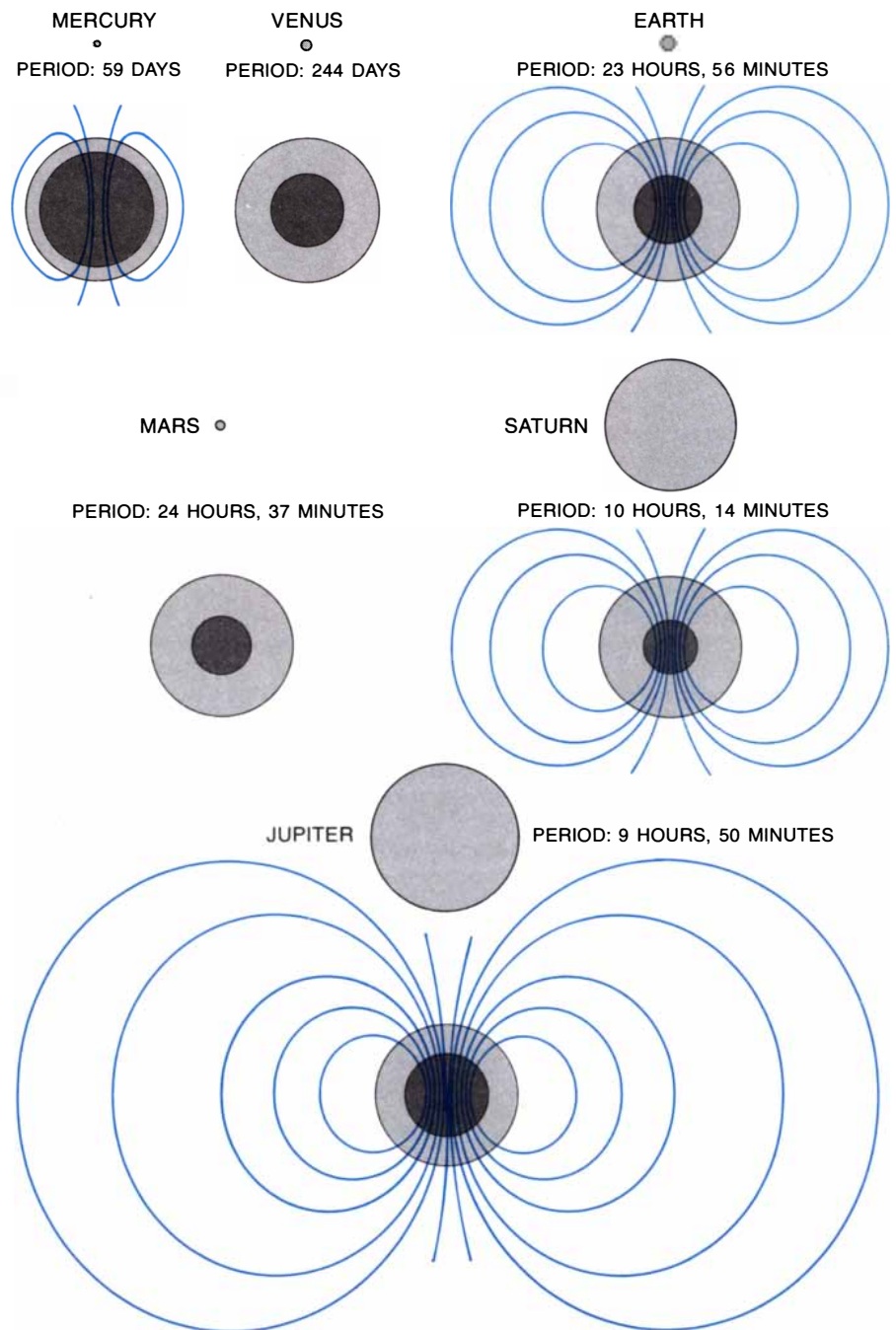
There is much less information to go on in the case of planets other than the earth. Thus the discussion of planetary fields is limited to basic questions. Why, for example, does the small planet Mercury have an intrinsic magnetic field, whereas Venus, which is roughly the size of the earth, has none? Why does Mars have no field?

The rate at which a planet generates a magnetic field is defined by what is known as the dynamo number. The dynamo number is the product of three quantities—the rotary velocity of the convective cells, the velocity gradient produced by nonuniform rotation and the volume of the core—divided by the resistivity of the body. If the dynamo number is below a threshold value, the body is incapable of generating a magnetic field as fast as the field decays. If the dynamo number is above the threshold, the rate of generation exceeds the rate of decay and the mode or spatial complexity of the growing field depends on the value of the number. For instance, a steady dipole field is generated by a body with a dynamo number lower

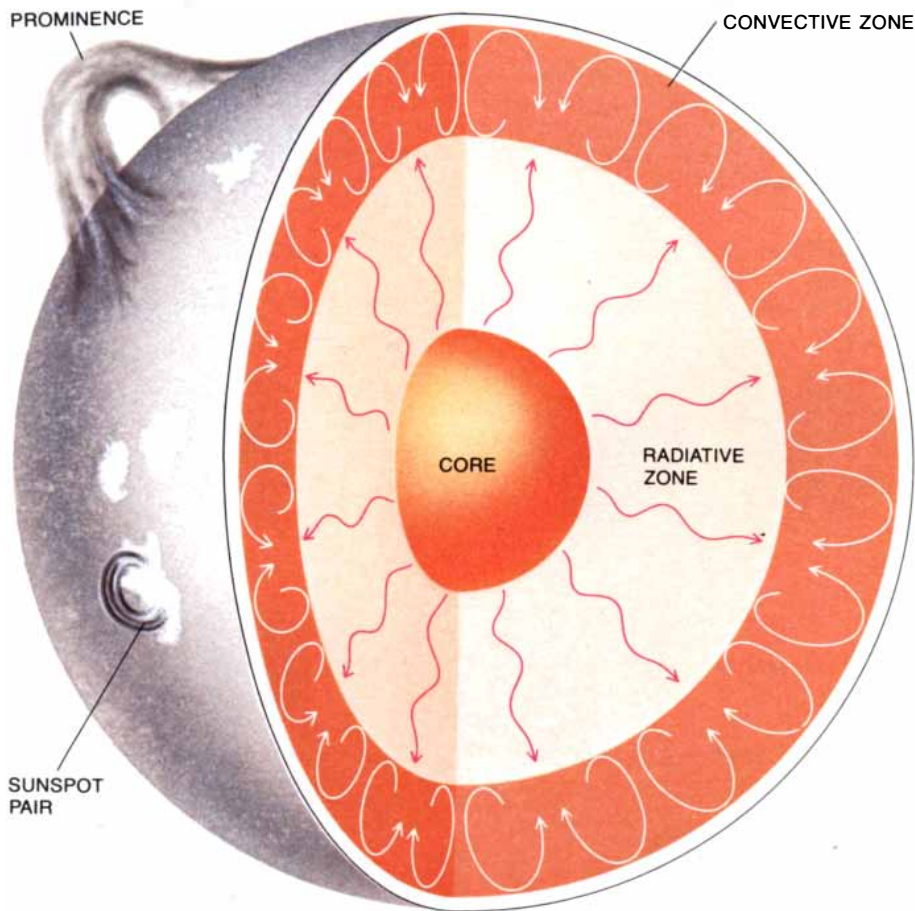
than that of a body generating an oscillating quadrupole field.

Mercury is a small but dense planet. It has a radius about a third the radius of the earth, but its density indicates it has a relatively large metallic core with a radius about half the radius of the

earth's core. On the other hand, Mercury rotates relatively slowly (with a period of 59 days). Because of the low rotation rate Mercury was not expected to generate a field. It came as a surprise, therefore, when the *Mariner 10* spacecraft registered a dipole magnetic field



STRENGTH OF A PLANET'S MAGNETIC FIELD is generally correlated with the absolute size of the planet's conductive core or layer and the planet's rate of rotation. Venus, for example, has no field because it rotates very slowly (with a period of 244 days). Both Jupiter and Saturn have strong fields because they rotate rapidly (with periods of about 10 hours) and have interior layers of metallic hydrogen and helium that are about as conductive as copper at room temperature. Saturn's surface field may be weaker than Jupiter's either because Saturn's layer of convecting metallic hydrogen is overlain by a static, perhaps rigid, layer of metallic hydrogen or because convection is weaker in Saturn than it is in Jupiter. In view of the presence of a field on Mercury the absence of a field on Mars is something of a puzzle. The size of Mars's core is not known for certain. The core may be very small or it may be nearly as large as Mercury's. Mars rotates more than 50 times faster than Mercury, yet Mars generates no field. It may be that Mars has a core too small to generate a magnetic field or that the convection in Mars's core is unusually strong and so compensates for the small core and slow rotation.



SOLAR DYNAMO is the convective zone of the sun, which has a radius about a fourth that of the sun. This thin-shell dynamo cannot generate a meridional field fast enough to maintain a large-scale field stretching across the entire sun; the field is therefore an oscillating dipole rather than a steady one. With this qualification the operation of the solar dynamo is similar to that of the terrestrial dynamo. Twinned sunspots and solar flares are manifestations of strong local azimuthal fields. High-latitude prominences are aligned with the local meridional field lines.

as it flew by Mercury in 1974 and again in 1975.

The weak irregular fields near Venus and Mars appear to be nothing more than the sun's field carried to and pressed against these otherwise non-magnetic spheres of rock by the wind of electrically charged particles flowing outward from the sun. Venus, often called the sister planet of the earth owing to the similarities in size and internal structure, apparently generates no field because it rotates very slowly (with a period of 244 days). Mars rotates at essentially the same rate as the earth does (with a period of 24 hours 37 minutes). The low density of Mars, however, suggests that if it has any metal core, it is a small one. The rate at which a magnetic field dissipates varies inversely with the square of the radius of the magnetic body, and so it may be that any field generated in the core of Mars is lost faster than it can be regenerated.

Comparisons among these four planets raise some interesting and as yet unanswerable questions. The field at the surface of Mercury is only a thousandth as strong as the field at the surface of the earth, but the very existence of the field

is puzzling. If Mercury can maintain a steady dipole field, the earth, which rotates 59 times as fast and has a core twice as large, should be able to sustain more complicated fields. Perhaps some factor inhibits the generation of such fields in the earth's core, or perhaps fluid convection in Mercury's core is more vigorous than it is in the earth's and makes up for the smaller core and slower rotation. Mars has no field, although it rotates more than 50 times as fast as Mercury. If Mars's core is comparable in size to Mercury's, as some workers have argued on the basis of the mean density of the planet, the absence of a field on Mars and the presence of one on Mercury is baffling. Evidently Mars's core is smaller or less convective than Mercury's, and it may be much smaller and less convective. Given the meagerness of the information about the convection in planetary cores, questions such as these are likely to go unanswered for some time to come.

The sun's dynamo is a hollow spherical shell rather than a full sphere like the dynamos of the planets. The dynamo is the sun's convective zone, which

begins just below the visible surface and extends a fourth of the distance to the center of the sun. The sun's period of rotation is from 25 to 35 days. The rates of rotation of solar features such as sunspots show that the surface of the convective zone rotates nonuniformly, the surface at the poles having an angular velocity about two-thirds the velocity at the equator. The convection is visible as granulation, the rice-grain markings seen in photographs of the sun's surface. Because the sun is rotating, the convection is undoubtedly cyclonic.

The solar dynamo operates much like the terrestrial one. The sun's azimuthal field, unlike that of the earth, is directly observable because the top of the convective zone can be seen. Twinned sunspots, and bipolar magnetic regions in general, are local bulges in the azimuthal field pushed to the surface of the sun by magnetic buoyancy. The azimuthal field below the surface is at least as strong as 200 gauss and may be as strong as 10,000 gauss. The meridional field generated from the azimuthal field is visible only at high latitudes, where it shows up in coronal streamers and polar prominences. It is much weaker than the azimuthal field, having a strength of perhaps five to 10 gauss.

The sun has a magnetic cycle with a period of 22 years. The coordinated waxing and waning of the azimuthal and meridional fields in the course of the cycle cause the field to reverse polarity about once every 11 years. It can be shown that this magnetic dance is a consequence of the characteristics of the solar dynamo. A thin-shell dynamo such as the sun's loses field too rapidly for the generated field to reach diffusive equilibrium over the full length (in latitude) of the shell. The low-latitude field has not had time to reach equilibrium before an opposite field is generated at high latitudes. Hence a thin-shell dynamo has no steady dipole mode; the lowest available self-sustaining mode is an oscillating dipole.

The dynamo mechanism as the explanation for the fields of some of the magnetic stars and of the galaxies does remain challenged by the rival hypothesis that the fields are primordial, trapped in the stars and galaxies since the formation of the universe. Stars are formed by the gravitational collapse of interstellar gas. The interstellar gas is threaded by the galactic magnetic field, so that a volume of gas carries some of the field with it when it collapses to form a star, compressing the field by the same factor as the surface area of the gas is compressed. Much of the trapped primordial field escapes from the star in the course of the vigorous mixing that accompanies the final stages of collapse, but it is not unreasonable to suppose some significant amount is retained.

The planets were formed in a similar

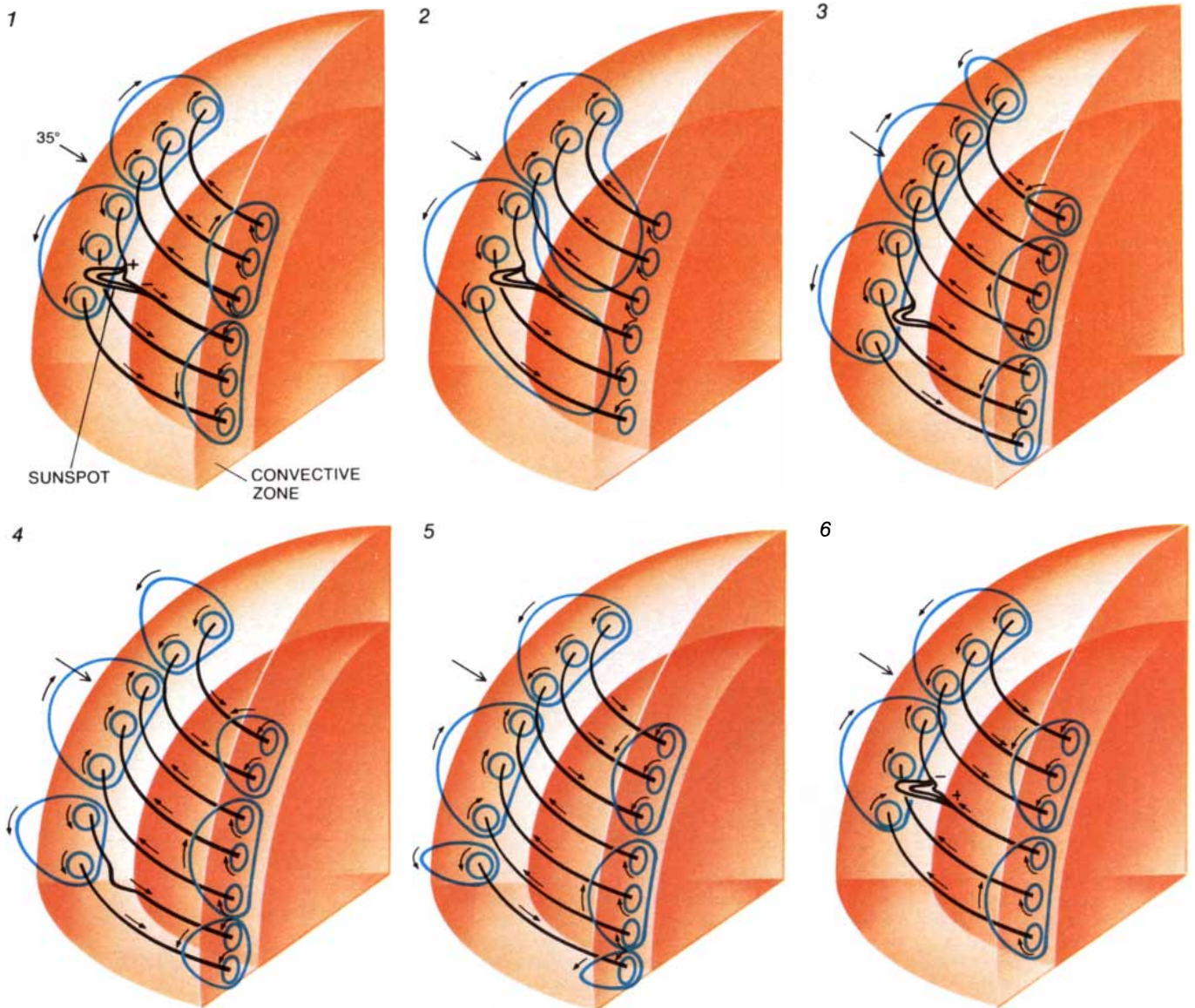
way, but in their case there is a simple argument against primordial fields. As I have mentioned, the magnetic relaxation time of the earth's core is about 30,000 years, far shorter than the earth's lifetime. The earth was formed more or less at the same time as the sun, some 4.5 billion years ago. Left to itself a primordial field would have disappeared long ago; the dynamo mechanism must be invoked to explain the presence of a field today. In the case of the stars the argument is less conclusive. As Cowling pointed out more than 30 years ago, the resistive-decay time of a star such as the sun is typically five billion years, comparable to the age of the sun. The sun's

magnetic cycle can only be explained on the basis of a dynamo, but the possibility that there is also an undetected primordial field as strong as a million gauss trapped in the stable core of the sun cannot be excluded.

The hypothesis of primordial fields is more interesting in the case of stars with fields much stronger than the sun's. One class of stars with strong fields consists of the magnetic *A* stars. (*A* stars are stars with a surface temperature of about 10,000 degrees Kelvin.) They were discovered in the late 1940's by Harold D. Babcock and Horace W. Babcock of the Mount Wilson and Palomar Observatories, who measured their

fields with a sensitive magnetometer they had developed. The fields of these stars range in strength from a few hundred gauss to 34,000 gauss.

It has since been shown that the field of a magnetic *A* star is a dipole that rotates more or less rigidly with the star itself. Moreover, the axis of the field tends to be perpendicular to the axis of rotation instead of approximately parallel as it is in planets and other stars. In addition the stars have a modest rate of rotation (a period of a month or so), and the more rapidly rotating stars tend to have weaker fields than the slower stars do. The lack of correlation between the strength and the mode of the field, be-



SUN'S MAGNETIC CYCLE begins with two bands of azimuthal field of opposite sign in each hemisphere (1). The band at high latitudes shows up as the "ephemeral" bipolar magnetic regions visible in magnetograms of the sun. The lower band shows up as twinned sunspots, the entry and exit points of an azimuthal field bulging above the surface of the convective zone. The meridional field generated from each band of azimuthal field by convection is stretched by the nonuniformly rotating gas of the convective zone into new azimuthal field (2). The new field reinforces the original field at the southern

edge of the bands and cancels it at the northern edge (3). As a result the bands of azimuthal field migrate toward the equator and a new band of opposite sign is created at high latitudes. At the equator the bands of azimuthal field are canceled by bands of opposite sign arriving from the other hemisphere (4). The sunspots associated with each band of field decrease in number and strength as the band slowly decays (5). The recovery from sunspot minimum to a new maximum is relatively rapid (6). This sequence of drawings shows one reversal of polarity, or one-half of the roughly 22-year solar magnetic cycle.

tween the strength of the field and the rate of rotation and between the orientation of the field and the axis of rotation makes it unlikely that the fields of these stars are generated by the dynamo mechanism.

It has also been suggested that the galactic magnetic field is primordial because the resistive-decay time of a body as large and as conductive as the galaxy is longer than the galaxy is old. The effect of the electrons circulating around the galactic field lines in interstellar space on the plane-polarized radio waves from distant sources provides a measure of the field strength. Observations of the variation of the plane of polarization with frequency (the Faraday rotation effect) show that the field is between 2×10^{-6} and 3×10^{-6} gauss. Presumably all that would be needed to account for the present field is that the diffuse gas from which the galaxy formed was threaded with a field of about 10^{-9} gauss. This field would be compressed when the gas collapsed to form the galaxy. It has been argued that the compression and subsequent stretching of the lines of force in the nonuniformly rotating galaxy could

have produced a field of the observed strength. Recent work by Jacques P. Vallée of the National Research Council of Canada, however, places an upper limit of 10^{-11} gauss on the present intergalactic field, a field strength much lower than the 10^{-9} gauss generally assumed in arguments for the primordial origin of the galactic field.

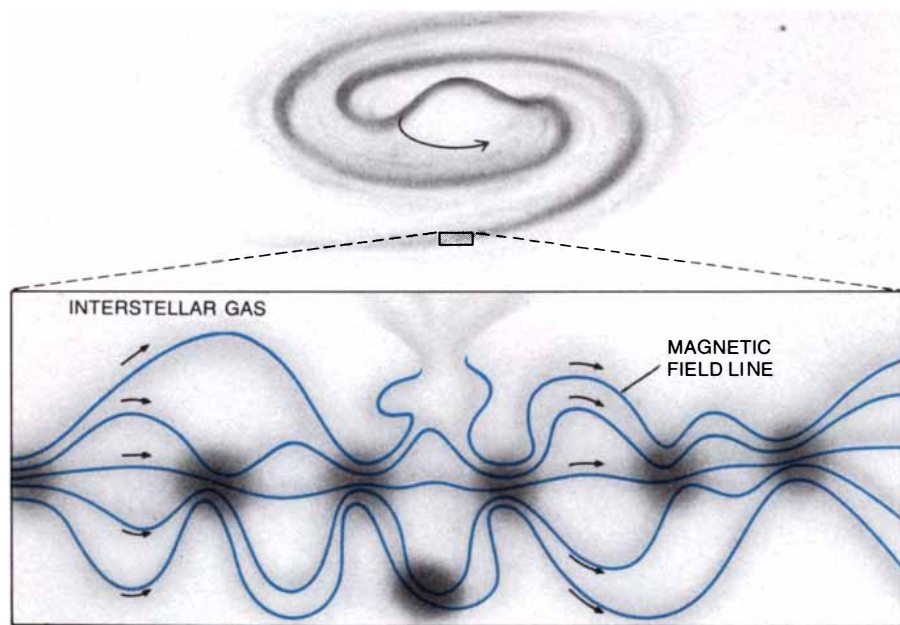
What is more, the system composed of the interstellar gas, the cosmic rays and the galactic magnetic field is highly unstable. The gas tends to gravitate toward the central plane of the galactic disk and the magnetic field tends to billow upward, away from the plane. Although there is not enough information to build a quantitative model of the turbulent loss of field, such calculations as are possible indicate that the vigorous small-scale undulations of the field cause it to wiggle out of the gas in the disk at a rate fast enough to preclude the possibility that a primordial field has persisted to the present day. Turbulent loss reduces the decay time of the field to about 300 million years, whereas the galaxy is 10 billion years old.

The dynamo mechanism successfully explains what is known about the galactic field. The dynamo equations recast in

the appropriate form for a thin disk with the effective conductivity of the turbulent interstellar gas show the galaxy could regenerate a magnetic field at a rate that would balance the loss by turbulent escape and resistive decay. The gaseous disk of the galaxy is known to rotate with an angular velocity that decreases outward from the center. Supernova explosions and the sudden formation and lighting up of massive luminous stars produce violent turbulence or convection. Because of the rotation, the turbulence is cyclonic. The primary field of the galaxy is azimuthal, extending around the disk of the galaxy. The meridional components of the galactic field cannot be observed because they are lost in the strong local fluctuations of the azimuthal field.

One last question deserves consideration even though it is, by virtue of the lack of evidence, more a philosophical question than a scientific one. The dynamo mechanism does not provide for the outright creation of a magnetic field but only for the reproduction, that is, the amplification, of existing fields. Where then did the fields the dynamos amplify come from? It may not be necessary to resort to the idea of a primordial field, the existence of which would in turn need to be explained. In 1941 Ludwig F. Biermann of the Humboldt University of Berlin pointed out that there are "atomic battery" effects in the moving ionized gas of the stars. For example, the isotherms, or surfaces of equal temperature, in a rotating star do not coincide with the oblate level surfaces at which the gravitational and centrifugal potentials are balanced. As a result of the unbalanced forces free electrons slide along these surfaces, creating weak meridional currents. If a star was initially free of magnetic field, the Biermann battery effect would over an extended period build up a weak azimuthal field. The battery effect alone cannot account for fields of the strength observed today, but it does guarantee that if all else fails, the stars and galaxies are seeded with weak magnetic fields on which the powerful hydromagnetic mechanism can then operate. Similar effects due to slight radioactivity or to chemical or thermal inhomogeneity would seed the planets with weak fields.

When all the evidence is considered, it looks as though magnetic fields are generated by fluid motions on all scales from the 2,000 kilometers of the liquid-metal core of a planet to the million kilometers of a star to the 10^{18} kilometers of a galaxy. Much remains to be learned both about the dynamos and about the unusual bodies that have fields too strong to be explained by the known dynamo effects, but the existence of these tantalizing unsolved problems is what makes the subject as interesting as it is.



GALACTIC MAGNETIC FIELD is predominantly azimuthal, that is, parallel to the plane of the galaxy, although the direction of the field fluctuates widely. The field's direction was originally determined from the polarization of light from distant stars. Dust grains in the interstellar gas are set spinning by collisions both with one another and with hydrogen atoms. Each dust grain carries an electric charge and hence tends to line up with its long axis perpendicular to a magnetic field line around which it spins. The polarization of the light absorbed and re-emitted by the dust lying between a star and the solar system is therefore determined by the orientation of the galactic magnetic field. The strong local fluctuations of the magnetic field are caused by its interaction with the interstellar gas and with the cosmic rays it confines and by which it is inflated. It is thought supernova explosions of massive stars blow the interstellar gas and cosmic rays outward from the plane of the galaxy, creating the spherical shells of gas visible on maps of the radio emission of atomic hydrogen, one component of the gas. The magnetic field threaded through the gas is carried with it and compressed in the periphery of the shell. If the magnetic field is not strong enough to balance the outward pressure of the gas, the gas and the entrained field are swept off into space. Such turbulence contributes significantly to the rate at which the galactic field dissipates. The rate of loss makes it unlikely the field is primordial and supports the argument that the field is regenerated by rotating and convecting gas.

Museum piece, circa 1987.

The clapstick may be taking an early retirement. This and other conventional tools of the filmmaking trade may find themselves relegated to crates and carted off to museums.

What will replace them? Kodak film with Datakode magnetic control surface! This radical advance in film manufacture, which unites chemical and magnetic imaging technologies, gives film the ability to "converse" with computers. This achievement could substantially reduce the time and costs associated with film postproduction.

The Datakode magnetic control surface is a thin, transparent layer of iron oxide particles (approximately 9 billion per square inch) coated across the entire back of the film during manufac-

ture. Less than 8 microns thick, this transparent layer provides the means to record machine-readable information (100 bits of binary data in a single track on each frame of film), and makes possible a uniform frame-indexing code that can be used with both film and videotape. Best of all, it accomplishes this interface without altering the *quality* or characteristics of the final image.

In the not-too-distant future, the use of discs, video displays, time-code synchronization, and automated printing equipment will speed filmmakers through all the noncreative, repetitive steps associated with film postproduction.



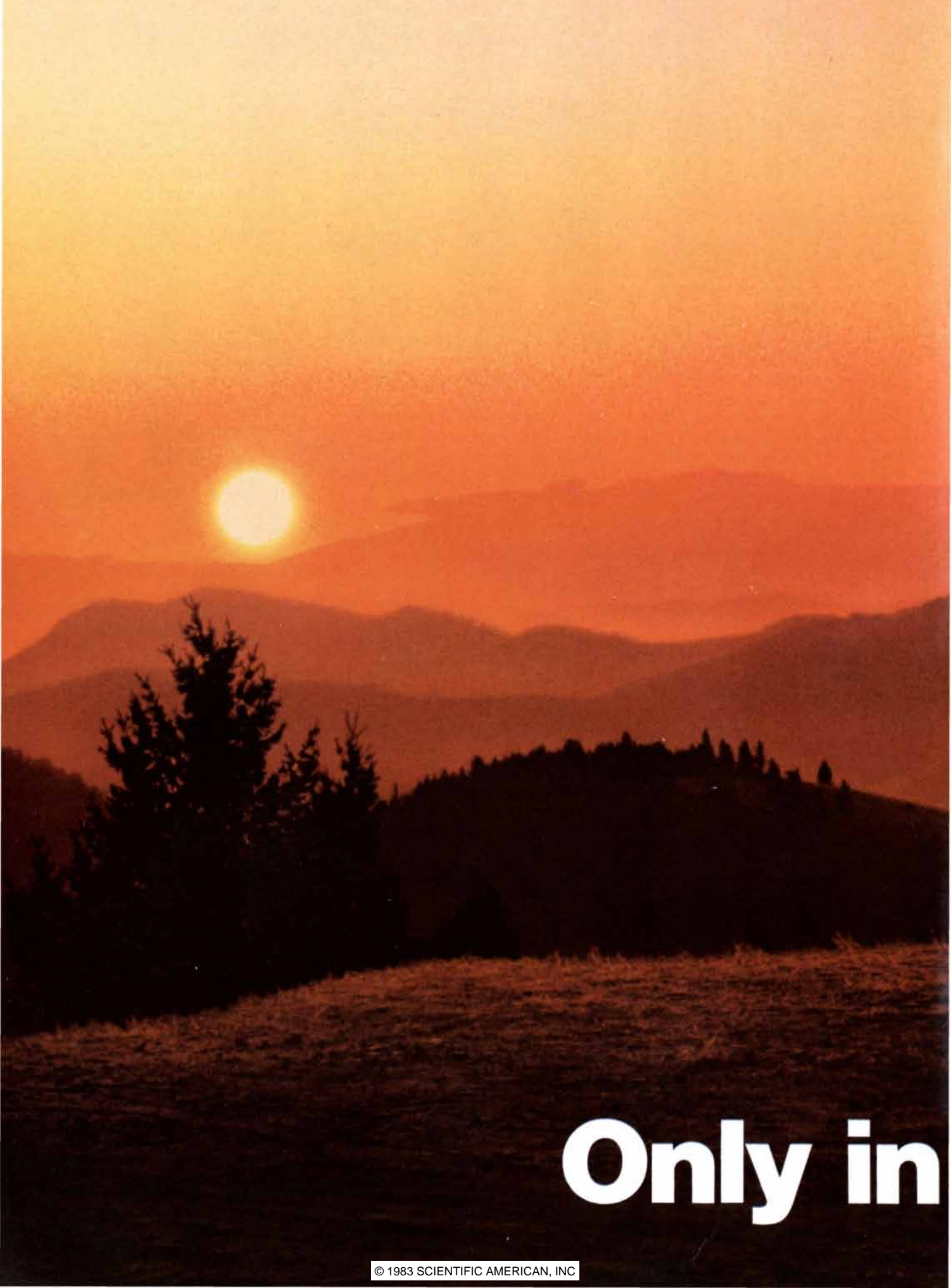
The new Datakode magnetic surface is more than just another milestone for Kodak and a boon to filmmakers. It represents the most significant development in motion picture technology in the last 50 years. It's a victory for science. And for the chemists and researchers who met the challenge to add to film's capabilities.

For more information on the new Datakode surface, send your request to: Eastman Kodak Company, Department GBSA-7, 343 State Street, Rochester, NY 14650.



Kodak. Where technology anticipates need.





Only in



a Jeep.®

Jeep  CJ



SCIENCE AND THE CITIZEN

From MIRV to MHV

Perhaps the most poignant aspect of the report by the President's Commission on Strategic Forces (the Scowcroft commission) is the implicit acknowledgment that in retrospect the U.S. decision of a decade or so ago to begin deploying multiple independently targetable reentry vehicles (MIRV's) on the nation's land-based strategic missiles was a mistake. The destabilizing effect on the strategic balance of highly accurate, "counterforce-capable" multiwarhead missiles has come to be generally recognized by U.S. military planners only after the U.S.S.R. followed the American lead and installed MIRV's on its long-range missiles, thereby threatening the land-based component of the U.S. deterrent forces. The response of the Scowcroft group and others is to advise that the U.S. now plan to revert ultimately to a single-warhead missile (nicknamed Midgetman) and try to persuade the Russians to do the same. In the words of former national-security adviser McGeorge Bundy, the introduction of MIRV technology ranks as "the United States' worst single contribution to the nuclear arms race."

The mistake is about to be repeated, in the view of many long-term students of the arms race. This time, however, the major technological initiative in question is the imminent introduction of a new generation of antisatellite, or ASAT, weapons. According to a petition prepared early this year by Richard L. Garwin and Carl Sagan and signed by more than 40 prominent American physicists, space scientists and defense experts, "the testing or deployment of any weapons in space [including antisatellite weapons] significantly increases the likelihood of warfare on earth." The petition, sent to the leaders of "all space-faring nations with an independent launch capability," calls for the prompt adoption of a treaty prohibiting the extension of the arms race to space, noting that "if space weapons are ever to be banned, this may be close to the last moment in which it can be done."

At present neither the U.S. nor the U.S.S.R. has a reliable ASAT system. For the past 15 years or so the Russians have apparently been testing a rather primitive ASAT weapon consisting of a non-nuclear explosive device launched into low earth orbit by an SS-9 intercontinental ballistic missile (ICBM). The system has reportedly succeeded in destroying an orbiting target satellite in about half of some 20 attempts. Most U.S. military satellites, including those responsible for long-range communications and early warning of a strategic-missile attack,

are in very high, geosynchronous or semisynchronous orbits and hence are considered invulnerable to the current Russian ASAT system.

Meanwhile the U.S. has developed a quite different ASAT weapon. It consists of a small, nonexplosive interceptor called the Miniature Homing Vehicle (MHV), which is mounted on a two-stage rocket designed to be launched into space from a high-flying F-15 fighter plane. The MHV is equipped with a combination of infrared telescopes, a laser gyroscope and an array of small jets that will enable it to seek its target and destroy it by direct impact at very high velocity. In its present form the MHV cannot reach satellites in geosynchronous orbits; unlike the U.S., however, the U.S.S.R. currently has almost all its "space assets" in either low or highly elliptical orbits, where they are considered more vulnerable. With a three-stage booster the U.S. ASAT weapon should be able to reach geosynchronous orbits, a capability the Russians could attain with their present system only by recourse to a new, much larger ground-based rocket.

The first test launch of the U.S. interceptor is expected soon, with tests against space targets scheduled for later this year or early next year. It is estimated that the MHV system could be operational as early as 1987. According to the Air Force, the projected total cost of the ASAT program is \$3.6 billion; unofficial estimates of the ultimate cost of the system by the General Accounting Office run as high as "tens of billions."

Bilateral negotiations to forestall the impending arms race in space are currently in abeyance, having been suspended by the Carter Administration in 1979 after the Russian intervention in Afghanistan. In 1981 the U.S.S.R. submitted to the United Nations a draft treaty calling for a "prohibition on the stationing of weapons of any kind in outer space" but making no reference to ground-based systems. Recently Foreign Minister Andrei A. Gromyko appeared to broaden the terms of the Russian proposal, offering to ban "in general the use of force both in space and from space against the earth." So far the Reagan Administration has not responded officially to these initiatives, in spite of several resolutions introduced in the Senate and the House of Representatives calling for renewed negotiations and for a moratorium on the flight testing of weapons in space.

In the absence of any apparent interest on the part of the administration in a negotiated ASAT agreement a private group, formed in March under the auspices of the Union of Concerned Scien-

tists, has drafted a "model treaty" intended "to focus attention on the issues that such negotiations are likely to encounter." Unlike the Russian draft treaty, the treaty proposed by the U.S. scientists would prohibit the testing of all ASAT systems, including ground-based ones, against objects in space.

Testifying before the Senate Foreign Relations Committee in May, Kurt Gottfried of Cornell University, chairman of the scientists' group, stated that he and his colleagues "have compared the status and potential capabilities of both ASAT systems and conclude that an immediate moratorium on ASAT space tests would not jeopardize our security if negotiations were to fail, or if a treaty were eventually abrogated or circumvented by the Soviet Union." Noting the "striking parallel" between the present competition in ASAT weapons and the history of MIRV technology, Gottfried concluded that the lesson of the MIRV episode "applies directly to antisatellite weapons. The Soviets have been both foolish and reckless to spend some 15 years nurturing a clumsy threat against a rather small portion of our satellites. Their major accomplishment has been to provoke us into building a far more sophisticated system. Our ASAT, if deployed, will give us a temporary advantage. But as with ballistic missiles, an ongoing competition in space weaponry will, inexorably, reduce the security of both sides."

The Basalts of Venus

A milestone in the exploration of the cloud-shrouded surface of Venus came in March of last year, when two Russian spacecraft landed on the surface, sampled Venusian rock and determined its chemical composition. *Venera 13* landed on an "upland rolling plain," *Venera 14* landed on "lowland." (According to radar mapping by the *Pioneer* Venus probe, these terrains make up 92 percent of the surface.) Writing in *Journal of Geophysical Research*, a group led by Yu. A. Surkov of the V. I. Vernadsky Institute of Geochemistry and Analytical Chemistry in Moscow now gives details of what the spacecraft found.

Each spacecraft carried "a miniature drilling rig" that took a sample of the surface. The sample was conveyed through an "air lock" to a "measuring cell." There it was bathed in radioactivity emitted by plutonium 238 and iron 55. The former isotope elicited the emission of X rays at characteristic frequencies from magnesium, aluminum and silicon atoms; the latter elicited the emission of X rays from potassium, calcium and titanium. The X rays were

analyzed by on-board spectrometers. Some initial calibration spectra were made while the measuring cell was empty. Then, 224 seconds after each landing, Venusian soil arrived in the cell. *Venera 13* succeeded in measuring 38 spectra; *Venera 14* measured 20. The investigators compared the spectra with those of 200 terrestrial and lunar "rocks of well-known elemental composition."

The rock of the rolling upland proved to be "close to the composition of potassium alkaline basalts of the earth's crust," an uncommon type of volcanic rock found in islands and along rift zones in the Mediterranean. Such zones mark the divergence of crustal plates as new ocean basins form. The Venusian rolling upland is thought to be quite different. It is high and smooth and is pocked by myriad craters; thus it is taken to "represent the ancient crust of Venus." The rock of the lowland proved to be similar to "basalts of the oceanic crust of the earth." The Russian investigators surmise "that Venusian lowlands are covered by basaltic lava flows."

The investigators offer some "general ideas that arise from looking at the data." They note that the earth, Venus and the moon have two main types of terrain, "ancient crust (continents on the earth, rolling uplands on Venus and highlands on the moon) and younger formations (the oceanic crust on the earth, lowlands on Venus and maria on the moon)." Nevertheless, Venus differs notably from the earth and the moon. The rocks are not as diverse. "Perhaps a considerable part of the surface of Venus is covered by basaltic rocks that have originated from different depths and at different times of the formation of the crust."

Doubtful Diets

Many physicians maintain that the health of Americans would be markedly improved if only people would modify a few simple behaviors rather than depending on new kinds of medical intervention. The trouble is that people need information on what behavior to modify and how, and (except in the case of cigarette smoking) the signals from public-health experts have been uncertain. The prevention or correction of high blood pressure by dietary change is a case in point. From 35 million to 60 million Americans (depending on who is making the estimate) have hypertension, and perhaps 95 percent of them have "essential" hypertension: elevated arterial blood pressure that cannot be attributed to any identifiable disease or condition. For many years correlations have been reported between blood pressure and dietary factors ranging from an excessive intake of calories (obesity) to a deficiency of magnesium, in some cases suggesting cause-

FLY THE LEADER.



The 757 is the most fuel-efficient jetliner in the sky. In passenger comfort, it's superior to any other plane its size. Here's the perfect match of advanced technology and passenger conveniences. It's one way Boeing is helping to keep air fares one of the world's best travel values.

BOEING
Getting people together.

SPEAK FRENCH Like a diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel*, that's who.

Now you can learn to speak French just as these diplomatic personnel do — with the Foreign Service Institute's Basic French Course.

The U.S. Department of State has spent thousands of dollars developing this course. It's by far *the most effective* way to learn French at your own convenience and at your own pace.

The Basic French Course consists of a series of cassettes and an accompanying textbook. Simply follow the spoken and written instructions, listening and repeating. By the end of the course, you'll be learning and speaking entirely in French!

This course turns your cassette player into a "teaching machine." With its unique "pattern drill" learning method, you set your own pace — testing yourself, correcting errors, reinforcing accurate responses.

The FSI's Introductory Basic French Course comes in two parts, each shipped in a handsome library binder. Part A introduces the simpler forms of the language and a basic vocabulary.

Part B presents more complex structures and additional vocabulary. Order either, or save 10% by ordering both:

- Basic French, Part A.** 12 cassettes (15 hr.), and 194-p. text, \$125.
- Basic French, Part B.** 18 cassettes (25 hr.), and 290-p. text, \$149.

(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: 1-800-243-1234.

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's French course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn French, return it and we'll refund every penny you paid. Order today!

81 courses in 36 other languages also available. Write us for free catalog. Our 10th year

**Audio-Forum
Suite 19K
On-the-Green,
Guilford, CT 06437
(203) 453-9794**



AUDIO-FORUM®

Or visit our New York sales office: 145 E. 49th St., New York, N.Y. 10017 (212) 753-1783

and-effect relations and implying that a particular modification of diet might be prophylactic or therapeutic.

Such findings have now been reviewed and evaluated in 35 papers published as a supplement to *Annals of Internal Medicine*. The signals are still unclear. As Harriet P. Dustan of the University of Alabama School of Medicine puts it in an *Annals* editorial, "we are far from identifying a dietary factor important in the pathogenesis of hypertension." The idea of diet modification nonetheless has attractions. Epidemiological evidence does show differences in the diet of populations with differing prevalences of hypertension, and diet is in theory readily subject to modification by an individual on his own.

A further reason for interest in the dietary approach by some physicians is disenchantment with aggressive drug therapy for mild hypertension. Within the past 10 years or so, as new antihypertension agents have become available, it has become almost axiomatic that any degree of hypertension should be treated. For most physicians this means prescribing drugs for patients with a diastolic pressure of 90 or more. Norman M. Kaplan of the University of Texas Health Science Center at Dallas takes issue with the practice, arguing that the data do not in fact support drug therapy for asymptomatic mild (90 to 104) hypertension and that any drug treatment has some risks. Nondrug therapies, including dietary modifications, are indicated for such patients, he says.

What modifications? The doctors disagree. Consider reduction of salt (sodium chloride) intake, perhaps the commonest prescription. James C. Hunt of the Center for the Health Sciences of the University of Tennessee holds that "excessive sodium consumption by the American public is a justified cause for concern." While conceding that hypertension is more directly related to sodium intake in some people than in others, he cites cross-cultural studies showing that societies whose sodium intake is elevated have a high frequency of hypertension and of stroke, one of its major sequels. On the other hand, within a particular society many studies have failed to show a correlation of elevated sodium intake with hypertension, leading other investigators to doubt the value of sodium restriction for the general population. Dustan implies that the case for cutting back on salt is not proved. The consensus nonetheless appears to be that since reduction of sodium may prevent the development of hypertension in some people and since a high-salt diet is almost certainly not beneficial, reduced salting of food and reduced consumption of salty snack foods is probably a good idea.

The situation is clearer with regard to excessive alcohol consumption and obe-

sity. Both are shown to contribute to hypertension, and to do so independently of each other. Dustan notes in particular that "obesity is so prevalent and so closely associated with hypertension" that more attention should be given to its control and to understanding why it causes hypertension.

Unraveling Knot

It is well known that in industrial countries the institution of marriage is in a state of change. How much has it changed and what is its future? Writing in *Bulletin of the American Academy of Arts and Sciences*, Kingsley Davis of the University of Southern California examines demographic data on marriage in the most prosperous industrial countries since World War II. He concludes that marriage is losing its connection to its two most significant functions: providing domestic intimacy and offering a framework for raising children. Further, he predicts that the relation between marriage and childbearing will become increasingly tenuous without fundamental social reforms.

From 1945 through 1960 marriage gained in the industrial countries. In general there was a decrease in the age at which men and women first married and an overall increase in the fraction of the population that was married. Beginning in the 1960's, and particularly in the 1970's, both trends were reversed. In 1960, 14 percent of all women aged 30 through 34 in the U.S. were unmarried; by 1980 the fraction was 27 percent. Other industrial nations show a similar increase in the fraction of single women. The change is due both to a rise in the divorce rate and to a tendency for young people to postpone marriage. After a steady decline since 1945 the age of first marriage began to rise in about 1970.

Changing social customs have resulted in a large group of young, single people. Unmarried people are not, however, forgoing the pleasures of cohabitation. From 1977 through 1980 the number of married couples in the U.S. changed little but the number of unmarried couples living together rose by 63 percent. In Norway in 1977 among all women who were 18 or 19 years old and were living with a man 43 percent were not married. The fraction of women living in such "consensual unions" decreases with age, but among younger women informal unions are competing vigorously with formal ones.

Some have suggested that a consensual union is simply a marriage without certain more or less trivial legal trappings, but Davis notes that there are significant differences between the two. Informal unions are made by younger people, are less stable and result in fewer children than marriage does. In any event the fraction of the population in

any kind of union, legal or not, is decreasing; thus the spread of consensual unions has not compensated for the decline in the marriage rate.

Not only is marriage getting competition from other kinds of relationships in providing domestic intimacy; it is rapidly losing its monopoly on reproduction and child rearing. Married couples are becoming steadily less fertile. From 1945 through 1949 the net reproduction rate in 19 of the most economically developed countries was 1.27. (The net reproduction rate indicates whether the population is replicating itself; the replacement level is 1.) From 1975 through 1979 the rate in the 19 countries was .87, well below replacement level. In countries such as the U.S., the U.K. and Sweden the rate would have been considerably lower if it had not been for the large number of immigrants from poorer countries, who have relatively high fertility.

Furthermore, the reduction in fertility among married women is much greater than the change in the reproduction rate suggests. The reason is that it has become much commoner for an unmarried woman to have a child. In 1950 a married white woman in the U.S. was 17 times as likely to have a child as an unmarried one; in 1970 the figure had dropped to four times as likely. Among U.S. blacks an unmarried woman is currently about as likely as a married woman to have a child. The fertility of consensual unions is quite low, hence most births to single women are to women who are living alone.

If the children of divorced parents are added to children of single mothers, it can readily be seen that a substantial fraction of children in developed countries are growing up in a setting other than the traditional marriage. In 1970, 15 percent of all U.S. children under 18 did not live with both of their biological parents; in 1981 the fraction was 24 percent. Fewer children are currently being raised in intact marriages. Conversely, fewer marriages include children: in 1981, 49 percent of all U.S. married couples had no children of their own under 18 living with them.

The data accumulated by Davis show that marriage is being separated from its traditional functions. Conjugal relations are more frequent in an informal union and childbearing is more frequent outside marriage, indeed outside any kind of union. In accounting for the changes in marriage and in particular the declining fertility of married couples Davis cites several factors. One of the most significant is the increasing likelihood that a married woman will work outside the home. About 60 percent of married women in the U.S. currently have such jobs, only slightly less than the rate of 65 percent for single women. In the absence of affordable arrangements for child

care the greater the fraction of working women is, the lower the fertility rate is likely to be. According to Davis, if the accustomed functions of marriage are to be preserved in industrial societies, provisions must be made to reduce the conflict between work and motherhood. He concludes: "Although people value children very highly, it seems likely that they will not exceed replacement-level fertility unless incentives for childbearing and child care are systematically improved and disincentives alleviated, which in large part means reforming the institution of marriage. Failing such a development, it seems likely that the population of the industrial nations will increasingly be sustained by immigrants from the Third World."

Nonfarming for Farmers

What does a good farmer do with fertile land he undertakes not to farm? The question comes up because many farmers in the U.S. will be withdrawing land from commercial production this year under the Farm Acreage Reduction Program, the Federal effort to deal with chronic surpluses of certain crops and the substantial amounts of farm products in storage from previous harvests. Participating farmers will receive cash payments for withdrawing 20 percent of their acreage and payments in kind from stored or current harvests for additional withdrawals. Advice on what to do with the withdrawn land is proffered in *Crops and Soils Magazine*, a publication of the American Society of Agronomy, by D. R. Hicks, N. P. Martin and E. A. Oelke of the University of Minnesota.

The program for withdrawal covers mainly corn and wheat. Hicks and his colleagues address their advice to corn farmers, but the recommended practices would be similar for producers of other crops. The advice is based on regulations issued by the Agricultural Stabilization and Conservation Service of the United States Department of Agriculture, which is administering the withdrawal program. The regulations offer the corn farmer three options: (1) to establish annual or perennial crops of grasses or legumes or both; (2) to plant a small crop of grain that would be made unharvestable before it ripens, and (3) to leave on the ground the residue of the 1982 corn crop and control the weeds that grow this year. Two other stipulations are that the diverted acres cannot be mechanically harvested this year and cannot be grazed between April 1 and September 1.

If the farmer needs forage, Hicks and his colleagues say, he would do well to establish a long-term stand of alfalfa, perhaps with a companion crop. The alfalfa could not be harvested this year for hay, but it could be grazed after Septem-



Passengers rate the Boeing 767 as a superb flying experience. It has two wide aisles and is the only wide-body with seven-abreast seating, so nearly everyone has either a window or aisle seat. Pilots rate the 21st Century flight control system as the most advanced in the world.

BOEING
Getting people together.

A BERLITZ® LANGUAGE COURSE

in a handsome briefcase
for only \$125
if you pay for it
BUT



...wouldn't your company
be interested in having you
speak a second language?

These days a second language can help you help your company and so help you earn more. The many hispanic nationals now in our work force respond more readily to someone who speaks their language. Also, selling today is international. Whatever language it would pay you to speak, Berlitz has a do-it-yourself course in a handsome briefcase you can take anywhere. Contents—90 minute introductory tape, 40 lessons on five 60-minute cassettes, 6 illustrated word and phrase books, rotary verb finder to help you find verb tenses instantly. If not satisfied, return within 10 days and get your money back. Learn from the leader!



For your convenience on credit card orders dial toll-free

1-800-228-2028 ext. 35

and refer to Dept. 3163

**24 hours a day,
7 days a week.**

Why not give us a ring—right now! (our expense)

(Nebraska residents dial 402-571-4900)

Berlitz Publications, Inc., Dept. 3163
3490 Lawson Blvd., Oceanside, NY 11572

Send Berlitz Comprehensive Cassette Course(s) checked. \$125 each in briefcase plus \$4 for shipping and insured delivery.

French 86100 German 86101 Italian 86102 Spanish 86103

Enclosed check _____ money order _____ payable to Berlitz.

Or charge my AMEX _____ Diners Club _____ VISA _____

MasterCard Interbank # _____ Exp. Date _____

Card # _____ Exp. Date _____

Name _____

Address _____

City _____ State _____ Zip _____

N.Y. residents add sales tax. Allow 4 weeks for delivery

ber 1; the companion crop would have to be made unharvestable. "We believe intensive grazing with 10 to 20 mature cows per acre between September 1 and September 15 would provide for higher-quality forage the following year," the Minnesota group writes.

If the farmer does not need forage, they say, he still should manage the diverted acres to control erosion by wind and water and to keep down weeds. "The simplest method would be to leave the 1982 corn stubble (if it is still there) and control the weeds with herbicides or by mowing." Since research at the University of Minnesota has shown that corn yields are reduced when corn is grown following a year when no cover crop was planted, the diverted acres "would be best managed by planting a legume to provide 1983 cover and improve the nitrogen status of the soil for the 1984 corn crop."

Another possibility is a grass crop. It would not add nitrogen to the soil as legumes do, but it would improve the physical condition of the soil when it was plowed under. Neither legumes nor any other crop serving as a "green manure crop" that is plowed under will add nutrients (except nitrogen). "The phosphorus, potassium and other elements in the plant growth were taken from the soil in the first place but will be available again as the green manure decomposes." Winter rye, winter wheat and winter barley are also good cover crops if they are seeded in the spring. Then they would not have to be mowed to prevent the formation of heads, that is, to be made unharvestable.

Talkwriters

The development of a machine that can automatically transcribe human speech has been painfully slow and piecemeal; the fundamental difficulty is that pronunciation varies enormously from person to person and varies widely even for a single speaker, depending on the linguistic context of a given sound, the speaker's mood, the quality of the microphone used for transmission and so on [see "Speech Recognition by Computer," by Stephen E. Levinson and Mark Y. Liberman; *SCIENTIFIC AMERICAN*, April, 1981]. Investigators at several major laboratories are studying the problem from a generalized perspective: To what extent is it possible to build a machine that understands language the way people do? A much less ambitious goal, however, may have wide practical application: the development of a voice-actuated typewriter. Now Raymond Kurzweil, a computer scientist formerly at the Massachusetts Institute of Technology, has formed a company that will attempt to build a voice typewriter, capable of transcribing speech drawn from a vocabulary of 10,000 words, by

1985. The company, Kurzweil Speech Systems, Inc., is independent, but it will benefit from a large investment in development by the Xerox Corporation.

Any machine for recognizing speech that is likely to be available in the next 10 years will be a compromise among cost, technical capability and the buyer's acceptance; the transcription of fast, connected speech by many speakers over a large vocabulary remains technically infeasible. Kurzweil's strategy calls for building a \$5,000 machine that can respond to discrete dictation, in which utterances must be parsed by the speaker with a very brief pause between each word. Nevertheless, the machine would take dictation as fast as 150 words per minute. It would be trained to the individual characteristics of a speaker's voice for about an hour before its ordinary operation could begin. The refinement of the templates used for pattern recognition would continue, however, even after the training period, and the words that made up the ultimate 10,000-word vocabulary would be the ones most often used by the speaker. The recognition system is designed to be compatible with many existing personal computer work stations, and so the editing of a text could also be done by voice command.

Several other companies are now manufacturing voice-recognition systems for special purposes, but none so far has publicly entered the race to make a voice-actuated typewriter. Threshold Technology, Inc., which declared bankruptcy last year, is still making voice recognizers intended chiefly for noisy, industrial environments. The machines can recognize 340 words, spoken with a short pause between each word, at up to 180 words per minute. The company expects soon to announce a device that can recognize up to 1,500 words spoken in isolation, and the machine will recognize connected utterances of any length drawn from a preselected vocabulary of about 50 words.

One of the most sophisticated devices commercially available for recognizing English is made in Japan by the NEC Corporation. The machine has to be trained by the individual speaker, but it can recognize connected utterances up to four seconds long drawn from a 150-word vocabulary. Other machines are being developed in Japan that are specially adapted to the Japanese language, and their capabilities cannot be readily compared with those of machines designed for recognizing English. There are only 120 syllables in Japanese, compared with some 10,000 in English, and the Japanese speaker is much more regular in pronunciation, with fewer glides from syllable to syllable, than the English speaker. Moreover, the incentive for the development of a voice-actuated typewriter is probably stronger in Japan

than it is in the U.S. and Europe because the potential benefit is much greater: the Japanese typist must select from more than 3,200 characters.

Workers at Bell Laboratories have almost completed the development of several speech-recognition products that will be announced at the end of the year. Bell is known to be designing systems that can respond to many speakers as well as systems that must be individually trained. Speaker-independent systems would enable any person to control a computerized data bank, place an order or respond to a programmed series of questions over the telephone without the intervention of another person. The tradeoff is that the recognized vocabulary must be smaller than it is on a machine designed to recognize one speaker at a time.

At the International Business Machines Corporation the emphasis has been on developing an office dictation system that can handle connected speech drawn from a vocabulary of at least 5,000 words. According to Frederick Jelinek, who heads the speech-recognition project at IBM's Thomas J. Watson Research Center, the aim is to develop a machine that can be trained to the voice pattern of the individual in less than 15 minutes. The approach to pattern matching is a statistical one: the spectrum of the speech event is sampled every 10 milliseconds and each sample is assigned to the most similar of 200 spectral patterns digitally stored in the memory of a computer. Words are recognized by finding the word that most closely matches a particular string of spectral patterns and also has a high probability of following the two preceding words; the probabilities related to three-word sequences are derived from extensive analysis of business correspondence. According to Jelinek, IBM is not yet developing any speech-recognition machine for commercial use, but the company is clearly in the race to perfect a voice-actuated typewriter.

There is good reason, in spite of the commitment of other companies, to take Kurzweil's entry seriously. In 1975 he built the first working model of a machine that can now recognize printed words in more than 200 type fonts and then read them aloud. The most important use for the device so far has been as a reading machine for the blind. A blind person can place a book, newspaper or any other printed text on top of a glass window and then listen as the machine automatically reads the text in a synthesized voice. The voice can be speeded up to 225 words per minute, and the user can ask the machine to repeat a phrase, spell out a word or skip sections of text. Kurzweil Computer Products, Inc., a subsidiary of Xerox, now markets the reading machine primarily to libraries and other institutions for about \$30,000.

Although optical character recognition and the recognition of vocal sounds make quite different technical demands, the underlying problems are the same: an enormous number of patterns must be compared with a limited number of internal models in a short time.

Chemical Pruning

The fruit produced by an orchard tree represents an investment of energy by the tree, and the less energy the tree has to put into new vegetative growth, including leaves, the more energy it can put into fruit and the better fruit it produces. Moreover, when an orchard tree has too many leaves, sunlight cannot reach some of the places where flowers (the first stage of fruiting) might appear and cannot reach the fruit effectively enough to bring the fruit to its full ripe color. Orchardists currently deal with the excess-foliage problem by pruning, which is both labor-intensive and costly. Now there is a chemical to do the job. It should be available commercially within a few years as a result of work by plant physiologists at the Fruit Research Laboratory of the Agricultural Research Service of the United States Department of Agriculture in Wenatchee, Wash.

The chemical has the proposed common name of Paclobutrazol but at present is designated ICI pp333. It was developed to control the growth of grass so that farmers whose product is grass seed would get a higher yield. Applied to the ground around a fruit tree, the chemical holds down the growth of new leaves on the tree. The result is a tree that looks as though it had been skillfully pruned by hand. Because of the relative lack of foliage, the tree has a large crop of fruit of uniform size and good color. The work is characterized by Max W. Williams, director of the project: "We are using chemicals to accomplish what genetics has so far failed to do, and for the first time ever we can have control of an apple or pear tree's vegetative and fruit growth throughout the life of the tree."

Williams and his colleagues began testing the compound on fruit trees as a foliage spray in 1978. In 1979 they started applying it to the ground at the rate of 2.5, 5, 10 and 20 grams of active chemical ingredients per tree. Williams foresees that in commercial practice the rate of application might be from one gram to five grams per tree. The ground treatment is superior to the foliage spray and avoids chemical contamination of the fruit. It has worked successfully with several kinds of apple and with Anjou pears, and Williams believes it will succeed with peaches and cherries. The chemical treatment may also help in the control of insects and diseases that prey on the trees, since leaves can sustain a variety of insects and fungi.

FLY THE LEADER.



World travelers choose the 737 because it has one of the best on-time records. The 737-300 is the latest addition to the family. This new jetliner, with added passenger amenities, advanced flight controls and quieter engines, will make flying an even better travel experience.

BOEING
Getting people together.

In 1982, Porsche won the World Championship of Makes. Audi won the World Rally Championship for Manufacturers and SCCA Pro Rally Championship.

What we learn from our competition cars, we put into our production cars.

And now you can arrange to drive your new Porsche or new Audi where it was built, tested, and refined—with our Delivery In Europe Program.

Simply visit your local Porsche Audi dealer and order the model of your choice: the Porsche 928S, 911SC, or 944. Or the new Audi 5000S, the 4000S, Coupe GT, or all-wheel drive Quattro. Then take delivery in Europe. And after your stay, we'll deliver your car to a U.S. port free of charge.

For more information, see your Porsche Audi dealer: the home of champions.

Visit Europe And Drive Home A World Champion.



PORSCHE + AUDI

The world's densest computer is now the heart of the world's most: 32-bit computer.

Our 32-bit CPU.

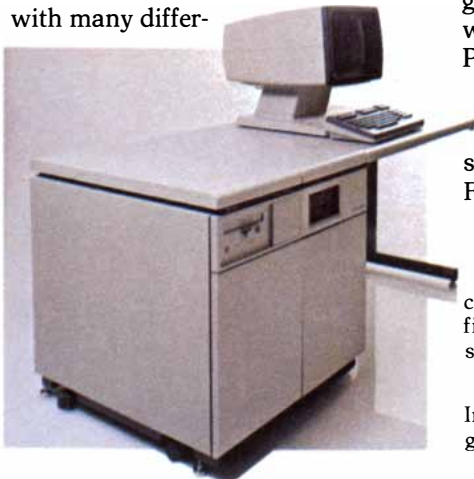
From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

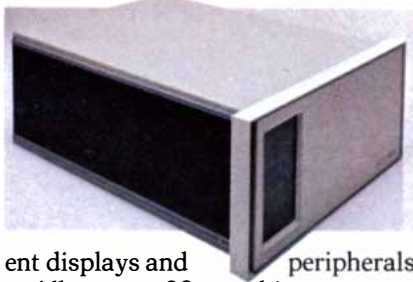
The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits—these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many differ-



As a minicabinet, it can handle multiple users.

A rack-mountable version is available, too.



ent displays and peripherals. All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and

The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



chip fordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet™, and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals.

Links to central computers

are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

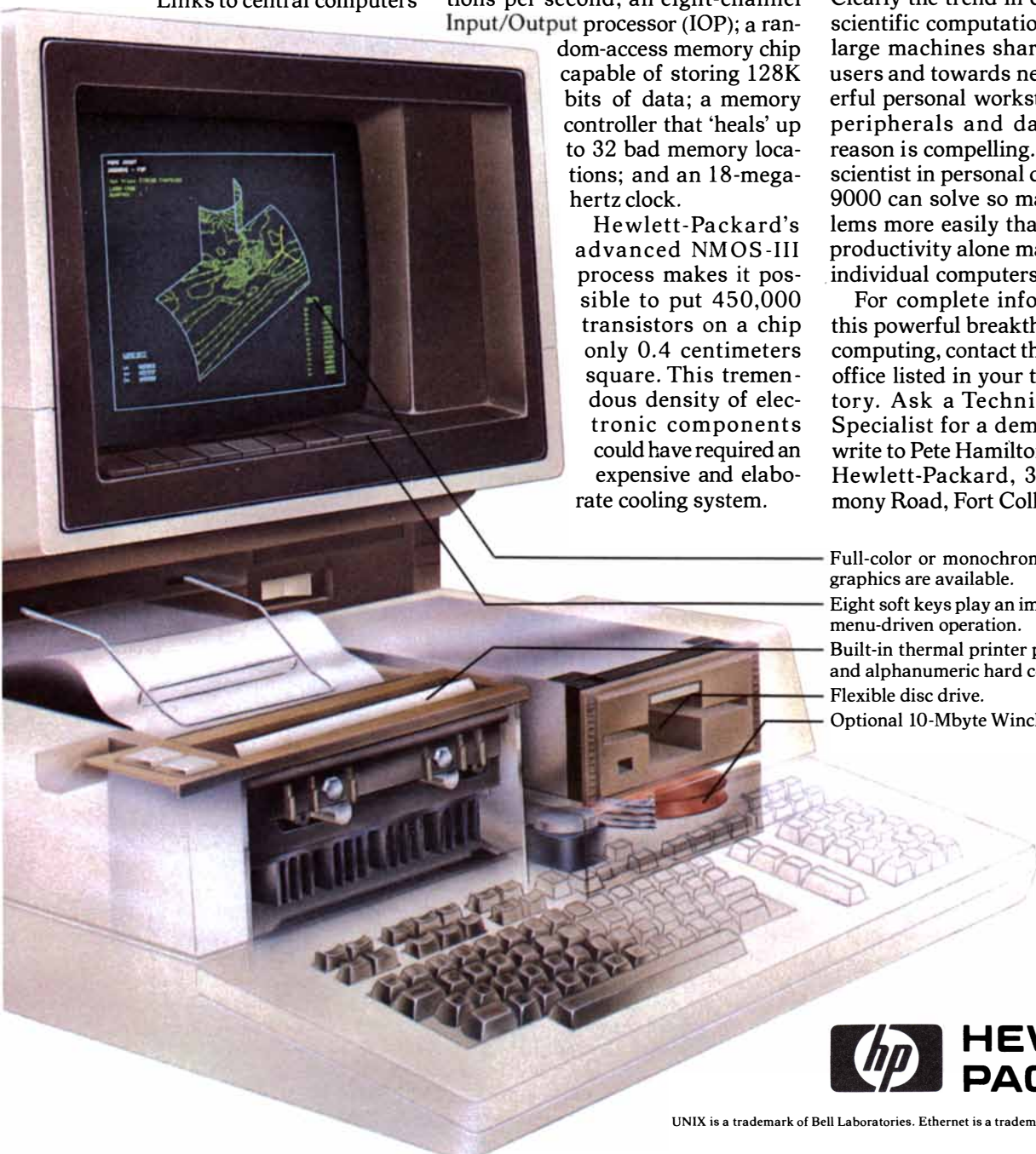
Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 41151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.



- Full-color or monochromatic display. 3-D graphics are available.
- Eight soft keys play an important role in the menu-driven operation.
- Built-in thermal printer produces graphics and alphanumeric hard copy.
- Flexible disc drive.
- Optional 10-Mbyte Winchester disc.



UNIX is a trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

Interstellar Matter in Meteorites

Carbonaceous chondrites, the most primitive meteorites, incorporate material originating outside the solar system, including matter expelled by supernovas and other stars

by Roy S. Lewis and Edward Anders

The solar system is much younger than the universe (only 4.5 billion years compared with 10 to 15 billion years), and so it must have formed from older matter that had a previous history. Before 1969, however, no such relict matter had ever been found, either in meteorites or in planets. And since theorists had proposed a hot origin for the solar system, which would have caused all preexisting solids to vaporize, it seemed that the solar system started out with a clean slate.

How could one recognize presolar matter? Perhaps by its age, but more reliably by its isotopic composition. Stars continuously rework the chemical elements and then eject them back into interstellar space, where they eventually form the next generation of stars. Both the elemental and the isotopic composition of the ejected matter vary from star to star, depending on the star's mass, temperature and stage of evolution. The isotopic composition is a particularly durable hallmark, since it can be changed by few processes (short of nuclear reactions) and then only to a limited, predictable extent. Therefore one looks for matter of unusual isotopic composition that cannot be explained by known or plausible processes within the solar system.

The first hint of presolar matter came, then, in 1969, when David C. Black and Robert O. Pepin of the University of Minnesota studied the isotopic composition of the noble (chemically inert) gas neon in carbonaceous chondrites: dark gray, rather nondescript-looking stones that are the most primitive meteorites known. Hidden by large amounts of normal neon was a small component of neon greatly enriched in the isotope neon 22. Given the rarity and oddity of noble gases, a local origin could not be ruled out. Then in 1973 Robert N. Clayton, Lawrence Grössman and Toshiko Mayeda of the University of Chicago found that certain minerals in the Allende carbonaceous chondrite were enriched in the isotope oxygen 16 by as

much as 5 percent with respect to normal oxygen. After ruling out a local origin they concluded that some oxygen made by nuclear reactions in stars other than the sun had found its way into the meteorite.

Other discoveries followed, and today some 19 chemical elements in carbonaceous chondrites are known to show isotopic anomalies. Evidently the slate was not quite clean; some presolar matter did survive in the cooler precincts of the early solar system. The challenge is to find these bits of presolar matter and to decipher their record of stellar nucleosynthesis and interstellar chemistry.

Carbon has yielded a particularly rich crop of presolar components; at least four have been recognized so far. All are well hidden and were found only because they are "tagged" with anomalous noble gases or hydrogen. Let us follow the four trails of discovery, after first describing the meteorites in which the anomalies are found.

Carbonaceous Chondrites

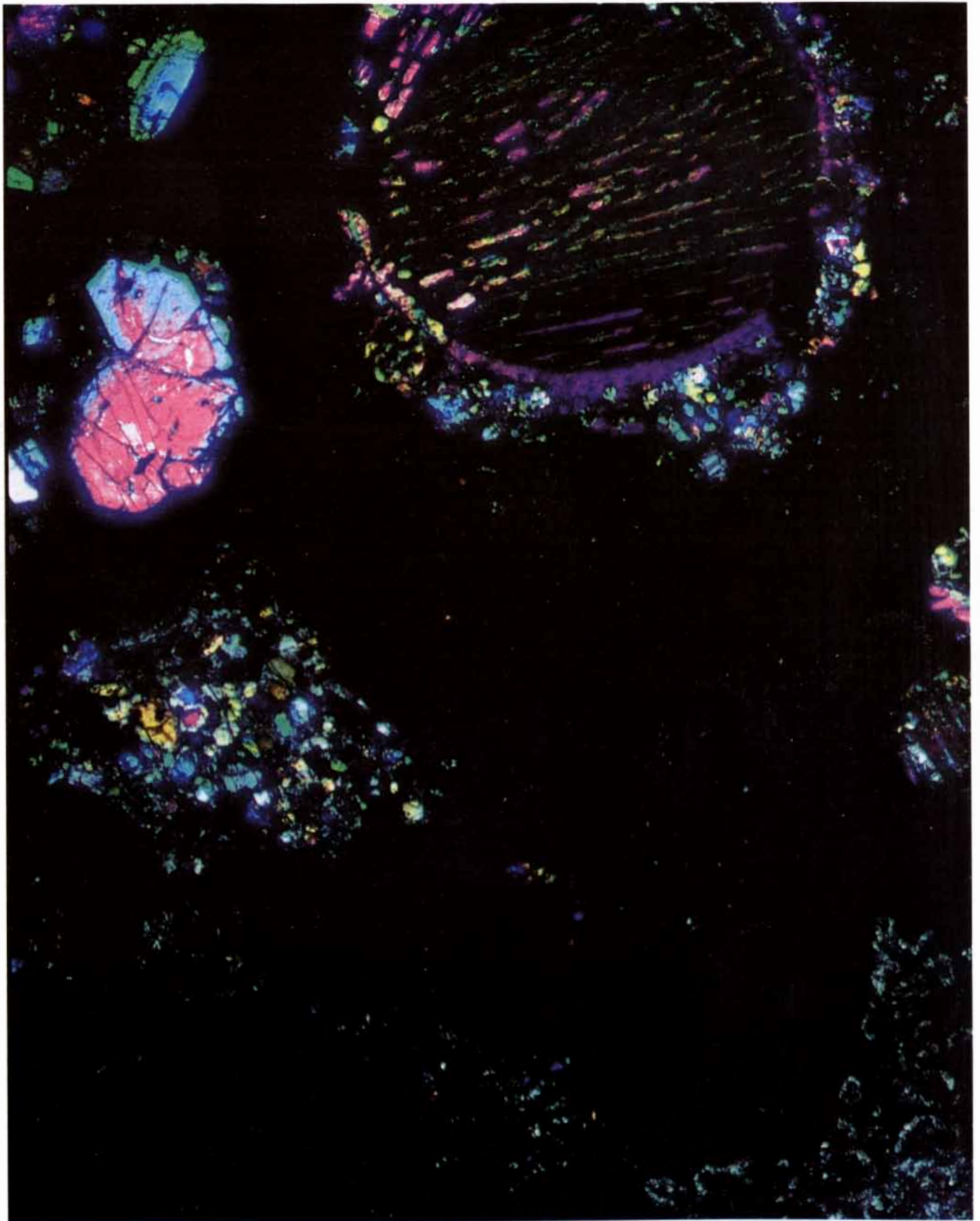
About 40 carbonaceous chondrites are known; they represent some 4 percent of all the meteorites observed to fall. They are divided into three main classes, designated C1, C2 and C3, on the basis of mineralogy and declining content of volatiles such as water, carbon and nitrogen. Their formation temperature (that is, the temperature at which they ceased to react with the gas in the solar nebula, the cloud from which the solar system condensed) increases from C1 to C3, so that if the temperature increased toward the center of the nebula, the C1's come from farthest out. (They formed at a temperature of 360 degrees Kelvin. C3's formed at a temperature some 60 degrees higher.) All carbonaceous chondrites were altered in their parent bodies, probably asteroids. C1's and C2's were exposed there to liquid water, and C3's were somehow reheated to a temperature of perhaps 600 degrees K.

Texturally the carbonaceous chondrites consist of a fine-grained matrix in which coarse-grained particles are embedded. The matrix particles range in size from 10^{-4} to 10^{-6} centimeter and are rich in volatiles; they consist of silicates and carbonaceous matter. The larger particles range in size from .01 centimeter to one centimeter. They are poor in volatiles. Some of them are chondrules: rounded particles looking much like bird shot. Others are more irregular but consist of the same minerals: olivine $[(Mg,Fe)_2SiO_4]$, pyroxene $[(Mg,Fe)SiO_3]$, troilite (FeS) and nickel-iron. Still others are refractory (heat-resistant) inclusions of minerals rich in calcium, aluminum and titanium but poor in silicon. In retrospect it is not surprising that isotopic anomalies are found in the matrix, which has been heated very little, and in the refractory inclusions, which are highly resistant to heat. The anomalies we shall now discuss are found chiefly in the matrix.

Neon E

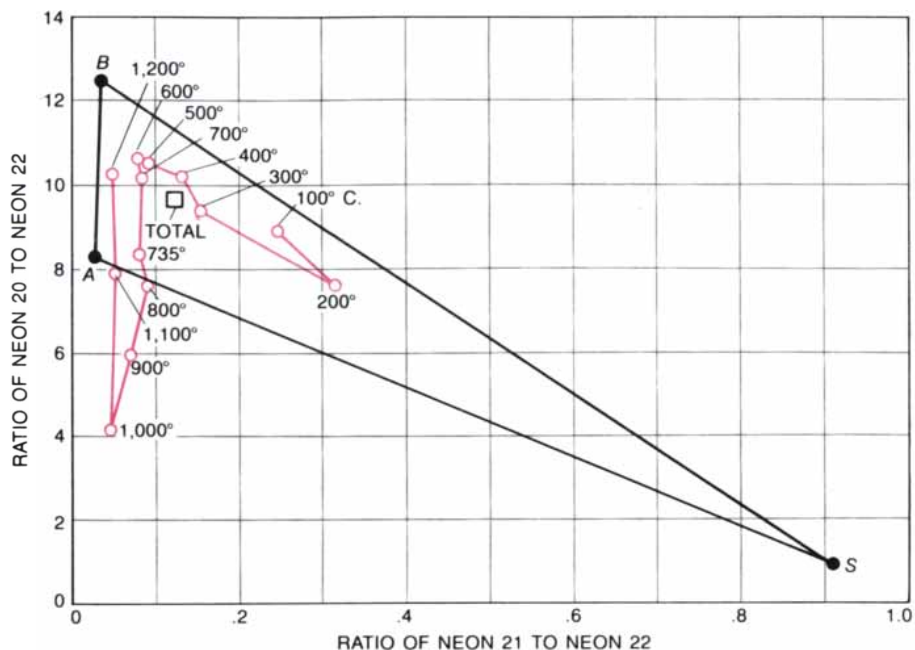
We shall begin with the noble gases. Among the elements in meteorites they are unique. Being highly volatile and unreactive, they did not condense in even the most primitive meteorites and hence are present at only a minute fraction of their proportion in the sun, ranging from about 10^{-5} for xenon to 10^{-9} for helium and neon. This tiny amount of gas, however, is tightly bound in the meteorite, coming free only at high temperatures when its host mineral begins to melt or decompose.

Black and Pepin tried to untangle three types of neon known to be present in primitive meteorites: primordial or planetary neon, also called neon A, which was trapped from the solar nebula; solar neon, also called neon B, which consists of solar-wind neon ions implanted in meteorites that happen to have been at the surface of their parent body; and cosmogenic neon, also called neon S, formed when cosmic rays pass-

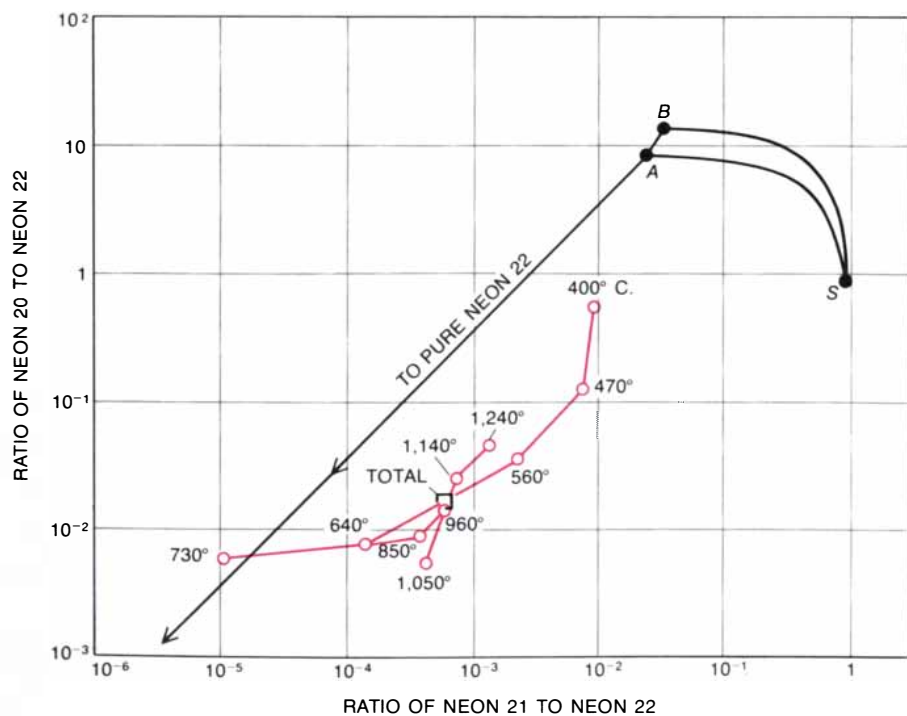


THIN SECTION of the Allende meteorite photographed under polarized light shows the characteristic texture of carbonaceous chondrites. The field of view is about a millimeter across. At the upper right is a chondrule, a solidified silicate droplet. Elsewhere in the field are irregular bits of silicates, sulfides and nickel-iron. At the lower right is a refractory (high melting point) inclusion of the min-

eral olivine. The rest of the field, black in the photograph, is a fine-grained matrix consisting of silicates and carbonaceous matter. Proportions of isotopes unlike those of the solar system's matter have now been found for some 19 chemical elements in carbonaceous chondrites, mostly in inclusions and in the matrix. The photograph was made by Glenn J. MacPherson of the University of Chicago.



THREE-ISOTOPE CHART reveals anomalous neon in the Orgueil meteorite. A sample of the meteorite has been heated to progressively higher temperatures, and the fraction of gas released at each temperature has been analyzed to determine its proportions of the isotopes neon 20, 21 and 22. The first two fractions are well within the triangular area representing mixtures of the three types of neon expected in meteorites: primordial neon (*A*), which comes from the solar nebula, the gas cloud that formed the solar system; solar neon (*B*), implanted in the meteorite by the wind of ions flowing outward from the sun, and cosmogenic neon (*S*), manufactured by collisions between cosmic rays and atomic nuclei inside the meteorite. Then, however, a sequence of fractions veers out of the triangle. Evidently something in the Orgueil meteorite releases a type of neon greatly enriched in neon 22. It was later called neon *E*. The experiment was done in 1969 by David C. Black and Robert O. Pepin of the University of Minnesota.



IN A LATER EXPERIMENT M. H. A. Jungck and Peter Eberhardt of the University of Bern applied stepped heating and three-isotope charting to a tiny mineral sample greatly enriched in neon *E* that they prepared from the Orgueil meteorite. The neon *E* proved to be essentially pure neon 22 (more than 99 percent for the fractions released from 640 through 1,050 degrees Celsius). Apparently neon *E* formed from the radioactive decay of sodium 22 in grains around a star, perhaps a nova. When the solar system condensed, the grains were incorporated into the meteorite. The chart is logarithmic; the triangle defined by *A*, *B* and *S* is at upper right.

ing through the meteorite spall, or shatter, atomic nuclei in their path. Each type has different proportions of the three isotopes of neon.

Black and Pepin employed the technique of stepped heating. Here a sample of the meteorite is heated to progressively higher temperatures, and the gas released is analyzed in a mass spectrometer. With luck the different gas components emerge one by one as their host minerals melt, decompose or become permeable. More often the gases emerge as a mixture that must be further resolved on the basis of isotopic composition. This can be done most easily by plotting two isotopic ratios against each other, both with the same denominator. In the case of neon one plots the ratio of neon 20 to neon 22 against the ratio of neon 21 to neon 22 for each temperature step. On such a "three-isotope plot" all mixtures of two types of neon lie on a straight line joining the two, mixtures of three types lie within a triangle whose corners are the three, and so on.

All meteoritic neon samples measured up to the time of Black and Pepin's work had fallen within the triangle bounded by neon *A*, neon *B* and neon *S*, as expected for mixtures of those three. When Black and Pepin analyzed six C1 and C2 chondrites by stepped heating, they found that the fractions released between 800 and 1,100 degrees Celsius consistently fell below the triangle. Evidently a new neon component was present, with a ratio of neon 20 to neon 22 that was less than 3.4. They named it neon *E*, the letters *C* and *D* having been preempted by two minor components of less extreme composition. Black and Pepin pointed out that the neon *E* could not have derived from solar neon by processes of mass fractionation such as diffusion through a solid or gravitational escape from a planet. The initial reservoir of neon would have been absurdly great: some 10^{10} times greater than the estimated content of neon in the solar nebula.

An alternative possibility is the intense irradiation of dust grains by protons (hydrogen nuclei) in the early solar system. Spallation reactions between protons and magnesium in the grains could make not only the isotopes of neon in the proportions of neon *S* but also sodium 22. If the grains were heated during the irradiation or right after it, the neon *S* would escape. The sodium 22 would then decay, with a half-life of 2.6 years, to neon 22. Donald D. Clayton and his co-workers at Rice University have noted, however, that proton irradiation should also have produced excesses of argon 36 and krypton 80. Such excesses are not observed.

That leaves stellar nucleosynthesis. Stars that have exhausted their hydrogen build neon and neighboring elements by thermonuclear reactions in-

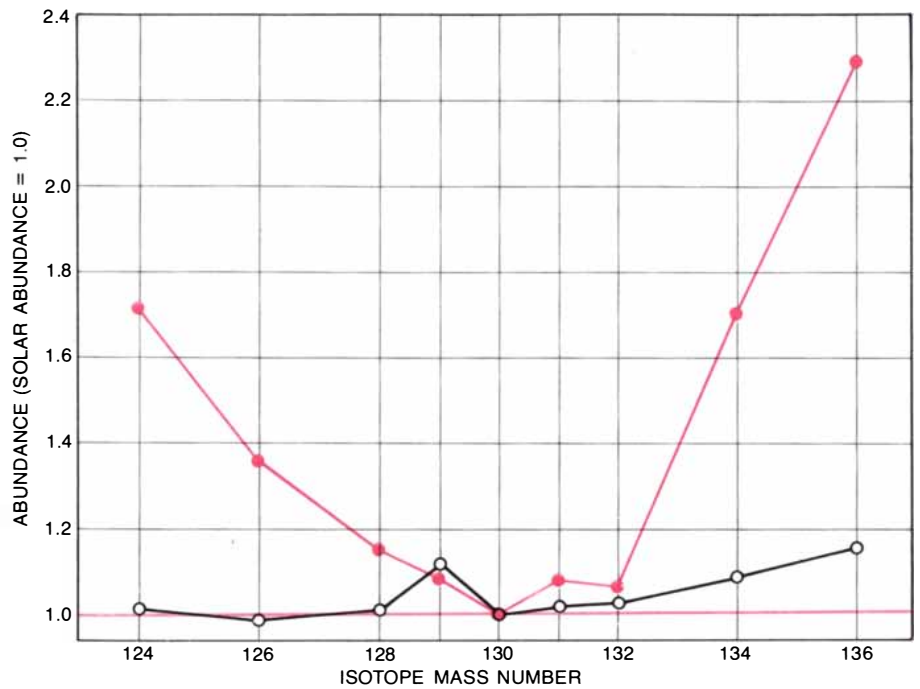
volving helium. For example, helium 4 combines with oxygen 16 to make neon 20. M. Arnould of the Free University of Brussels and H. Nørgaard of the Nordic Institute for Theoretical Atomic Physics (NORDITA) in Copenhagen have shown that neon *E* could form in stars, either directly (under special conditions that greatly favor neon 22 over the other isotopes of neon) or indirectly (as sodium 22 that preferentially condenses on dust grains after ejection from the star and then decays into neon 22).

The Carriers of Neon *E*

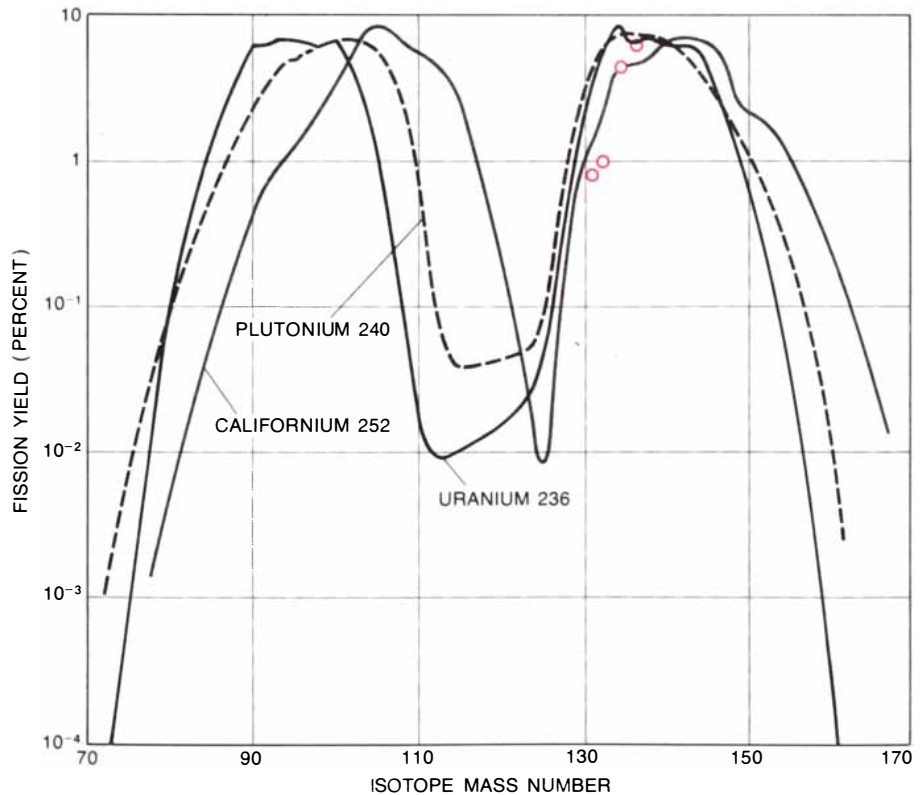
Here was a chance to study "star-dust," and perhaps even to learn what type of star it came from. The obvious next step was to isolate the host mineral (the "carrier") of the neon *E* and learn whether very pure neon *E* contained only neon 22 or the other two isotopes of neon as well. This problem was solved mainly by Peter Eberhardt of the University of Bern and his students F. Niederer, M. H. A. Jungck and F. O. Meier. Starting at Chicago in 1974, Eberhardt systematically took the Orgueil meteorite apart by various physical methods of separation. The mineral magnetite, for example, could be removed with a magnet. Fine-grained silicates could be removed by letting them form a colloidal suspension in concentrated salt solutions. At each stage Eberhardt checked to see where the neon *E* had gone. In this way he found that it concentrated in a coarse-grained, non-magnetic fraction. Further separations showed that the neon *E* had at least two carriers, one of low density and release temperature (less than 2.2 grams per cubic centimeter and 600 degrees C.), the other of higher density and release temperature (3.5 grams per cubic centimeter and 1,100 degrees). The corresponding types of neon *E* were named *L* for low and *H* for high.

Further experiments at Bern and Chicago (in which Leo B. Alaerts took part) established the nature of the carriers. The carrier of neon *E(L)* is carbonaceous. The carrier of neon *E(H)* consists of two minerals: spinel ($MgAl_2O_4$) and apatite [$Ca_2PO_4(OH,F)$]. Meanwhile the compositional limits for neon *E* were being tightened from year to year, strengthening the possibility that neon *E* is pure neon 22 and hence derived from sodium 22. The issue was clinched by Jungck and Eberhardt, whose purest sample was nearly 99 percent neon 22.

Now that we know the carriers of neon *E*, what can be said about its origin? The short, 2.6-year half-life of sodium 22 means it formed quickly and was trapped quickly, before it decayed to neon. Donald Clayton has pointed out that these conditions are met by exploding stars: novae and supernovae. Both synthesize elements explosively and



ANOMALOUS XENON was discovered in the Renazzo meteorite by John H. Reynolds and Grenville Turner of the University of California at Berkeley; in a later effort its isotopic composition in the Allende meteorite was analyzed by B. Srinivasan and the authors at Chicago. First a sample of carbon grains from Allende was analyzed; its xenon (black) differed only slightly in composition from solar xenon. Then the investigators etched the surface of the grains with nitric acid. The etching removed some 95 percent of the xenon. The remainder (color) was grossly enriched in both heavy and light isotopes. Evidently the xenon in the sample was a mixture of primordial xenon on the surface of the grains and two types of anomalous xenon inside. One type (xenon *H*) is enriched in heavy isotopes, the other (xenon *L*) in light isotopes.



FISSION OF HEAVY ELEMENT inside meteorites was the first proposal for the origin of xenon *H*. It was known that the nuclei of uranium and other heavy radioactive metals split asymmetrically, yielding the two-hump distributions of fission products shown in the illustration. The heavy isotopes of xenon discovered by Reynolds and Turner (colored dots) fit this general trend but do not match any one known curve in detail. Reynolds and Turner therefore suggested that xenon *H* might come from the fission of some unknown superheavy element now extinct in the solar system. Xenon *L*, however, would have formed some other way.

eject them almost instantaneously. Interestingly, the first minerals predicted to condense from the gases ejected by these two types of stars are respectively carbon and spinel, the carriers of neon *E(L)* and neon *E(H)*. Only one part in 10^9 of the mass of each grain would have to be sodium 22 in order to account for the neon *E* today. The carrier apatite, however, remains unexplained. It is not an expected condensate in a stellar

explosion, or in the solar nebula for that matter.

Xenon HL

When William Ramsay, having discovered xenon in 1898, named it after the Greek word for stranger, he could not have foreseen how well the name would fit the xenon in meteorites. At least three strange types of xenon

are present in carbonaceous chondrites. Two are abundant but controversial, the third is rare but straightforward.

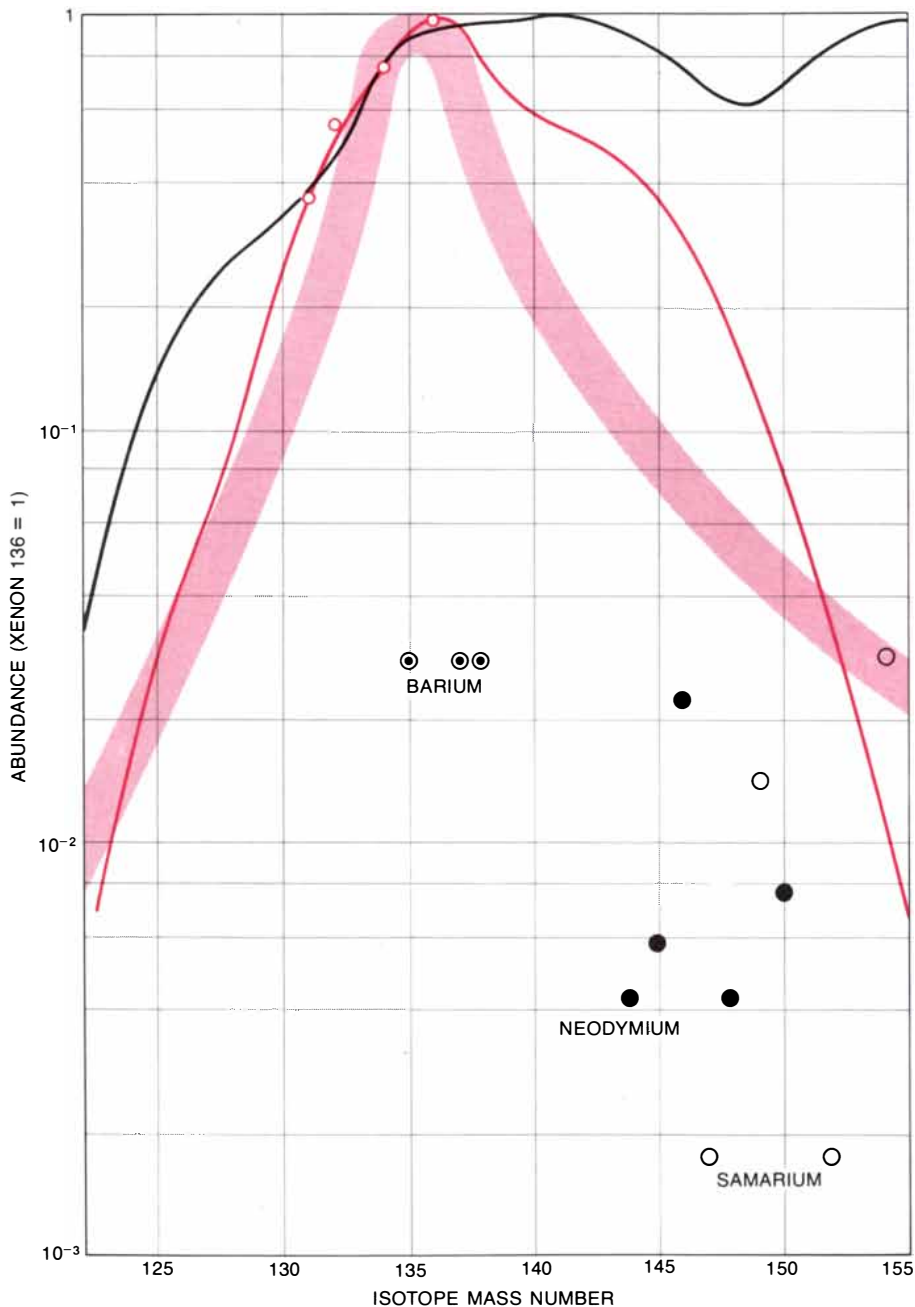
The story begins in 1964, when John H. Reynolds and Grenville Turner of the University of California at Berkeley examined the C2 chondrite Renazzo. They were looking for xenon 129 from the radioactive decay of iodine 129, an isotope discovered by Reynolds whose half-life of 16 million years means that it is extinct in the solar system. (In the 4.5-billion-year history of the solar system it has decayed through about 280 half-lives and has therefore been depleted by a factor of 10^{85} .) To get rid of the large amounts of normal planetary xenon they employed the technique of stepped heating.

They indeed found the xenon 129. In addition, however, they noticed that in the fractions released between 600 and 1,100 degrees C. the heavy isotopes of xenon, ranging in mass from 131 through 136, were enriched by as much as 6 percent with respect to primordial xenon.

The enrichment increased from isotope 131 to isotope 136, as it does in the xenon formed by the fission of uranium and other heavy elements. Reynolds and Turner therefore suggested that the new xenon component came from the fission of some extinct heavy element that had once been present in the meteorite. They also noted a problem: the rare light isotopes xenon 124 and 126 were also enriched in the meteorite. These isotopes do not form by fission, and so their enrichment "would have to be explained by some additional process." Apparently not one but two new xenon components are present in the meteorite. They were later named *H* and *L* for heavy and light. Although of different origins, they have proved inseparable in meteorites, and we shall therefore refer to their mixture as xenon *HL*.

A Superheavy Element?

To be sure, none of the known fissionable elements heavier than uranium yields xenon with the isotopic pattern of xenon *H*. Moreover, the extension of the periodic table to elements beyond 100 showed that their half-lives get shorter with increasing atomic number (that is, the number of protons), and this suggested that the synthesis of new chemical elements might not be feasible beyond element 106 or 107. In 1966, however, W. D. Myers and W. J. Swiatecki at Berkeley tried to calculate the properties of still heavier elements. They came up with a surprising result: the predicted half-lives grew longer rather than shorter. According to their calculations, the next "magic numbers" (which correspond to particularly stable configurations of neutrons or protons in a nucleus) would come at 114 protons and 184



COMPETING HYPOTHESES for the origin of xenon *H* are diagrammed. The hypothesis that it formed by the fission of a superheavy element inside the meteorite is represented by a fission-yield curve calculated for element 114 (black). A competing hypothesis is that the xenon was made in a supernova, either directly, when lighter nuclei captured neutrons (light colored band), or indirectly, when short-lived superheavy nuclei in the supernova decayed by fission (solid color). Both hypotheses account more or less well for the isotopes in xenon *H* (colored dots). In both hypotheses, however, the elements barium, neodymium and samarium should be made in similar quantity. Actually they are not; their upper limits (black dots) in an Allende residue rich in xenon *H* fall far below the theoretical curves. The supernova hypothesis escapes this predicament by assuming that the barium, neodymium and samarium expelled from a supernova condense on grains at a time when xenon, which is more volatile, cannot yet be trapped.

neutrons, giving rise to an "island of stability" in the periodic table centered on element 114. Efforts to synthesize such "superheavy" elements in accelerators or to identify them in cosmic rays are continuing. Their predicted half-lives, although highly uncertain, range up to 10^8 years. Thus in 1969 three groups of workers in the U.S. and Poland proposed that an extinct superheavy element in the island of stability might be the parent of xenon *H*.

Soon the idea was challenged. O. K. Manuel and his colleagues at the University of Missouri examined xenon from three carbonaceous chondrites. In every one they found the puzzling correlation of light and heavy isotopes that Reynolds and Turner had noted. That was a major embarrassment for the fission hypothesis. If xenon *H* was made by the fission of a heavy element inside the meteorite, why was it always accompanied by the same proportion of xenon *L*, which must have a different origin?

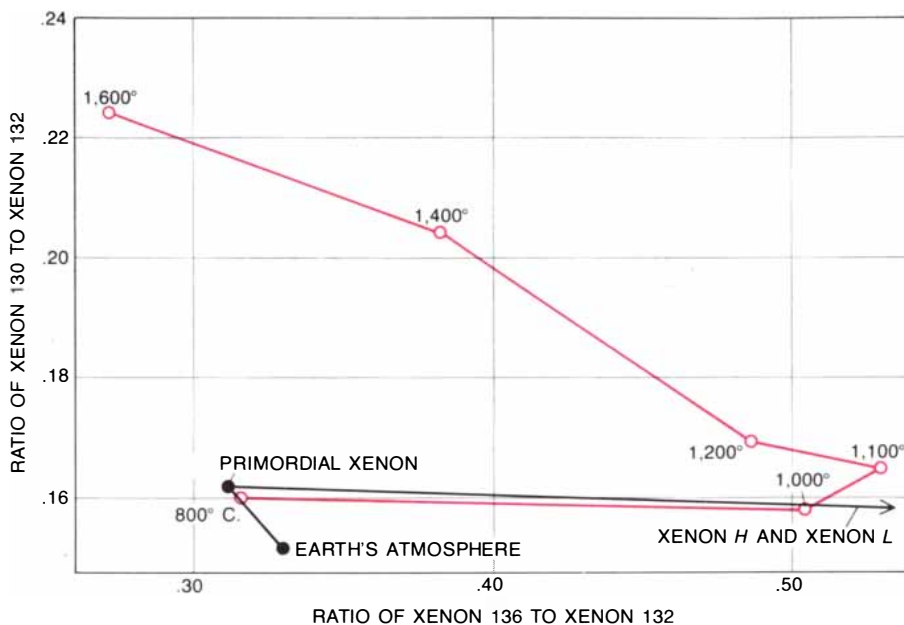
Manuel and his colleagues concluded that the two types of xenon were made together before meteorites ever formed and then were trapped in meteorites pre-mixed. They suggested three possible origins, one of which was nucleosynthesis in different zones of a supernova. The xenon *H* would be made by the *r* process (the capture of neutrons by nuclei on a rapid time scale) and the xenon *L* by the *p* process (proton capture or some other process that gives rise to isotopes poor in neutrons).

Stumbling on the Carrier

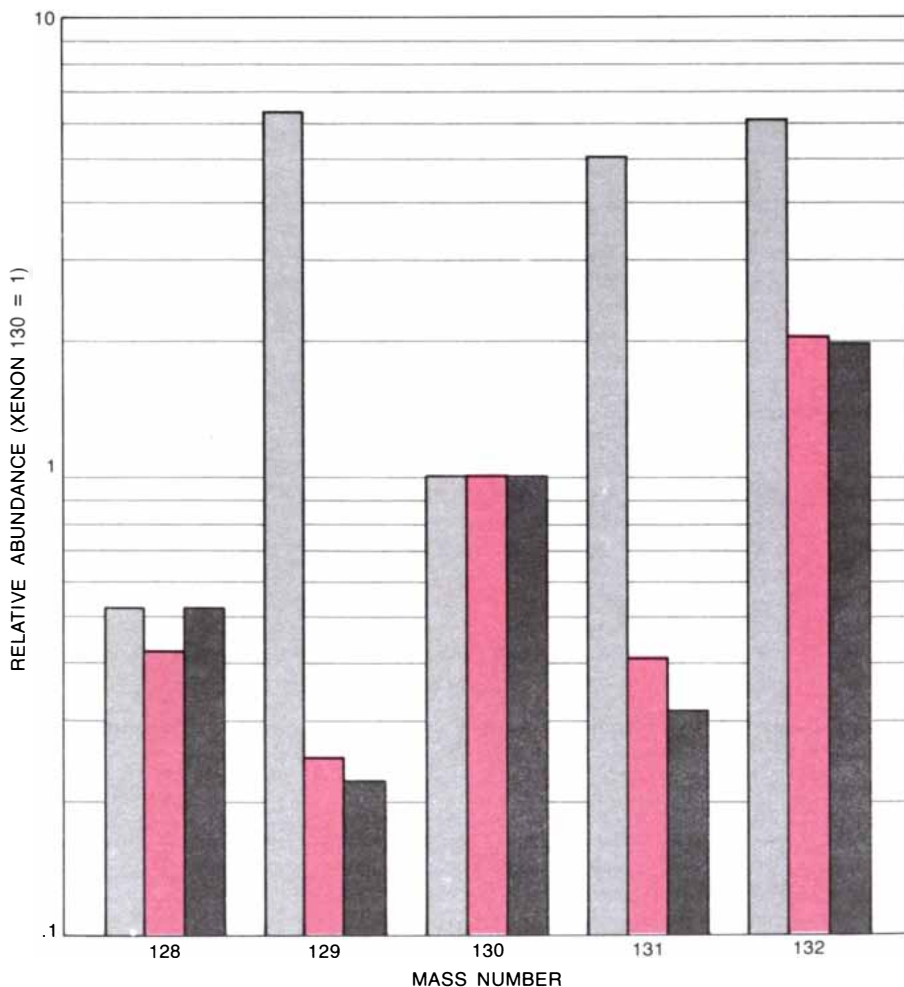
To test the hypotheses (fission inside the meteorite on the one hand, stellar nucleosynthesis on the other) it was necessary to isolate the carrier and study it in detail. If xenon *H* is exotic, its carrier should also be exotic and should show isotopic anomalies of its own. If, however, xenon *H* was made by fission in the meteorite, its carrier ought to show evidence of fission, such as other fission products. In collaboration with B. Srinivasan, who is now at Washington State University, we undertook the project.

As so often in our work, we did the right experiment for the wrong reasons. On the mistaken hunch that xenon *H* is located in the mineral pentlandite $[(Fe,Ni)_9S_8]$ we dissolved a sample of the Allende carbonaceous chondrite in hydrochloric acid and hydrofluoric acid, which supposedly do not dissolve pentlandite. A black residue remained, representing about .5 percent of the mass of the meteorite, and it indeed contained most of the xenon *H*. But pentlandite was only a small and, as it turned out, irrelevant part of the residue. Ultimately the residue proved to consist mostly of amorphous carbon, chromite (Cr_2FeO_4) and spinel.

Annoyingly, the residue also con-



ANOTHER TYPE of anomalous xenon emerged from the stepped heating of a residue of the Murchison meteorite. Srinivasan and one of the authors (Anders) had treated it with reagents that oxidize organic polymers. The intent was to remove primordial xenon. The first two fractions of gas still were close to the line representing mixtures of primordial xenon and the xenon Reynolds and Turner discovered. Then the fractions veered away from the line. The aberrant fractions proved to contain a type of xenon (*s* xenon) consisting mainly of middle isotopes.



ISOTOPIC COMPOSITION of *s* xenon (colored bars) differs from that of solar xenon (light gray bars) but matches that due to the *s* process (slow neutron capture) in red-giant stars (dark gray bars). *S*-process pattern was calculated by D. D. Clayton and R. A. Ward of Rice University.

tained abundant planetary xenon, which thoroughly masked the xenon *H*. We now made a second lucky mistake. Still thinking that at least one of the xenon components was trapped in a sulfide, we etched the residue with nitric acid, which is known to dissolve all sulfides. Sure enough, the primordial xenon in the residue was lost, whereas the xenon *HL* stayed behind, in much purer form than ever before. Actually, however, none of the xenon is trapped in a sulfide; the experiment succeeded for quite a different reason. The primordial xenon is trapped at the surface of grains and is lost when nitric acid etches the surface away. Xenon *HL* is trapped inside the grains and stays behind.

At this point, then, we had managed to concentrate the carrier (or carriers) of xenon *HL* in a residue representing a minute fraction of the mass of the meteorite. Now we could look for isotopic anomalies due either to stellar nucleosynthesis or to fission. The first tests were negative: both carbon and osmium were normal. Then Robert Clayton and Mark H. Thiemens at Chicago and Urs Frick and Pepin at Minnesota tested oxygen and nitrogen. Both were distinctly anomalous. Since the residue still contained a mixture of carbon, chromite and spinel, however, it remained unclear whether the anomalies were associated with the actual xenon carriers. The mixture of minerals was analogous to a set of nested Russian dolls. If xenon *HL* was in the innermost doll, it would do no good at all to measure the isotopic composition of other elements in the outer

dolls. The problem was to find the innermost doll and make all measurements on that.

Next spinel was eliminated, on the ground that pure spinel extracted from the residue had almost no trapped gases in it. Carbon and chromite remained. But a detailed study of those minerals, done in part by Frick and Ulrich Ott of Berkeley and Sherwood Chang of the Ames Research Center of the National Aeronautics and Space Administration, showed that the major part of the carbon and chromite is poor in trapped gases. Nearly all the xenon *HL* is thus in a small part of the carbon (and possibly chromite), which may or may not be the innermost Russian doll.

Last year P. K. Swart, M. M. Grady and C. T. Pillinger of the University of Cambridge determined the isotopic composition of the carbon. They found that the ratio of carbon 12 to carbon 13 was 93, a number within the range for terrestrial carbon (88 through 93). Taken at face value this result supports a local origin for the carbon and hence the xenon, since a ratio of 93 is quite uncommon in the galaxy. Values in stars range from less than 10 to several hundred, and the average for interstellar clouds is 62.

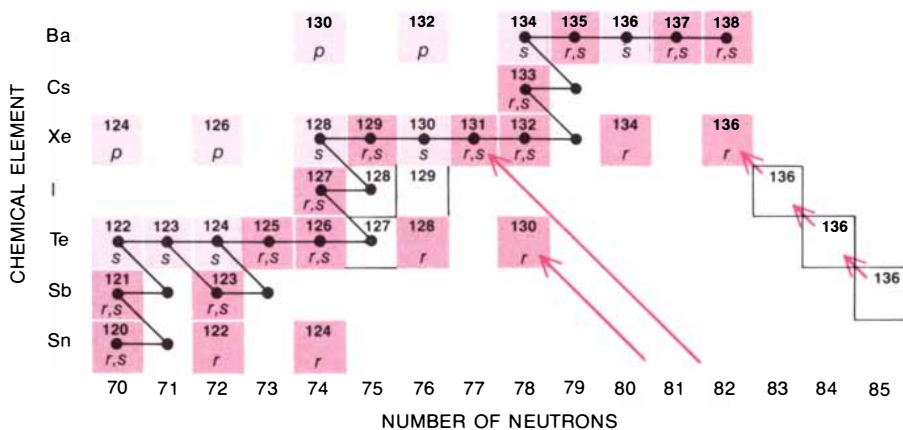
Troubles for Fission

Tests for fission were inconclusive, and so a stalemate persisted for more than 10 years, until two crucial experiments were done at last in 1982. In the first one we, working with I. P. Wright,

S. J. Norris and Pillinger of Cambridge, measured the isotopic composition of the nitrogen in two residues from the Murchison carbonaceous chondrite, one of which, the more fine-grained one, was enriched nearly 600-fold in xenon *HL*. This residue proved to contain very unusual nitrogen, enriched in nitrogen 14 by nearly 30 percent. Such an anomaly could have resulted only from stellar nucleosynthesis, not from fission, and so if the fine-grained sample is indeed the innermost Russian doll, xenon *HL* must have also been made in a star. We are left with the puzzle of why the carbon in the sample is so perfectly normal. One marvels at nature's sense of humor in choosing carbon of terrestrial isotopic composition as a package for highly anomalous nitrogen and xenon.

In the second experiment we, working with G. W. Lugmair and Tadashi Shimamura of the University of California at San Diego, tested the Allende meteorite for isotopic anomalies in barium, neodymium and samarium. Any process that made xenon *HL* should also have made comparable amounts of at least some isotopes of these neighboring elements, in proportions that would vary from one process to another. Surprisingly, we found no anomalies whatever (except for enrichments of neodymium 142 and 143, both attributable to radioactive decay). That finding seems at first an equal embarrassment for both hypotheses about the origin of xenon *H*, stellar nucleosynthesis and fission inside the meteorite. For nucleosynthesis, chemistry offers a possible way out of this predicament. Barium, neodymium and samarium are quite involatile and thus would be among the first elements to condense in the expanding, cooling gas expelled by a supernova. Presumably they would condense on oxide minerals such as perovskite (CaTiO_3). Xenon, being a noble gas, would condense much later, at lower temperatures and on a different substrate, presumably carbon.

No such excuse is available for the fission hypothesis, and so we are left with stellar nucleosynthesis. One possibility is some variant of the *r* process during explosive carbon "burning," which may occur in a supernova when a shock wave passes through a carbon-rich zone of the dying star. Calculations by Dieter Heymann and Marlene Diczkaniec of Rice suggest that the neutrons and gamma rays (high-energy photons) released at a temperature of two billion degrees rework the isotopic composition of heavy elements in the zone so that xenon emerges in the right isotopic proportions. Another possibility is a faster *r* process that builds nuclei all the way up to short-lived superheavy elements of mass near 280, which then fission into lighter elements including xenon. Calculations by E. P. Steinberg and B. D. Wilkins of the Argonne Na-



NEIGHBORHOOD OF XENON on a chart of the isotopes shows how different isotopes derive from different processes: the *s* process (*s*), the *r* process (*r*) and the *p* process (*p*). Each involves characteristic moves on this nuclear chessboard. In the *s* process stable atomic nuclei slowly capture neutrons (horizontal black lines). Whenever a capture makes a nucleus unstable, that is, radioactive (white boxes), a beta decay (diagonal black line) intervenes: the nucleus emits an electron, converting one of its neutrons into a proton. The *s* process cannot make the two lightest and the two heaviest isotopes of xenon; thus it fits the pattern of the xenon discovered by Srinivasan and Anders. In the *r* process—rapid neutron capture—neutron-rich nuclei (mostly off scale to the lower right) are built up so fast that even the short-lived radioactive ones have little chance to decay. When the supply of neutrons is gone, a series of beta decays converts them into stable isotopes. One series, leading to xenon 136, is shown (string of diagonal colored arrows). The radioactive nuclei made by fission decay along similar paths. Some stable isotopes (light colored boxes) cannot be made by beta decay. They are "shielded": a stable nucleus richer in neutrons (dark colored boxes) ends the sequence. The shielded isotopes form only by the *s* process or the *p* process: proton capture or the ejection of neutrons by gamma rays.

tional Laboratory show that the fission would help to explain the high solar-system abundance of the rare-earth elements terbium, dysprosium, holmium and erbium. It also accounts for the xenon *H* data rather well.

Whether or not short-lived super-heavy elements exist may ultimately be decided by efforts to synthesize them at the major nuclear research centers at Berkeley, Dubna in the U.S.S.R. and Darmstadt in Germany. Prospects for a long-lived superheavy element have dimmed greatly, however, as a result of the latest work on xenon *H*.

Dust Grains from Red Giants?

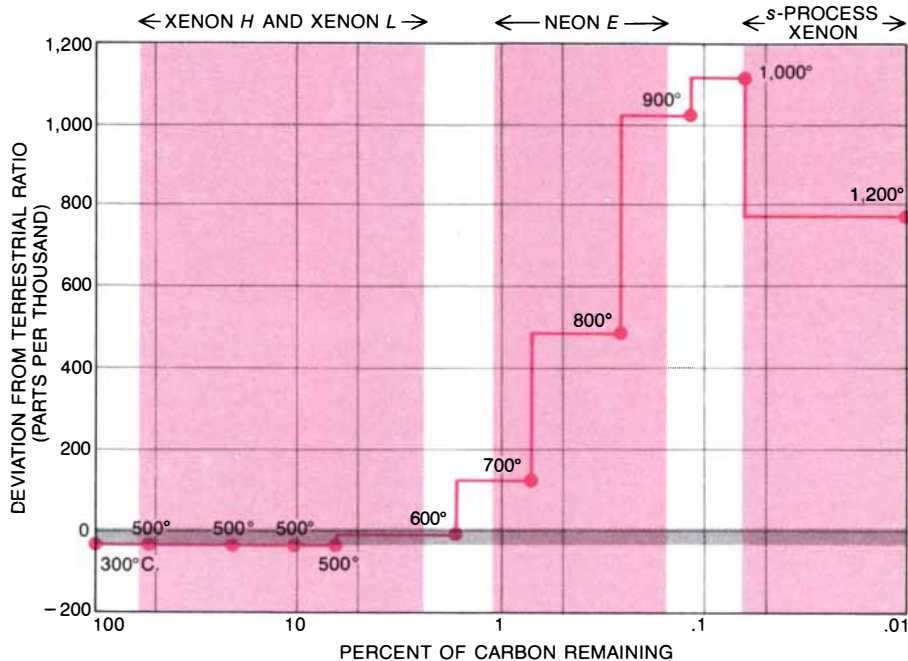
While struggling with xenon *HL* we accidentally discovered a third anomalous type of xenon. It is rarer than the others and also better hidden, but its isotopic composition points quite plainly to a specific place of origin.

Again we did the right experiment for the wrong reason. Srinivasan and one of us (Anders) were trying to characterize the xenon *HL* in the Murchison meteorite in the hope that it might differ from the xenon in Allende in its ratio of xenon *L* to xenon *H*. (Murchison is a C2 carbonaceous chondrite, Allende is C3.) To our disappointment the etched residues from Murchison still contained a lot of primordial xenon. Presumably the nitric acid treatment had failed to remove all the organic polymer. (Organic polymer contains primordial gases and is present in C2 chondrites but not in C3.) We therefore treated the residues with NaOCl and Na₂O₂, compounds that oxidize organic polymers. Then we examined the samples by stepped heating.

The first two data points showed what we were looking for. Primordial xenon came off at 800 degrees C., followed at 1,000 degrees by xenon *HL* purer than any Murchison residue had ever yielded before. The next four points were surprises. Beginning at 1,100 degrees they veered sharply away from the line on our chart representing mixtures of primordial xenon and xenon *HL*. Apparently the sample contained a new xenon component lying at or beyond the 1,600-degree point. It constitutes less than one part in 10⁴ of the total xenon, so that we had missed it in all previous experiments.

The new component proved to be enriched mainly in the even-numbered isotopes xenon 128, 130 and 132 and grossly deficient in the lightest and the heaviest isotopes xenon 124, 126, 134 and 136. This pattern was highly characteristic: it clearly pointed to the *s* process. By coincidence the theoretical pattern for *s*-process xenon had just been calculated by Donald Clayton and R. A. Ward; it matched our new component perfectly.

The *s* process, like the *r* process,



COMBUSTION OF A CARBON RESIDUE from the Murchison meteorite shows that it hides three exotic types of carbon; each one carries anomalous neon or xenon. The sample was heated in steps in the presence of oxygen; in each resulting fraction of carbon dioxide the ratio of carbon 13 to carbon 12 was determined. The first fractions came from the carrier of xenon *H* and xenon *L*. By chance their carbon had isotopic proportions like those of carbon on the earth (gray band). Later fractions came from the carriers of neon *E* and *s* xenon. The experiment was done by P. K. Swart, M. M. Grady and C. T. Pillinger of the University of Cambridge.

builds heavy elements by neutron capture, but much more slowly: intervals of years rather than milliseconds intervene between successive neutron captures by any one nucleus. Hence any short-lived radioactive nuclei formed will tend to decay before capturing another neutron. The buildup of nuclei thus follows an "s-process path" involving neutron captures punctuated by decays. Take tellurium 122, a stable isotope. The *s* process converts it sequentially into tellurium 123, 124, 125 and 126, all of which are stable, or nonradioactive. Then comes tellurium 127, a radioactive isotope. It decays into iodine 127. (Specifically, a neutron in the nucleus of tellurium 127 emits a beta particle, or energetic electron, and is thereby converted into a proton.) The capture of another neutron yields iodine 128, which is radioactive; then xenon 128 through xenon 132; then xenon 133, which is radioactive; then cesium 133, and so on. Note that the *s* process does not make the lightest and heaviest xenon isotopes, the very ones that are missing in the new meteorite component.

The *s* process occurs in red giants: stars that have exhausted the hydrogen in their core and shine by thermonuclear reactions that convert helium 4 into carbon 12. They also burn the hydrogen that remains in a gaseous shell farther outward. The site of the *s* process is below the hydrogen-burning shell, where neutrons are freed by reactions such as the one fusing neon 22 and helium 4 into magnesium 25, allowing a neutron to

escape. The neutrons then are captured by iron 56 and other heavy nuclei; thus the *s* process begins.

At one time, therefore, the *s*-process xenon, *s* xenon for short, must have been inside a red-giant star. When and how did it escape? Perhaps at the red-giant stage; red-giant stars are known to emit a wind of particles by which they lose mass at rates as great as 10⁻⁷ times the mass of the sun per year. Perhaps at some later time, when a red-giant star evolves into a nova or supernova. Matter is ejected at all the stages of a red giant's evolution and cools by expansion and radiation until its temperature is sufficiently low (less than 2,000 degrees K.) for grains to condense. Stable isotopes of xenon and other heavy elements, once made by the *s* process, undergo no further change in a red giant until the star becomes a supernova and temperatures rise above two billion degrees. In contrast, the lighter chemical elements such as carbon, nitrogen, lithium, helium and neon continue to react as the star evolves and heats up. Therefore we turned to the lighter elements. Their isotopic composition might provide some clues to the precise stage at which the grains bearing *s* xenon were ejected.

Heavy Carbon

We soon found that the *s* xenon in meteorites is in carbon. We treated our residue of the Murchison meteorite with perchloric acid, which oxidizes carbon

into carbon dioxide while leaving most minerals unaffected. The *s* xenon disappeared.

There was reason to expect this carbon to be isotopically anomalous, because red giants and their successor stars have ratios of carbon 12 to carbon 13 that vary more than a hundredfold. The problem was how to resolve this exotic carbon from the much larger amount of ordinary carbon that presumably was still in the sample. We used the fine-grained Murchison residue that had been demineralized for nitrogen. It had been demineralized with acids, etched with alkaline oxidizing agents to remove the polymer and finally ultrafiltered to remove material coarser than one micron. It was analyzed by Swart, Grady and Pillinger.

To resolve the various carbon components they used a stepped-combustion technique invented by David J. Des Marais of the Ames Research Center, in which a sample is heated to progressively higher temperatures with oxygen and the carbon dioxide formed at each step is analyzed in a mass spectrometer. The different types of carbon in the sample burn off according to their grain size and crystallinity and with luck can be analyzed one by one.

The major part of the carbon proved to belong to the carrier of xenon *HL* and had its characteristic composition, some 3 to 4 percent lighter than standard carbon. Above 600 degrees C., however, the carbon got increasingly heavy, reaching a 110 percent enrichment in carbon 13 at 900 to 1,000 degrees. Such an extreme composition is unprecedented. Apparently a new, isotopically heavy type of carbon is present in the meteorite. If the extremest enrichment of 110 percent represents this heavy car-

bon in pure form, our Murchison residue contains .45 percent of heavy carbon and the bulk Murchison meteorite contains about five parts per million.

Actually the heavy carbon is probably a mixture of two types of carbon: the carrier of *s* xenon and that of neon *E(L)*. The carrier of xenon *HL* is also carbon. This gives a total of three types of exotic, presumably presolar, carbon in primitive meteorites.

A Tracer of Interstellar Molecules

The last trail we shall follow involves deuterium (hydrogen 2) and leads to a fourth type of exotic carbon. The first hint came in 1953. Giovanni Boato, working in Harold C. Urey's laboratory at Chicago, found that several carbonaceous chondrites were enriched in deuterium by up to 31 percent (with respect to ocean water). For many years no attempt was made to trace the enrichment to one of the principal chemical forms of hydrogen in carbonaceous chondrites: clay minerals containing OH groups and organic matter containing CH groups (mainly the organic polymer mentioned above).

The key experiment was done at last in 1979 by Y. Kolodny, J. F. Kerridge and I. R. Kaplan of the University of California at Los Angeles. Stimulated by the work of F. Robert and his colleagues at the Saclay Nuclear Research Center (CENS), they analyzed samples of carbonaceous chondrites by stepped heating before and after the organic matter in the samples had been burned off in a plasma (an ionized gas) of oxygen; they inferred the deuterium content of the organic matter from the difference between the two analyses. The deuterium enrichment turned out to be lo-

calized in the organic matter, with excesses ranging up to 160 percent. Later measurements by Robert, Richard H. Becker, Jiyoung K. Yang and Samuel Epstein of the California Institute of Technology extended the range to 310 percent and showed that of the various types of organic matter it is the polymer that shows the greatest enrichment.

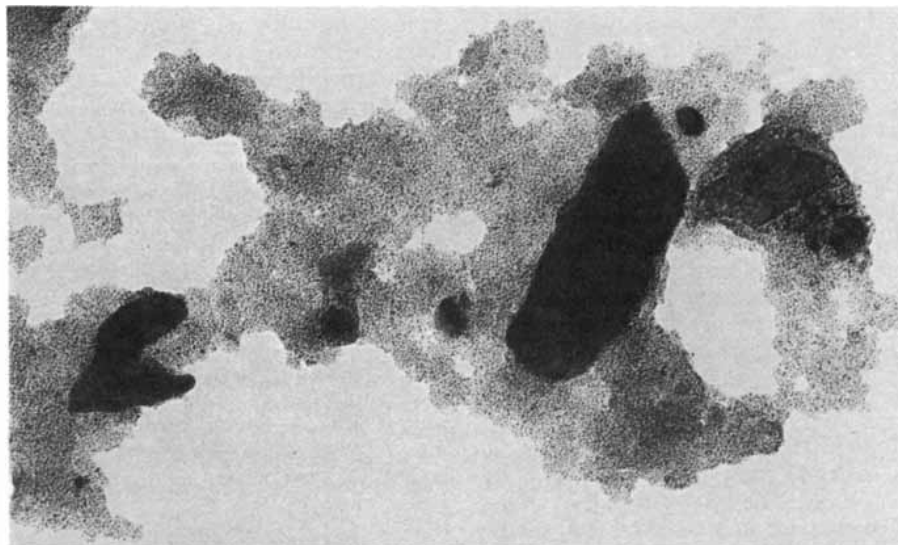
Deuterium is the most fractionation-prone of all isotopes, having twice the mass of its sister isotope hydrogen 1. An enrichment of 310 percent, or a factor of 4.1, is nonetheless hard to explain. In fact, the enrichment is even greater, because seawater itself is enriched eightfold with respect to galactic hydrogen. Thus the overall enrichment of the meteoritic polymer is 32-fold.

In principle the enrichment could be produced by isotopic exchange reactions such as the one that transfers a deuterium atom (D) from molecular hydrogen to methane (CH_4): $\text{CH}_4 + \text{HD} \rightleftharpoons \text{CH}_3\text{D} + \text{H}_2$. According to theoretical calculations, this reaction enriches the deuterium content of methane 32-fold at a temperature of about 130 degrees K. Similar enrichments occur when the organic matter forms from small molecules. The catch is that reaction rates at 130 degrees are ludicrously low. The half-time for the formation of methane from carbon monoxide and molecular hydrogen is 10^{30} years.

J. Geiss of Bern and Hubert Reeves of Saclay have developed a better explanation from a suggestion they first made in 1972. Molecules in interstellar space such as HCHO and HCN are enriched in deuterium by factors of up to 10^5 , according to their radio spectra. Perhaps the meteorites contain a trace of such interstellar molecules.

Ion-Molecule Reactions

Why are interstellar molecules enriched in deuterium? The question was answered in 1973 by William D. Watson of the University of Illinois at Urbana-Champaign. It was known at the time that interstellar molecules form mainly by a series of ion-molecule reactions, which are fast even at the lowest temperatures. Typically a carbon ion made by cosmic rays reacts with a hydrogen molecule to form the molecular ion CH_2^+ . The molecular ion then reacts with other hydrogen molecules to form a more complex ion. Finally it reacts with a free electron to form a stable, neutral molecule such as CH_4 . Watson noted that the chemical bond between carbon and deuterium is slightly stronger than the bond between carbon and hydrogen 1, causing deuterium to concentrate in ions during exchange reactions. For the same reason ions shedding hydrogen will tend to lose hydrogen 1 rather than deuterium, causing deuterium to concentrate still further in the molecular ion and ul-



ELECTRON MICROGRAPH of a Murchison sample containing the three exotic types of carbon was made by Mitsuo Ohtsuki of the University of Chicago. Only one of the types can be pinpointed: the fine-grained material carries xenon *H* and xenon *L*. The other two types make up a few percent of the larger, darker grains, which range up to a micrometer in size.

timately the neutral molecule. The extent of the enrichment depends on the particular molecule but rises with falling temperature.

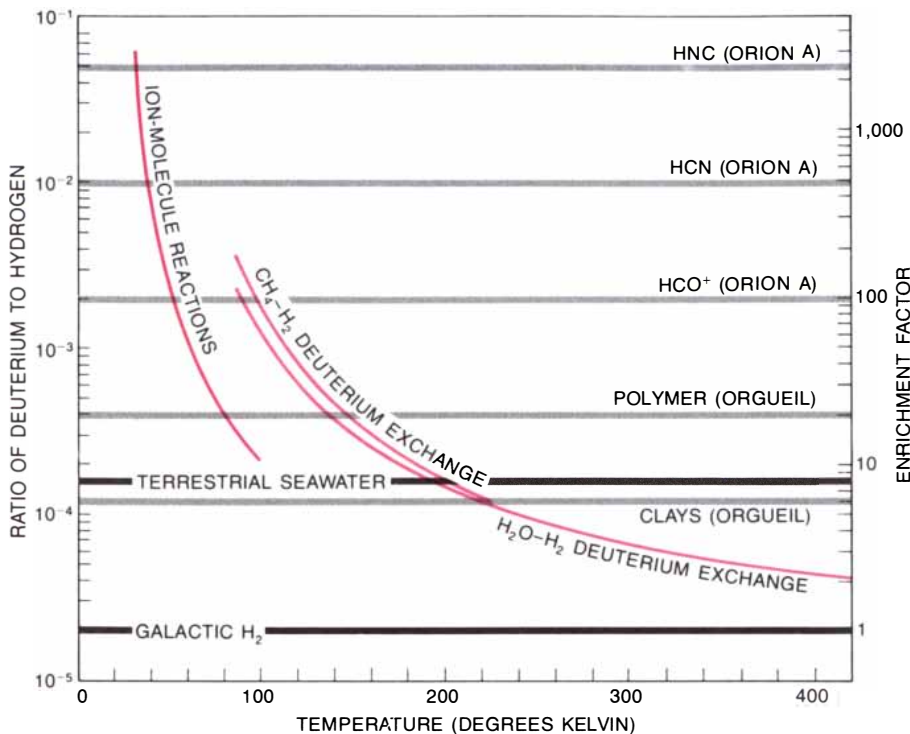
The deuterium-rich matter in meteorites is present as a complex organic polymer, not as pristine small molecules such as those detected by radio astronomers. This is not surprising, since these molecules are too volatile and also too reactive to persist unchanged. The carbon in the polymer is isotopically unexceptional (1.6 percent lighter than standard carbon), suggesting either that the carbon in interstellar molecules is unexceptional or more probably that the interstellar carbon is greatly diluted by "local" (primordial) carbon.

The local polymer, in any case, has a complex structure, consisting of aromatic ring systems with up to four fused rings joined by hydrocarbon bridges $[(CH_2)_n]$. That tells us little about its origin; coal formed from plant matter is similar. Apparently many kinds of organic matter form aromatic polymers on sustained heating. The deuterium-rich, interstellar part of the polymer need not have the same structure, because many interstellar molecules, including HNC, $H(C\equiv C)_nCN$ and HCHO, are quite reactive and may have polymerized on the surface of grains at low temperatures. Perhaps the difference will make it possible to separate the local and exotic polymers by chemical means.

Solar-System Carbon

In the course of this article four types of exotic carbon in primitive meteorites have been discussed; three are forms of elemental carbon carrying neon $E(L)$, xenon HL and s xenon, and the fourth is a polymer enriched in deuterium. Each is hidden in local carbon. How did carbon end up in such a wealth of chemical states?

We can try to answer the question in terms of thermodynamics, which predicts the carbon compounds that should have formed in the solar nebula at various temperatures and pressures. Let us therefore see what happens when a solar gas cools from high temperatures at a pressure of 10^{-5} atmosphere, a reasonable value for the region of the future asteroid belt. (The model we shall present is based on work Ryoichi Hayatsu and one of us, Anders, have done with M. H. Studier of Argonne.) Initially carbon is present mainly as gaseous carbon monoxide, the dominant form of carbon in interstellar space. It remains in that form as the gas contracts and heats up to form the solar nebula. On cooling it ought to hydrogenate into methane. In the absence of catalysts, however, the rate of reaction is very low. The same is true of the reaction that becomes feasible at somewhat lower temperatures



ANOMALOUS DEUTERIUM reveals a fourth exotic type of carbon. In the chart the ratio of deuterium to light hydrogen (that is, hydrogen 2 to hydrogen 1) is compared for a number of sources of hydrogen. The ratio in our galaxy serves as a standard. The earth's seawater proves to be enriched by a factor of eight. Clays from the Orgueil meteorite are enriched slightly less than that; carbon polymer from Orgueil is enriched markedly more. Molecules in interstellar clouds such as Orion A are enriched up to 100,000-fold. In principle, reactions that transfer deuterium from hydrogen molecules to other molecules (light color) could cause the polymer's enrichment; the reactions, however, are slow. A more likely possibility is that the polymer includes a trace of interstellar molecules, whose enrichment comes when molecular ions such as CH_2^+ react with hydrogen molecules (dark color). The illustration was devised by J. Geiss of the University of Bern and Hubert Reeves of the Saclay Nuclear Research Center (CENS).

TYPE OF CARBON	CLASS OF METEORITES	CONCENTRATION OF CARBON (PARTS PER MILLION)	NOBLE GASES OR OTHER MARKERS	ORIGIN	RATIO OF CARBON 12 TO CARBON 13
ORGANIC	C1, C2	ABOUT 30,000	PRIMORDIAL	LOCAL (CO + H ₂)	91
ELEMENTAL	C3	2,000	PRIMORDIAL	LOCAL (2CO → C + CO ₂)	91
CARBONATE	C1 SOME IN C2	2,000 OR LESS		LOCAL (CO ₂ + MgO)	83-85
CARBON-ALPHA	C1, C2 SOME IN C3	5 OR LESS	NEON E	NOVA?	80 OR LESS
CARBON-BETA	C1, C2 SOME IN C3	5 OR LESS	s-PROCESS XENON	RED GIANT	42 OR LESS
CARBON-DELTA	C1, C2, C3	200	XENON H, XENON L, AND LIGHT NITROGEN	SUPERNOVA	92
ORGANIC	C1, C2	3,000 OR LESS	DEUTERIUM	INTERSTELLAR MOLECULAR CLOUD	?

TYPES OF CARBON in carbonaceous chondrites are summarized. Three types are local: their properties can be explained by processes in the early solar system. Four types are exotic: their interstellar origin is revealed by the isotopic composition of the carbon or by isotopic anomalies in "markers" that they carry. Each exotic type is discussed in the text of this article.

To preserve your copies of SCIENTIFIC AMERICAN

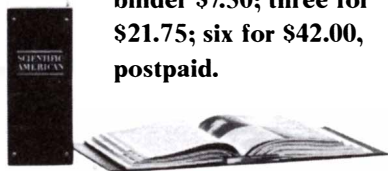
A choice of handsome and durable library files—or binders—for your copies of SCIENTIFIC AMERICAN. Both styles bound in dark green library fabric stamped in gold leaf.

Files Each file holds 12 issues. Price per file \$5.95; three for \$17.00; six for \$30.00, postpaid.



(Add \$2.50 each outside U.S.A.)

Binders Each binder holds 12 issues. Issues open flat. Price per binder \$7.50; three for \$21.75; six for \$42.00, postpaid.



(Add \$2.50 each outside U.S.A.)

To: Jesse Jones Box Corp.
P.O. Box 5120
Philadelphia, PA 19141

Send me _____
SCIENTIFIC AMERICAN

Files Binders
For issues dated through 1982
 1983 or later.

I enclose my check or money order for \$ _____ (U.S. funds only).

Name _____ (please print)

Address _____

City _____

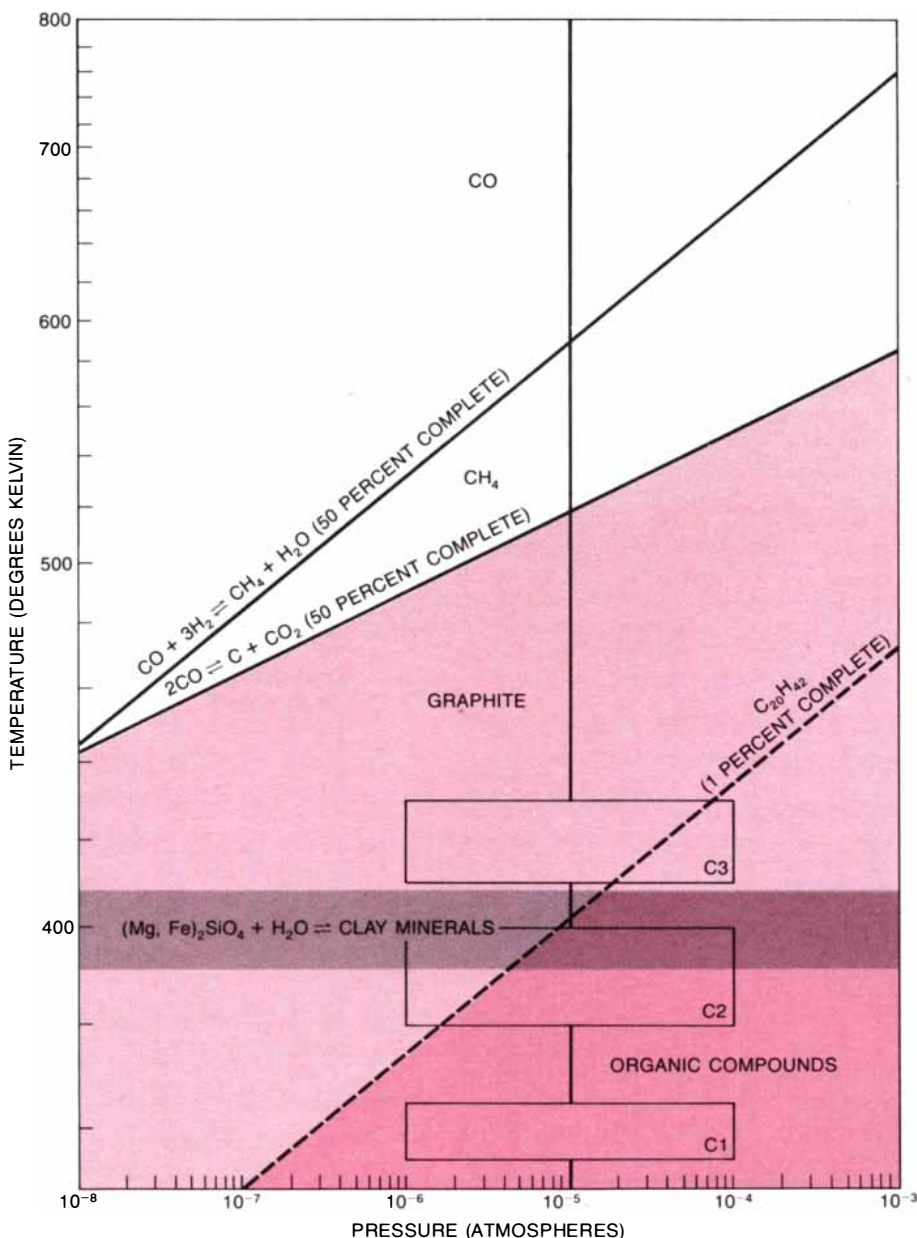
State _____ Zip _____

NOTE: Satisfaction guaranteed or money refunded.
Allow four to six weeks for delivery.

and converts carbon monoxide into carbon dioxide and elemental carbon.

Below 400 degrees K. the situation changes drastically. Clay minerals and magnetite form, both of which are excellent catalysts for the hydrogenation of carbon monoxide, but at these lower temperatures methane, still the most stable hydrogenation product, is no longer the only possible one. Higher hy-

drocarbons (C₂₀H₄₂, for example) can form along with many other organic compounds. Actually they do, as the reaction pathway favors the growth of long hydrocarbon chains. The reaction is essentially the industrial Fischer-Tropsch process: the conversion of carbon monoxide and hydrogen into hydrocarbons, alcohols and other compounds in the presence of a catalyst.



CHEMISTRY OF CARBON in the solar nebula explains the various states of the local carbon in carbonaceous chondrites. At high temperatures carbon monoxide (CO) is the dominant form of carbon in the nebula. Then the nebula cools. The CO ought to be converted into methane (CH₄); a sloping line shows where the reaction is half complete. In the absence of catalysts, however, the conversion is slow, and so most of the CO survives. Next carbon dioxide (CO₂) ought to appear along with elemental carbon (C) in the form of graphite or amorphous carbon grains. Again the conversion is slow. Near 400 degrees Kelvin a dramatic change begins. Silicates react with water, forming clay minerals that catalyze the formation of complex organic compounds. A line charts the combinations of temperature and pressure at which the conversion of carbon monoxide and hydrogen into C₂₀H₄₂, a typical complex molecule, is 1 percent complete. Three rectangles in the illustration show the ranges of temperature and pressure in which carbonaceous chondrites presumably formed. C3 chondrites such as Allende formed at temperatures greater than 400 degrees K.; they contain small amounts of mostly amorphous carbon. C1 chondrites such as Orgueil and C2 chondrites such as Murchison formed at lower temperatures. They contain organic compounds and also the clay minerals that were catalysts.

The industrial process was developed in 1923. The meteoritic hydrocarbons show the signature of the reaction: predominance of straight carbon chains.

The solar-nebula sequence explains the diversity of the local carbon found in carbonaceous chondrites. C3 chondrites, which according to the calculations of J. W. Larimer of Arizona State University formed at temperatures between 410 and 430 degrees K., should contain mainly amorphous carbon but no organic compounds, and that is the case. C1 and C2 chondrites, which formed at temperatures of less than 400 degrees, should contain mostly organic carbon, and they do. The two forms of organic carbon found in C1 and C2 chondrites (soluble compounds and polymer insoluble in standard organic solvents) may represent different lengths of contact between the organic matter and the mineral catalyst. Fischer-Tropsch reactions in the laboratory at first yield only soluble compounds, but after a period of six months these compounds are transformed into complex insoluble matter resembling the meteoritic polymer. The small amount of carbonate in meteorites may have formed when oxides of magnesium, calcium and iron (by-products of the formation of clay minerals) reacted with carbon dioxide (a Fischer-Tropsch by-product).

For presolar carbon one further possibility is available: direct condensation of graphite grains at high temperature. Thermodynamic calculations show that such condensation is feasible only in a gas whose ratio of carbon to oxygen is greater than .9, well above the solar ratio of .6. Red giants and their successor stars tend to have such ratios. In addition there are indications that the dust shells surrounding red giants do contain graphite grains. Such grains ought to be better crystallized and thus more resistant to chemical change than the amorphous carbon that forms at lower temperatures from carbon monoxide. This may well be the reason we were able to purify presolar carbon by techniques of partial combustion and chemical oxidation. In the process, however, we probably lost more reactive types of carbon.

Given only quanta of light and the laws of physics, astronomers and astrophysicists have inferred the nature and inner workings of a wealth of astronomical objects, including red giants, novas, supernovas and interstellar clouds. Now new clues are becoming available: bits of tangible matter from the objects themselves, each containing a rich record of stellar nucleosynthesis and interstellar chemistry. The potential yield of knowledge is great, since laboratory measurements can reveal much detail that is inaccessible to astronomical techniques. Fifteen years ago there seemed no prospect of ever studying stardust. Today the stardust is in hand.

How to tame your data.

To soothe the savage information beast, a businessperson could use the IBM Personal Computer XT.

Because with XT's 10-million-character fixed disk drive and IBM data management software *specifically* designed to complement the hardware, you can whip thousands of names and numbers into more manageable shape. (Helping you get a better shot at the lion's share.)

Use IBM PFS:FILE* to generate a "form" on the screen. Customize it by putting pertinent data in the blank spaces provided.

Then use IBM PFS:REPORT to sort, organize, search, update, store and print the facts with ease.

To learn more about how the IBM Personal Computer XT can help you more efficiently handle many high-volume applications, visit your authorized IBM Personal Computer retail dealer.



The IBM Personal Computer A tool for modern times

For more information on where to buy the IBM Personal Computer, call 800-447-4700. In Alaska or Hawaii, 800-447-0890. *PFS is a registered trademark of Software Publishing Corporation

The Chemical Defenses of Termites

Soft-bodied and blind, termites are subject to predation. In defense termite soldiers attack intruders with an array of sophisticated irritants, toxins, anticoagulants and glues

by Glenn D. Prestwich

For more than 100 million years the insect world has had a chemical arms race. Insect predators attempt to subdue their prey with toxic venoms or to lure them with seductive perfumes. Insect prey species counter by ejecting irritating, sticky, hot or poisonous substances or by being inedible. Unique among the chemically defended insects is the termite caste of soldiers, whose heads and bodies are so thoroughly modified into weapons that they can neither feed themselves nor reproduce. No other insect species, not even the army ants, have such a totally specialized and dependent full-time army. Nor does any other insect order exhibit such a diverse arsenal of chemical weapons and delivery systems.

Termites evolved some 150 million years ago from an ancestral stock resembling the modern roaches. Today there are more than 2,000 species of termites around the world, 95 percent of them in the Tropics of the Old World and the New. They make up the order of insects known as the Isoptera ("equal wings"), which are the most primitive of the social insects.

Termites, often incorrectly called "white ants," evolved quite separately from the social ants, bees and wasps of the Order Hymenoptera. Termite social organization evolved around the need for food sharing to exchange the bacterial and protozoan symbionts necessary to digest cellulose. Termites have rigidly structured societies in which morphologically specialized individuals execute specific tasks: the king and queen reproduce, the workers forage, build the shelter and care for the young, fertile winged pairs fly off to form new colonies and the soldiers defend. Communication among individuals of the colony is based on the exchange of chemical signals, both by smell (olfaction, using the antennae) and by taste (contact chemoreception). Building, food finding, nest-mate recognition, trail following, alarm and defense all involve specific chemical cues.

Termite soldiers can be sterile males

or sterile females, and their heads are much different in shape and size from those of termite workers. The soldiers' primary role in the activities of a termite colony is to protect it from intruders. This they do continuously from maturity to death. A dead soldier is to the colony what dead epidermal cells are to the human body; the loss increases the probability of genetic transfer to the next generation. As a result termite soldiers are walking weapons and can execute their defensive role by tactics that include biting, snapping, hole plugging, squirting, oozing, daubing and even self-destructive defecation.

Here I should like to discuss certain termite soldiers representative of both the "lower" and "higher" termites. I shall be placing particular emphasis on those genera of both groups that have supplemented their mechanical weapons (enlarged mandibles) with chemical ones. Many different kinds of disagreeable secretion can be applied to predatory intruders in many different ways. I shall also be showing how the analysis of the specific chemical composition of these different secretions helps in tracing the evolutionary pathways followed by different termite lines, in some instances for as long as 70 million years.

Although the weaponry of termite soldiers can be subdivided into the mechanical and the chemical, the full range of defensive tactics is even broader. For example, the soldiers of some species puncture the cuticle of an intruding insect with their powerful mandibles and then coat the wound with an anticoagulant secretion from their frontal gland. Neither the wound nor the applied secretion would in itself be fatal to the intruder, but the combination of a gashed cuticle and a chemical that either prevents clotting or is internally toxic means the wounded insect will eventually succumb.

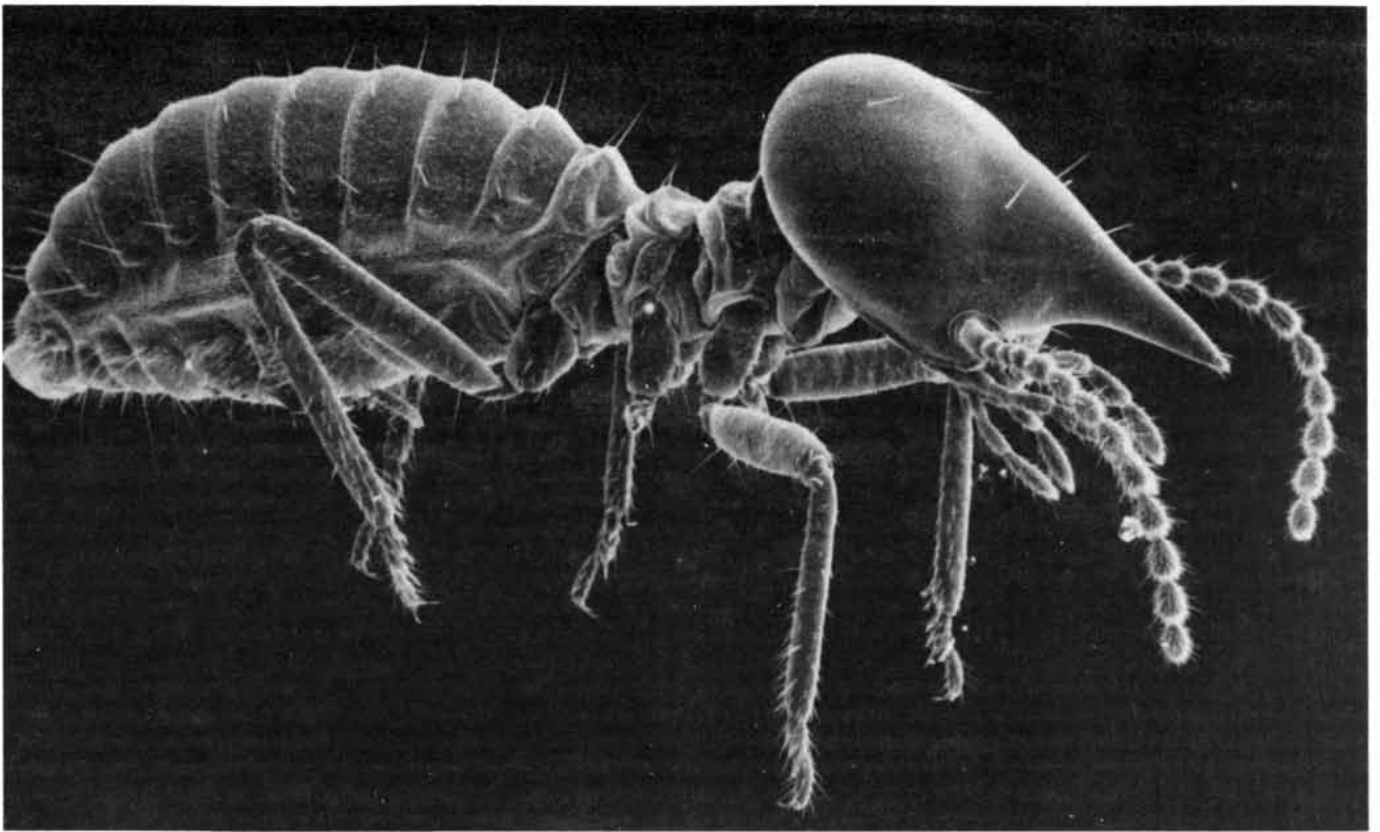
At the same time at least two other defensive tactics depend on neither mandibles nor secretions. The head of the *Cryptotermes* soldier is shaped some-

thing like a drain plug, and the soldiers' reaction to an alarm is to disperse to each of the colony's narrow entrances and literally seal off the cylindrical holes with their pluglike heads. Colonies of *Anoplotermes* do not even have a soldier caste, but each worker in the colony has an abdomen ringed with a specialized constricting muscle. When the worker is confronted by an intruder, it contracts the muscle, rupturing its abdomen and bathing the attacker with a mixture of feces and other gut contents.

What has most attracted the interest of entomologists in recent years is the remarkable variety of termite defensive secretions. Two investigators who have studied the morphology and glandular structure of termite soldiers' chemical weapons are André Quennedey of the University of Dijon and Jean Deligne of the Free University of Brussels. They recognize three main methods of chemical defense. The first is biting, with the simultaneous introduction of an oily or toxic material into the wound. The second is daubing, with a contact poison being applied to the cuticle of an attacker with an enlarged labrum, or upper lip, resembling a paintbrush. The third is "glue squirting": the soldier sprays at the aggressor an irritating, viscous entangling agent.

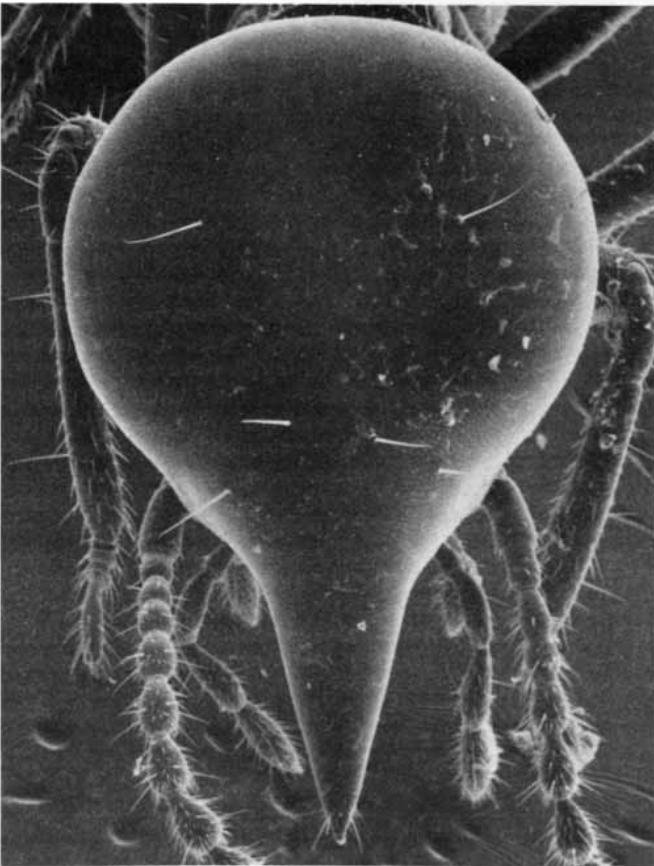
The first of the three methods has evolved independently a number of times in the family of the higher termites, the Termitidae, and also in one of the six families of the lower termites, the underground-dwelling Rhinotermitidae. More than 40 "biting/injecting" species have now been studied in my laboratory at the State University of New York at Stony Brook and by investigators in the Chemical Entomology Unit at the University of Southampton. Something of the variety in chemistry and morphology that has evolved along these lines can be illustrated by considering three biting/injecting genera.

Colony defense in one of these genera, the higher-termite genus *Macrotermes*, is the task of a soldier caste of sterile females with a dual membership. Small

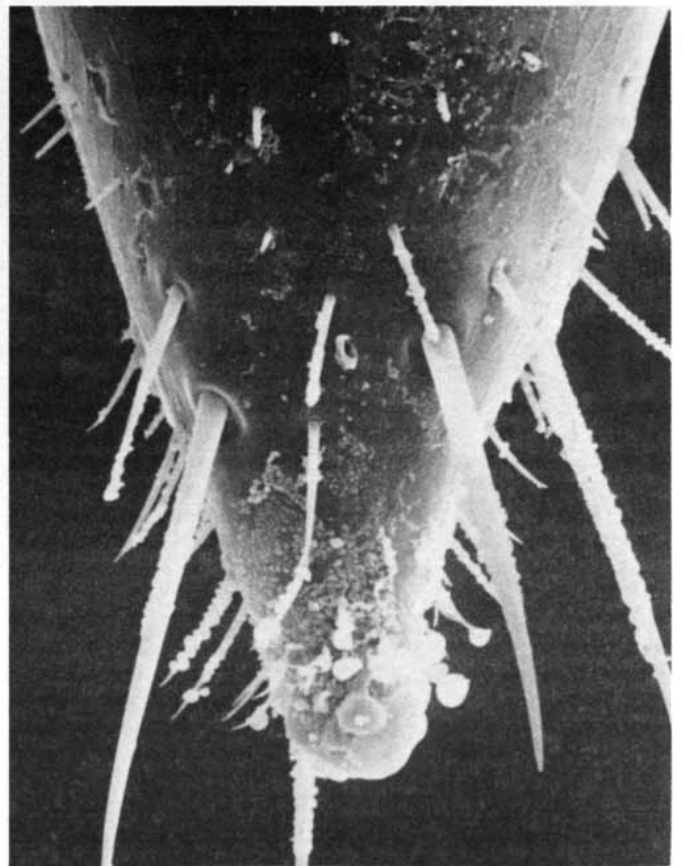


SOLDIER TERMITE of the species *Nasutitermes corniger* is magnified 50 times in this scanning electron micrograph. *N. corniger* is one of the more than 500 species in the higher termite subfamily, the Na-

sutitermitinae. The subfamily is named for the elongated nasus, or snout, extending from the soldiers' forehead. With this bazooka-like proboscis the nasute termites squirt a gluelike substance on intruders.



TOP OF THE HEAD of a soldier of the same species of termite appears in this scanning electron micrograph magnified 70 diameters.



TIP OF THE SNOOT of a soldier of the same termite species is further magnified 1,100 diameters in this scanning electron micrograph.

Macrotermes soldiers escort the large sterile male workers as they gather food and repair the colony's mound. Large soldiers guard the inner hive, where the king and queen and the larvae are sheltered. In soldiers of both sizes the frontal gland of the head secretes an oily substance and exudes it through the fontanelle, a glandular opening in the head cuticle.

The secretion is chemically the same for both kinds of soldiers, but the large soldiers carry 500 times more of it than the small ones. Indeed, it amounts to almost 10 percent of the large soldiers' dry weight. The secretion is a hydrocarbon similar to mineral oil or paraffin: it consists of long-chain alkanes and alkenes [see illustration on opposite page]. The chains, made up of two-carbon acetate units linked together, are from 21 to 35 carbon atoms long; the exact composition of the molecule varies with the species of termite and the location of the colony.

The commonest raiders of termite colonies are ants. When *Macrotermes* soldiers fight an ant, they bite furiously with their mandibles in an effort to snip off the intruder's legs. The heat of the activity turns the waxy secretion in the frontal gland into a free-flowing oil that oozes from the fontanelle, runs down

the soldier's rostrum, or forehead, and spreads out onto the labrum overlying the mandibles. Thus every successful bite applies a liberal coating of the oil to the ant's punctured cuticle. In the absence of a puncture the nontoxic hydrocarbon is harmless. When it is applied to injured cuticle, however, it appears to soften the punctured area. Hence the application interferes not only with the coagulation of the ant's hemolymph but also with the resclerotization, or natural repair, of cuticle damage.

The second biting/injecting genus among the higher termites is *Cubitermes*, one of the soil-eating termite genera found in Africa. The soldiers have saberlike mandibles and squarish orange heads. They too secrete a hydrocarbon. The *Cubitermes* secretion, however, is not a simple straight-chain molecule but belongs to the family of terpenes [see illustration on opposite page]. Terpene molecules are grouped according to the number of five-carbon isoprene units in them. Therefore two-isoprene units (10 carbons) are referred to as monoterpenes, four-isoprene units (20 carbons) as diterpenes and so on. Diterpenes are secreted by very few insects. In fact, except for some of the higher termites no insects have been shown to be capable of synthesizing them. *Cubitermes* soldiers, however, are in the diterpene

manufacturing business on a large scale. As postdoctoral associates of Jerrold Meinwald at Cornell University, David Wierner and I identified three new diterpenes secreted by this termite genus. Of the more than 16 diterpenes found among seven species of *Cubitermes* I now know at least five that are unique to these termites. No other invertebrates, vertebrates or even plants are known to synthesize these novel compounds.

The third biting/injecting genus of higher termites is *Armitermes*, a soil-eating termite found in Central and South America. The soldiers of this genus have sharp-pointed mandibles resembling a pair of tongs and have a nozzlelike protrusion on their forehead. With the nozzle they apply droplets of an oily secretion to the puncture wounds they inflict on intruders. Thanks to diligent collecting in the rain forest of Guyana by one of my collaborators, Margaret S. Collins of Howard University, I have had access to soldiers of several *Armitermes* species for analysis of their secretions.

The *Armitermes* secretions consisted of modified fatty acids with from 22 to 36 carbon atoms to a molecular chain. One end of the chain bears a carboxylic acid group and the other end a hydroxyl group. Similar chains are found in the lanolin of wool, but what is unusual in these secretions is that the two ends of



SOLDIERS OF ANOTHER SPECIES (*Nasutitermes kempae*, found in Africa) cluster around the entrance to a foraging tube that has been opened to allow for photography. The snouts of four of the soldiers

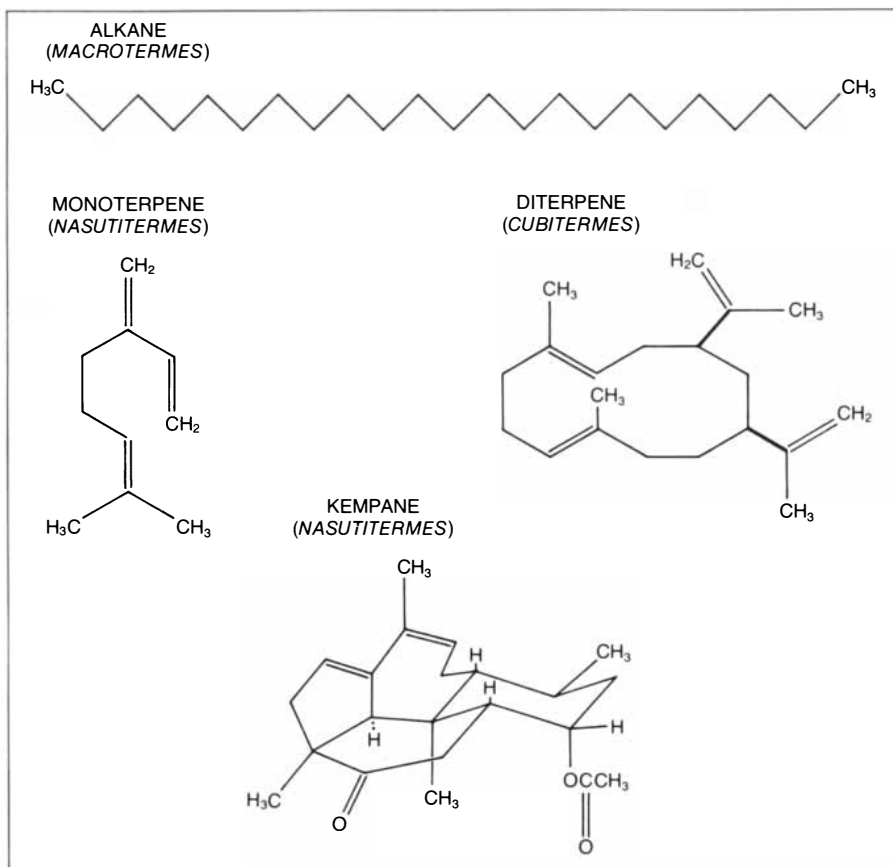
point in the general direction of the camera. This orientation is not accidental. Although termites are blind, the soldiers have oriented in response to air currents caused by the photographer's movements.

the chain are bound together in an ester linkage, forming a molecular loop. Similar loops with fewer carbon atoms, called macrolides, are found among the sex-attractant substances of mammals and substances that line the nests of solitary bees. As far as is known, however, macrolides with from 22 to 36 carbon atoms are unique to *Armitermes* and a related biting/injecting higher-termite genus, *Rhynchotermes*. Barbara L. Thorne of Harvard University and James F. A. Traniello of Boston University have found that when the secretion is applied to an intruder's puncture wound, it acts as a toxin rather than simply inhibiting the healing of cuticle.

The second of the three main methods of chemical defense among termite soldiers, as established by Quennedey and Deligne, involves daubing an intruder with a toxic secretion of chemically reactive lipids derived from fatty acids. Soldier termites of the lower-termite family Rhinotermitidae are morphologically adapted for this mode of defense. Their labrum has been enlarged so that it forms a kind of bristly paintbrush the soldier presses against the intruder, spreading the pungent, oil-soluble contact poison over its cuticle. Unlike the biting/injecting higher termites, these species do not store their secretions in the head alone but have developed a large backup reservoir in their abdomen. In some soldiers the combined capacity of these reservoirs exceeds 35 percent of their dry weight.

Quennedey and his associates at Dijon first described the poison of an African genus, *Schedorhinotermes*, as a vinyl ketone in 1973. At about the same time Ján Vrkoč and his colleagues in the Czechoslovak Academy of Sciences reported that the poison of the pantropical genus *Prorhinotermes* was a nitroalkene. Since then my colleagues and I have confirmed and expanded both identifications in related species. We have also reported a third group of poisons, beta-ketoaldehydes, secreted by the soldiers of two New World Rhinotermitidae: the genera *Rhinotermes* and *Acorhinotermes*.

Although each of these contact poisons has a different molecular structure, all three have two key features in common. First, they are derived from fatty acid molecules with from 14 to 16 carbon atoms, and they all take the form of long carbon chains that are soluble in fat. Second, all three have a chemically reactive electrophilic (electron-seeking) group of atoms at one end of the chain. They might be likened to a poisoned arrow. The fat-soluble shaft of the "arrow" facilitates its passage through the intruder's waxy cuticle. Once inside the cuticle, the poisoned tip of the "arrow," the electrophilic group, does internal chemical damage. Such electrophilic groups are a common feature in natu-



TYPES OF MOLECULES found in termite secretions are represented by their backbones. Termites of the genus *Macrotermes* secrete alkanes: straight-chain hydrocarbons (top). Termites of the genera *Nasutitermes* and *Cubitermes* secrete more complex terpene hydrocarbons (middle). Even more complex terpenes are secreted by the most advanced *Nasutitermes*.

ral defensive secretions. For example, they are found in the secretions of many fungi and higher plants. They can act as antibiotics, tumor inhibitors and insect repellents.

How do these termites avoid poisoning themselves with the toxins they secrete? They must have evolved biochemical adaptations that enable them to detoxify their own electrophiles; in the absence of some method of detoxification the termite soldiers could survive neither the synthesis nor the storage of their chemical weapons, and the termite workers could not survive their deployment. Until recently, however, no one had determined the exact biochemical basis of the termites' immunity. It was speculated that glutathione (GSH), a sulfur-containing tripeptide, and a group of detoxifying enzymes, the glutathione S-transferases, were involved. GSH S-transferases make electrophiles more water-soluble and less chemically reactive, thereby facilitating their excretion. Both vertebrates and invertebrates, and plants too, are known to rely on GSH S-transferase to rid their system of toxic substances. Moderate levels of GSH and GSH S-transferase are also found in termites.

One of my students, Stephen Spanton, and I discovered, however, that the workers of the Florida termite *Prorhinotermes simplex*, whose soldiers synthesize toxic nitroalkenes, and the workers of the African termite *Schedorhinotermes lamanius*, whose soldiers synthesize toxic vinyl ketones, detoxify their soldiers' secretions in another way. They reduce the molecules' electron-poor double bonds, converting the nitroalkenes into nitroalkanes and the vinyl ketones into saturated ethyl ketones. The reduced products are only a tenth as toxic as the unreduced secretions.

We found that the initial detoxification was accomplished by an enzyme that specifically reduces the electrophilic alkene of that species and that it required NADPH, a reduced nucleotide cofactor, as a source of hydride. Among *Schedorhinotermes* workers 50 percent of the vinyl ketone was converted into ethyl ketone in less than an hour. Then all the ethyl ketone was catabolized (broken down) into acetate within two days. The *Prorhinotermes* workers enzymatically reduced their soldiers' toxin in an analogous way. When these workers were exposed to the toxin of the *Schedorhinotermes* soldiers, however, they could not tolerate doses as large as those



**BREAKTHROUGH:
ADD SOUND TO
JET ENGINES TO MAKE
AIRPLANES QUIETER.**

Two airstreams exit from a jet engine in operation—a hot, fast central core and a cooler, slower outer layer. When they mix, they become chaotic. This turbulence causes an unpleasant low-frequency rumble.

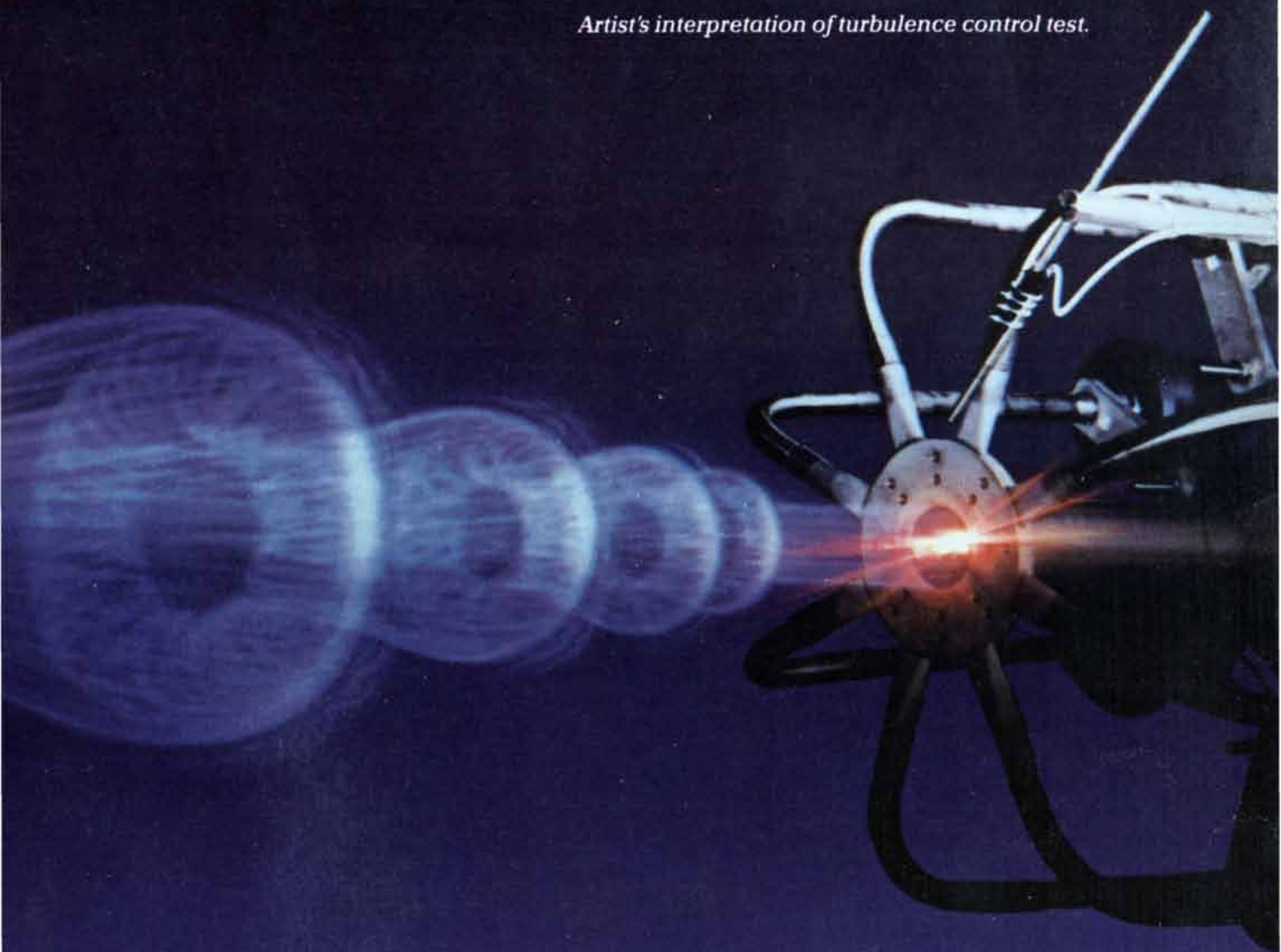
We're finding new ways to control the turbulence by adding sound to smooth the flow. The smoother the flow, the quieter the airplane. The gentle symmetry you see here could mean a quieter ride in the future—quieter for the people aboard the plane, quieter for the people on the ground below.

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

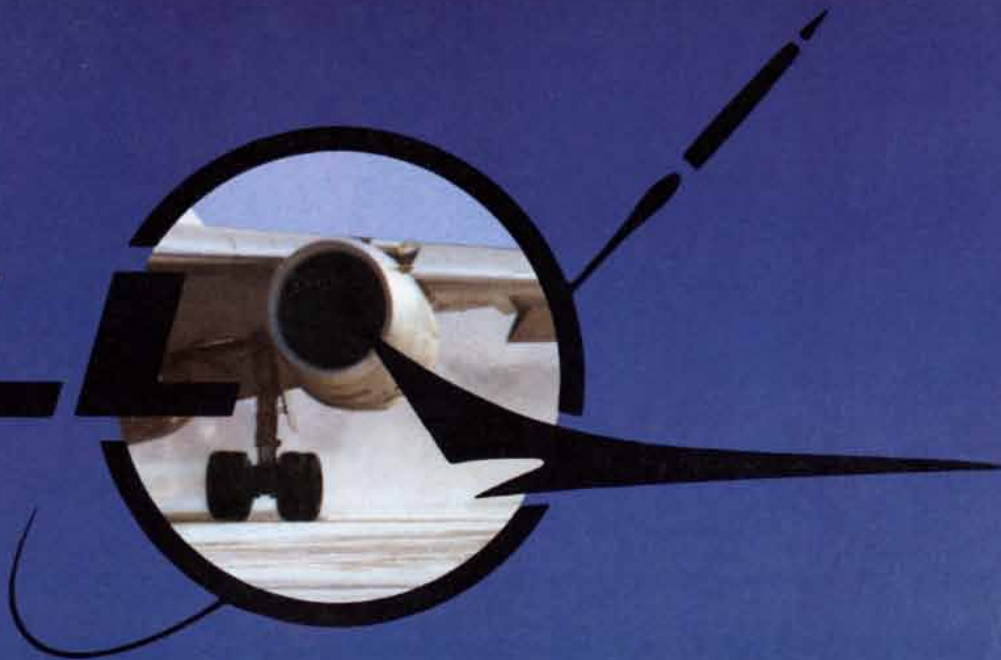
We're McDonnell Douglas.

**MCDONNELL
DOUGLAS**

Artist's interpretation of turbulence control test.



WELLS
AS



they survived when they were exposed to their own soldiers' toxin.

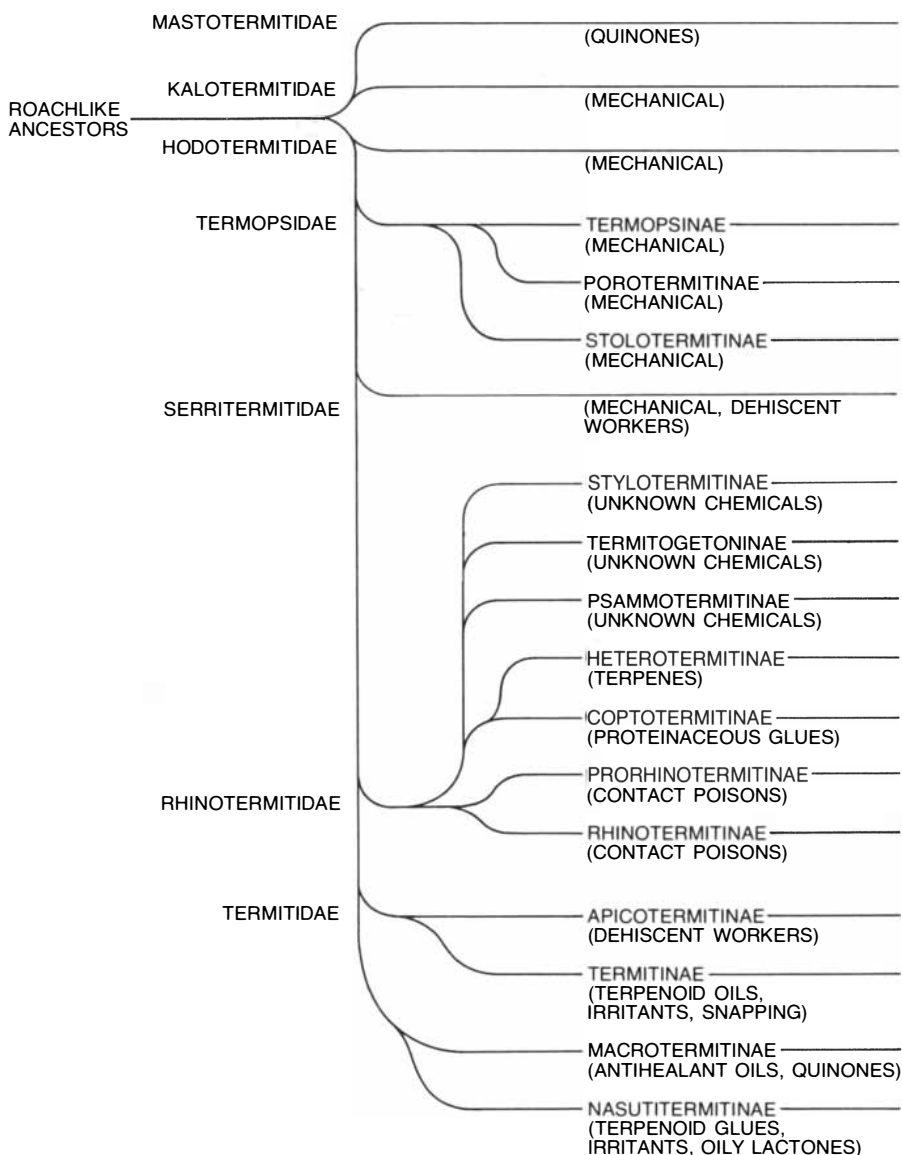
Except among these two termite species the reduction-catabolism pathway for detoxifying electrophilic poisons is apparently a rarity. This may be because termites in general are under evolutionary pressure to develop other detoxification pathways in order to conserve nitrogen for protein synthesis: their cellulose-rich foods are poor in nitrogen. For example, termites practice such nitrogen-conserving strategies as selective foraging and cannibalism.

Cathy J. Potrikus and John A. Breznak of Michigan State University have shown that the termites' own symbiotic bacteria can fix atmospheric nitrogen

and recycle the nitrogenous compound uric acid; both activities contribute importantly to the termites' nitrogen supply. Barbara L. Bentley of the State University of New York at Stony Brook and I have shown that the *Nasutitermes* genera of Costa Rica fix enough molecular nitrogen to double their nitrogen supply in less than six months. The Floridian and African electrophile-resistant termites, however, appear to avoid nitrogen loss during detoxification by recycling, rather than excreting, both the nitrogen-rich amino acids used for detoxification and the energy stored in the carbon chain of the defense secretion.

The last main method of termite chemical defense, and in an evolution-

ary sense the most advanced of the three, calls for still another modification of the soldiers' anatomy. Instead of daubing an intruder with a toxin the soldiers of some 500 species in one abundant pantropical termite subfamily rely on a modified forehead with a nasus: a snoutlike tube. That subfamily is the Nasutitermitinae, named for the nasus. When a nasute nest is attacked, the alarmed soldiers charge from the interior to the point of intrusion and squirt an irritating glue-like secretion from their nasus to entangle the intruder. The workers join the action, biting at the intruder with their sharp mandibles and trying to clamp onto its legs. Ants, spiders and other insectivorous animals, including anteaters, generally avoid the malodorous, distasteful and potentially lethal barrage of the nasute soldiers.

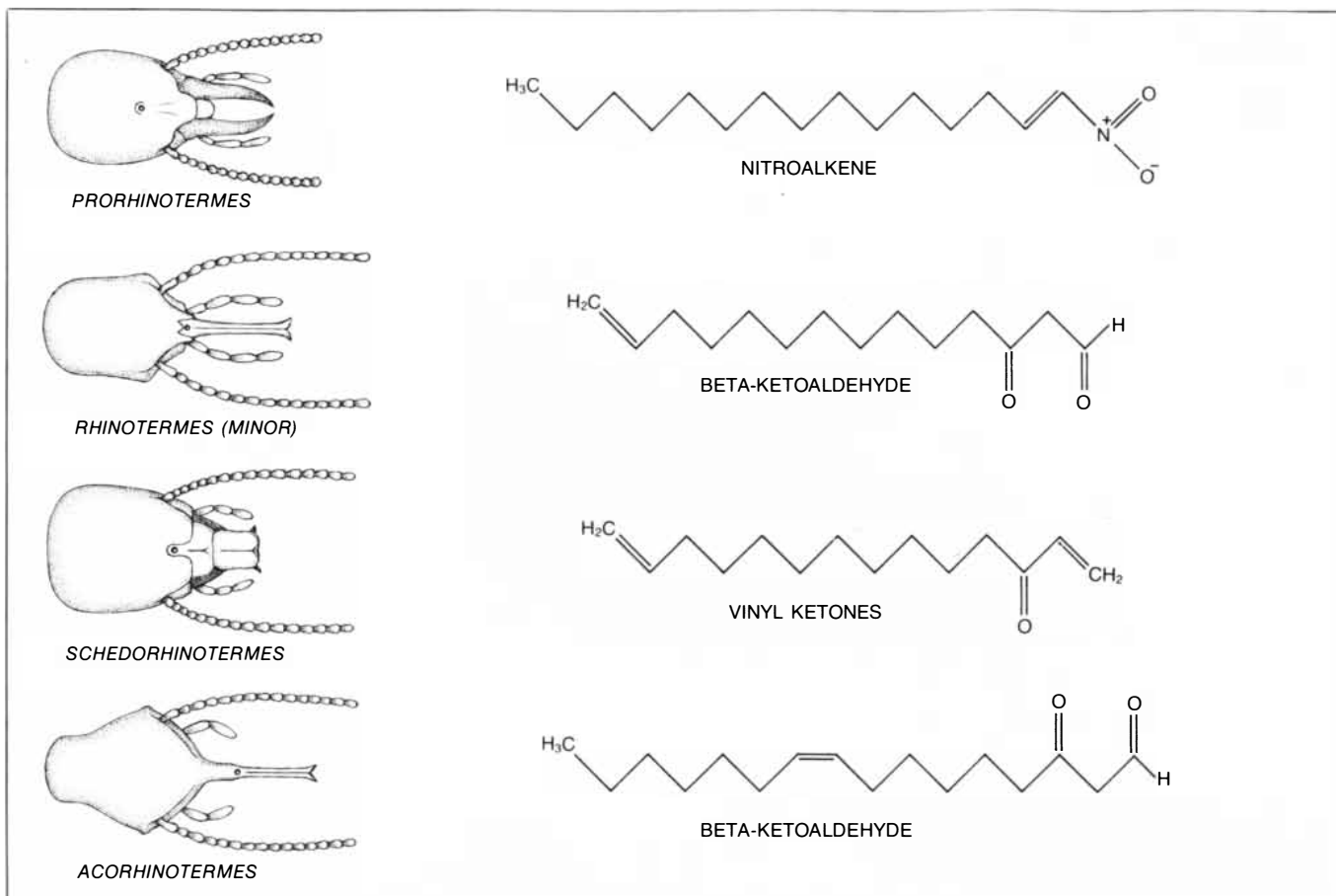


SEVEN FAMILIES OF TERMITES arose (left) from a roachlike ancestral stock. The six families known collectively as the lower termites depend mainly on mechanical defenses: the powerful jaws of their soldiers. In this group, however, the Rhinotermitidae include seven subfamilies, six of them possessing soldiers that use chemical defenses either with biting or without. The seventh termite family, the Termitidae (bottom), includes the four subfamilies of higher termites. One subfamily, the Apicotermitinae, has no soldier caste. Colony defense is undertaken by workers that are dehiscents: they explode and shower the intruder with gut contents. The Nasutitermitinae are found in the Tropics of both the Old World and the New. They evidently evolved some 70 million years ago. Defenses of each group are given in parentheses.

It has been known for some time that the viscid defensive secretion of the nasutes was analogous to pine sap: a mixture of monoterpene hydrocarbons (the solvent) and other isoprenoids of higher molecular weight (the resin). The termite secretion differs from pine sap, however, in containing none of the usual resin acids. In 1974, when my colleagues and I began to investigate this nasute glue, we had many questions. What exactly were the heavy compounds? Were they merely sticky or were they toxic as well? Did the soldiers actively biosynthesize the substances or were they acquired from food? The nasute termites make up the most numerous and diverse termite subfamily and are worldwide in their distribution; how did the composition of the glue vary from species to species and from place to place? Now, nine years later, we are able to answer some of these questions.

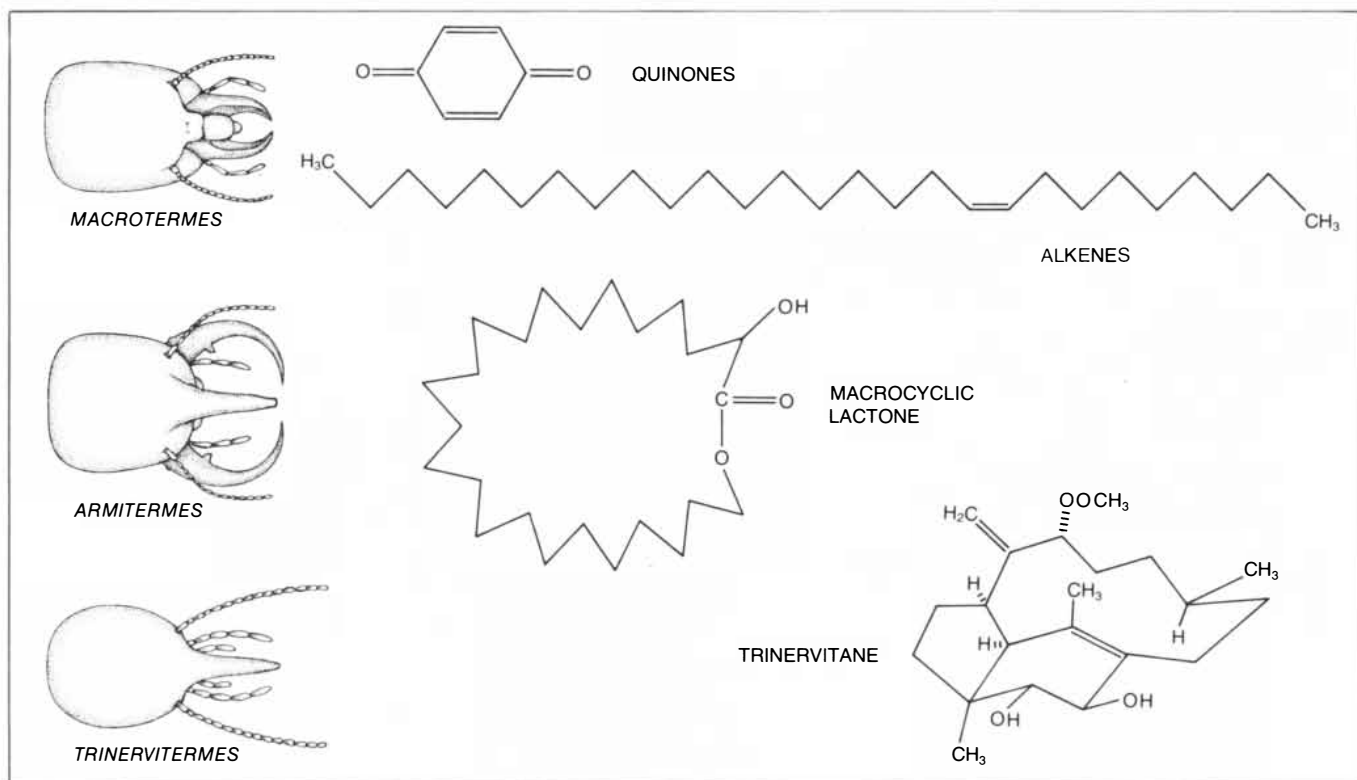
Our first success came with the isolation of several glue components in purified form and the elucidation of their structure. This required the combination of my work in Africa with the efforts of Jon C. Clardy and his colleagues at Iowa State University and Koji Nakanishi and his colleagues at Columbia University. Because thousands of nasute soldiers of the African grass-harvesting genus *Trinervitermes* were the source of the research material, the first glue components were named trinerivitane. They are diterpene molecules consisting of three fused rings with their skeleton of carbon atoms in a dome-shaped arrangement. The carbon skeletons do not resemble any other known natural molecular structure.

Since then we and others have isolated one bicyclic (two-ring) and three tetracyclic (four-ring) diterpenes, known respectively as secotrinerivitane, kempane, rippertane and longipane. Together with some 60 derivative substances they are secreted by the soldiers of other nasute-termite genera. All exhibit the same



MODIFICATIONS OF THE HEAD in the soldiers of four termite genera in the family Rhinotermitidae are shown from the top at the left. The molecular structure of the defensive secretions that each

soldier applies to an invader of its colony is at the right. The long-chain hydrocarbons secreted by the four genera range from a simple nitroalkene to a vinyl ketone and a more reactive beta-ketoaldehyde.

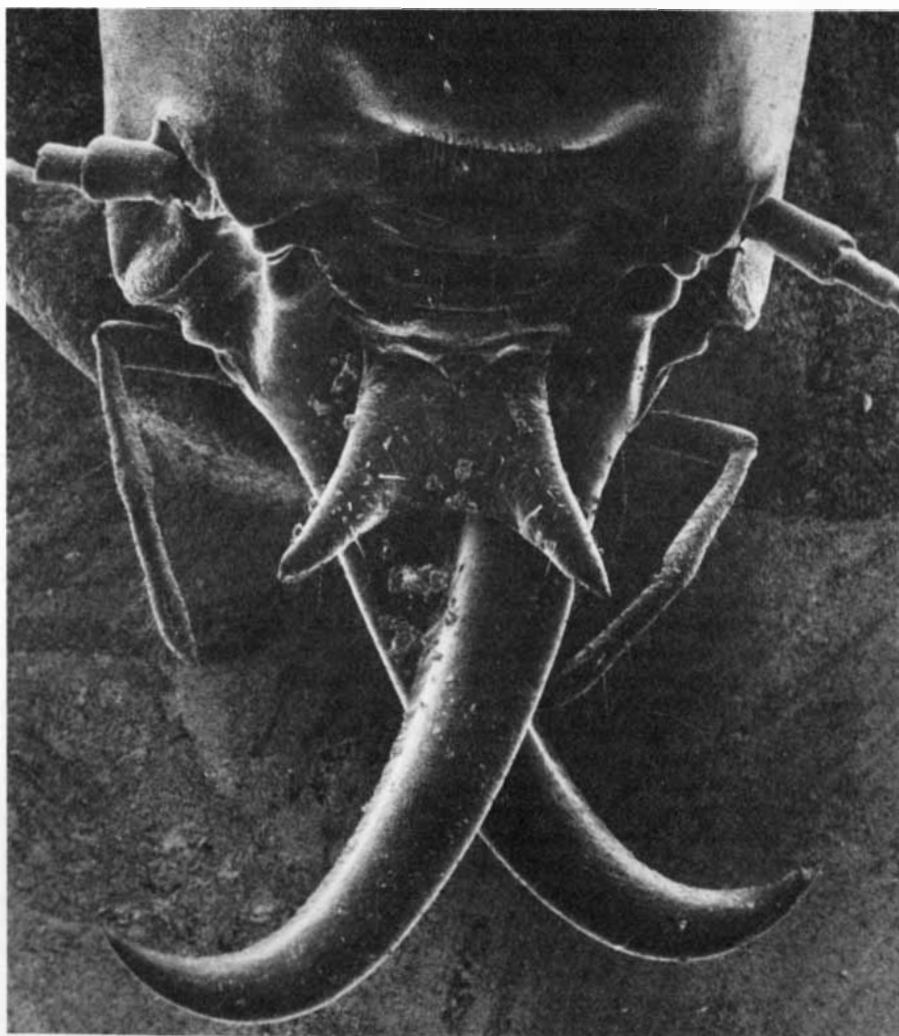


SIMILAR MODIFICATIONS in the heads of soldiers have arisen in three higher-termite genera. At the bottom is the modification typical of the nasute genus *Trinervitermes*. The *Macrotermes* secretions

can be simple quinones, alkanes or alkenes. The nasute genus *Armitermes* secretes molecular loops called macrolides. The *Trinervitermes* secretion is a mix of diterpenes unknown elsewhere in nature.



SQUARED HEAD is the adaptation of the *Cryptotermes* soldier. When these soldiers are alarmed by an intruder, they plug the entrance to the colony with their heads, presenting an array of mandibles to discourage other intruders. These termites have no chemical defenses.



SABERLIKE MANDIBLES of a *Cubitermes* soldier will gash the cuticle of an attacking ant while a secretion from the soldier's frontal gland is spread on the wound. The secretion seems to inhibit the intruder's blood coagulation and cuticle repair, and it will soon bleed to death.

gross structural features. Each molecule has the same biosynthetic progenitor: a monocyclic diterpene, cembrene-A. Each is dome-shaped, with oxygen-containing water-attracting groups protruding from the convex surface and a water-repelling region on the concave surface. Defense against intruders is thus achieved with a substance rather like pine sap: a viscous solution of mixed diterpenes in association with monoterpene solvents. The mixture has superior wetting abilities when it is applied to the normally water-repellent cuticle of insects and other arthropods.

We have also found that the secretions vary among individual soldiers in a single colony, among soldiers in related populations of the same species, among geographically isolated populations of the same species and among species of the same genus. This kind of chemical variation, which is commoner in the plant kingdom than it is in the animal one, can be useful in taxonomic studies.

Exactly what does the nasute glue do? Thomas Eisner and his colleagues at Cornell have shown in a series of elegant laboratory experiments that it functions as an entangling agent, as an irritant (a property that promotes its spread when the sprayed intruder grooms itself) and as a topical poison. Whether the soldiers synthesized the glue or obtained it from food remained a question my colleagues and I decided to pursue with radioactive-tracer experiments. One cannot feed a labeled precursor to termite soldiers because the only nourishment they will accept is the fluid that workers regurgitate for them. We bypassed the problem by resorting to a micropipette to inject radioactively labeled precursor molecules (sodium acetate or sodium mevalonate) directly into the soldiers' abdomen. Because the termite abdominal wall is very flexible, we were able to inject about half a microliter into each subject, roughly doubling the size of its abdomen.

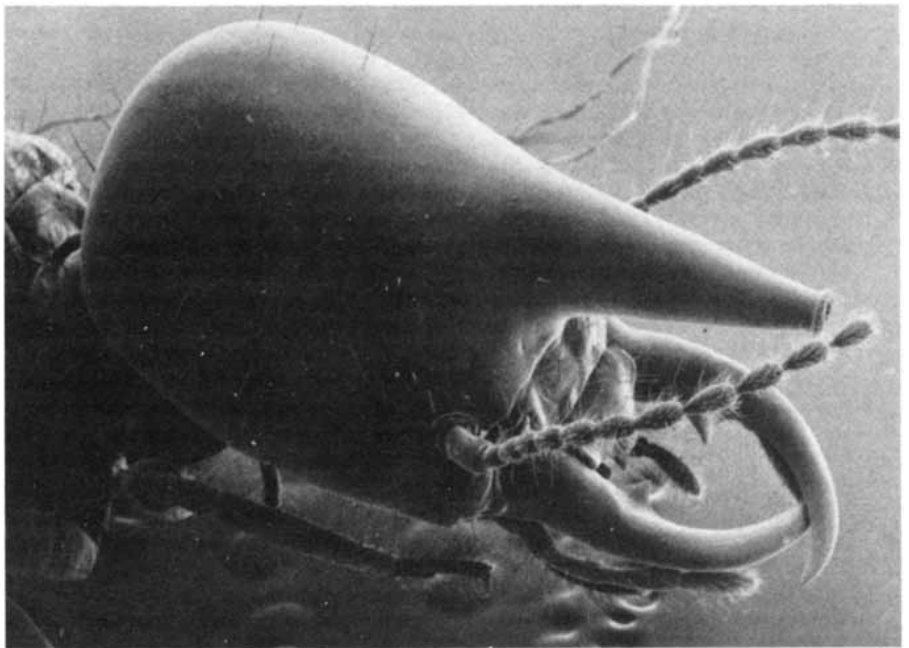
The soldiers survived for some hours after the injection. By then each precursor had been incorporated into the monoterpenes and diterpenes stored in each soldier's head. When we injected workers (which do not secrete diterpenes) with the same labeled precursors, the precursors were not found in their terpenoid compounds. After careful purification of each component we concluded that the nasute soldiers were indeed synthesizing their defensive secretions *de novo*. Their ability to do so sets them apart from all other insects (except perhaps for certain scale insects). None of the rest can synthesize any terpenes larger than a 15-carbon-atom sesquiterpene.

Another experiment demonstrated that the diterpenes secreted by a nasute soldier were not affected by the soldier's

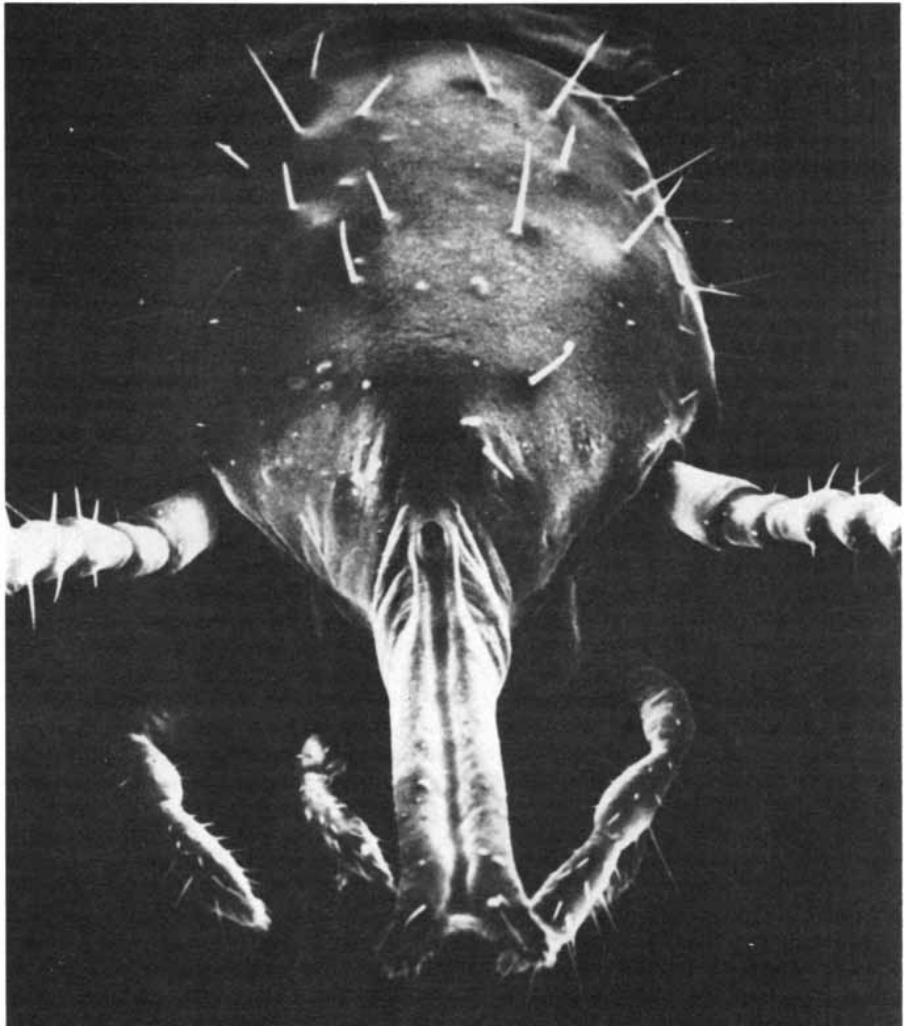
diet. The method involved raising two chemically distinct populations of a single *Trinervitermes* species, beginning with separate pairs of winged male and female reproductives after their nuptial flight. Both pairs and the colonies they gave rise to were given identical food-stuffs: dried grass from one of the collection sites. In spite of their common diet the soldiers that reached maturity six months later in each new colony produced diterpenes chemically identical with those of their parent colonies.

The chemical analyses of the nasute soldiers' secretions have illuminated a particular aspect of termite evolution. It has been customary to consider the pantropical nasute subfamily, the Nasutitermitinae, a leading example of parallel evolution. This is to say that the nasute genera of Africa, Asia, Australia and South America arose from ancestral termites whose soldiers were armed with mandibles alone. In the course of thousands of generations the primitive soldiers (now found only in South America) underwent a regression of their mandibles and a modification of their forehead into the bazooka-like nasus. It seems to be asking too much of a parallel evolutionary process, however, to go on to say that the capacity to synthesize defensive diterpenes, a feat of chemistry unique to the advanced nasute termite soldiers, also evolved independently. I suggest that instead a common ancestral stock of diterpene manufacturers existed in West Gondwana before plate tectonics divided that land mass into a proto-Africa and a proto-South America in Cretaceous times.

In support of this contention one can point to the fact that many existing "primitive" nasute genera can be fitted into an ancestral scheme in which the mandibles did shrink and the nasus did elongate. Chemical examination of the secretions of the intermediate genera, however, fails to show any of the four unique diterpenes that characterize the secretions of the advanced nasutes. The South American nasutes with large mandibles cannot squirt their secretions; they can only ooze them out in droplets, and what is oozed consists of fatty-acid-derived macrolides and mono- and sesquiterpenes. The more "advanced" South American nasutes, on the other hand, do not have functional mandibles but can squirt diterpene solutions as their African, Asian and Australian relatives do. The identical three-dimensional chemical structure of all the nasute diterpenes worldwide supports the hypothesis of a common evolutionary origin. Thus it seems that what appeared to be an excellent example of parallel evolution must be written off the books. It also seems likely that further study of the defensive chemistry of the termites in general will yield still further surprises.



SNOUTED TERMITE SOLDIER of the genus *Armitermes* is a similar chemical defender. After piercing an attacker's cuticle with its hooked mandibles it uses its snout to apply droplets of a greasy secretion to the wounded area. The mixture is toxic when it enters the wound.



LONG UPPER LIP of a soldier in the genus *Rhinotermes* is used to daub the cuticle of an intruder with a toxic ketoaldehyde stored in its head and abdomen. The fat-soluble poison then penetrates the intruder's waxy cuticle and fatally interferes with metabolic processes.

Rational Collective Choice

Axiomatic analysis of voting systems has probed the compatibility of several desirable properties of an ideal method. Compromises among rationality, decisiveness and equality seem unavoidable

by Douglas H. Blair and Robert A. Pollak

Can a system of voting be devised that is at the same time rational, decisive and egalitarian? Studies of this question by philosophers, political scientists and economists (including the two of us) suggest that the answer is no. These characteristics of an ideal system are in fact incompatible. A method of voting may avoid arbitrariness, deadlock or inequality of power, but it cannot escape all three. The continuing analysis of this dilemma has led to a deeper understanding of existing voting systems and may lead in time to the discovery of better ones.

The axiomatic analysis of rational voting procedures was initiated some 33 years ago by the economist Kenneth J. Arrow of Stanford University. He advanced five intuitively appealing axioms that any procedure for combining or aggregating the preferences of individuals into collective judgments should satisfy, and he proved that the only procedures obeying all of them concentrate all power in the hands of a single individual. No nondictatorial method satisfying all Arrow's axioms can be found, not for want of ingenuity but because none exists. In part for this work Arrow shared the Nobel prize in economics in 1972.

Over the past 15 years investigators have reexamined Arrow's axioms in an effort to circumvent his "impossibility theorem" by relaxing his requirements. The problem has attracted widespread interest because it is closely linked with

central questions in philosophy, political science and economics. Philosophers face it, for example, in analyzing the practical implications of utilitarianism, the ethical doctrine holding that the rightness of actions depends on their consequences for people's happiness and hence requiring a method for aggregating the preferences of individuals. Political scientists encounter it in designing or evaluating rules of voting for committees or legislatures. Economists confront it in analyzing rationing and other nonmarket methods of allocating resources. This task is an important one in normative economics, because in determining the appropriate scope for intervention by the government in the operations of a free-market economy it is crucial to understand the potential performance of the alternatives to laissez faire.

Majority rule deserves first consideration among procedures for aggregating individuals' preferences; its virtues include simplicity, equality and the weight of tradition. Majority rule is essentially a procedure for ranking pairs of candidates or alternatives. When more than two alternatives must be ranked, however, majority rule encounters a difficulty the Marquis de Condorcet recognized nearly 200 years ago.

The difficulty pointed out by Condorcet is now known as the "paradox of voting." Suppose a committee consist-

ing of Tom, Dick and Harry must rank three candidates, x , y and z . Tom's preference ranking of the candidates is x , y , z . Dick's is y , z , x and Harry's is z , x , y . Majority voting between pairs of candidates yields a cycle: x defeats y , y defeats z and z defeats x , all by two votes to one. This voting cycle is the simplest example of Condorcet's paradox of voting.

Political scientists have identified many historical cases of voting cycles. For example, William H. Riker of the University of Rochester argues that the adoption of the 17th Amendment, providing for the direct election of U.S. senators, was delayed for 10 years by parliamentary maneuvers that depended on voting cycles involving the status quo (the appointment of senators by the state legislature) and two versions of the amendment.

When more than two alternatives are feasible, some new principle is needed for generating choices from pairwise rankings. The preference configurations that induce the paradox of voting create difficulties for each natural approach. The simplest method chooses an alternative that is undefeated by any other. In a paradox-of-voting situation, however, no such alternative exists, because each alternative or candidate loses to another.

A second method for proceeding from pairwise rankings to choices is to specify an agenda, listing the order in which pairs of alternatives will be taken up. For example, the agenda might call for an initial vote on x v. y , followed by a second stage in which the winner is matched against z . Under this agenda our three-member committee would first vote for x over y and at the second stage z would defeat x . It is easy to verify that under each of the three possible agendas in this situation the alternative taken up last emerges as the victor: the agenda determines the result. Voting cycles therefore present substantive difficulties as well as aesthetic ones. When a cycle occurs, the choice of an ultimate winner is at best arbitrary (if the agenda is selected randomly) and at worst deter-

COMMITTEE MEMBERS	PREFERENCE AMONG ALTERNATIVES	OUTCOME OF MAJORITY VOTING
TOM	x, y, z	
DICK	y, z, x	
HARRY	z, x, y	

PARADOX OF VOTING can arise under majority rule when voters with conflicting preferences must choose among more than two alternatives. The paradox is depicted for a three-member committee considering three candidates or alternatives, x , y and z . The outcome is cyclic: x defeats y , y defeats z and z defeats x , all by two votes to one. Such problems led Kenneth J. Arrow of Stanford University to propose five axioms that a voting procedure should satisfy.

mined by the machinations of the agenda setter.

Further opportunities for strategic maneuvering arise if a voter can alter the agenda by introducing new alternatives. Suppose (with the same committee) z represents the status quo and y is an alternative embodied in a motion that has been introduced. With only these two alternatives available y will defeat z , and Harry (who prefers z to y) will be disappointed. If he can introduce an amendment x to the motion y , however, x will defeat y on the initial vote, and at the second stage x will lose to z . Harry thus will obtain the enactment of his favored alternative.

Even if new alternatives cannot be introduced and the agenda cannot be manipulated, opportunities may still exist for voters to profit by misrepresenting their preferences. Consider again the agenda in which z is taken up last. If each member of the committee votes his true preference on each ballot, the winning alternative, z , is the least desirable one for Tom. Suppose, however, Tom votes for y instead of x on the initial ballot; then y prevails, going on to defeat z in the second stage. By this stratagem he has blocked the choice of the alternative he liked least.

Cyclic collective preferences present problems of arbitrary outcomes and strategic behavior. These difficulties arise whether the preferences are generated by majority rule, as in our example, or by some other voting procedure. Arrow was therefore led to ask: Do inconsistent collective preferences arise only under majority rule and closely related methods or are they inherent in all voting systems? To answer the question he might have assembled a list of voting procedures and for each procedure checked whether any configurations of individuals' rankings gave rise to cycles or to collective preferences with some other unacceptable feature. The difficulty is that he would have had to consider an immense number of aggregation procedures, differing widely in the roles they assign to particular voters and in the criteria they employ in ranking particular pairs of alternatives.

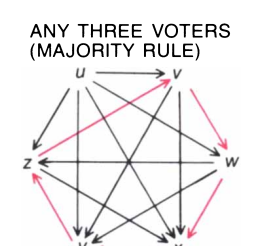
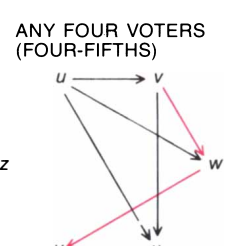
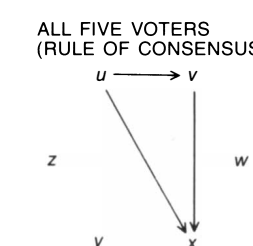
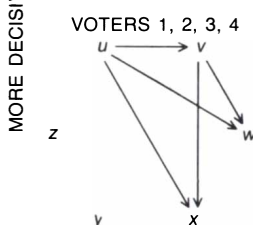
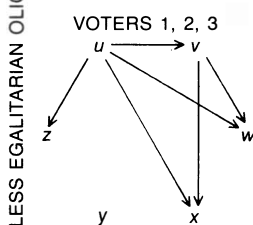
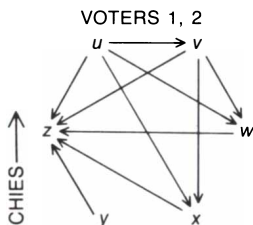
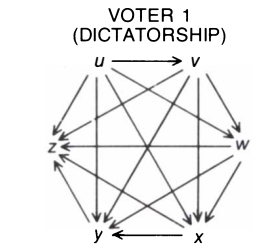
Of necessity Arrow chose an axiomatic approach instead. He formulated the problem as the choice of a constitution, that is, a rule assigning a collective ranking of the alternatives to each configuration of individuals' rankings. A constitution specifies whether each alternative stands as preferred, inferior or indifferent to every other one. (Two alternatives are indifferent if society regards them as being equally attractive.) Arrow narrowed the field of possible constitutions by imposing five requirements that (he argued) are necessary properties of any ethically acceptable method of aggregation. He then characterized the class

of constitutions that satisfy all five properties.

The first of Arrow's axioms, Universal Scope, requires that a constitution be capable of aggregating every possible configuration of voters' preferences. Since one cannot predict all the patterns of conflict that will arise over the life of a voting rule, Arrow argued, a society

should not adopt a constitution that will break down when certain configurations of voters' preferences arise. He contended that the society should instead insist on a constitution sufficiently general to resolve all possible controversies.

Arrow's second axiom, Unanimity, governs the operation of a constitution when there is no disagreement among



MORE DECISIVE, LESS EGALITARIAN OLIGARCHIES ↑

↑ MORE DECISIVE, LESS RATIONAL EGALITARIAN RULES →

PREFERENCES AMONG ALTERNATIVES

VOTER

- 1: u, v, w, x, y, z
- 2: y, u, v, x, w, z
- 3: u, z, v, x, w, y
- 4: z, u, v, w, x, y
- 5: w, y, z, u, v, x

EACH CONSTITUTION IS LABELED WITH A DESCRIPTION OF ITS DECISIVE SETS.

$x \rightarrow y$ MEANS x IS STRICTLY PREFERRED TO y UNDER THE CONSTITUTION.

ALTERNATIVES NOT CONNECTED BY ARROWS ARE INDIFFERENT.

COLLECTIVE-PREFERENCE RANKINGS corresponding to a particular set of preferences held by five voters are shown for seven constitutions, or voting rules. The rankings typify the tradeoffs faced by the designer of voting rules among rationality, inequality of power and decisiveness (the infrequency of collective indifference). Majority rule gives voters symmetric roles and is highly decisive, but for the preference rankings of these five voters a five-way cycle (color) and several shorter cycles arise. The rule of four-fifths is less decisive but yields no cycle here. It fails, however, to be P -transitive (consistent in ranking strict preferences) in that (to employ a shorthand notation) vPw (alternative v is strictly preferred to alternative w) and wPy but vIy (v is indifferent to y , meaning that society views them as being equally attractive). Oligarchies are always P -transitive but are less decisive as they get larger and more egalitarian.

voters. It specifies that for preference configurations in which every individual prefers x to y the collective ranking must put x above y . If one accepts the view that a society's ranking should reflect its members' preferences, it is difficult to quarrel with the Unanimity condition, which resolves what surely are the easiest problems of preference aggregation.

Arrow's third axiom, Pairwise Determination, requires that society's ranking of any pair of alternatives depend only on individuals' rankings of those two alternatives. No matter how the preferences of individuals for other alternatives may change, as long as each individual's ordering of x and y remains invariant the collective ranking of x and y does also. This condition implies, for example, that the collective ranking of Ronald Reagan and Jimmy Carter is independent of how individuals rank Edward Kennedy with respect to those two or to Walter Mondale.

A constitution satisfying Pairwise Determination limits the information about individuals' rankings required to determine the collective ranking of a pair of alternatives. In particular, information about the preferences of individ-

uals for unavailable options is irrelevant to the collective ranking of the available ones; this is an advantage when it is difficult or costly to elicit individuals' preference rankings. Without the condition of Pairwise Determination the constitution must specify what other alternatives are relevant to the determination of the collective ranking of x and y and how the preferences of individuals for those alternatives affect the collective ranking of x and y .

One common procedure, rank-order voting, violates Pairwise Determination. (It is the system normally employed by newspaper wire services to determine the ranking of college athletic teams.) When there are three alternatives, this constitution assigns each individual's first choice three points, his second two points and his third one point; the collective ranking is then found by summing the scores for each alternative and ranking them according to their total scores. In the three-member committee we have described each candidate receives a score of six, and so the committee is indifferent among the three candidates.

Suppose, however, Tom's preference ranking changes from x, y, z to x, z, y . Although no voter has changed his ranking of x and y , rank-order voting now yields a collective ranking of x over y , since x still receives a score of six but y now gets five. Hence under rank-order voting the collective ordering of x and y depends not only on how individuals rank them but also on the relative positions of other alternatives such as z .

Arrow's fourth and fifth axioms are best discussed using a shorthand notation. P denotes a strict collective preference (analogous to the relation "greater than" between a pair of real numbers), I denotes collective indifference (analogous to equality) and R denotes a weak collective-preference relation (analogous to "greater than or equal to"). Thus the expression xRy stands for "x is collectively at least as good as y," that is, either xPy or xIy .

Arrow's fourth axiom is Completeness: for every pair of alternatives x and y it must be true that xRy or yRx (or both, in which case x and y are indifferent). This axiom compels the aggregation procedure to rank every pair of alternatives. As long as the constitution has the option of declaring any pair of alternatives indifferent, Completeness seems to be a relatively innocuous requirement.

The fifth axiom, R -transitivity, requires that a weak collective preference be transitive: formally, if xRy and yRz , then xRz . Transitive relations involving pairs of real numbers include "greater than" ($>$), "equal to" ($=$) and "greater than or equal to" (\geq). Hence if a number x is greater than y and y is greater

than z , x must be greater than z . In economic analysis Completeness and R -transitivity are conventional assumptions, and individuals whose preferences obey these axioms are said to be "rational." By extension Arrow employed the term "collective rationality" to describe constitutions satisfying both Completeness and R -transitivity. He imposed R -transitivity to ensure that the chosen alternative would be independent of the agenda or path by which it is reached.

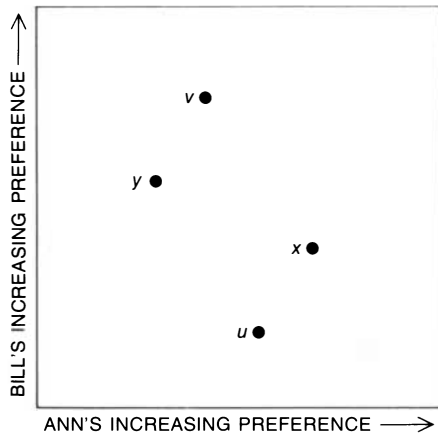
Having defined and defended this set of five desirable properties, Arrow proved that the only constitutions satisfying all of them share a simple and startling defect: each constitution is dictatorial. A dictator is a person with the power to impose on the society his strict preference over any pair of alternatives. Arrow stated his theorem in a slightly different way. He added a sixth axiom, nondictatorship, and proved that no constitution exists that obeys all six axioms. For this reason Arrow's result is often described as an "impossibility theorem."

Thus the designer of voting procedures for legislatures, committees and clubs who accepts these conditions as necessary properties of constitutions is simply out of luck. Arrow's apparently modest requirements have powerful and unpalatable implications. As his impossibility theorem demonstrates, the five axioms are highly restrictive; although they are attractive singly, they are pernicious in combination. Voting theorists have devoted much effort to reexamining the axioms, seeking a way around Arrow's unhappy conclusion.

A plausible argument can be made that Universal Scope is too ambitious a requirement. Not every logically possible configuration of preference rankings is equally likely. Since some configurations may be extremely unlikely, requiring a constitution to aggregate consistently every logically possible configuration into a collective ranking seems unnecessarily strong.

The commonest strategy in relaxing this requirement has been to focus on a particular procedure, usually majority rule, and to look for restrictions that rule out preference configurations implying intransitive collective preferences. For example, if only preference configurations in which there is no disagreement among individuals could arise, the Unanimity axiom would determine collective preferences and the problem of intransitivity could not arise. The best-known nontrivial restriction is single-peaked preferences, discovered in the 1940's by the British economist Duncan Black.

Single-peakedness arises when all individuals evaluate alternatives according to some single criterion and, in any pairwise choice, each individual votes



P-TRANSITIVE CONSTITUTIONS are voting rules in which if xPy and yPz , then xPz . All such constitutions satisfying Arrow's other axioms are neutral, that is, they rank pairs of alternatives by the same criteria. Neutrality means that if a particular configuration of voters' rankings of u and v implies that u is collectively preferred to v , the same configuration of rankings of x and y implies that x is collectively preferred to y . A proof with two voters is depicted. Suppose that for the alternatives u and v Ann prevails over Bill under the constitution when she prefers u to v . On x and y they differ. For this configuration xPu by Arrow's Unanimity axiom, uPv by the assumption that Ann prevails for u against v and vPy by Unanimity. Under P -transitivity it follows that xPy for this particular preference configuration. Arrow's Pairwise Determination condition, however, allows the more general conclusion that Ann prevails for x against y , regardless of the positions of u and v in her ranking or in Bill's. (Pairwise Determination states that the collective ranking of any pair of alternatives depends only on the preferences of voters regarding those two choices.)

for the alternative closer to his own most preferred position. For example, each voter might rank candidates according to how close they were to his own position on the political spectrum from liberal to conservative. Hence if x is more liberal than y and y is more liberal than z , a society with single-peaked preferences that contained liberals (x, y, z), conservatives (z, y, x), and moderates (y, x, z or y, z, x) could not contain individuals for whom the middle alternative is ranked below both extremes (x, z, y and z, x, y). If single-peakedness could be expected to hold in practice, the case for majority rule would be compelling. Usually, however, people rank alternatives by multiple criteria and so single-peakedness will fail.

More generally, the strategy of imposing restrictions on preference configurations can be fruitful only if the restrictions are plausible in terms of a theory of preference formation or preference structure. Social scientists, however, have not succeeded in formally modeling either the role of socialization in the development of tastes and values or the degree of similarity of preferences needed for social stability. Therefore, notwithstanding a great deal of effort, no characterization of possible patterns of rankings has yet been formulated that is broad enough to encompass voters' actual preferences and at the same time narrow enough to evade the dictatorship conclusion.

The possibility of abandoning the Unanimity axiom has generated little enthusiasm. On reconsideration, Unanimity still seems to be a mild requirement to impose on mechanisms for aggregating individuals' preferences into collective rankings. Furthermore, Robert Wilson of Stanford University has shown that the only additional constitutions satisfying Arrow's other axioms but violating Unanimity are even less appealing than dictatorships. In particular there are two new possibilities. The first possibility is universal indifference, the rule that makes every pair of alternatives perpetually indifferent regardless of how individuals rank them. The second is inverse dictatorship, a rule under which some particular individual's preference ranking is reversed to form the collective ranking. This method would serve an orderly society only if a voter with infallibly bad judgment could be found.

Pairwise Determination drew heavy fire in the first decade after Arrow published his work, but criticism of this axiom has subsided. Arrow's original defense of the condition was that constitutions satisfying it can be implemented without the burden of collecting large amounts of preference information. To rank the alternatives x and y it is never necessary to ascertain the position of z in individuals' rankings. Constitutions

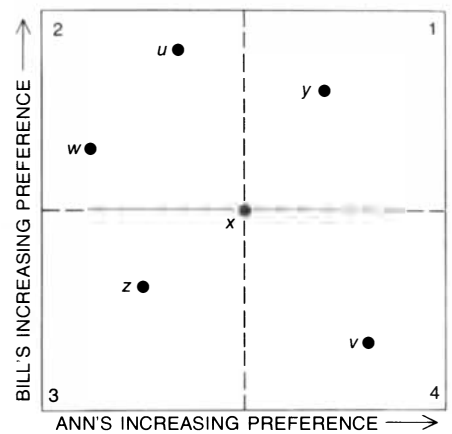
that violate this axiom are generally cumbersome, at least when there are many alternatives, because much preference information must be obtained to rank even a small number of feasible alternatives. Moreover, since the collective ranking of x and y under constitutions violating this axiom is sensitive to individuals' rankings of third alternatives, it is often possible for voters to manipulate the outcome on x and y by misrepresenting their preferences about other alternatives.

Arrow's least defensible requirement is probably R -transitivity. To avoid Arrow's conclusion theorists of voting have examined the consequences of several less restrictive conditions. It is not difficult to show that R -transitivity is equivalent to the conjunction of two weaker conditions, P -transitivity (the transitivity of the strict collective-preference relation P) and I -transitivity (the transitivity of collective indifference). Therefore a straightforward way to weaken Arrow's rationality requirement is to retain one of these conditions while abandoning the other.

I -transitivity is particularly vulnerable to criticism, since studies by psychologists have shown that individuals in experimental situations often exhibit intransitive indifference. For example, an individual who expresses indifference between x and y and between y and z will often prefer x to z . Thus the analogy between individual preference and collective preference yields little support for requiring collective rankings to be I -transitive.

The economist and philosopher Amartya K. Sen of the University of Oxford showed that abandoning I -transitivity while retaining P -transitivity provides an escape from Arrow's dictatorship result. He offered an example of a nondictatorial constitution satisfying Arrow's first four axioms and P -transitivity. His procedure, which might be called the rule of consensus, yields xPy if and only if every individual ranks x as being at least as good as y and at least one individual strictly prefers x to y . Therefore when two individuals disagree over x and y , the result is xIy . Each individual has a veto that enables him to block a strict collective preference opposite to his own. To see that collective indifference need not be transitive consider a committee with two members, one with a preference y, x, z and the other with a preference x, z, y ; the rule of consensus gives xIy and yIz but xPz .

The phenomenon of intransitive individual indifference may reflect the inability of individuals to distinguish among alternatives that are close together. For example, in judging political conservatism candidates x and y may seem equally conservative to an individual because their policy positions are in-



DICTATORSHIP THEOREM of Arrow is proved for two voters. One possible configuration of preference rankings for Ann and Bill is depicted; the quadrant in which each alternative appears depends on how the rank it with respect to the alternative x . Quadrant 1 contains alternatives (such as y) both prefer to x , 2 the ones Bill prefers to x but Ann does not, 3 the ones neither prefers to x and 4 those Ann prefers but Bill does not. The neutrality property of transitive constitutions implies that every alternative in a quadrant must be ranked collectively in the same way with respect to x . It also implies that the argument does not depend on the particular alternative defining the quadrants. Quadrant 2 cannot be indifferent to x for then both uIx and xIw . Thus by I -transitivity uIw , contradicting Unanimity, since both voters prefer u to w . Similarly, quadrant 4 cannot be indifferent. Quadrants 2 and 4 cannot both be preferred to or inferior to x , since Neutrality implies that if uPx , then xPz , because the voters' rankings of u and x and of x and z are the same. Since Arrow's Unanimity axiom implies that quadrant 1 is preferred to x and 3 is indifferent, Arrow's five axioms are consistent with only two constitutions: if 4 is preferred to x and 2 is inferior, Ann is a dictator, whereas if 2 is preferred and 4 is inferior, Bill is a dictator. The proof is due to Charles Blackorby, David Donaldson and John Weymark.

distinguishably close, and the same may seem true of y and z . Yet candidates x and z may be far enough apart for the individual to perceive that x is more conservative than z and hence to prefer one to the other.

The psychologist R. Duncan Luce of Harvard University has proposed a notion of consistency called the semiorder to model situations entailing such thresholds of perception. Semiordered preferences exhibit P -transitivity but allow intransitive indifference. If collective choice is seen as a process of aggregating the policy judgments of individuals to form collective policy judgments, the notion of imperfect discrimination can be applied to collective preferences as well as to individual ones.

The semiorder, however, is a stronger rationality requirement than P -transitivity alone, a fact that has important consequences for preference-aggregation procedures. As we have demon-

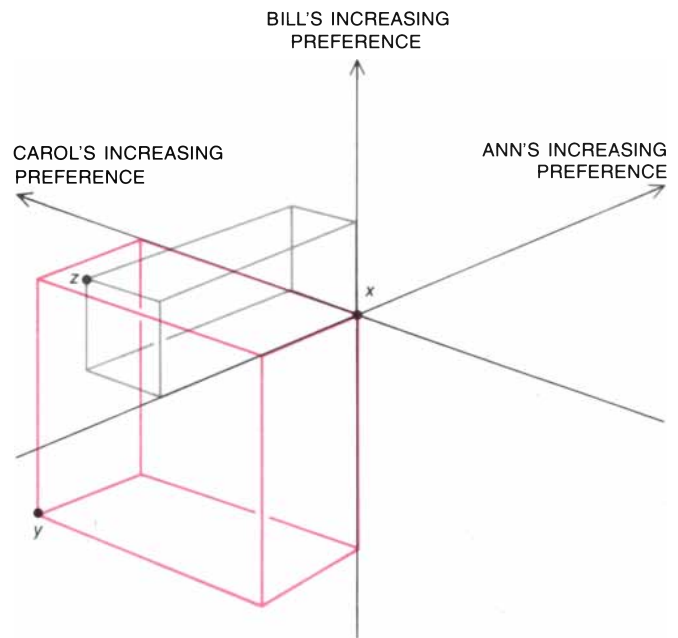
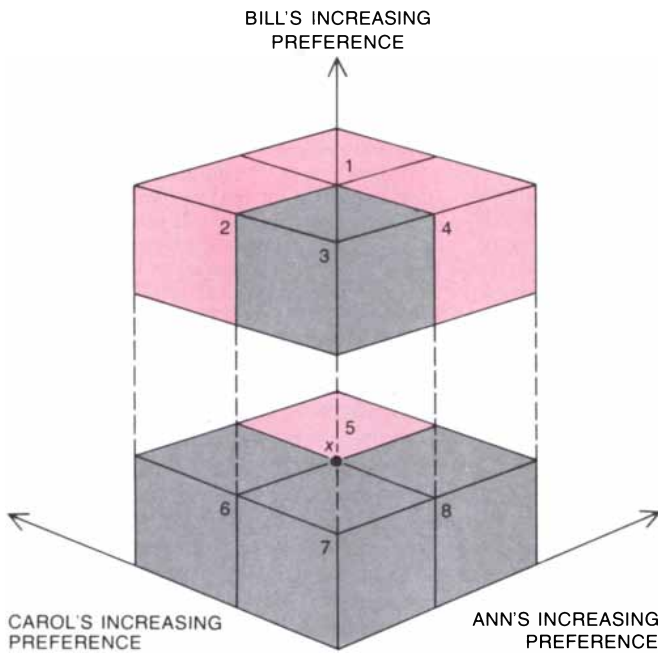
strated elsewhere, requiring that a constitution yield semiordeered collective preferences and satisfy Arrow's other axioms still implies dictatorship. Thus the perception-threshold justification for weakening *R*-transitivity, although it is appealing and plausible, does not avoid Arrow's dismal conclusion.

Arrow's principal justification for *R*-transitivity was that the choice from some set of alternatives should be independent of the agenda or path by

which the choice is made. Remarkably, the desire for path independence leads directly to an argument for *P*-transitivity alone as the appropriate rationality requirement for collective-preference rankings. Charles R. Plott of the California Institute of Technology has proposed a formal definition of path independence and has shown that all constitutions satisfying this condition are *P*-transitive. Furthermore, any *P*-transitive constitution satisfies path independence. Therefore, although Arrow's

original collective rationality condition guarantees path independence, a less restrictive condition would accomplish the same objective.

Sen's rule of consensus shows that nondictatorial *P*-transitive constitutions exist. As Sen recognized, however, the rule of consensus is not an appealing solution to Arrow's problem because it is so often indecisive. Whenever any two individuals have opposing strict preference rankings of a pair of alternatives—surely a ubiquitous form of conflicting



NEUTRAL CONSTITUTIONS for three voters can be represented with eight-octant diagrams of voters' preferences analogous to the four-quadrant diagram for two voters. By neutrality, which is assumed for acyclic, or noncycling, constitutions and is a necessary property of *P*-transitive ones, all the alternatives in an octant must bear the same relation (preferred, indifferent or inferior) to *x* in the collective ranking. If one octant, such as 2, is collectively preferred to *x*, its opposite (8) must be inferior, and if an octant is collectively indifferent to *x*, its opposite is also indifferent. Since the Unanimity axiom compels a constitution to prefer octant 1 to *x* and *x* to octant 7, all possible neutral constitutions are fully described by specifying the collective preferences (with respect to *x*) for octants 2, 3 and 4. Since each of the three octants can be preferred, indifferent or inferior to *x*, there are 27 possible three-person neutral constitutions. Under *P*-transitive constitutions satisfying Arrow's other axioms it can be shown that additional support for *x* with respect to *y* cannot worsen *x*'s position with respect to *y* in the collective ranking. For example, suppose Ann changes her ranking so that *z* moves from octant 3 to octant 4, that is, instead of preferring *x* to *z* she now prefers *z* to *x*. Then the collective ranking of 4 with respect to *x* must be at least as favorable as 3; if 3 is indifferent, 4 must be either indifferent or preferred. Some constitutions that are not *P*-transitive fail to satisfy this property of positive association of the collective ranking with individuals' rankings. Since such a constitution is unattractive, positive association can be made an additional requirement for constitutions that violate *P*-transitivity. Because the movement of an alternative from 3 to 2 or 4 reflects an improvement in some voter's evaluation of the alternative with respect to *x*, positive association requires that if 3 is preferred to *x*, 2 and 4 are too. If 3 is indifferent, 2 and 4 must be either preferred or indifferent. If 2 is inferior, 8 must be preferred; hence 4 must be preferred by positive association. Such arguments reduce the number of three-person constitutions satisfying neutrality, positive association and Arrow's first four axioms from 27 to 11. Majority rule is the one depicted here, with preferred octants in color and inferior ones in gray. No octant is indifferent under a majority-rule constitution.

***P*-TRANSITIVE AND ACYCLIC CONSTITUTIONS** can be characterized by means of the configuration of individuals' preferences shown here. There are 11 three-voter constitutions that satisfy neutrality, positive association and Arrow's first four axioms; they are represented in the illustration at the left and the one on the opposite page. Every acyclic constitution must have an individual with veto power. In other words, no octant can be collectively preferred to *x* if the vetoer prefers *x* to that octant. Each of the 11 constitutions except majority rule has at least one vetoer. Showing that majority rule can cycle proves the result. Under majority rule octants 1, 2, 4 and 5 are collectively preferred to *x*, whereas 3, 6, 7 and 8 are inferior. For the configuration of individuals' preferences indicated here majority rule cycles: xPy (because *y* is in octant 6), yPz (since *y* bears the same relation to *z* as octant 5 does to *x*) and zPx (because *z* is in octant 2). Thus acyclicity entails a vetoer. Allan Gibbard of the University of Michigan proved that *P*-transitive constitutions satisfying Arrow's first four axioms are oligarchic, meaning they empower a unique group of voters to impose their unanimous strict preferences on the society and grant those voters the right as individuals to veto strict collective preference counter to their own. All the lower seven constitutions represented in the illustration on the opposite page are oligarchic. To prove Gibbard's result with three voters, then, it is necessary to show only that the top three constitutions can violate *P*-transitivity. Since they differ from one another only in the naming of the voters, the argument is given only for the second of them. Because *z* is in preferred octant 2, zPx . Alternative *y* is in octant 6, opposite the preferred octant 4, so that xPy . *P*-transitivity would require zPy . Yet *z* bears the same relation to *y* as octant 3 (now indifferent) does to *x*, contradicting the neutrality property of *P*-transitive constitutions and proving the result. The requirement of *P*-transitivity therefore entails an oligarchy. Strengthening that requirement to *R*-transitivity, which calls for weak collective preference to be transitive, eliminates oligarchies with more than a single member, leaving only the three dictatorships. The boxlike shapes in this diagram serve to locate *y* and *z* in the appropriate octants in the illustration at the left.

interests—the rule of consensus declares the two alternatives to be indifferent. A constitution that yields collective indifference whenever individuals' rankings conflict is virtually useless.

Can more attractive P -transitive constitutions be discovered? The philosopher Allan Gibbard of the University of Michigan has shown that they cannot. Gibbard proved that under every P -transitive procedure obeying Arrow's remaining axioms there exists a privileged set of individuals he called an oligarchy. This oligarchy as a group has the power to impose on the entire society its unanimous strict preference over any pair of alternatives. Moreover, each member of the oligarchy as an individual has the power to veto strict collective preference opposite to his own: whenever any oligarch strictly prefers x to y , yPx is impossible. Thus a dictator is an oligarchy of one, whereas the rule of consensus implies an oligarchy consisting of the entire society.

Not all oligarchic constitutions distribute power unequally, as the rule of consensus demonstrates. As more members are installed in the oligarchy the distribution of power becomes more nearly equal. Yet a large oligarchy increases the probability of indecisiveness, since conflicts between oligarchs imply collective indifference. Because path independence requires P -transitivity, Gibbard's theorem implies that the attractive property of path independence can only be purchased at the cost of indecisiveness or inequality. To escape the dilemma of choosing between the inequality of small oligarchies and the indecisiveness of large ones, collective rationality must be weakened beyond P -transitivity.

P -transitive constitutions satisfying Arrow's other axioms have an additional drawback. They are inflexible in the sense that, except when all members of the oligarchy are indifferent, they cannot impose different requirements for strict collective preference on different pairs of alternatives. Instead they must treat pairs of alternatives in a neutral or symmetric fashion. Hence if Tom but neither Dick nor Harry has a stake in a change from policy y to policy x , it may be appropriate to give Tom the right to veto the ranking xPy . Yet it may be undesirable to give him such power over other pairs. P -transitive constitutions satisfying Arrow's other axioms, however, cannot exhibit this flexibility.

Under the oligarchic P -transitive constitutions satisfying Arrow's other axioms a group (or an individual) with the power to impose its will over one pair of alternatives must have the same power over all pairs. Similarly, an individual or a group with veto power over one pair of alternatives must also have that power over every pair. If a particular configuration of preferences between x and y

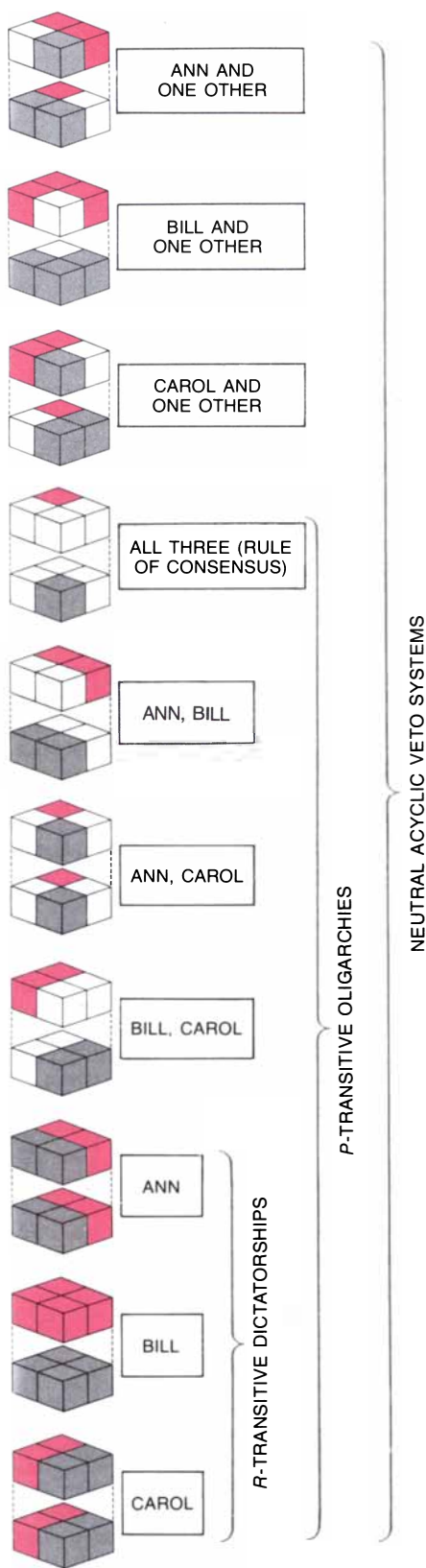
implies that x is collectively preferred to y (Pairwise Determination guarantees that no additional information about individuals' preferences figures in determining the collective ranking of x and y), the same configuration of individuals' preferences between z and w implies that z is collectively preferred to w . It is possible to treat pairs of alternatives in asymmetric or non-neutral ways under constitutions satisfying weaker rationality conditions than P -transitivity.

A less restrictive requirement than P -transitivity is acyclicity, the absence of cycles of strict collective preference regardless of their length. For example, with three alternatives the collective preference ranking xPy, yPz, xIz is acyclic (since it has no cycles of strict preference), but it violates P -transitivity (since the first two rankings would require xPz). Among the collective preferences ruled out by acyclicity is the three-alternative cycle encountered in the paradox of voting.

Acyclicity is an attractive property to demand of a constitution, particularly when the procedure adopted for converting pairwise collective rankings into choices is to select the alternatives that are undefeated by all others. Acyclicity guarantees that at least one such alternative always exists in every finite feasible set, and it is the least restrictive condition that does so. In the acyclic example above, for instance, alternative x is not defeated by either of its competitors, whereas with the cyclic collective preference xPy, yPz and zPx each alternative loses to another one. Without acyclicity the appropriate collective choice is far from clear.

Acyclic constitutions satisfying Arrow's other axioms need not be neutral; they allow pairs of alternatives to be treated in an asymmetric fashion. Non-neutral procedures are quite common. In the U.S. Senate, to take one (sometimes cyclic) example, an ordinary bill passes with a simple majority, but a motion to limit debate requires three-fifths and a proposed constitutional amendment requires two-thirds. Whether or not neutrality is a desirable feature of a constitution depends on the nature of the alternatives. When candidates for office are being ranked, a neutral rule is more appealing than one favoring a particular candidate. When alternatives have asymmetric consequences, however, non-neutral rules biased against more drastic outcomes may be advantageous. The criminal-justice system provides examples at both ends of the spectrum: jury trials are not guaranteed in minor cases, and death sentences are automatically reviewed by appellate courts.

Relaxing the rationality requirement from P -transitivity to acyclicity makes possible many constitutions without ol-



THREE-VOTER neutral constitutions that satisfy various rationality conditions are portrayed. Each eight-octant figure illustrates the ranking of alternatives in relation to a fixed alternative x under a particular constitution. Alternatives in colored octants are preferred to x , those in gray octants are inferior to x and those in white octants are indifferent to x . The numbers in the boxes indicate which voter or group of voters is decisive in each situation.

igarchies. All the new constitutions, however, share one of the principal drawbacks of oligarchic rules. All acyclic constitutions satisfying Arrow's other axioms grant some group or individual extensive veto power.

Typical of these additional acyclic constitutions was the voting rule in the United Nations Security Council until 1965. The Security Council then consisted of five permanent members and six nonpermanent ones. A motion succeeded if it received at least seven affirmative votes and no negative vote from any permanent member. Thus each permanent member of the Security Council had a veto; no motion could pass if any permanent member of the council opposed it. These five nations did not constitute an oligarchy, howev-

er, since additional support from non-permanent members was required for strict collective preference.

Under acyclic constitutions an individual or a group may have veto power over some pairs of alternatives but not over others. The extent of veto power under such rules must therefore be described in terms of the number of pairs over which some individual or group exercises a veto. As we have recently shown, when the number of alternatives is large with respect to the number of individuals, at least one individual must be able to veto a large number of pairs of alternatives. More precisely, as the ratio of alternatives to voters increases without limit, the proportion of pairs of alternatives over which some

particular individual must have a veto approaches unity.

Even when there are fewer alternatives than there are individuals, their ratio is critical. *R*-transitivity and *P*-transitivity are rationality conditions applying to triples of alternatives. Acyclicity, in contrast, rules out cycles of every possible length: the absence of cycles with three alternatives does not imply the absence of cycles with four. As the number of alternatives increases, longer cycles become possible.

With only two alternatives, *x* and *y*, majority voting cannot yield a cycle. A cycle would require both that more than half of the voters prefer *x* to *y* and that more than half of the voters prefer *y* to *x*; consequently at least one individual would prefer *x* to *y* and *y* to *x*, which is clearly impossible. When there are three or more alternatives, as the paradox of voting shows, no individual need agree with all the links in the collective-preference cycle, and so the cycle does not contradict the transitivity of any individual's preference ranking.

A five-member committee further illustrates the critical role of the ratio of alternatives to individuals. Consider the constitution requiring four affirmative votes for strict preference—a constitution intermediate between majority rule and the rule of consensus. Could the cycle xPy, yPz, zPx arise under the four-fifths-majority constitution? It could not, since each individual has transitive preferences and at least four voters must agree with each link in the collective ranking. With a five-member committee at least one member would have to agree with each of the three links in the collective-preference cycle, which is impossible.

With a five-member committee and three alternatives four-fifths-majority rule has advantages over both majority rule and the rule of consensus. Unlike majority rule, no cycle can occur. Unlike the rule of consensus, no member has a veto. These advantages, however, are won at a price. The four-fifths rule is less rational than the rule of consensus: although collective preferences are acyclic, they are not *P*-transitive and so are not path-independent. The rule is less decisive than bare majority rule: although no individual has a veto, any group of two voters can block a collective ranking of two alternatives that stands opposite to their own.

Under an acyclic constitution the size of the smallest group with veto power depends on the relative numbers of alternatives and individuals. When there are only two alternatives, as the case of majority rule makes clear, the smallest veto group needed is half of the electorate. When there are as many alternatives as there are individuals, under any acyclic rule there must exist at least one individual who, standing alone, has veto

SECOND ALTERNATIVE IN PAIR →

	x_1	x_2	x_3	...	x_{n-1}	x_n	x_{n+1}	...	x_m
x_1		$n-1$	$n-1$...	$n-1$	$n-1$	$n-1$...	$n-1$
x_2	$n-1$		$n-1$...	$n-1$	$n-1$	$n-1$...	$n-1$
x_3	$n-1$	$n-1$...	$n-1$	$n-1$	$n-1$...	$n-1$
⋮									⋮
x_{n-1}	$n-1$	$n-1$	$n-1$...		$n-1$	$n-1$...	$n-1$
x_n	n	n	n	...	n		n	...	n
x_{n+1}	n	n	n	...	n	n		...	n
⋮									⋮
x_m	n	n	n	...	n	n	n	...	

← FIRST ALTERNATIVE IN PAIR

EXTENSIVE VETO POWER is characteristic of all acyclic constitutions that satisfy Arrow's first four axioms. The constitution represented here, which is defined for n voters and m alternatives, can be described in terms of the minimum number of voters whose agreement is necessary for strict collective preference between elements of an ordered pair of distinct alternatives. This number appears in the cell corresponding to the ordered pair. Unanimous consent is necessary for strict collective preference for any of the $(m-n+1)(m-1)$ pairs in the colored region of the table; hence each voter has a veto over each of these pairs under this rule. To see that this constitution cannot yield cyclic collective preferences suppose, on the contrary, it can. At most one voter disagrees with each of the links in the collective-preference cycle that lies in the black region of the table. Thus even if a different voter disagrees with each black link, some voter must agree with all of them, since at most $n-1$ pairs in the cycle lie in the region. The same individual must also agree with each colored link in the cycle. Since he agrees with every link in the cycle, his preferences are cyclic. The contradiction proves that the constitution cannot in fact cycle. When m exceeds n , this rule gives each voter veto power over $(m-n+1)(m-1)$ pairs of alternatives. When the ratio of m to n is large, the rule is unattractive because it is highly indecisive. Yet as the authors have recently demonstrated, every acyclic rule satisfying Arrow's first four axioms gives someone veto power over at least this number of pairs when m exceeds n . Even though the rule is unattractive, a better one cannot be found in the circumstances.

power over some pairs of alternatives. In intermediate cases, as the constitution stipulating a majority of four-fifths illustrates, some minority groups have veto power. Constitutions that grant veto power to a large number of small groups are likely to lead to deadlock rather than to decision.

An uncompromising egalitarian would argue that since some individuals must have veto power, all individuals should have it. The larger the set of vetoers, however, the more frequent the incidence of collective indifference, since indifference occurs whenever two individuals with veto power in appropriate directions rank x and y in opposite ways. The palatability of acyclic constitutions thus depends on the ratio of alternatives to individuals. When the number of alternatives is only slightly smaller than the number of individuals, "small" groups must be endowed with extensive veto power, although there need not be an individual with veto power. As the number of alternatives increases, the size of the smallest vetoing group required for acyclicity grows smaller. With as many alternatives as there are individuals, at least one individual must be able to veto some pairs. As the number of alternatives increases further, the proportion of pairs over which the individual has veto power approaches unity.

The "impossibility theorems" that began with Arrow's famous proposition define constraints on a society's choice of a rule for collective decision making. The constraints are severe. Three widely shared objectives—collective rationality, decisiveness and equality of power—stand in irreconcilable conflict. If society forgoes collective rationality, thereby accepting the necessary arbitrariness and manipulability of irrational procedures, majority rule is likely to be the choice because it attains the remaining goals. If society insists on retaining a degree of collective rationality, it can achieve equality by adopting the rule of consensus, but only at the price of extreme indecisiveness. Society can increase decisiveness by concentrating veto power in progressively fewer hands; the most decisive rule, dictatorship, is also the least egalitarian.

There is little comfort here for those designing ideal procedures for collective choice. Nevertheless, every society must make collective choices and devise voting procedures, however imperfect they may be. Axiomatic analysis, pursuing the line of investigation begun by Arrow 33 years ago, has yielded a richer understanding of existing voting methods and may eventually yield better ones. It also demonstrates that the opportunities for improvement are severely limited. Stark compromises are inevitable.

BIG BOOM FOR REAL ESTATE INVESTORS.

Thousands of people like you find new opportunities for income, appreciation and tax advantages through the CENTURY 21® network.

Tax laws have changed. And now real estate is an even better investment than ever before.

At CENTURY 21 offices all over the country, we're showing people how to take advantage of the latest tax laws. How they can use leverage. How they can now shelter more of their income. Even income from salary, stock dividends and other sources.

A CENTURY 21 Investment Professional can show you, too. Through real estate, you may be able to get a greater annual cash flow, after taxes, than through other traditional types of investment. And you don't have to invest a lot of time and effort. Because a CENTURY 21 office can also handle the details of property management for you.

The following chart illustrates the potential advantages of real estate over other investments:

	REAL ESTATE	MONEY MARKETS	STOCKS
Cash Flow	✓	✓	✓
Depreciation	✓		
Loan Amortization	✓		
Appreciation	✓		✓

The opportunities are impressive. And that's one of the reasons investment business is booming at CENTURY 21 offices. Last year, we helped investors buy or sell real estate worth over \$4 billion. We're North America's leading real estate sales organization. And now, we're leading the way in real estate investment, too.



**Call this toll-free number
for a CENTURY 21 Investment Professional.
1-800-228-3399**

Nebraska Residents please call 1-800-642-8788. In Alaska and Hawaii call 1-800-862-1100.

© 1983 Century 21 Real Estate Corporation, as trustee for the NAF ® and TM-trademarks of Century 21 Real Estate Corporation Printed in U.S.A. Equal Housing Opportunity

EACH OFFICE IS INDEPENDENTLY OWNED AND OPERATED.

The Stave Churches of Norway

In the 10th century Norwegian builders blended pagan and Christian elements in wood churches. Some are still standing, demonstrating that with sound design and maintenance wood buildings can be permanent

by Petter Aune, Ronald L. Sack and Arne Selberg

There is ample evidence that in antiquity large structures were built not only of stone, brick and mortar but also of wood. Yet virtually all the large structures that have come down to us from antiquity and even from medieval times were built of stone. The reason is almost self-evident: stone, brick and mortar are far from indestructible but under most conditions they are more durable than wood, which is vulnerable to decay, pests, fire and neglect. An exception is the stave church of Norway, a type of wood structure that dates back to the 10th century. The exception is a meaningful one.

The staves of the stave churches are stout wood columns. "Stave church" is translated from the Norwegian *stavkirke*, and *stav* is related both to the stave of a barrel and to a wood staff. Stave churches were built in large numbers in Norway from the 11th century to the 14th. In the construction of the churches pagan and Christian elements were combined. The result is a design of primitive grandeur, with an exotic exterior and a serene interior that is distinctly human in proportion. The most intriguing aspect of the stave church, however, is the durability of its structure. How were these wood buildings made so that some of them could have survived for more than 800 years?

We are three structural engineers who have surveyed more than half of the stave churches still in existence. We have found that there are three reasons the buildings have been able to last so long. First, extraordinary care was taken in the preparation of the wood; the unique seasoning methods of the early Norwegian builders yielded materials that are superior to modern lumber. Second, the overall design is an excellent one, well suited to the exposed sites on which many of the churches were built. The structure is capable of sustaining the vertical forces generated not only by the mass of the church but also by heavy loads of snow. In addition it provides the combination of strength and stiffness that is needed to withstand the lateral

forces exerted by high winds. Third, small but significant structural innovations have served to protect the wood from deterioration. The lesson is that with appropriate design and with special care taken in preparing the building material large wood structures can be permanent. Indeed, the history of the stave church suggests that the frequent assumption that wood is suitable mainly for small buildings or for buildings not expected to have a long life might well be reexamined.

The stave church incorporates two main types of structural assemblies made out of wood: truss structures and beam structures. In an ideal truss structure each member is subject either to tensile stress or to compressive stress but not simultaneously to both. Furthermore, there are no other types of stress on the members. A simple example of a truss structure is a triangular assembly with the apex of the triangle at the top; such a structure might be found in a roof. A downward vertical force applied at the peak puts the two diagonal members into compression and the bottom member into tension. Such a downward force could be supplied by the weight of the roof or by a load of snow.

In a beam structure, on the other hand, tensile and compressive stresses can both be present in a particular cross section of one member. There can also be other kinds of stress in the structure, such as those exerted by shear or bending. A simple example of a beam structure is an assembly with a horizontal beam attached to the upper end of two vertical columns. Such an assem-

bly could form part of the frame for a house. The weight of the top beam and the loads it supports provide a downward force that compresses the vertical members. The vertical members, however, tend to curve outward; as a result they are subject not only to compressive forces but also to bending stresses.

Beam structures go back to prehistory. The design can be executed simply by laying a horizontal beam on two supports. The truss design, however, also has a long history. Roman architects are remembered primarily for their work in brick, mortar and marble, but they also did superb work in wood. Trajan's Column in the Imperial Forum of Rome was erected in A.D. 114 to commemorate the emperor's conquest of the Dacians, who lived in what is now Romania. Among the figures on the column is a group showing that in A.D. 104 Apollodorus of Damascus, the imperial architect, built a wood bridge across the Danube; the design of the bridge incorporated trusses. The way truss elements and beam elements are combined in the stave church does much to explain why the churches are so sturdy.

Wood continued to be employed as a building material in Europe after the collapse of the Roman empire. The fabled ships built by the Vikings demonstrate that techniques of construction in wood were highly developed in the Nordic countries by the early Middle Ages. Moreover, the Viking ships kept the Nordic peoples in contact with the centers of culture and hence made available to them the building-construction technology that had evolved elsewhere in Europe.

FANTOFT STAVE CHURCH is a well-preserved example of the early stave churches. Built in about 1200 at Fortun, a village at the head of Sogne Fjord, it was dismantled and reconstructed in 1883 at Fantoft near Bergen. The stave church is named for the wood columns that are its primary supports. In the exterior of the church the builders combined pagan and Christian motifs. The dragon heads on the gables, similar to those on the prow of a Viking ship, overshadow the crosses below. The building consists of a rectangular nave (in the foreground) and a smaller rectangular chancel (at the rear). The Fantoft church has a basilican design: the central nave is raised above the aisles on each side. Of the three large roofs, the highest covers the nave. The middle roof covers the aisles. The lowest roof covers a walkway called the ambulatory.



The influence of such technology in Norway was strengthened by the introduction of Christianity. The Christian religion was slower to reach Norway than it was to reach most of the rest of Europe. The first king to attempt to bring Christianity to Norway was Haakon the Good, who ruled from A.D. 935 to 961. Haakon had three churches built in the Møre district. The churches were soon burned and the priests whom Haakon had imported from England were killed. The missionary efforts of several later kings were also fiercely re-

sisted. It was not until the reign of Olav Kyrre, who ruled from 1066 to 1093, that a lasting official Norwegian Church was established. Olav set up episcopal sees at Trondheim, Bergen and Oslo, and he probably also instituted a large program of church building.

Thus in the 10th and 11th centuries the number of churches in Norway grew, albeit not without setbacks. The early missionary churches were simple buildings consisting of a large rectangular nave with a small chancel attached to

its east end. The nave provided room for the congregation, whose members generally stood during the service; the chancel was for the altar and clergy.

Some of the missionary churches of Norway have been excavated in modern times. The remains of structural members found buried at the sites suggest that many of the nave-and-chancel churches were built with the ancient method of palisade construction. This method called for hewing tree trunks in half lengthwise and sinking the timbers into the ground next to each other to form the walls. In the refinement known as posthole construction the corner columns were sunk into the ground and horizontal foundation beams were laid between them resting on the ground. Wall planking was erected on the foundation beams.

Posthole construction was an advance over the palisade method, but the foundation members, being in contact with the ground, remained susceptible to rotting. That is why few specimens of posthole buildings survive, even though archaeological excavations show they were common throughout Europe. A further innovation extended the life of the foundation beams and made larger structures possible, namely the building of a complete foundation above ground level. On a layer of flat stones was set a rectangular frame of heavy members called sill beams; the main vertical columns were put at the corners of the frame where the sill beams intersected.

Vertical wall planks were inserted into a groove in the sill beam. In some instances the narrow space between planks was closed with a tongue-and-groove joint; in others it was covered with a thin wood spline. Although it is the larger corner posts that are called staves, the vertical wall planking does call to mind the staves of a barrel. The stave style, with its upright wall planks, stands in contrast to the commoner method of wood construction in Norway, which is called *laft* construction. In *laft* construction the wall members are laid horizontally and are held together at the corners by interlocking notches.

After the first stave churches were built in the middle of the 10th century, the design was intensively refined over a period of several centuries. The church at Urnes on Luster Fjord, a branch of Sogne Fjord, provides a convenient illustration of the process of development. Built in about 1150, it is thought to be the oldest existing stave church in Norway. Findings made in an excavation under the church floor in 1956 suggest, however, that two earlier churches had been built on the same site.

The first Urnes church was probably put up in the time of the transition to Christianity. It was built in the palisade style: all the structural members were embedded in the ground. The sec-



GOL STAVE CHURCH was built between 1200 and 1250. In 1884 it was moved to the Norwegian Folk Museum at Bygdø near Oslo. The view in the photograph is from the nave into the chancel, where the altar stands. The Gol church also has a basilican design. The central nave is defined by the interior staves: 14 wood columns eight meters high. Six of the staves have been terminated above the floor. The curved brackets under the shortened staves allow the loads to be transferred to the full-length corner staves. The corner staves transmit the loads to the substructure. The stave assembly is given rigidity by sets of curved braces called *arcading* (visible above the head of the crucifix) and by the intersecting diagonal struts called *St. Andrew's crosses* (visible behind the body of the crucifix). The staves visible in the rear support the chancel.

ond church was put up in the latter half of the 11th century. It was built by the posthole method. Both of the early churches were small structures of a nave-and-chancel design.

In the middle of the 12th century the second Urnes church was torn down. Some of its parts were utilized for the building of the third church, the one that is still standing. Thus three churches were built at Urnes in less than a century and a half. The design of the third church was a dramatic departure from that of its predecessors. The central section of the nave is raised above the rest of the building on a framework consisting of 16 large staves, which define the nave and the surrounding aisles. The roof of the central compartment is about two meters higher than that of the aisles. The striking design of the third Urnes church spread throughout Norway and was followed in the construction of many later stave churches.

A significant influence on the new design appears to have been the cathedrals in the basilica style that were going up elsewhere in Europe at the time. Most of the basilican cathedrals were built of stone, but it is possible the Norwegian builders saw ways of adapting the design to wood. Most basilican cathedrals have at one end a large semi-circular apse, a structure adjacent to the chancel. The large rectangular nave is separated from the two side aisles by two parallel rows of columns. The builders of the stave churches had examples of the basilican design close at hand: stone cathedrals of this type were built at Oslo, Hamar, Stavanger, Bergen and Trondheim. Because of its similarity to the basilican cathedral, we shall refer to the kind of stave church that first appeared at Urnes as the basilican stave church.

Although the basilican cathedrals could well have been the decisive influence on the builders of the church at Urnes, other models were culturally even closer at hand. Kenneth J. Conant of Harvard University notes that excavations at Gamla Uppsala in Sweden have revealed the remains of a pagan temple with a plan much like that of the basilican stave church. According to Conant, in the temple "there was a square central compartment with corner timbers more than two-thirds of a meter in diameter and a smaller post between [each pair of larger timbers]. The outer wall was supported by light posts, and [was] relatively low, while the large corner timbers indicate a towerlike proportion for the central square." The Gamla Uppsala temple was built in the middle of the 11th century.

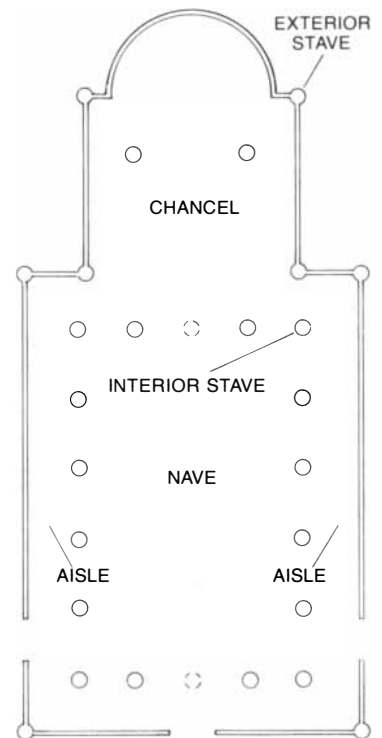
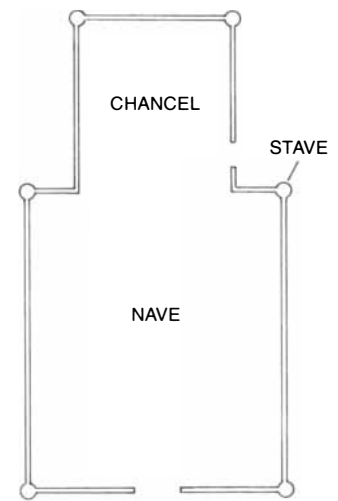
Moreover, excavations at Lund in Sweden have uncovered the remains of churches dating from about the same time. These early churches were con-

structed on the palisade design. Some of them also had freestanding interior staves, as the basilican stave churches do. Such old missionary churches are generally regarded as the forerunners of the stave churches.

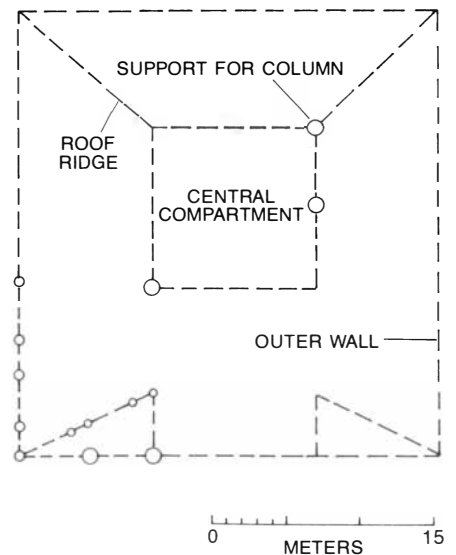
Hence in converting from the simple nave-and-chancel church to the grander basilican design the builders in the Norwegian villages probably drew on both Christian and pagan sources. The conclusion is reinforced by the decorations with pre-Christian motifs that appear on the outside of stave churches. In many of the structures the exterior is adorned with elaborate carvings of dragons and other zoomorphic figures in an exuberant tangle. The designs have been traced to several possible sources, including the carved runic stones of Scandinavia and certain kinds of manuscript illumination then being done in Ireland. Whatever the source of the carvings, they have a distinctly pagan spirit. Only after 1100 did Christian symbols such as plant tendrils begin to appear among the dragons. In addition to the zoomorphic carvings the stave churches have on their gabled roof dragon heads much like those on the prow of a Viking ship. They are often larger and more dramatic than the crosses on the roof.

The stave design was clearly popular with the Norwegian villagers. It has been estimated that in A.D. 1300 there were about 1,000 stave churches in the country. Many of them were built in the area north of Bergen along the tributaries of the Sogne Fjord, but there were also stave churches from the south coast up to Nidaros (the modern Trondheim). Moreover, although there are no surviving stave churches in Sweden and Denmark, recent excavations suggest that such churches were also built there.

As is customary with Christian ecclesiastical buildings, most stave churches are built with the long axis of the nave laid out from east to west and the chan-



FLOOR PLANS show the development of the basilican stave church, the first of which was built in about 1100. Before the 12th century stave churches had a simple nave-and-chancel design (top). Large horizontal sill beams laid on a foundation of flat stones made up the footings. At the corners were six large staves. The vertical wall planks between the corner staves were inserted in the sill beam at their base. In the basilican churches the interior staves defining the raised central compartment are laid on four large raft beams that form a rectangle (middle). One influence on the basilican church was the design of pagan temples such as the one at Gamla Uppsala in Sweden (bottom). The plan of the temple has been reconstructed from the few column supports that were excavated intact (circles). The temple had low outer walls and a towerlike central compartment. The central compartment was supported at the corners by columns more than two-thirds of a meter in diameter.



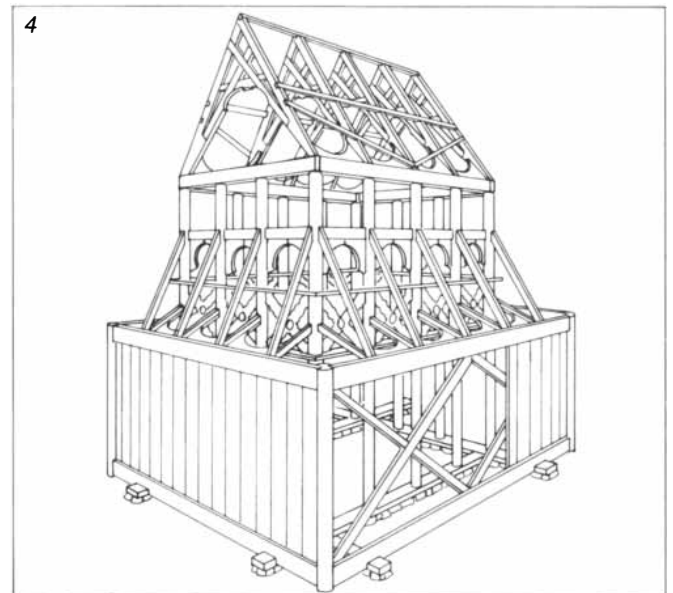
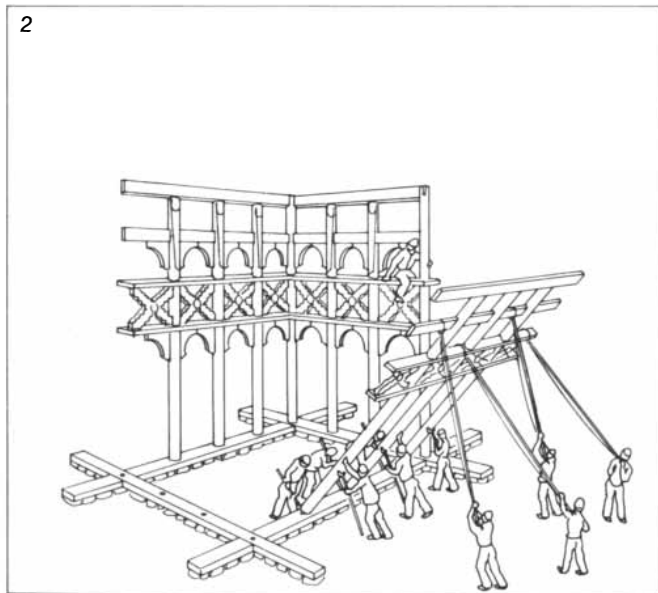
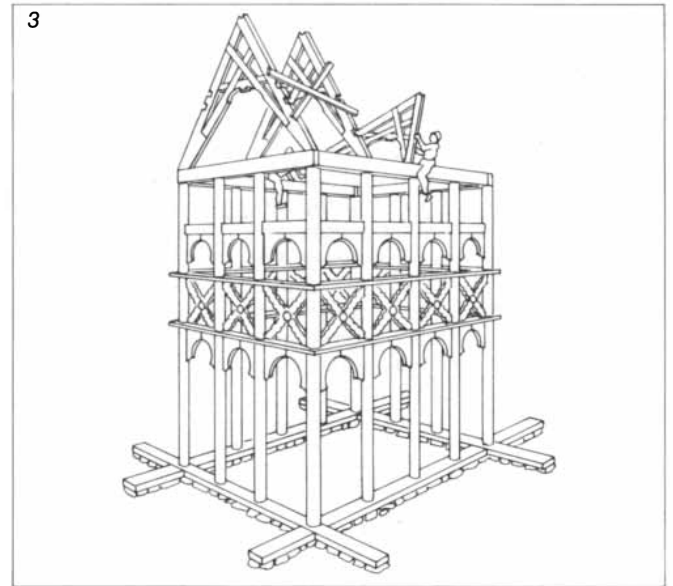
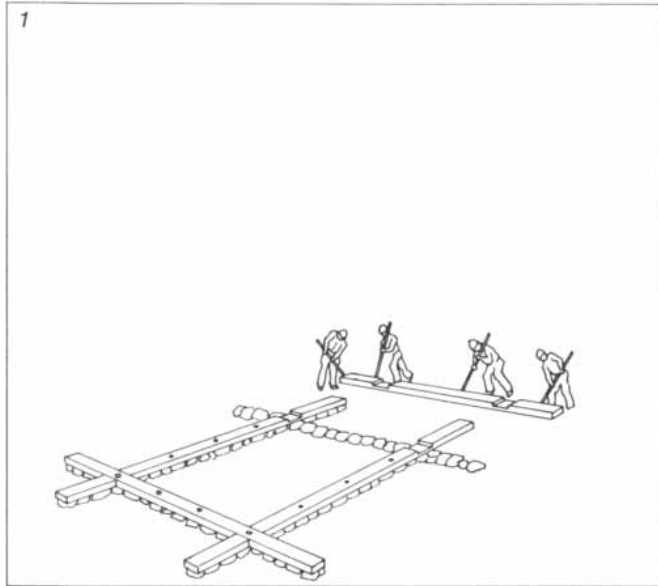
cel at the east end. The churches were generally built on visually prominent sites and are therefore exposed to the full effect of sun, wind, rain and snow. In order to construct a tall wood building on such a site several significant engineering problems must be overcome. The foundation must provide a stable footing for the building and protect the lower members from rot. The vertical load imposed by snow and the mass of the structure must be conveyed to the footings in such a way that the building is stable. The structure must be able to sustain the substantial lateral forces exerted by high winds. Provision must be made for protecting exposed joints from deterioration and for replacing parts that are subject to rot. How these

problems were solved emerges from a survey we have done of 16 of the surviving stave churches. The basilican stave churches are the most complex of the group, and in what follows we shall be concentrating on them.

The main interior staves in the basilican stave church are supported by four horizontal members called raft beams. The raft beams are laid on rows of flat stones set on the ground and are joined at the corners by mortises (meshing notches). They are arranged in two parallel pairs that intersect to form a square with the beams extending beyond the corners in an arrangement much like the symbol designating a number (#). The beams are notched at

their outer ends to accommodate the sill beams. Depending on the exact design of the church, other foundation beams can be added to the interlocking grid of raft beams.

The foundations of the nave and the chancel are often constructed so that they function independently of each other. The interaction of structural components resulting from this system makes it possible for forces to be transmitted through the building in a complex way. The configuration has the distinct disadvantage of introducing large stresses into the structure if one of the units settles into the soil more than the other. Our survey shows, however, that there is almost no such "settlement distress" in the stave churches. Therefore it



CONSTRUCTION OF A SMALL STAVE CHURCH with the basilican design was carried out with methods much like those employed to erect a small frame house. The raft beams were laid on a bed of flat stones and joined by mortises (1). Each assembly of staves corresponding to one face of the central compartment was raised on the

raft beam and held in place by temporary bracing (2). The tenon on the bottom of the stave was tapered so that it would slide easily into the mortise in the raft beam. The assemblies were joined at the corners by the arcading. When the four sides of the central compartment were up, the roof was put on (3). The aisles were then added (4). The

appears that the soil conditions in the areas where the stave churches survive are favorable to the design.

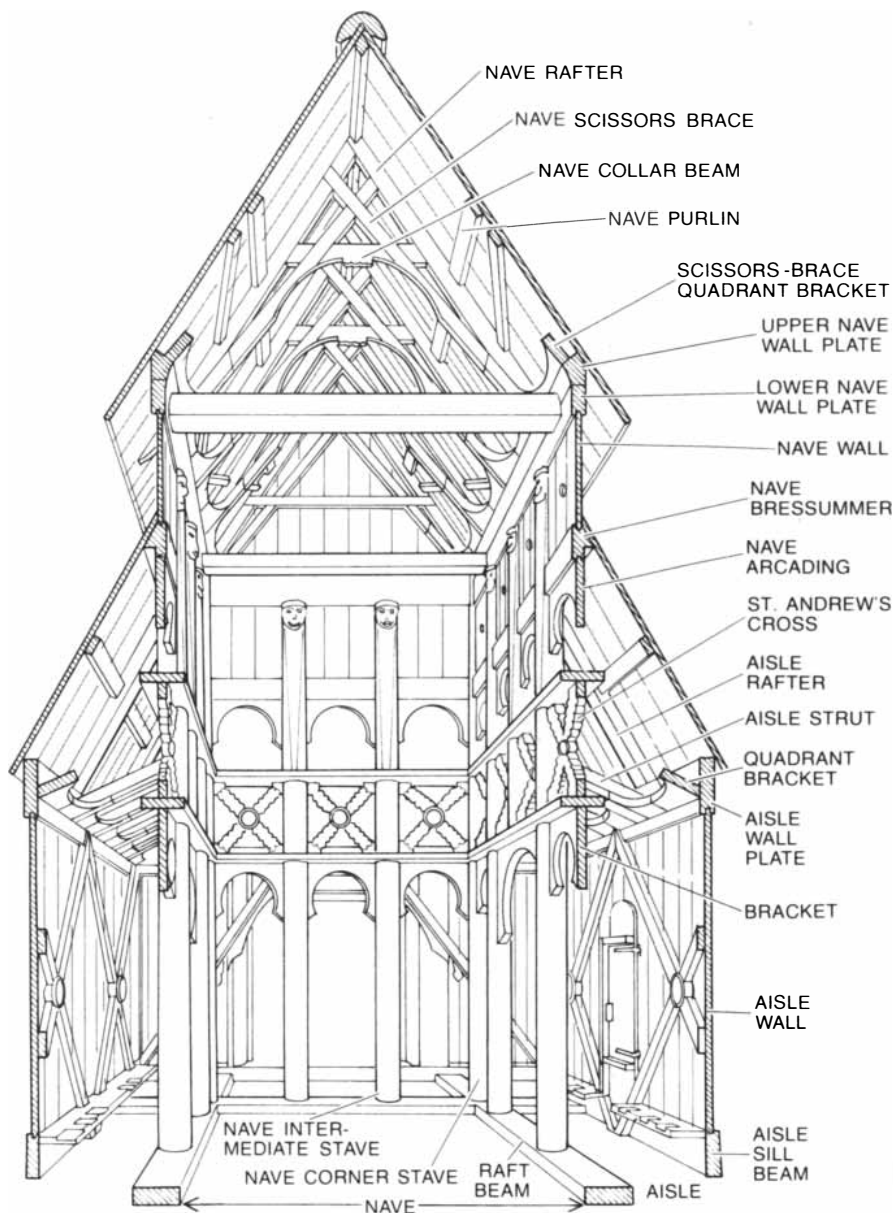
The primary structural members that transfer the load from the roof and walls to the footings are the staves. The interior staves can be as long as 11 meters; their diameter varies from 30 to 40 centimeters. A tenon, the projection that fits into a mortise, connects the base of the stave to a corresponding mortise in the raft beam. At the top of the stave is a structure called a wall plate, made up of two smaller beams; the rectangle composed of the nave wall plates provides the base for the roof of the church. Thus the central compartment of a basilican stave church has the form of a parallelepiped. The top of the parallelepiped is

defined by the nave wall plates, the bottom by the raft beams. The four sides of the compartment are defined by the interior staves, and there are generally from three to five staves to a side.

The rectangular central compartment must be braced to keep it from collapsing. Rigidity in several directions is required. Each side must be braced to prevent the staves from falling over like a row of dominoes. Resistance to forces in the plane of the staves is provided partly by horizontal beams called bressummers, which are placed from 1.25 to two meters below the top of the stave. As we shall see, the bressummer has several significant functions in addition to bracing the staves. In the right angle between the bressummer and the stave is a

curved wood bracket that supplies additional stiffness; the brackets are collectively termed arcading.

Near the arcading in many stave churches is a bracing assembly made up of two intersecting diagonal members between each pair of staves. Such assemblies, called St. Andrew's crosses, first appeared in about 1200. Many of the churches put up before then were fitted with crosses long after they had been built. The crosses were probably intended to give horizontal strength to the assembled staves; in many churches, however, the crosses are too loose to provide much resistance. Indeed, climbing in the interior of the Torpo stave church, one of us (Sack) narrowly escaped injury when a cross almost pulled loose from its moorings. On some very long staves a second tier of arcading has been put below the crosses. The brackets and crosses are fastened to the staves and to the horizontal beams by round wood pegs about 25 millimeters in diameter. Some of the pegs have an enlarged head, but in many instances the peg was fixed in place by a wood wedge driven into its outer end.



planks making up the aisle wall were inserted into the sill beam at their base. The sill beam fits into a mortise in the raft beam. The completed church is shown at the right. The drawing is based on the church at Borgund, which was built in about 1200 and is one of the most beautiful and best-preserved of the stave churches. The view is from the chancel through the nave toward the main entrance. The central compartment is defined by 14 interior staves 6.7 meters high.

As the builders of the stave churches gained experience they began to change the configuration of the interior staves in order to open up the nave, allow easier entry and provide a better line of sight from nave to chancel. This was done by shortening some of the staves. The staves at the corner of the central compartment were left intact but some staves between the corners were cut off about halfway up, so that the bottom of the staves rested not on the footings but on a wood arch formed by a pair of brackets. This feature was introduced with increasing boldness. In some early churches one stave at the west end and one at the east end were shortened. In the church at Heddal, which was built in about 1300, short interior staves alternate with full-length ones. In the church at Hurum and the one in Lomen only the corner staves continue to ground level.

Such modifications were probably possible only after much experience with stave construction. Shortening the interior staves calls for a sure sense of the strength of the structure. Moreover, the modification makes it necessary to have a large set of brackets to carry the load from the shortened staves to the corners where the intact staves convey the load to the footings.

The builders probably raised the inner compartment of a small stave church the way a small frame house is erected today. Each assembly of three to five staves corresponding to one wall of the central unit was put together on the ground at the building site. After the raft beams had been laid down the tenon at the bottom of the stave was inserted into

the mortise in the raft beam; the tenon was tapered so that it would slide easily into the mortise when the stave was lifted. The flat assembly was then raised and held up by temporary supports.

The process was repeated until all four sides of the central compartment were up. The stave assemblies were connected by the curved corner brackets. In addition each corner was generally fitted at the level of the bressummer with a horizontal "knee bracket," which gave the central unit strength to resist horizontal deformations. The central compartment of the stave church is geometrically similar to a cardboard box with the top and bottom removed; very strong winds blowing at a skew angle to the church could flatten the compart-

ment much as cardboard boxes are flattened before being stacked in a supermarket. The knee brackets contribute resistance to such flattening.

When the central compartment was up, the roof was put on. The roof of the stave church is sharply peaked, generally with a rise of 3:2, or an angle of 56 degrees. The rafters are braced by a pair of beams; the beams cross under the peak of the rafters. Such a "scissors brace" often has a rise of 1:1, or an angle of 45 degrees. A horizontal beam called a collar beam is included in the roof bracing near midheight; it is generally connected to the rafters and to the scissors braces. Light secondary members called purlins, which run parallel to the roof ridge, transfer roof loads to the

rafters and the roof bracing. In the latter stave churches diagonal bracing was incorporated into the plane of the roof to keep the structural components from collapsing.

The roof assembly of rafters, scissors brace and collar beam resembles an orthodox truss structure but does not function as an ideal truss. An assembly found in many European wood' buildings is the German truss, or scissors truss. It consists of nine subunits: four diagonal members at the top, two lower ties, two interior struts and a vertical member [see bottom illustration on opposite page]. As in other truss structures, each member is either in tension or in compression: under a downward vertical load the four top members and the two interior struts are in compression whereas the two lower ties and the vertical member are in tension.

In the scissors braces of the stave church roof, however, the vertical member is conspicuously absent. In its place is the collar beam. As a result the brace does not operate as an ideal truss; an analysis of the structure shows that many of the beams are subject to bending forces in addition to tension and compression. It is not surprising that the 12th-century Norwegian builders did not construct an ideal truss. Truss design was not formalized until the beginning of the 16th century, and before that time local solutions to engineering problems were worked out by trial and error. The collar beam nonetheless serves an important function: it redistributes the forces in the assembly and thereby minimizes the horizontal components of the forces delivered to the walls. If the collar beam were absent, the tops of the walls could be pushed apart under a heavy snow load.

Once the rafters and the roof bracing were up light wood planks were added running down from the ridge to the eaves. Exterior planks were then laid parallel to the roof ridge; the two layers of planking were usually fastened together with wood pins. In some of the churches the two layers were covered by a third layer made up of wood shakes, or rough shingles. The shakes were often cut in an unusual form that gives the roof the appearance of fish scales.

When the roof was up, the aisles were added. The roof of the aisles, being lower than that of the raised central compartment, emphasizes the vertical lines of the building and gives the exterior profile the look of a set of inclined steps. The upper end of the sloping rafters that support the aisle roof is attached to the interior stave with wood pegs. The lower end rests in a notch cut in a horizontal beam called the aisle wall plate, which is at the top of the aisle wall.

Under the aisle rafters is a crossbeam with an incline less steep than that of



PAGAN IMAGERY dominates the carvings found on the exterior of many stave churches, as can be seen in this section of the carving from the church at Hopperstad, which was built in about 1180. The zoomorphic designs in the panel are related to carvings on Scandinavian runic stones and to Irish manuscript illumination. The motifs include the dragon and the griffin. Christian motifs such as plant tendrils were inserted among the pre-Christian designs.

the rafter. Called the aisle strut, it is attached with pegs to both the aisle wall and the interior stave. It can therefore transmit both compressive and tensile forces. In contrast, the aisle rafter is attached only to the interior stave. At the aisle rafter's other end, where it meets the wall plate, it merely rests in a notch. It can therefore transmit only compressive forces.

Between the aisle struts is a large curved brace known as a quadrant bracket. Connected to the aisle wall plate, it stiffens the entire aisle structure considerably. It is particularly effective in resisting horizontal forces parallel to the long axis of the nave.

One of the final steps in construction was to fit the vertical wall planks between the sill beam and the aisle wall plate. The sill beam and the wall plate were grooved to accommodate the tenons that project from the top and bottom of the wall planks. The gap between a pair of planks was sealed with a tongue-and-groove joint or a spline.

The aisle assemblies that flank the nave make a notable contribution to the lateral strength of the entire church. Consider a wind blowing at a right angle to the long axis of the nave. The forces on the windward aisle wall are transmitted from the planking to the aisle wall plates and the interior stave through the aisle strut and the quadrant brackets. The structure will tend to deform in the direction of the wind. In the deformation the windward aisle rafter tends to respond as a tensile member. Since the rafter is attached at one end only, it is ineffective in transferring the load.

The lateral forces are transmitted across the central compartment of the church by the interior staves, the roof structure, the arcading and the St. Andrew's crosses. On the lee side of the building the aisle rafter and the aisle strut act as a small truss to transfer the load to the aisle wall and the exterior staves. Thus the aisle assemblies fulfill a stabilizing function much like that of the flying buttress in a Gothic cathedral. In the Gothic cathedral, however, it is the great mass of the buttress that provides stability. In the stave church it is the capacity of the entire cross section to transfer forces efficiently from one side of the building to the ground on the other side that is the stabilizing factor.

The builders of the stave church relied on an interesting combination of beam members and truss structures to provide resistance to vertical and horizontal loads. The roof is a modified truss assembly. The interior compartment, which must resist both vertical and horizontal forces, is made up primarily of beam members. The aisles, which offer crucial lateral resistance, act as small trusses. The combination of truss and beam components is highly effective.

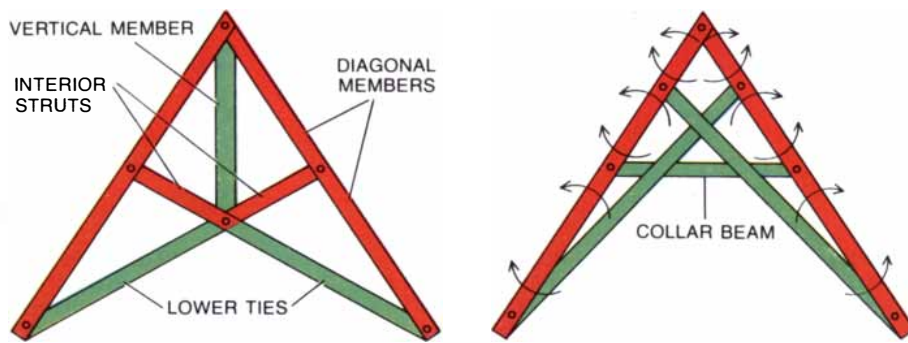


DRAIN IN A SILL BEAM helps to protect the vulnerable footings of the stave church against rot. At the left in this view of the footings of the church at Øye is the bottom of an exterior corner stave. The large horizontal member is the sill beam; above it are the aisle-wall planks. Water running down the wall can collect in the joint between the planking and the sill beam. To drain the joint the builders put small holes every few meters along the sill beam. The outlet of one such drain appears at the right. As a result of this feature the footings in most stave churches are in excellent condition. The churches are also painted with tar every two to five years.

The materials employed in building the stave church are almost as significant as the structure itself. In the selection and preparation of wood the church builders appear to have been extraordinary craftsmen. The staves, wall planks and many other structural components are Scotch pine (*Pinus sylvestris*), a tree that is abundant in Norway. Great care was taken in selecting and seasoning the large main staves. From the many available trees the builders selected a few from which the tops were removed while the tree was still standing. The tree

was left on the stump to dry for five to eight years before it was felled and the large structural members were cut to size. The outer sapwood of the tree was taken off; only the durable inner heartwood went for the staves. The wood for many of the curved brackets was from a Norwegian species of birch (*Betula verrucosa*). The wood for many of the pins and other connectors was common juniper (*Juniperus communis*), a dense softwood.

Our survey of the stave churches shows that the seasoning practices of the



ROOF BRACING of the stave church is a modified truss structure that functions differently from an ideal truss. In an ideal truss structure each member is either in tension (green) or in compression (red) but not in both at once. There are no other kinds of force in the assembly. At the left is the structure known as the German truss, which operates as an ideal truss. In the scissors bracing of the stave church roof, shown at the right, the vertical member is replaced by a horizontal collar beam. As a result some of the members are subject to bending forces (arrows). The collar beam helps the structure resist the lateral force exerted by a heavy snow load.

medieval builders yielded lumber that is superior to what is available to builders today. In most of the 16 churches we visited the tall interior staves were in remarkably good condition. In the church at Urnes one 800-year-old stave was still exuding pitch. The wood of the original staves was sound and relatively free of longitudinal cracks. Indeed, the worst case of longitudinal cracking was in two interior staves that had been in-

stalled about 40 years ago; both had cracks some 40 millimeters wide running from top to bottom. As is to be expected, the exterior staves had generally deteriorated more than the interior ones, but even the exposed staves appeared fundamentally sound.

The stave design helped the churches to withstand fire. Large wood members such as the primary structural staves can survive a fire because it takes time

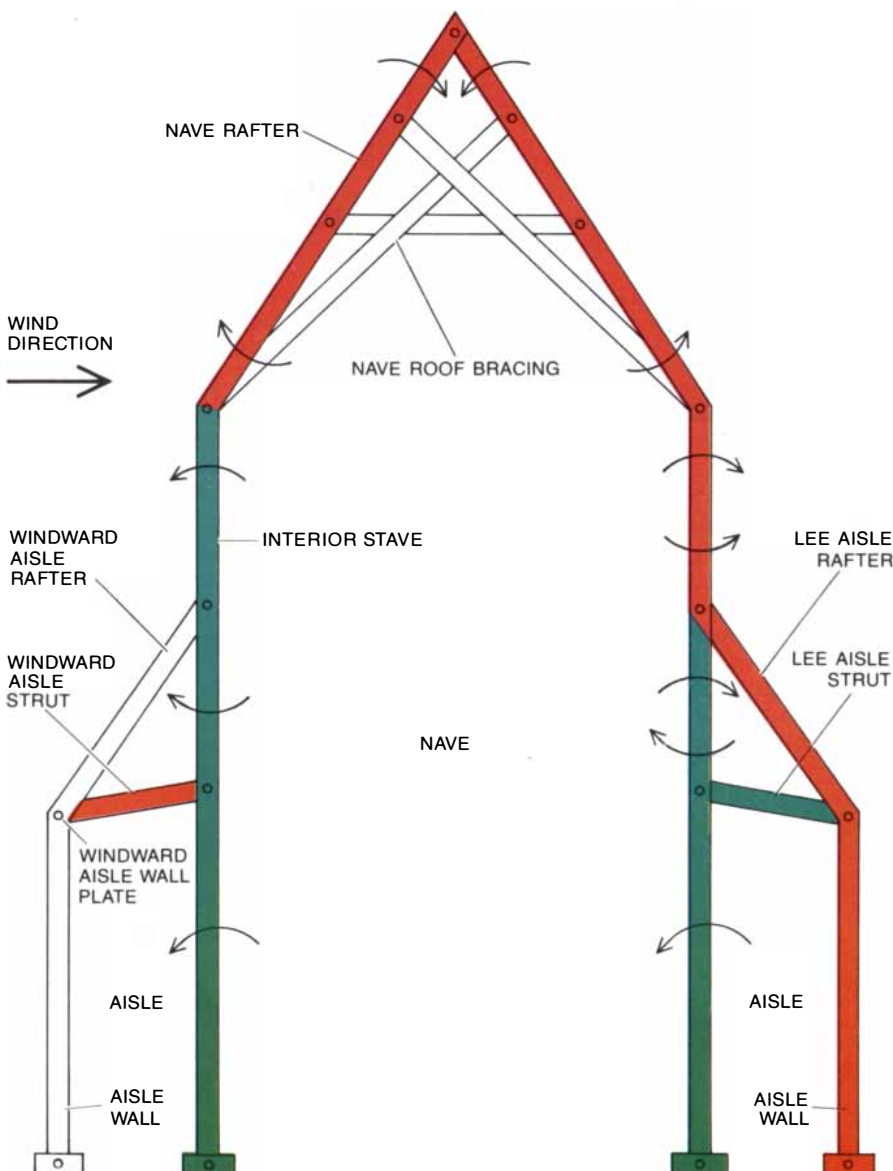
for them to burn through. Thus a stave church could last through a fire without collapsing, and when it was repaired, the charred members could be replaced. If the fire were intense enough, however, even the large structural members would burn through. A number of stave churches are known to have been destroyed by fire. A program is under way to protect all the remaining stave churches with sprinkler systems.

The resistance of the stave churches to the ravages of climate and time has been aided by a number of structural details. Along the bressummer and the outside of the sill beam the potential for rotting is great. Water can run down the vertical wall planks until it is stopped by the horizontal beam, where it can collect and penetrate the joint. To counteract the problem the builders made small drain holes at intervals of a few meters along the bressummer and the sill beams to allow water to escape from the groove in which the wall planks rest. The strategy has been effective: we noted only one sill beam in an advanced state of decay.

In many of the churches a covered walkway called the ambulatory runs around the building. The ambulatory does not serve any primary structural function, but it shields the lower part of the church from the weather. In one church bulbous wood covers were put around the base of the staves. The coverings are replaceable; if rot sets in, new ones can be installed. Most of the churches are to this day painted with tar every two to five years, a practice that gives them a characteristic smell.

Two factors of a more general nature also contribute to the durability of the stave churches. First, many of the structural members are stressed at a very low level. In most buildings put up today the supporting members are under a stress that corresponds to a large fraction of their capacity to sustain stress. This is to save money and to reduce weight and cross section. Our calculations show, however, that in the stave church the staves are under a stress that is only about a tenth of their capacity. Such a great reserve means that it is unlikely the members will fail under overload conditions. Second, the detailed shaping of all the pins and connecting elements in the building was quite accurate. As we have noted, the capacity of the cross section of the stave church to transfer loads laterally is a significant element in its overall structural integrity. Only if the grooving, notching and shaping of all the connecting members is accurate is the load transferred efficiently with no distress to the building.

Over the centuries the stave churches have been more successful in withstanding decay and structural failure than they have been in meeting another



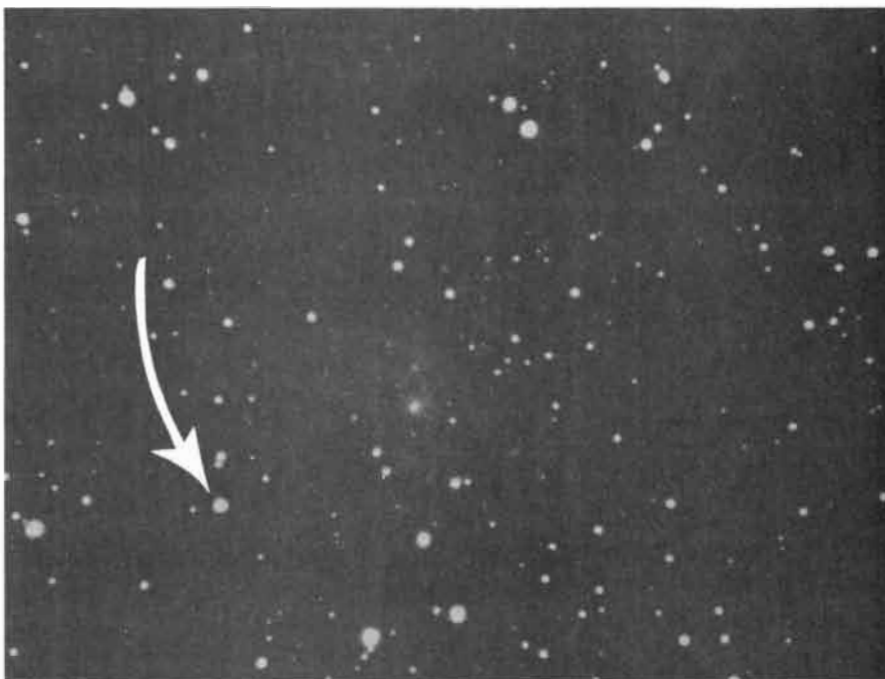
RESISTANCE TO LATERAL FORCES in the stave church is provided by the entire cross section of the church, which functions as a single assembly. In this drawing of a transverse section across the long axis of the nave the structural members in tension are indicated in green and the members in compression are indicated in red. Bending forces are designated by arrows. A strong wind is blowing from left to right. In such a wind the aisle wall plate on the windward side transfers the load to the interior staves by means of the quadrant bracket and the aisle strut. (In the computer simulation on which the drawing is based the arcading was omitted.) The staves, the roof structure and the roof bracing transfer the load across the central compartment. On the lee side the aisle rafter and the aisle strut function as a small truss that shifts the load to the leeward aisle wall. Thus the aisles have a stabilizing function like that of the flying buttress in a Gothic cathedral. In the Gothic cathedral stability is provided, however, by the mass of the buttress. In the stave church it is provided by the response of the entire cross section.

threat: the changing needs and tastes of the communities in which they were built. In about 1350, when the number of stave churches was at its highest, the bubonic plague struck Norway, as it did much of the rest of Europe. The population of the country was greatly reduced and church construction stopped. When it resumed, it was under the influence of new religious attitudes; the Lutheran religion was introduced into Norway in 1536. The stave churches were considered too small, too dark and too cold. Many were replaced or remodeled; others were not properly maintained and were destroyed by decay. By 1800 there were only about 100 left.

In 1814 Norway, which had been ruled by Denmark for some 400 years, became an independent nation. Independence was accompanied by a revival of national pride and interest in religion. In 1851 a law was passed requiring that all churches be able to hold 60 percent of the congregation, and as a result many older churches were razed. At the same time there was a countervailing interest in the country's cultural heritage, and a few Norwegians began to try to save the surviving stave churches.

One of the leading rescuers was the painter Johan Christian Dahl. When the Vang church in the district of Valdres was about to be torn down, Dahl made a valiant effort to raise the funds needed to buy it. Failing to do so, he persuaded King Frederick William IV of Prussia to buy the church, which was dismantled and reconstructed in a small village in Silesia (now part of Poland). There the church still stands. The episode aroused much interest among the Norwegian people, and in 1844 the Society for the Preservation of Ancient Monuments in Norway was founded to preserve the remaining stave churches. There are currently 29 of them left, many heavily modified or moved to a site other than the original one.

Among the stave churches that stand unmodified and on their original site, however, are two of the most beautiful: the church at Urnes, the one that may represent the genesis of the basilican style, and the church at Borgund. Their survival, along with the results of our survey, shows that the stave churches were well designed and well constructed. They provide confirmation of the fact that it is possible to build wood structures that will last indefinitely if certain conditions are satisfied. The wood must be carefully selected and cured, meticulous design practices must be applied, attention must be given to details that forestall decay, construction methods must be of high quality and the structure must be maintained continuously to minimize deterioration. The evidence shows that if these conditions are met, wood buildings can be permanent.



An example of the interesting things that might turn up unexpectedly while you are casually sweeping the deep sky is this supernova in NGC 6946, in a portion of a photograph taken by Hubert Entrop on November 4, 1980, with his Questar. This was 7 days after the official discovery of the nova, as described in *Sky & Telescope* in the January, 1981, issue.

Great news from Questar. . . The Wide-Sky Telescopes

The comet is coming. And Questar is ready for it with two brand-new instruments for all dedicated comet watchers and other deep-sky enthusiasts.

Our emphasis in developing these new designs was on low magnification and optimum field of view—magnification low enough and field of view wide enough to sweep the sky for richfield observing without need of finder or equatorial mount.

We also wanted to dispel the popular misconception that short focal lengths ($f/3$ to $f/5$) are essential for effective deep-sky observing and optimum image brightness. Image brightness is a function of aperture and magnification, and has nothing to do with focal length. We continue to use long focal ratios, even in our new wide-sky models. This rules out the comatic aberrations of short-focal-length reflectors and the spurious

chromatic effects of short-focal-length refractors, as well as the below-par imagery of short-focal-length eyepieces. Imagine, then, the pleasure of having that familiar sharp Questar® resolution, with no aberrations, available in a telescope with a field of view of 3°.

Specifications of Questar Wide-Sky Telescopes

Questar Wide-Sky 3½ (shown left): aperture 89mm; focal length 700mm; focal ratio $f/7.8$; magnification 22× with 32mm wide-angle eyepiece; field of view 3°; faintest visible star 12th magnitude (visual), 12.8 with special coatings.

Questar Wide-Sky 7: aperture 178mm; focal length 2400mm; focal ratio $f/13.5$; magnification 44× with 55mm wide-angle eyepiece, 75× with 32mm wide-angle eyepiece, field of view 1.15° with 55mm eyepiece, 1° with 32mm wide-angle eyepiece; faintest visible star 13th magnitude (visual) with standard coatings, 14th magnitude (visual) with special coatings.

Both instruments can be equipped with various accessories to provide for tracking and photography.

Special uses for the Wide-Sky Questars are comet seeking and general sky scanning, monitoring telescopic meteors, observing the occultation of stars and planets by the Moon, observing lunar and solar eclipses, observing the deep-sky phenomena — nebulae, star clusters and galaxies — and variable star observing.

Photographic accessories are available for use in these applications, in which the Questar becomes an astrographic camera. All Questars, of course, can be used with a video monitor.

Let us send you our price list.

©1983 Questar Corporation

QUESTAR

Box C, Dept. 206, New Hope, PA 18938
(215) 862-5277



QUESTAR, THE WORLD'S FINEST, MOST VERSATILE TELESCOPE, IS DESCRIBED IN OUR BOOKLET IN COLOR, WITH PHOTOGRAPHS BY QUESTAR OWNERS. SEND \$2 TO COVER MAILING COSTS ON THIS CONTINENT, BY AIR TO SO. AMERICA, \$3.50; EUROPE AND NO. AFRICA, \$4; ELSEWHERE, \$4.50.

Digital Typography

Most type is now produced not by casting metal or by photography but by computer. The digital typesetter can create new letterforms with the flexibility of a scribe at up to 15,000 characters per second

by Charles Bigelow and Donald Day

The type on this page of *Scientific American* is set by a machine whose operation is digital. It is formed on the screen of a cathode-ray tube by an electron beam moving in a vertical pattern consisting of 800 lines to the horizontal inch; the motion of the beam is fast enough to generate a page of text in about 15 seconds. The resolution of the screen is much higher than that of a standard television receiver, but in other respects the two devices operate in the same way. The tiny individual picture elements—pixels—along each scan line can be made to fluoresce or remain dark by turning the beam on or off as it scans. The resulting pattern of discrete vertical strokes, which would be visible at moderate magnification, is perceived by the unaided eye as a page of smooth letterforms.

Although a digital computer is needed to control the on-off pattern of the electron beam, the type itself is digital because it is made up of discrete elements. These elements can be line strokes, pixels, colors, shades of gray or any other graphic unit from which a letterform can be constructed. Hence digital typography is not new: mosaic tiles, embroidered samplers and arrays of lights on theater marquees have long represented alphabetic characters as relatively coarse discrete arrays. These digital letterforms, however, are typographic curiosities, far from the mainstream of traditional type design and composition. The traditional letter is not digital but analogue; its final form varies smoothly with the continuous variation of some process employed in its creation, such as the pressure of a brush on paper or the contour of a punch that stamps the matrix or mold used to cast type in metal. With the development of the computer and digital electronics, typography in the past 15 years has been seeing a wholesale replacement of analogue text by digital text, which may rival the shift in the Renaissance from script to print.

It is estimated that each day in the U.S. about 10^{14} letters are reproduced; indeed, letterforms make up much of

the visual texture of civilized life. Although some letters are now destined primarily to be “read,” or decoded, by a machine, all letters must be read by people if the communication channel between writer and reader is to remain intact. Reading skills, however, are difficult and costly to acquire. In order to protect this educational investment the letters one reads as an adult must not be noticeably different from the letters one learned as a child, and the letterforms read by the current generation must not be significantly different from the ones created by previous generations. The transition to digital typography therefore presents a subtle question with far-reaching implications: How is it possible to take the fullest advantage of digital technology and still ensure that digital letterforms retain the quality of the traditional letters, whose beauty and legibility have contributed profoundly to literacy in our culture?

The advantages of digital typography are substantial: once letterforms are represented as discrete elements they can be efficiently encoded as discrete and distinguishable physical properties in any convenient medium, processed as bits of information by a computer, transmitted over great distances as pulses of current and decoded to reconstitute the letterforms for the person receiving the message. Indeed, once type is digitized it is effectively encoded in the binary language of the computer, and so the size, shape and subtler characteristics of letters can be readily modified by

a computer program. Since some varieties of digital type can be read by machines, the semantic content of the information represented by the letters can be manipulated by a computer as well as the letterforms. Unlike analogue information, digital information is highly resistant to noise or degradation introduced during the transmission of a signal. The digital receiving device need distinguish between only two states of the signal (say on or off) in order to decode and recover the information originally transmitted.

Digital typography can be adapted to a wide variety of output devices. The composition of most daily newspapers and national magazines in the U.S. is done by cathode-ray-tube typesetters. A new generation of high-resolution laser printers has now been introduced; a highly collimated laser beam replaces the electron beam as a writing instrument, and it can either expose a typographic image on a printing plate directly from digital information or make an intermediate photographic exposure on paper or film. Lower-resolution laser printers, which are also called text setters, are employed as output devices in data processing and in the publication of forms and documents in small batch quantities. In such devices the laser beam writes out the image of the text by setting up a pattern of electrostatic charge on a belt or drum. Fine pigmented particles called toner particles are attracted to the belt or drum where the charge is created, are transferred to paper like ink from a press and are then

SAMPLES OF DIGITAL LETTERFORMS, shown at decreasing resolution from the second row to the bottom row of the illustration, are each based on the designs in the top row by Kris Holmes and one of the authors (Bigelow). The letters in each row are formed by superposing a square grid of increasing coarseness on the model letterforms at the top. If some part of the model letter coincides with the center of a square on the grid, the entire square is blackened; otherwise the square is left white. Because digital type is constructed out of discrete elements, it is ideally suited for storage, transfer and manipulation by digital electronics systems and computers. Nevertheless, it imposes a new set of constraints on typographic design: the basic letterforms must be “well tempered,” that is, they must retain maximum legibility across a variety of digital displays, and so the designer must take into account many versions of the image simultaneously. The digital letters were generated by Autologic, Inc., of Newbury Park, Calif.

R N Q b a e g

R N Q b a e g

R N Q b a e g

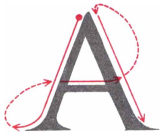
R N Q b a e g

R N Q b a e g

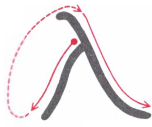
R N Q b a e g

R N Q b a e g

ROMAN
MONUMENTAL
CAPITALS
(FIRST CENTURY)



ROMAN
CURSIVE
(SECOND CENTURY)



ROMAN
UNCIAL
(FIFTH CENTURY)



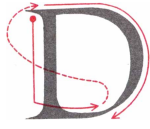
CAROLINGIAN
(NINTH CENTURY)



HUMANIST
MINUSCULE
(15TH CENTURY)



CHANCERY
CURSIVE
(16TH CENTURY)



DUCTAL LETTERS are handwritten letters whose basic topology is the result of a smooth series of movements of the writing tool in the plane of the writing surface. The path of the tool is called ductus. The

ductus of each letter is shown in color; the broken lines indicate the part of the ductus that is invisible in the final form of the letter, namely the path of the tool when it is not in contact with the writing sur-

ironed or fused onto the paper by heat.

There are also many devices in service whose output resolution is quite low, that is, far fewer pixels are available to approximate the form of the letter than are required for finely rendered alphabetic details. For example, the printer that is often attached to both large and personal computers is made up of a column of fine wires. The tips of the wires strike a ribbon impregnated with ink and impose a pattern of ink dots in a vertical column on the paper. As the column of wires moves across the page the tips of different wires in the column are actuated by solenoid magnets, and so the changing pattern of vertical dots on the page generates digital text. Ink-jet printers have also been developed that create letter images by directing the flight of electrically charged ink droplets toward the paper. The trajectory of the droplets is controlled by charging them electrically and then passing them through an electrostatic field [see "Ink-

Jet Printing," by Larry Kuhn and Robert A. Myers; *SCIENTIFIC AMERICAN*, April, 1979]. Soft-copy, or transient, images of digital type on cathode-ray tubes, liquid-crystal displays and light-emitting diodes have become increasingly important as the number of personal computer work stations and word-processing devices has grown.

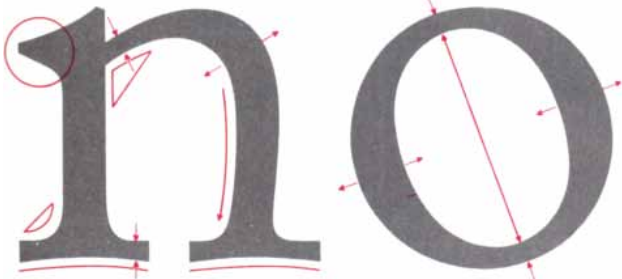
The capacity of all such devices to render typographic characters is best estimated by the number of pixels along the side of a square called an em square, which is equal to the printer's point size of the type in a given font. (There are approximately 72 printer's points to the inch.) The side of an em square, which is also called an em, is slightly more than the distance from the top of the ascender on a letter such as the lowercase *h* to the bottom of the descender on a letter such as the lowercase *y*. For the standard text size called pica, or 12-point type, most high-resolution digital typesetters have a resolution of from 100 to 300 lines per

em; electrostatic text setters have a resolution of from 33 to 80 lines per em, and low-resolution digital output devices can vary from 10 to 30 lines per em.

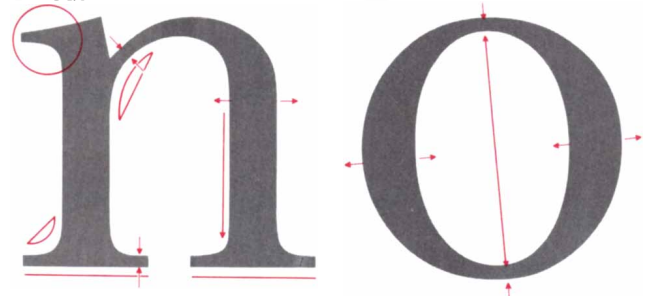
The speed, versatility and low cost of digital typography have made its proliferation irresistible. Moreover, if the letterform is mapped onto a raster, or grid, of digital pixels that is sufficiently fine, the differences between the original letter and the digital letter are virtually indistinguishable. In principle it seems possible, given a fine enough digital raster, to imitate any of the traditional printed or written letterforms, no matter how refined. In practice, however, digital typography, like any other technological innovation, carries with it special problems and a new set of selective pressures.

For example, not all digital typography can begin at high resolution. There are technical limitations on the size of the dot that can currently be reproduced

RENAISSANCE



BAROQUE



EVOLUTION OF GLYPHAL LETTERS reflects the fundamental changes in the technology of letter production brought about by the invention of printing from movable type. The letter image was no longer the result of a series of strokes written "on the fly" by a scribe;

instead each letter was painstakingly engraved as a single master copy onto the face of a steel punch. Early glyptal letters such as the Renaissance forms imitated scribal forms such as the Humanist minuscule, and many details (*color*) simulate the effects of a broad-edged

ROMAN
MONUMENTAL
CAPITALS
(FIRST CENTURY)

ROMAN
CURSIVE
(SECOND CENTURY)

ROMAN
UNCIAL
(FIFTH CENTURY)

CAROLINGIAN
(NINTH CENTURY)

HUMANIST
MINUSCULE
(15TH CENTURY)

CHANCERY
CURSIVE
(16TH CENTURY)



face. Alphabets such as the Roman capitals, for which a high proportion of the ductus is invisible, are called formal, whereas alphabets such as Chancery cursive, for which a high proportion of the ductus

is visible, are called cursive. Much of the evolution of the Latin alphabet is the result of the interplay of the opposing tendencies between formal and cursive writing. The calligraphy was done by Holmes.

by methods such as ink-jet printing, and the attainable resolutions are too coarse to escape visual detection. Moreover, the speed and cost advantages of digital typography are reduced as the number of digital bits needed to generate the letter is increased. The number of pixels per letter increases as the square of the linear resolution of the printing device: doubling the linear resolution of the device implies a fourfold increase in the amount of information, or number of bits, that must be transmitted and processed. Although there are computational methods for compressing the data in the bit map of a letterform, the general relation between cost and resolution remains valid.

Perhaps the most challenging problem that confronts the digital-type designer is to make effective use of the enormous flexibility inherent in digital technology. For example, the text of a document could first be written at a cathode-ray-tube terminal with a reso-

lution of 10 lines per em for pica type. A proof of the same text could then be pulled from a wire-matrix printer with a resolution of 20 lines per em, circulated for correction and commentary as the output of a laser text setter with a resolution of 50 lines per em and finally set for publication on a cathode-ray-tube typesetter with a resolution of 200 or more lines per em. Similarly, when digital letters must be represented at different sizes on a machine with a fixed raster, a different bit map is required for each letter size. It is obviously desirable in such circumstances for the type designer to create a single kernel, or underlying, letterform that can generate all the forms in which the letter occurs on different machines and in various sizes.

In order to appreciate the magnitude of the problem, consider the variability of letterforms that is reflected in a single superfamily of typeface designs. For each modern design one of each of three opposing features must be speci-

fied: whether the type is roman or italic, whether it is normal weight or boldface and whether it is serif or sans-serif. (A serif is a short finishing stroke to a major writing stroke, such as the small horizontal line at the bottom of the stem in the letter *T* as it appears on this page; sans-serif type has no serifs.) Taken together, the three features generate eight typeface designs. Furthermore, each type alphabet typically includes characters in 16 different sizes. The total number of glyphs, or individual bit maps, necessary to accommodate a single character for a minimum superfamily of type is therefore 128; the number of glyphs necessary for a complete superfamily, which may include 128 letterforms, is 128², or more than 16,000.

Although digital technology imposes a new set of problems on typographic design, such problems are not without precedent; it is instructive to explore the relative stability of the letterforms

NEOCLASSICAL



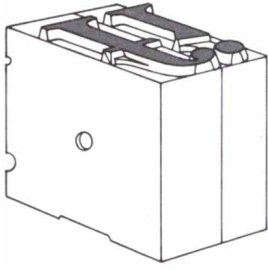
pen held at an angle of approximately 30 degrees to the horizontal. In general the contrast between thick and thin elements is relatively low. In the Baroque and Neoclassical versions of the glyptal letter symmetry, harmony of structure and contrast between thick and thin

INDUSTRIAL SANS SERIF

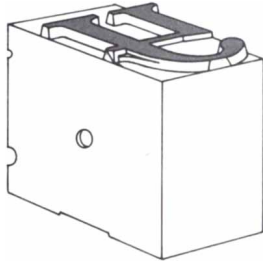


strokes become more important than the trace of the pen. More recent designs either eliminated serifs, as in the industrial sans-serif types, or converted the serifs into independent elements, as in the industrial slab serif (not shown). The drawings were done by Holmes.

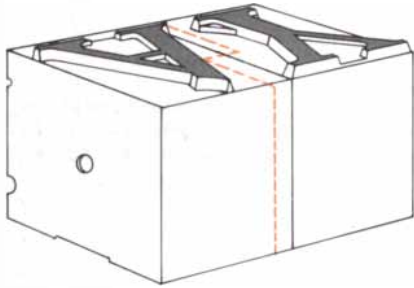
fi



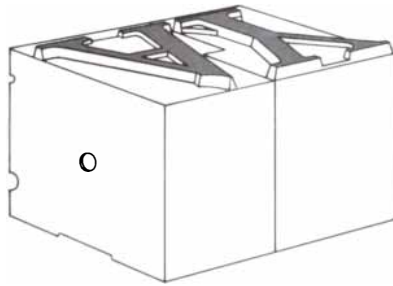
fi



AY



AY



METAL TYPE is cast on a body of fixed width, and so the variant forms and spacing for different combinations of letters that are hallmarks of the fine manuscript cannot easily be emulated. In the first decades after the invention of movable type some printers cast hundreds of variant letterforms, thereby nearly canceling the advantage in time and economy afforded by the new technology. Later forms were well-tempered designs, cast on a metal body in such a way that their assembly into words would automatically lead to proper spacing. A few combinations, such as *fi*, resisted tempering and were cast on a single body of metal. Other letter combinations such as *AY* could not be properly spaced unless the type body was mortised by hand.

6 POINT

RQEN baegnov

8 POINT

RQEN baegnov

14 POINT

RQEN baegnov

18 POINT

RQEN baegnov

SCALING THE SIZE of type cannot be done proportionally even within a single typeface if optimum legibility is to be retained. In the smaller sizes from six to 12 point the letters must be wider and their stems, hairlines and serifs must be thicker than their counterparts in larger sizes. Moreover, in smaller sizes ascenders such as the stem of the lowercase *b* and descenders such as the tail of the lowercase *g* must be shorter in proportion to the height of the lowercase *n*; the counter, or interior space of a letterform, must be more open and the space between letters must be greater. The type font in the illustration is Times Roman; it has been scaled photographically to the same height of the letter *n* so that differences in design can be compared.

through typographic history and their adaptations under the pressures imposed by previous technological shifts. Two stages in the evolution of type design can be recognized following the introduction of each new technology. First, there is a period of imitation, in which the outstanding letterforms of the previous typographic generation serve as models for the new designs. Second, as designers grow more confident and familiar with the new medium, innovative designs emerge that are not merely imitative but exploit the strengths and explore the limitations of the medium. Since type is ultimately intended for the reader, however, the technology of type production is not the only influence on the final letterform; typographic design remains an art, and the successful design subtly reflects the tension between imitation and innovation.

The shapes of letters have persisted longer than any other artifacts in common use. Letter designs are still in service that are more than 2,000 years old, and many common typefaces are replicas of designs popular in the 15th and 16th centuries. The text type in this magazine, for example, was originally designed by the British typographers Stanley Morison and Victor Lardent for *The Times* of London in 1931 and is called Times Roman. It is based on French and Flemish types designed in about 1570. The basic forms of our lowercase alphabet were established in the eighth century in the monasteries and chancelleries of Charlemagne's Frankish empire. The forms of our capital letters are substantially the same as those inscribed by the Romans in the reign of Augustus Caesar. Fully half of our capital letterforms are structurally unaltered from the inscriptional forms used in Periclean Athens in the fifth century B.C.

In the scribal era the contours of the letterform were created by a continuous sequence of movements of a brush or a pen as it moved across the plane of the writing surface. During the smooth sequence of movements the writing tool was pressed against the surface or lifted from it, thereby generating the strokes that make up the letter. The sequential pattern of movements is called ductus, and it defines a characteristic topological structure for each letter of the alphabet. The evolution of writing was impelled by changes in ductus [see top illustration on preceding two pages] and by variations in the shape and flexibility of the writing tool. The tool was responsible for the contrast between thick and thin strokes. When the tool was an edged pen, the contrast of the letter was determined by the angle between the edge and the direction in which the pen was moving in the plane of the writing surface. When the tool was a brush or a flexible pointed pen, the contrast in the letter was determined by variations in

*An incisive look into our
scientific and technological future*

Frontiers in Science and Technology

The National Academy of Sciences

The third in a series of reports to the American people, this landmark study examines science and technology's impact on today's most significant issues.

"Perhaps the most important theme that emerges from this volume is the remarkable fertility of American science and technology. New knowledge has provided new ways for understanding diseases. Such fundamental research areas as surface physics and turbulence theory have yielded new technologies and new controls for ancient problems. Assuming real growth in research support, the United States can expect to remain at the forefront of virtually all scientific frontiers in the 1980's and beyond."—Floyd E. Bloom, The Salk Institute, from the Introduction

CONTENTS The Genetic Program of Complex Organisms—Maxine F. Singer, National Cancer Institute. The Molecular and Genetic Technology of Plants—Joseph E. Varner, Washington University. Cell Receptors for Hormones and Neurotransmitters—H. Guy Williams-Ashman, University of Chicago. Psychobiology—Steven Hillyard, University of California, San Diego. Surface Science and Its Applications—Homer D. Hagstrum, Bell Laboratories. Turbulence in Fluids—Willem V. R. Malkus—Massachusetts Institute of Technology. Lasers—C. Kumar N. Patel, Bell Laboratories. The Next Generation of Robots—Jacob T. Schwartz, New York University, and David Grossman, IBM Watson Research Center.

Paperbound. 228 Pages. ISBN 0-7167-1517-1 \$14.95
Available at fine bookstores, or use the order form at right.

Frontiers in
Science and
Technology



Yes, please send me _____ copy (copies) of FRONTIERS IN SCIENCE AND TECHNOLOGY at \$14.95 each.

_____ I enclose a check for \$14.95 for each copy ordered, plus \$1.50 to cover postage and handling. (New York, California, and Utah residents please add appropriate sales tax.)

_____ Please charge my MasterCard Visa

Account # _____ Expiration Date _____

Signature _____

(All credit card orders must be signed.)

Name _____

Address _____

City _____ State _____ Zip _____



W. H. Freeman and Company
41 Madison Avenue
New York, NY 10010

ISBN: 0-7167-1517-1

SA

pressure as the tool described its ductus.

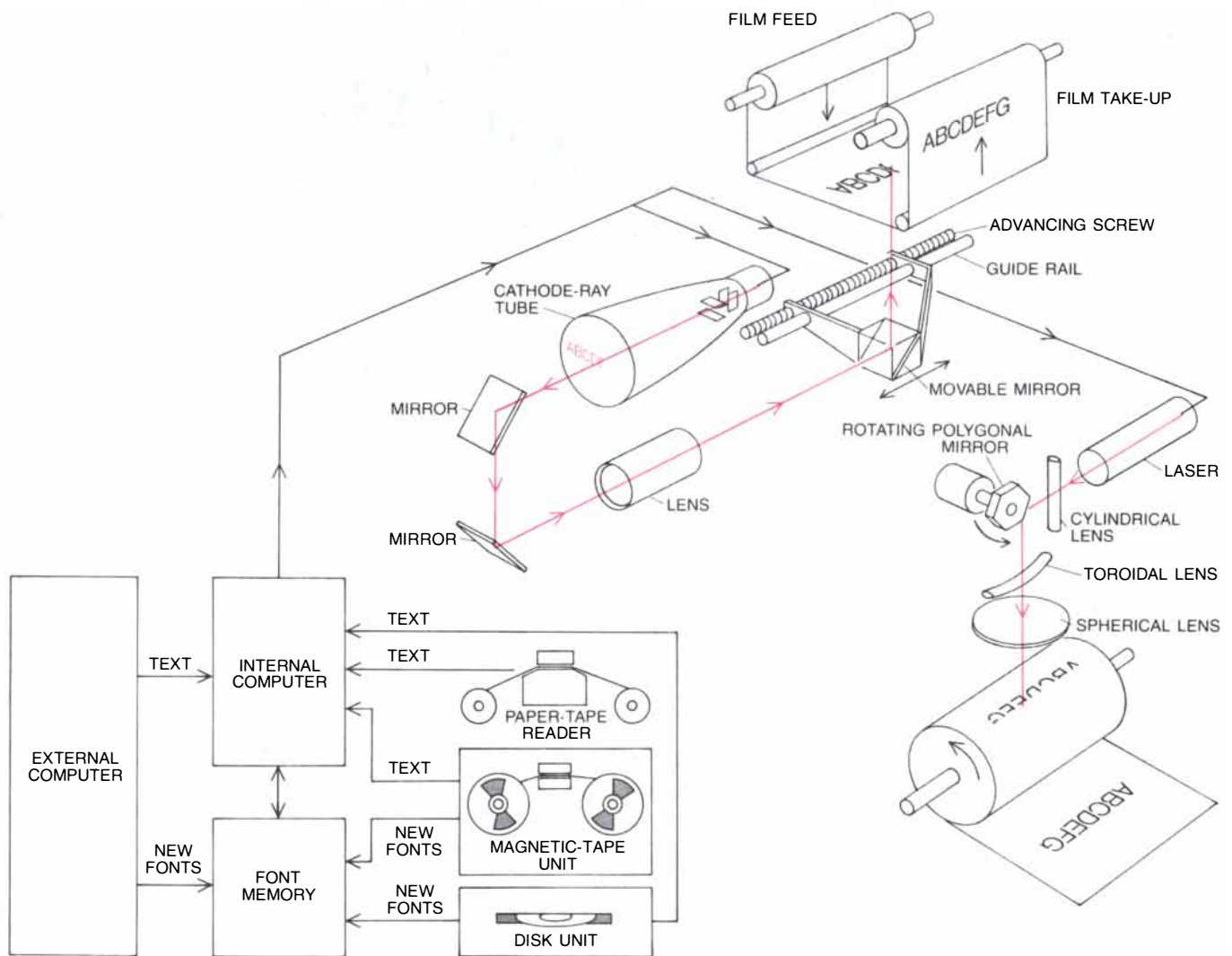
An important advantage of scribal lettering is the immediate feedback between the final form of the letter and the scribe. The accomplished scribe is expert in making minor changes in the form of each letter in order to harmonize the letter with the letters adjacent to it. Moreover, to justify lines of text the scribe could insert alternative or abbreviated forms of letters and words. The great disadvantage of scribal practice, of course, is that the design of each letter must be executed in "real time": any letter that is to be read must be generated from scratch each time by hand.

When movable type was invented in Europe, it was at first modeled on scribal forms current in the 15th centu-

ry, even though it was created by a radically different technology. An image of the letterform was engraved in relief onto the face of a steel punch. The punch was then hardened and struck into a blank of copper to form an intaglio, or recessed, matrix. The matrix was placed in an adjustable mold; when a molten alloy of lead, tin, antimony and copper was poured into the mold and the matrix, the design appeared in relief once again on the face of the cooled alloy. The face of the type was then inked and impressed directly onto paper. The letterform was thus created by a glyptal, or sculptural, process. The letter design was freed from the limitations of real-time execution but was constrained by the need to place it on a rigid rectangular solid. In order to imitate the scribal

variations in letterform for different letter combinations a few early hot-metal typographers maintained hundreds of variant letterforms; such forms were so expensive to cast as well as to compose, however, that they nearly negated the economic advantages of movable type.

The evolution of glyptal typeface contours was guided by conceptual and perceptual forces instead of by the needs of rapid handwriting. Greater attention was given to the shape of the spaces within letters and between them. Proportion, width, weight and construction were altered independently of the underlying topology of the letter rather than being partially determined by it, as they were in the ductal letter. The designers of glyptal letters were obliged to make a different engraving for each let-



DIGITAL TYPESETTING MACHINES can store enormous numbers of type fonts for almost instant retrieval and can generate characters in almost any size at rates of up to 15,000 characters per second. The data are entered into the machine by means of punch-coded paper tape, by magnetic tape or disk or as a stream of data from an external computer. In the most recent installations the external computer organizes the text into lines, columns and pages, justifies the text if that is wanted and automatically hyphenates words where it is necessary for good spacing. Once the font, size and position of each character on the page have been specified, the data are assembled

into a series of electronic pulses by an internal, "slave" computer, which drives a marking engine. In the schematic diagram two kinds of marking engine are shown. (The two devices would not both be incorporated into any real machine.) In the cathode-ray-tube typesetter an electron beam reconstructs each letter as a series of closely spaced vertical lines. The image is projected onto photosensitive film or paper. In the laser printer a laser beam takes the place of the electron beam; the laser scans horizontally across an entire page at once and generates a pattern of charge on a drum. Toner particles are attracted to the charged regions and ironed directly onto paper by heat.

TAKE THE 3-VOLUME HANDBOOK OF ARTIFICIAL INTELLIGENCE (A \$142.40 VALUE) FOR ONLY \$3.95

when you join the Library of Computer and Information Sciences.
You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

Just completed, the massive, 3-volume HANDBOOK OF ARTIFICIAL INTELLIGENCE promises to become the standard reference work in the growing AI field.

Conceived and produced by leading scientists and researchers at Stanford University, with contributions from universities and laboratories across the nation, the *Handbook* makes available to scientists, engineers, students, and hobbyists who are encountering AI for the first time the techniques and concepts in this rapidly expanding computer universe.

The 200 articles cover the emerging issues, technical problems, and design strategies which have been developed during the past 25 years of research. The *Handbook* has been written for people with no background in AI; jargon has been eliminated; and, the hierarchical organization of the book allows the reader to delve deeply into a particular subject or browse the articles which serve as overviews of the various subfields.

The 15 chapters (5 per volume) include: the history, goals, and current areas of research activity; the key concept of “search”; research on “natural languages”; the design of programs that understand spoken language; applications-oriented AI

The comprehensive HANDBOOK OF ARTIFICIAL INTELLIGENCE answers questions like:

- What is a “heuristic problem-solving program?”
- How do computers understand English?
- Can computer programs outperform human experts?

AND INCLUDES:

- Over 1,450 pages.
- More than 200 articles in 15 chapters.
- With numerous charts, tables, and schematics.
- Edited by Avron Barr, Paul Cohen and Edward Feigenbaum.

research in science, medicine, and education; automatic programming; models of cognition; automatic deduction; vision and learning research; and, planning and problem solving.

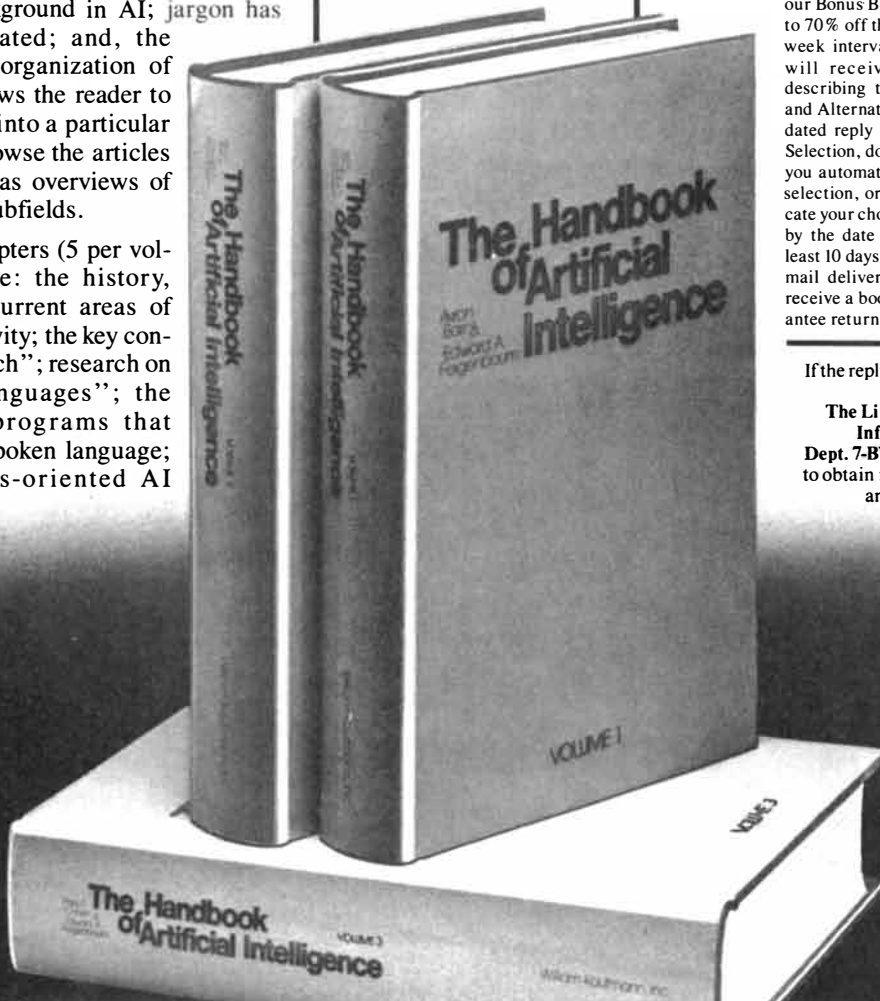
The Library of Computer and Information Sciences is the oldest and largest book club especially designed for the computer professional. In the incredibly fast-moving world of data processing, where up-to-date knowledge is essential, we make it easy for you to keep totally informed on all areas of the information sciences.

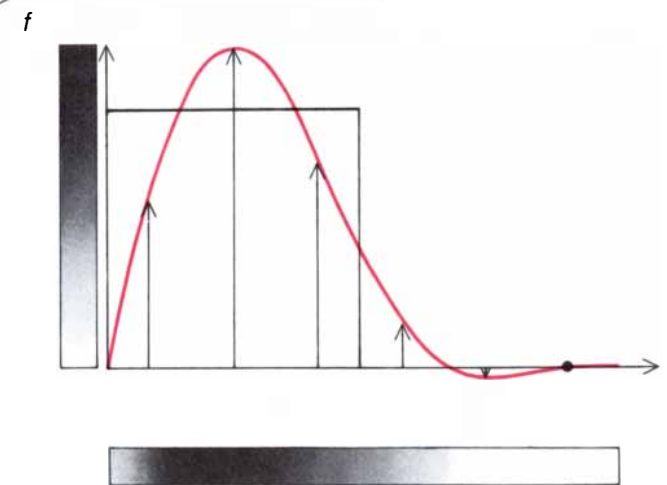
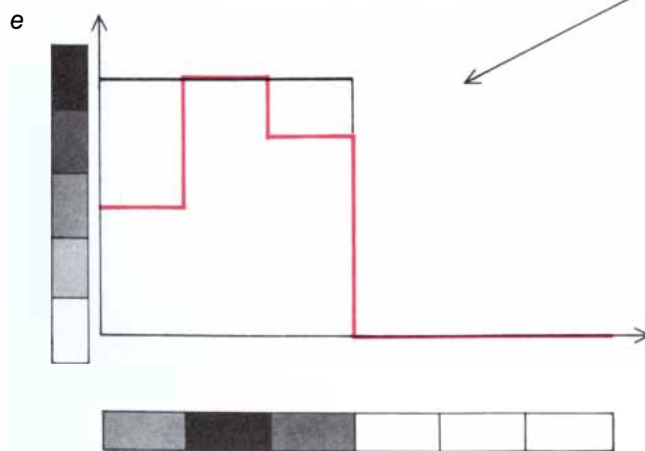
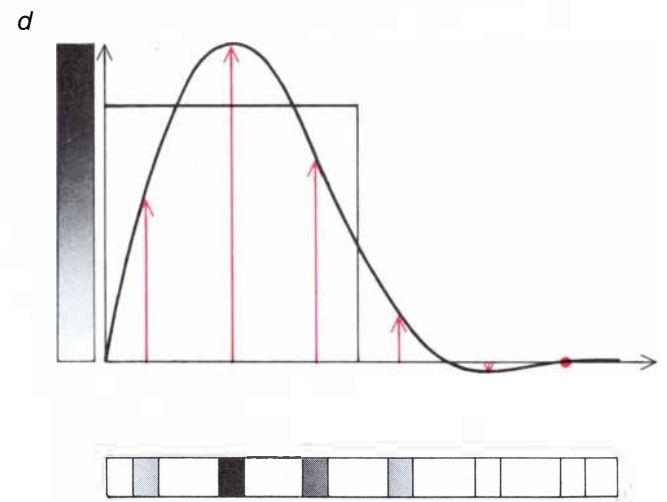
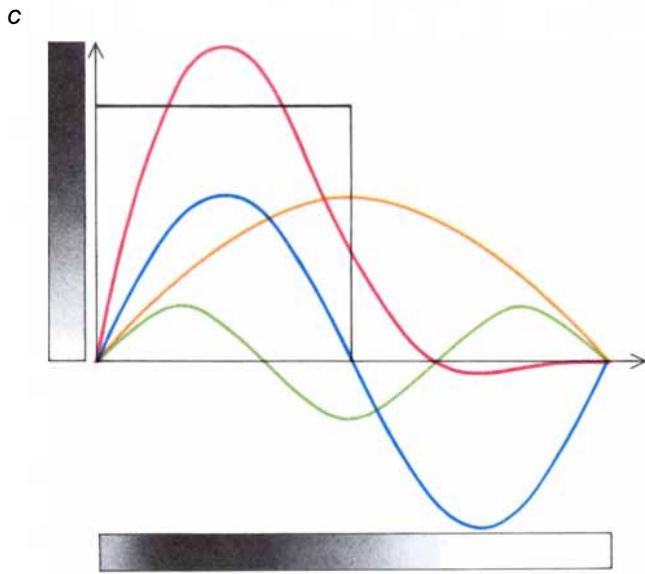
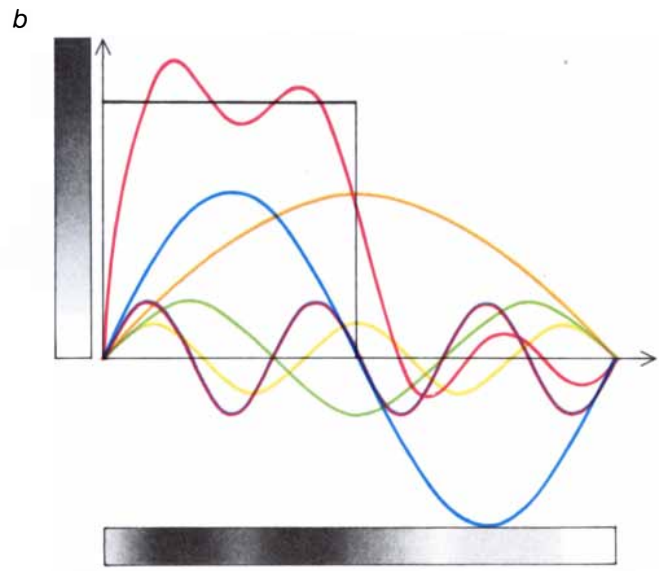
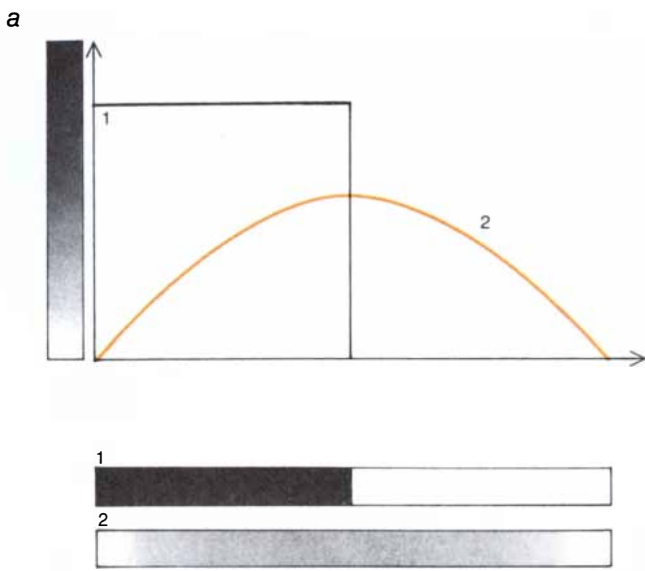
Begin enjoying the club's benefits today!

MEMBERSHIP BENEFITS: In addition to getting the 3 volume Handbook of Artificial Intelligence for only \$3.95, when you join, you keep saving substantially on the books you buy. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the coming Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If the reply card has been removed, please write to:

The Library of Computer and Information Sciences
Dept. 7-BV8, Riverside, N.J. 08075
to obtain membership information and an application.





ter size, and so, like their scribal forebears, they could make subtle adjustments in spacing and stroke width for different type sizes in order to accommodate the eye. Alternative letterforms were almost entirely eliminated, and the special demands of the new technology were met by letter designs that were felicitous in almost any combination. Nevertheless, the vestiges of the earlier ductal technology were still evident in type fonts cast in the mid-20th century: certain letter combinations such as *fi*, *fl*, *ffi* and *ffl* were cast as a ligature, that is, on a single body of metal. To achieve good spacing for combinations such as *TA* and *AY* the body of the type was often kerned, or mortised, by hand at considerable cost.

With the advent of photocomposition in the mid-1950's, typography was confronted once again by a new set of technological variables. In early phototypesetters a stroboscopic light was flashed through a negative master character mounted on transparent film or glass, and from there through a lens that projected an image of the master character onto photosensitive paper or film. The size of the letter image was controlled by interchanging several lenses or by employing a zoom lens of variable focal length. A printing plate was made from the image after the paper or film was photographically developed. Because the position of the character on the paper was controlled by the lens and by a prism or mirror that moved in fine increments across a line of type, the adjustment of interletter spacing for special letter combinations was much simpler than the kerning of metal type. On the other hand, a significant advantage of hot-metal typography was given up. The type design could not be varied with the size of the type without preparing a new master image. The cost of additional master-image carriers and the inconvenience of interchanging them discouraged their use.

SPATIAL-FREQUENCY ANALYSIS can be carried out for any two-dimensional image, such as a letter, just as the graph of almost any mathematical function can be approximated by a sum of sines and cosines. A sharp-edged rectangle, for example, can be represented as a straight line parallel to the horizontal axis of the graph; the density of gray or black in the rectangle corresponds to the height of the line above the horizontal axis. In order to approximate the rectangle a unique set of sine and cosine waves can be superposed. The first sine wave in the approximation is shown in *a*. The spatial wave, or variation from white through shades of gray and back to white, that corresponds to the sine wave is a rather crude approximation to the rectangle. The more sine and cosine components there are in the approximation, however, the better it becomes: the five colored sine waves of various heights and wavelengths are added together along each vertical line to give the smooth red curve in *b*. In order to eliminate high-frequency noise any shape can be electronically filtered to remove components above a certain frequency. In *c* the two highest-frequency components of the red curve in *b* have been eliminated, namely the yellow and purple sine waves, and the result is the red curve in *c*. The filtered curve can be sampled by finding its height, or gray-scale value, at evenly spaced intervals along the horizontal axis (*d*). A digital approximation to the filtered curve is made by assigning a discrete shade of gray across an interval that corresponds to each sample (*e*). It can be mathematically proved that the original filtered curve can be completely reconstructed from the samples by interpolating between the sample points, provided the frequency at which the samples are taken is greater than twice that of the highest-frequency component of the filtered curve (*f*).

The typographer working within the constraints of each of these technologies faced a set of problems quite similar to those now encountered by the digital type designer. For example, any typographic version of the letterforms familiar to the reading public is a successful communication system to the extent that it is able to balance two opposing characteristics: discriminability and similarity of alphabetic characters. To avoid confusing the reader each letter must be rapidly and unambiguously distinguishable from every other letter. On the other hand, the letterforms must also share many graphic features: if a letterform is too distinctive from the other letters in a font of type, it disturbs the flow of reading. Donald E. Knuth of Stanford University has succinctly stated the goal of text-type design: "A font should be sublime in its appearance but subliminal in its effect."

Throughout the history of writing economic pressures have opposed the pressures for more readable and more beautiful typography. As always, therefore, need must be matched to purpose. Letter quality can be sacrificed on certain documents if the tradeoff is faster, cheaper or more compact reproduction. A bill of sale, for example, can be cheap to prepare and to store because it will be read infrequently by few people; the slow and expensive production of an inscription is justified when it will be read frequently by the multitude. In the time of the Romans the basic alphabet was written rapidly in a semilegible cursive when it was employed in a papyrus document, but it was chiseled in clear imperial capitals on a monument. Today inexpensive dot-matrix printout is less readable than the costly typography in mass-market advertising.

The effects on the reader of complex variations in letter design have not yet been quantified, but the response of the visual system to simple spatial variations in light intensity has been studied for more than two decades. The simplest

variation to analyze is a visual analogue of an acoustically pure tone, whose periodic variation of intensity with time can be plotted as a sine or cosine wave. A train of spatial sine or cosine waves can be visualized as a ribbon compressed along its length so that its edge traces a series of ordinary sine or cosine waves. If the top side of the ribbon were shaded in such a way that the crests of the ribbon were black, the troughs remained white and the intermediate sections were various shades of gray, then the blurred, parallel bands, or grating, of light and dark that could be seen from above the ribbon would form a spatial sinusoidal wave train.

A pure musical tone is characterized by its amplitude, or loudness, and by its frequency, or pitch. The amplitude of a spatial sine or cosine wave is the maximum contrast, or deviation from neutral gray, that is found in the lightest or darkest parts of the wave train, and the frequency is the number of variations from light to dark and back again within a given distance. Psychophysicists have measured the ability of the visual system to distinguish sinusoidal bands of various contrasts and frequencies from a uniformly gray field. They have found that sensitivity to spatial variation of light and dark depends on the frequency of the variation; the sensitivity is greatest when the spatial frequency is approximately three cycles per degree of visual angle, and no contrast, no matter how strong, can be perceived under most conditions when the frequency is greater than 60 cycles per degree. (The detection of telephone wires against the sky is one of the relatively unusual circumstances in which spatial frequencies that are probably higher than 60 cycles per degree can be discerned.)

What is the importance of these findings for reading? Although the spatial variation from black to white in letterforms is not sinusoidal, fundamental rhythmic patterns in the letterforms are apparent. For example, as the lowercase letter *n* is scanned from left to right across its middle, the brightness of the image varies fairly smoothly from light to the dark stem of the first vertical stroke, then light again in the counter, or interior of the letterform, then dark on the second vertical stroke and finally light again to the right of the letter. Because reading is often done under rather poor lighting conditions, one might expect that the fundamental frequency of text letterforms has evolved to match the peak contrast sensitivity of the visual system.

Such a match is almost exactly what is found in the typesetting of English. In close reading the image given the most attention is projected onto the fovea, the most sensitive area of the retina. The

fovea subtends an angle of one or two degrees, and at a reading distance of 12 inches the subtended angle corresponds to a linear distance only slightly more than the length of a five-letter word set in 10-point type, the commonest type size. There are on the average about 10 spatial cycles across a five-letter word, and so the spatial frequency of the image received at the fovea is about five to 10 cycles per degree, only slightly higher than the most contrast-sensitive frequency of the visual system. There is good evidence that in fast reading word groups longer than five letters can be read in one eye fixation. The reading image is then partially projected onto the parafovea, the region of the retina surrounding the fovea.

We have tacitly assumed so far that the major patterns of black and white that make up letterforms are sinusoidal spatial waves, but it may not be apparent that the sine-wave model can support a more detailed analysis. It turns out, however, that just as the complex sound of a symphony orchestra can be analyzed as a sum of harmonics, or pure tones of various frequencies and

intensities, so can almost any form be analyzed as a combination of many spatial sine and cosine waves. The result follows from a theorem of the French mathematician Jean Baptiste Joseph Fourier. In 1807 Fourier proved that by superposing one-dimensional sine and cosine waves of various phases, amplitudes and frequencies the graph of almost any function can be approximated to any desired degree of accuracy [see illustration on page 114].

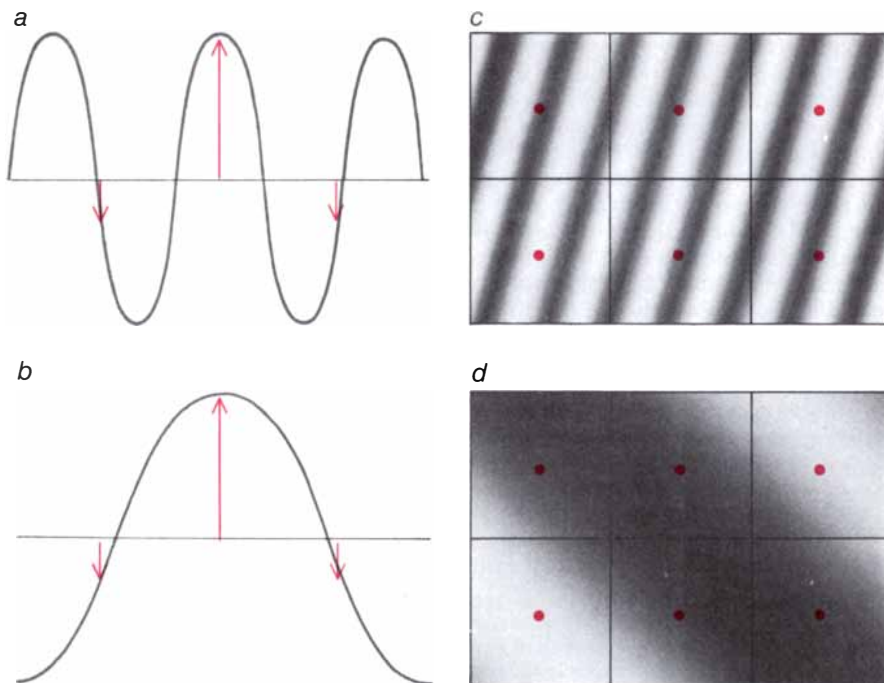
In two dimensions the spatial sinusoidal components have an additional degree of freedom, namely their orientation with respect to some fixed direction in the plane. Fourier's theorem shows it is possible to reconstruct almost any pattern such as a letterform to any desired accuracy by superposing spatial sine and cosine waves of the proper phase, amplitude, frequency and orientation. In general the high spatial frequencies of a letter image correspond to its edges and to such details as fine serifs and the tapering of the main strokes of the letter. As we have stated, the low spatial frequencies define the fundamental rhythm of the letter, or in other words its overall pattern of light and

dark. The spectrum of spatial frequencies required to represent a letter is called the bandwidth of the letter. Based solely on the maximum frequency at which the eye can detect contrast, namely 60 cycles per degree of visual angle, it might seem that completely adequate typographic reproduction could be achieved if all the spatial frequencies of a letterform above 60 cycles per degree were eliminated.

In practice high-quality digital typesetting requires that letterforms include spatial frequencies at least as high as 120 cycles per degree. One reason is that for fine detail such as the hairline at the end of a serif the acuity of the visual system may well exceed 60 cycles per degree. The main reason, however, is the relation between the spatial frequencies that make up the input letter and the process of digitizing the letter: reducing the letterform to an array of discrete units. In order to digitize a letter it must be sampled at various points. For example, if the edges of the letter were perfectly sharp, the letter could be sampled by superposing it on a grid of squares and then noting whether or not the point at the center of each square coincides with some point on the letter. If the center point of a square coincides with a point on the letter, the entire square is shaded black; if the center point does not coincide with any point on the letter, the square is left unshaded.

A perfectly sharp edge would require an infinite number of sines and cosines to represent its full spectral bandwidth. On the other hand, if the bandwidth of a letter is limited in such a way that it includes no sinusoidal component above a certain frequency, its edges must be slightly blurred. The superposition of a finite number of sinusoidal components yields a smooth transition from the black through shades of gray to white. There is nonetheless a mathematical advantage to the band-limited letter: according to results developed by Harry Nyquist at the Bell Telephone Laboratories in 1924 and 1928 and extended by Claude E. Shannon, also of Bell Laboratories, in 1949, if one knows the highest-frequency component of a band-limited signal and if samples are taken with equal spacing at any rate greater than twice the highest-frequency component, the original band-limited letter can be completely reconstructed from the sample points alone. Theoretically, therefore, a letter whose highest frequency is 60 cycles per degree of visual angle can be reconstructed solely from the measured gray values at evenly spaced sample points taken slightly more frequently than 120 times per degree of visual angle.

One difficulty with the theoretical sampling frequency is the small error introduced by the machines that scan



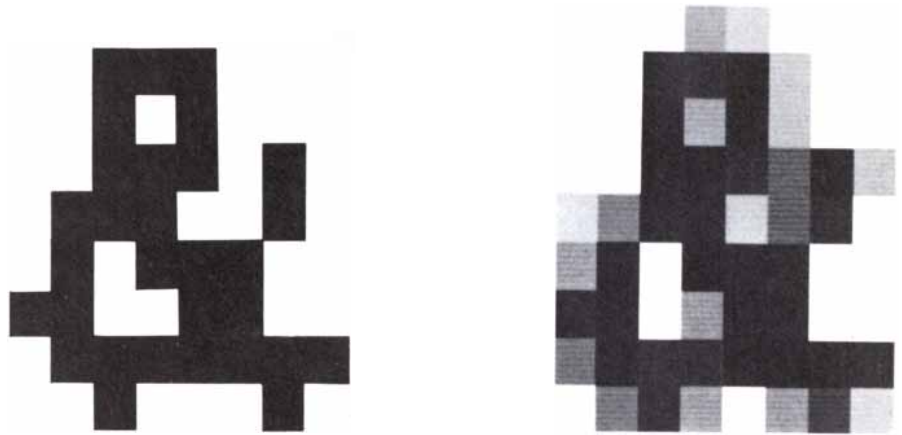
ALIASING is the commonest form of noise, or unwanted signal, found in digital typography. When the height of a wave or the gray-scale values of an image are sampled, information in the original curve or image may be lost. A reconstruction of the wave or the image from the sample points requires that the values of points between the sampled points be interpolated by finding sine and cosine components whose superposition coincides with the values of the sampled points. If the original wave or image is sampled at less than twice the frequency of its highest-frequency component, however, the highest-frequency components found by the reconstruction will have an alias frequency, lower than the highest-frequency component in the original image. In other words, spurious low-frequency components will replace the true high-frequency ones. For example, if a sine wave is sampled only 1.5 times per cycle (a), a reconstruction that matches the values of the samples (colored arrows) is a sine wave of half the original frequency (b). In the two-dimensional spatial wave the samples are gray-scale values taken at the centers of squares (c). A reconstruction of the image from the sampled values generates an alias not only of lower frequency than the original waves but also oriented in a different direction (d). Quantizing a letter image only compounds the errors introduced by aliasing.

and sample the input letters. Sample points may not be evenly spaced and the measured values of gray may be slightly inaccurate. A more serious problem is that, in practice, sharp band limiting is impossible. Hence in typography of the highest quality the samples are taken at a rate of at least 240 per degree of visual angle, which is theoretically sufficient to reconstruct a letter band-limited at 120 cycles per degree. Such a sampling rate is equivalent to a resolution of 200 lines per em for 12-point type, or 1,200 lines per inch.

If the letter image is sampled too sparsely, that is, at less than twice the frequency of its highest sinusoidal component, the gray-scale values of the samples can be identical with the values that would be given by a lower spatial frequency [see illustration on opposite page]. In such undersampling noise is added and information is inevitably lost. Low-frequency spatial waves that do not figure in the original wave spectrum of the letter replace the original high-frequency components; a reconstruction of the original letter from such samples would incorporate the spurious low-frequency components. The phenomenon is called aliasing, and it is the most obvious source of noise or distortion in digital typography. When the letter is converted into an array of pixels from the samples, the aliasing becomes manifest, and the amount of aliasing depends on the coarseness of the sampling.

At a high output resolution aliasing is evident only as a slight roughness in the letter contours; at medium resolution the contours become jagged, and at lower resolution the curves become polygonal and the diagonals develop dislocations. At still lower resolution the differences among straight, curved and diagonal elements is obscured and the letters become illegible. Moreover, as the resolution becomes coarser the diversity of possible letter designs is reduced; for very coarse resolutions, such as the five-by-seven or six-by-nine arrays of dots on many cathode-ray-tube terminals and dot-matrix printers, few variant designs are possible.

A less obvious but ultimately more serious consequence of undersampling is the loss of information about the proportional relations reflected in an alphabet design. In a design that has been finely tuned to the characteristics of the visual system there is a subtle interplay of proportion and thickness among the thick elements of a letterform and between thick and thin elements. The ratio of stem height to the height of the lowercase letter *x* and to the average width of a letter is also carefully adjusted. The outcome of such detail in design is barely perceptible in the individual letterform, but in a block of text the design becomes manifest as a visually harmo-

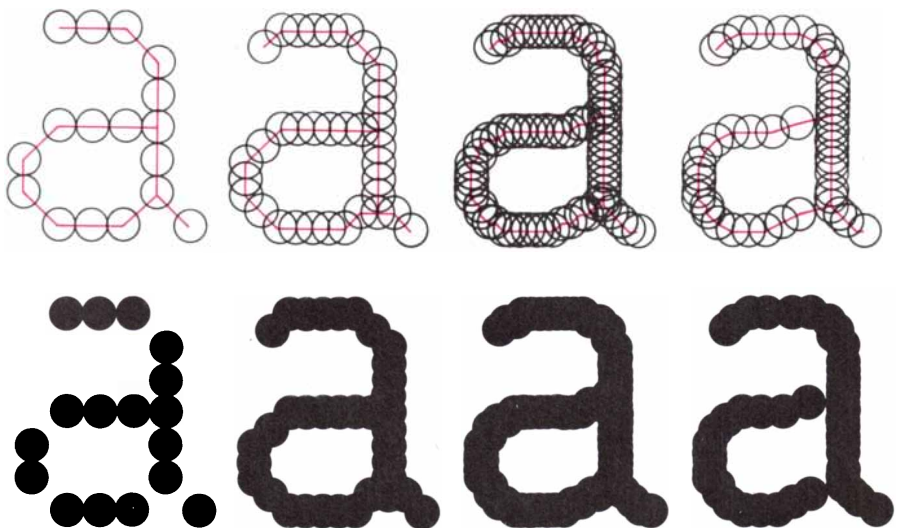


QUANTIZATION ERROR results when the shade of gray of a pixel, or minimum discrete picture element, does not match the shade of gray at the sampled point corresponding to the pixel on the original image of a letter. In most digital typography, for example, the pixel must be either black or white, which reflects the naive view that the edges of a master letterform are perfectly sharp. Actually the spatial frequencies too high to be perceived by the visual system should be filtered out before the letter is sampled and digitized. The edges of the filtered letter that results are not sharp; there is instead a continuous tone of gray that makes the transition from black to white. When the letterform is reproduced on a machine of relatively low resolution, such as a cathode-ray-tube terminal, the apparent degradation of the letter caused by the low-resolution sampling can be reduced if the pixels more accurately reproduce the shade of gray measured at each sample point. The pixels that make up the letterform at the left must be either black or white, and the design is almost illegible. The pixels that make up the same letterform can take on one of 16 gray values in the design at the right. By squinting and observing the design at the right from a distance of about 25 feet an ampersand can clearly be perceived. The digital designs were prepared by John E. Warnock of the Xerox Corporation.

nous pattern of black letterforms and white counterforms and a pleasing level of gray in the text as a whole. At a high digital resolution these proportional variations can be closely approximated, and the typographic texture appears only slightly less refined than it is in ana-

logue typography. As the digital resolution of a letterform is decreased, however, proportional variations are rounded off, typographic elements become homogeneous and the resulting textual pattern seems crude and awkward.

When cost or the technical limits of



SIZE OF A WRITING SPOT in a digital printer such as a dot-matrix printer determines the smallest resolvable element of the letterform, but the position of the spot can still be adjusted to improve the image. The parallel rows of wires on the printer head can be staggered or the head can be made to pass several times across the same character. At the left a lowercase *a* is reproduced without overlap of the writing spot, and the result is a letter of poor quality. The overlap of the writing spot is increased to 50 percent in the second letter from the left and to 75 percent in the third one; the resulting letterforms are more continuous and the edges of the letters are better defined. As the overlap increases, however, the horizontal lines become too thick with respect to the vertical lines and the joins become too blotchy. At the right certain spots have been removed in order to lighten the horizontals and streamline the joins.

the output device require that the letterform be stored or reproduced at less than ideal resolution, there are still several strategies that can be followed to improve the image of the letter. For example, if the shade of gray measured at each sample point were preserved across the entire pixel corresponding to the point, the jagged lines would be much less apparent in the final letter than they are when the pixel must be either black or white. Some devices that have low spatial resolution can reproduce gray-shaded pixels, and so what is called quantization error between the shade of the pixel and the shade of the sample point can be made quite small. The video-terminal screen, for example, can have 256 or more gray levels for each pixel, which can help to compensate for low sampling resolution.

A second strategy for improving the image of a letterform is to store the letter in the memory of a machine in such a way as to take advantage of the overall regularity in its design. Most letters, for example, are simple and connected forms; information about the gray-scale value of the samples inside one of the stems of a letter is redundant because all the samples are black. Hence the space in the memory of a computer that would be necessary to store an arbitrarily complex set of gray-scale values for the pixels making up the stem can be reassigned to store more information about its boundaries. Letters can also be structurally coded as combinations of parts that correspond to the pen or brush strokes in the

handwritten forms. More sophisticated coding of the regularity in letterforms requires less storage in computer memory but more effort to decode it later.

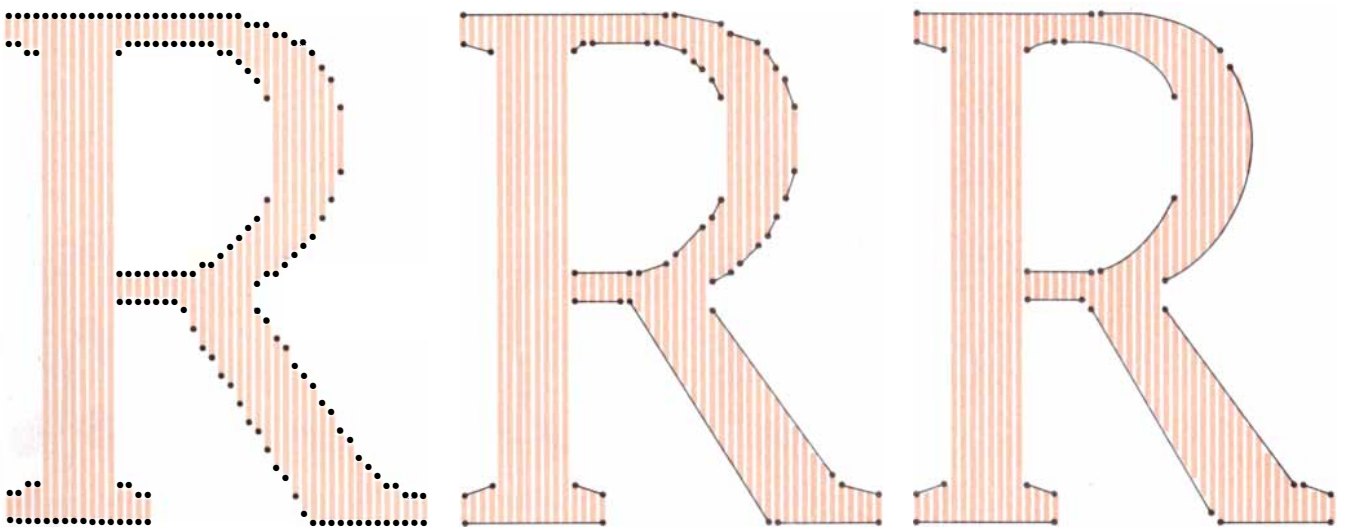
One of the most versatile ways to reduce the number of samples that must be stored is to record only the positions of a few selected points along the outline of a letter. The points can then be joined by mathematically constructing lines or curves called splines, which can be computed from the outline points when the letter is retrieved from the computer memory. If only linear splines are employed, the curves of the letter are rendered as polygons; the line segments must therefore be short enough for the polygonization not to be discerned. Curved splines derived from the graphs of quadratic equations such as the circle, the ellipse and the parabola yield a better approximation to the complex outlines drawn by hand than linear splines do. Higher-order splines such as those derived from the graphs of cubic equations can give even closer approximations, but they require more computation to construct.

When the letterform is to be stored as a set of spline knots, or points, the coordinates of the points can be encoded by a device called a digitizer tablet at resolutions as high as 8,000 lines per em, much higher than the resolution of any text-output device. The outline of the character given by the splines can then be employed as a template for generating an array of pixels for any scanning pattern. The stored set of spline knots can also serve as a generic letterform from which a computer can generate a wide

variety of particular versions of the letter. For example, higher-order splines can be converted into circular arcs or straight lines on devices that generate a letter outline rather than a bit map. The contours of the letter can be scaled up or down and stretched in any direction. If the position of each spline knot is supplemented by a descriptive label that states its role in the shape of the letter with respect to other spline knots, a typeface can be automatically fitted to a particular raster without the tedious job of turning pixels on or off by hand.

Several computer programs have been written that can manipulate spline-based letterforms. The Ikarus system, developed by Peter Karow of URW Unternehmensberatung in Hamburg, is widely used in the typographic industry. A precise drawing of the outline of a letter on which the spline knots have been marked is placed on an electronic grid, and the positions of the knots are entered into computer memory with a cursor. Letter elements such as stems and serifs are identified as they are entered. The Ikarus program computes the spline outlines, and it can automatically carry out a number of design variations, such as changing the size of a letter on a fixed scanning pattern, varying the thickness of the letter strokes from light or medium to boldface and interpolating between the differing forms of the same letter in two type fonts.

Another spline-based design system called Metafont has been developed by Donald Knuth. The language employed for programming in Metafont is based



STORAGE SPACE in the memory of a computer necessary to reproduce a letter at a given resolution can be reduced in several ways. One strategy is called run-length encoding, in which only the endpoint positions of each stroke of an electron beam or a laser beam are stored instead of the gray-scale value of each pixel (*left*). Spline encoding can further reduce the memory requirements: the outline of the letterform can be specified at several critical points called spline

knots. When the letter is needed, a computer can then interpolate straight lines (*middle*) or straight lines and circular arcs (*right*) between the spline knots in order to generate the endpoints for each stroke of the writing instrument. More elaborate curves such as logarithmic spirals can also be interpolated in order to more closely approximate the original letterform, but such curves trade a reduction in the computer memory for increased calculation "on the fly."

on ductal principles; once the topology of the letterform has been described the final form of the letter is determined by specifying the characteristics of a virtual "pen" that traces the skeleton of the letter. (The virtual pen must not be confused with any real pen; the computer simply represents the final shape of a letter as if it had been drawn with a real pen.) The size of the virtual pen, its angle with respect to the writing surface, the shape of its tip and other variables that can be independently specified generate quite different letter contours from the same skeleton. Other versions of Metafont are under development that describe the letterforms according to glyptal principles as well as ductal ones.

Several other systems have also been realized: FRED, a cubic spline program written by Patrick Baudelaire, and Prepress, a pixel editing program written by Robert F. Sproull, were developed for the Alto computer work station at the Palo Alto Research Center of the Xerox Corporation. A spline system based on spiral curves was developed by Peter Purdy and Ronald McIntosh of Purdy and McIntosh in Watford, England, and the ELF system, based on ductal principles, was written by David Kindersley and by Neil Wiseman of the University of Cambridge. A program and computer work station called the Letter Image Processor has been developed by the Camex Corporation in Boston.

In spite of its almost universal application and widely recognized flexibility, digital typography is only now beginning to move from the imitative to the innovative phase of its development. The sampling of letters and indeed the entire theoretical apparatus on which the sampling is based presuppose an already existing letter design. This model letter is an analogue form, and the success of the digital letter is still judged almost entirely on the degree to which it imitates the ductal or the glyptal letter. Nevertheless, the problems inherent in sampling and digitizing such a letter suggest it would be more productive to design new letterforms directly for digital technology. Moreover, recent advances in the study of vision create an opportunity for the digital-type designer to experiment with ways of adapting typography even more closely to the needs of the visual system. If letter images can be developed that more closely approximate the "language" of visual perception, the speed and efficiency of reading can be enhanced.

Creative design can best be accomplished on a synthetic system in which the type designer can interact rapidly with a computer and immediately see the effects of design changes on a screen, much as the traditional designer can immediately see and correct the mark of a pen or a brush. The system must be

precise, high in resolution and capable of instantly reproducing the design in many versions, such as a spline-based outline, a bit map, a gray-scaled pixel array or a simulation of the output of a particular printer. No such machine is yet at hand, but the emerging generation

of special integrated circuits for use in the graphic arts may soon make it possible to build one. When such systems become available, there will surely be a flowering of new letterforms as the digital era, like the ductal and glyptal eras before it, enters its creative phase.

a
Hamburgefons Hamburgefons

b
Hamburgefons Hamburgefons

c
a a a a a a

d
A pen of aspect 1/3 generated these letters.

A pen of aspect 2/3 generated these letters.

A pen of aspect 1/1 generated these letters.

e
The x-height and the heights of ascenders and descenders can be independently specified.

f
A 'slant' parameter transforms the pen motion, as shown in this sentence, but the pen shape remains the same. The degree of slant can be negative as well as positive, if unusual effects are desired. Too much slant leads, of course, to letters that are nearly unreadable. Perhaps the most interesting use of the slant parameter occurs when Computer Modern Italic fonts are generated without any slant.

g
The 'square root of 2' in these letters is 1.100.

The 'square root of 2' in these letters is 1.300.

The 'square root of 2' in these letters is 1.414.

The 'square root of 2' in these letters is 1.500.

The 'square root of 2' in these letters is 1.700.

VARIATIONS IN THE DESIGN of digital letterforms can be carried out automatically with the aid of several computer programs. In the Ikarus system, developed by Peter Karow of URW Unternehmensberatung in Hamburg, letters can be compressed or expanded without altering the width of their stems (*a*). The program can also correct the design automatically for different printing sizes; the required changes are much more complex than simply stretching the image in one direction (*b*). Smooth interpolations can be generated between a given letter in two different fonts. In *c* the letter at the left is Bembo and the letter at the right is Helvetica Black; the intermediate forms were constructed by the Ikarus program. Another design system called Metafont has been developed by Donald E. Knuth of Stanford University. A program written in the Metafont language alters certain characteristics of the letters by controlling the properties of a virtual pen, by means of which the computer represents the final form of the letter. For example, the shape of the tip of the pen can be controlled; it is an ellipse for which the aspect ratio, or the ratio of the vertical axis to the horizontal axis, must be specified (*d*). Samples of the effects of additional design variables on letterforms are also shown (*e, f, g*).

THE AMATEUR SCIENTIST

*In which simple equations show
whether a knot will hold or slip*

by Jearl Walker

A hitch is a knot intended to secure a line to another line or to a fixture. Are some hitches securer than others? Is there a limit to the load that can be placed on a hitch before it fails because the free end slips through the knot? Is there some way to alter a normal hitch so that it can withstand a heavier load? In exploring these questions I am guided by the work of Benjamin F. Bayman of the University of Minnesota, who has published an impressive theory about the strength of hitches.

The security of a hitch depends primarily on two features of the knot: how it is wrapped around an object (in what I call wraparounds) and how it passes over itself (wrapovers). Both features provide friction to maintain the hitch when the cord is put under tension. Without friction the cord slips through the knot, untying itself. My discussion is limited to cords of moderate diameter tied around fixed rods of larger diameter. What degree of friction in wraparounds and wrapovers will ensure that a hitch does not fail even when an arbitrarily large load is put on the cord?

I shall disregard here several practical aspects of knots. For example, some knots are better suited to a load that pulls perpendicular to the rod; others are designed for a load that pulls parallel to the rod. Some knots offer a certain advantage in holding for a long time even though they are slipping. Certain types of knots may be better if the load varies periodically in strength. All these factors are of secondary importance here. I shall also disregard any additional strength in a hitch that derives from the way adjacent sections of cord rub against each other (except at a wrapover). The principal sources of friction in a hitch are in the wraparounds and wrapovers.

To understand the role of tension in a cord under load imagine holding up a weight by a cord. Tension means that any small section of the cord is being pulled in opposite directions by forces from the adjacent sections. The tension

is uniform along the length of the cord. Hence the force applied at the upper end of the cord must match the weight attached to the lower end if the assembly is to be stationary.

Next imagine holding up the weight by passing the cord over a fixed horizontal rod. The section of the cord between the weight and the rod is under uniform tension but the tension of the section in contact with the rod varies, decreasing from the side with the weight to the other point where the cord leaves the rod. The tension of the section between the rod and your hand is uniform but is lower than it is on the other side of the rod because of friction between the cord and the rod. Because the cord rubs against the rod, you no longer have to apply as large a force to hold the weight in place.

The amount of friction on the cord depends partly on the roughness of the two surfaces in contact and also on how much of the cord touches the rod. You gain friction by wrapping the cord several times around the rod in the maneuver known to sailors and climbers as belaying. Each additional wraparound reduces the force you must apply at your end of the cord.

Friction is also generated in a hitch if the cord passes over itself. When the cord is put under tension, the bottom part of a wrapover is pressed against the rod by the top part. The squeeze opposes the tendency of the cord to slide. The tighter the squeeze, the greater the friction. The hitch holds fast provided the friction is large enough to nullify the force that tends to make the bottom part of the cord slide through the wrapover.

A hitch is designed so that the friction generated within the knot is great enough to hold against the tension generated by a load. Suppose the load increases steadily. Does the tension eventually rise to the point where the free end of the cord begins to slide through the knot, or does the increase in tension merely increase the friction so that the cord cannot slip at all? If the friction increases automatically, the knot is self-tightening and cannot fail under any

load. The cord breaks first. Whether the hitch fails or holds depends on its topology and on the friction provided by the wraparounds and wrapovers.

Consider a wraparound that is an integral multiple of a revolution. On one side the tension in the cord is large; call it T_2 . The tension is less on the other side; call it T_1 . The cord will not slip over the rod if T_2 does not exceed a certain multiple of T_1 . The multiple is an exponential function of two quantities: the angle (in radians) through which the cord is wrapped around the rod and the coefficient of friction between the cord and the rod. If the cord makes one revolution around the rod, the multiple is the exponential function of 2π radians multiplied by the coefficient of friction. For simplicity the function can be written as e . The cord will not slip if T_2 does not exceed $e \times T_1$. If T_2 is larger than this amount, the friction between the cord and the rod is overwhelmed and the cord slips.

If the cord is wrapped twice around the rod, the multiple is an exponential function: 4π times the coefficient of friction. A simple way to write the function is e^2 . The cord will not slip provided T_2 does not exceed $e^2 \times T_1$. This relation shows the advantage of an additional wraparound. T_2 can now be much larger than T_1 because the additional friction in the extra revolution of wraparound helps to hold the cord in place.

For any integral number of revolutions of the cord around the rod the relation between the tensions on the two sides of the wraparound is similar in form. Let n be the number of revolutions. The cord will not slip provided T_2 does not exceed $e^n \times T_1$.

The other method of gaining friction in a hitch is with a wrapover. Since the top part of a wrapover squeezes the bottom part against the rod, the bottom part resists being pulled through the wrapover to loosen the hitch. Consider the tensions on the two sides of the bottom section in a generalized wrapover. On the side with the large tension (T_2) the force tending to make the bottom part of the cord slide through the wrapover is equal to T_2 .

Opposing this tendency is the smaller tension (T_1) on the other side of the wrapover. An additional opposing factor is the friction provided by the squeeze from the wrapover. At most the friction can be a certain multiple (u) of the tension in the top part of the wrapover. The bottom section of cord does not slip through the wrapover if T_2 does not exceed $T_1 + uT$. If T_2 is too large, the friction from the pressure of the top section is not enough to hold the bottom section in place.

The value for u is difficult to calculate for a general situation because it involves the coefficients of friction between the rod and the bottom cord and



To: Gina
From: Bill
Subject: IBM Technology

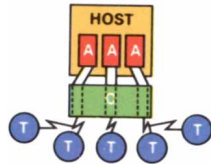
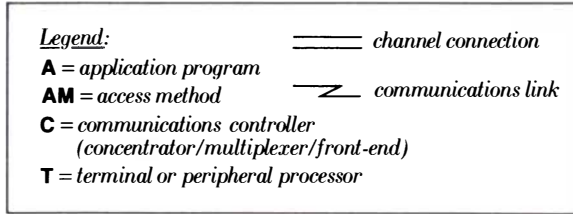
Here's the partial list I promised you of our past and present technological achievements. There are lots of things here that should be of real interest to the scientific, engineering and academic communities. What's your choice for the next topic in this series?

- Vacuum tube digital multiplier
- IBM 603/604 calculators
- Selective Sequence Electronic Calculator (SSEC)
- Tape drive vacuum column
- Naval Ordnance Research Calculator (NORC)
- Input/output channel
- IBM 608 transistor calculator
- FORTRAN
- RAMAC and disks
- First automated transistor production
- Chain and train printers
- Input/Output Control System (IOCS)
- STRETCH computer
- "Selectric" typewriter
- SABRE airline reservation system
- Removable disk pack
- Virtual machine concept
- Hypertape

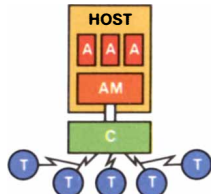
- System/360 compatible family
- Operating System/360
- Solid Logic Technology
- System/360 Model 67/Time-Sharing System
- One-transistor memory cell
- Cache memory
- Relational data base
- First all-monolithic main memory
- Thin-film recording head
- Floppy disk
- Tape group code recording
- Systems Network Architecture
- Federal cryptographic standard
- Laser/electrophotographic printer
- First 64K-bit chip mass production
- First E-beam direct-write chip production
- Thermal Conduction Module
- 288K-bit memory chip
- Robotic control language

*Bill -
SNA is becoming more
important every day.
Let's tell that story.
Gina*

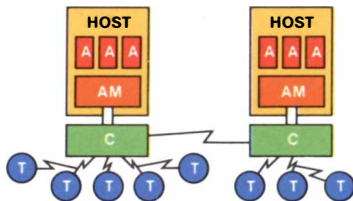
Figure 1. EVOLUTION OF SNA NETWORKS



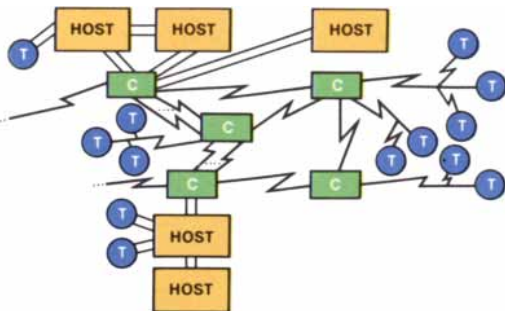
(a) In a typical pre-SNA network, communications links and terminals were dedicated to single uses or applications. All terminals on a link had to connect to the same application program, which included communications software. Usually, changing the terminal or link connections also forced the application programs to be changed.



(b) Early SNA introduced sharing of links among various application programs. A host access method permitted easy access from any terminal to any application program in the host processor. The connections could be readily changed without affecting the application programs.



(c) Subsequently, SNA configurations were enhanced to allow access between host processors for distributed processing and data-base sharing. Moreover, any terminal could access any application program at any host.



(d) Today, SNA networks can be fully meshed configurations. Parallel links between adjacent communications controllers allow increased network availability and traffic balancing. Access from host to host and terminal to host is permitted over multiple routes. The number of different types of network nodes has increased considerably, particularly among terminals and peripheral processors. SNA networks include open interconnection of both IBM and non-IBM nodes.

Advances in computing, processing and communications technologies have prompted increased interconnection of terminals, processors and communications facilities.

These various devices have been linked into networks for distributed access to processing and data-base resources.

A variety of networking applications has been developed for airline reservations, banking, store checkout, process control, remote job entry, office systems and personal computing.

Networks include a broad range of cost/function trade-offs and technologies, in such diverse components as analog/digital converters, specialized and general-purpose terminals, line concentrators and multiplexers, communications links and low- to high-capacity processors.

The networking environment requires a master interconnection strategy so that these diverse products and applications can share computational and communications facilities while interacting compatibly.

Since its introduction in 1974, IBM's Systems Network Architecture has provided the blueprint by which the capabilities of IBM networking products have evolved in an orderly fashion. SNA provides rules for all levels of interaction, from physical/electrical interconnection of computing devices and terminals to meaningful application-oriented processing.

Thus one uniform design now eliminates the complexity and inefficiencies inherent when each type of product had to have its own specialized agreement with each other type. SNA is now integrated into the whole range of IBM products — from large mainframe computers to terminals to personal computers.

By eliminating the chaos once caused by incompatible implementations, SNA allows a computer user to communicate from office to office or from continent to continent.

An important feature of SNA is the organization of functions into multiple layers. In the most basic sense, different products can be configured into networks simply by adapting them to the transmission and electrical characteristics of the media interconnecting them. But physical interconnection does not result in meaningful communication. The lower layers control only the basic transfer of bits, while the higher layers support meaningful exchange of messages and documents and allow application-

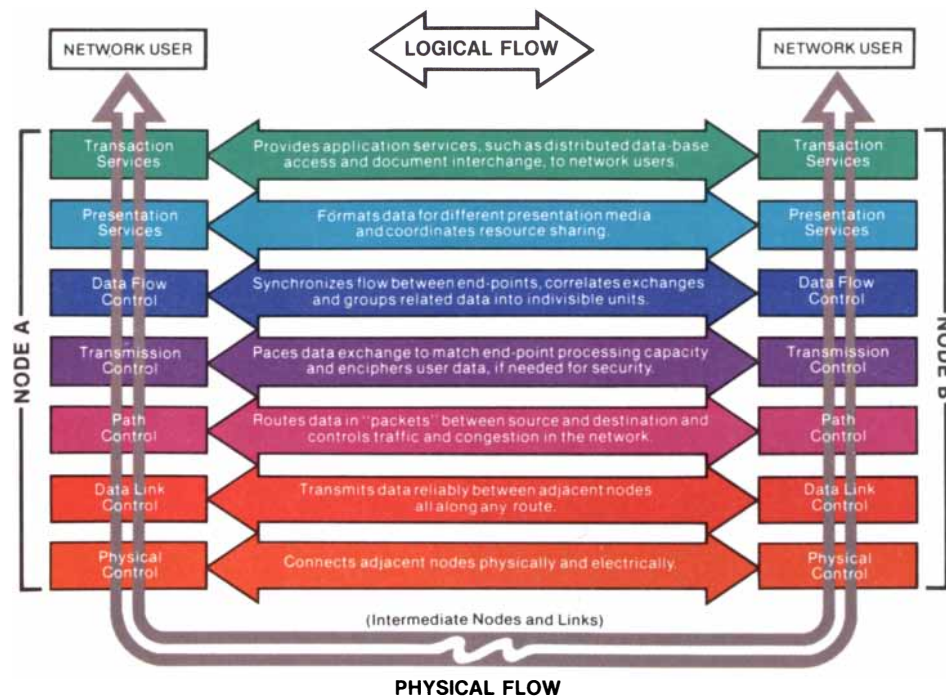


Figure 2. Each node in an SNA network separates functions into multiple layers. Logically, a given layer in one node communicates with the corresponding layer in another node. This peer-to-peer communication relies on lower layers to transport the data.

program interactions and data-base sharing. SNA's separation of independent functions into multiple layers means that changes in technology and capabilities can be confined to individual layers. This modular design eases adaptation to network evolution.

SNA includes a variety of functions at different layers of the architecture. For example, SNA's Synchronous Data Link Control offers increased efficiency over earlier techniques. State-of-the-art advances also have been made in traffic routing, congestion control and network availability. Additionally, SNA office systems provide document encoding uniformity and support distributed interchange, filing and retrieval services.

SNA has also incorporated protocols adopted by national and international standards organizations. This means SNA is compatible with standards such as X.25 public packet switching, High-Level Data Link Control and the Data Encryption Standard.

SNA management aids include product capabilities and software tools for planning, installing, changing, operating and maintain-

ing networks. In today's environment, where annual growth and change typically can involve 20-50% of a network's facilities, aids such as these are critical to reduce operational expense and to foster optimal levels of network availability and performance.

IBM scientists, programmers and engineers around the world have spent collectively thousands of years of development on SNA. They continue to improve SNA's usability, manageability and performance, and also to extend its capabilities. Recent studies have focused on local-area networking, more dynamic reconfiguration within networks and interconnection of independent SNA networks.

SNA's success in reducing customer cost, while promoting ease of development of network applications, is reflected by a recent milestone — more than 10,000 large-system installations now incorporate SNA networking technology.

Systems Network Architecture is one example of IBM's commitment to product and technological leadership. Last year IBM's total worldwide investment in research, development and engineering was \$3 billion.



For free additional information on SNA, please write:
 IBM Corporation, Dept. 605E/002
 P.O. Box 12195, Research Triangle Park, NC 27709

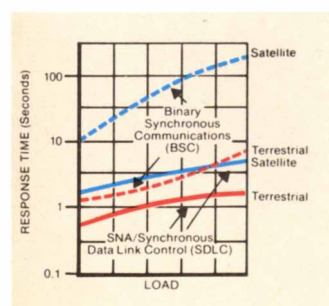
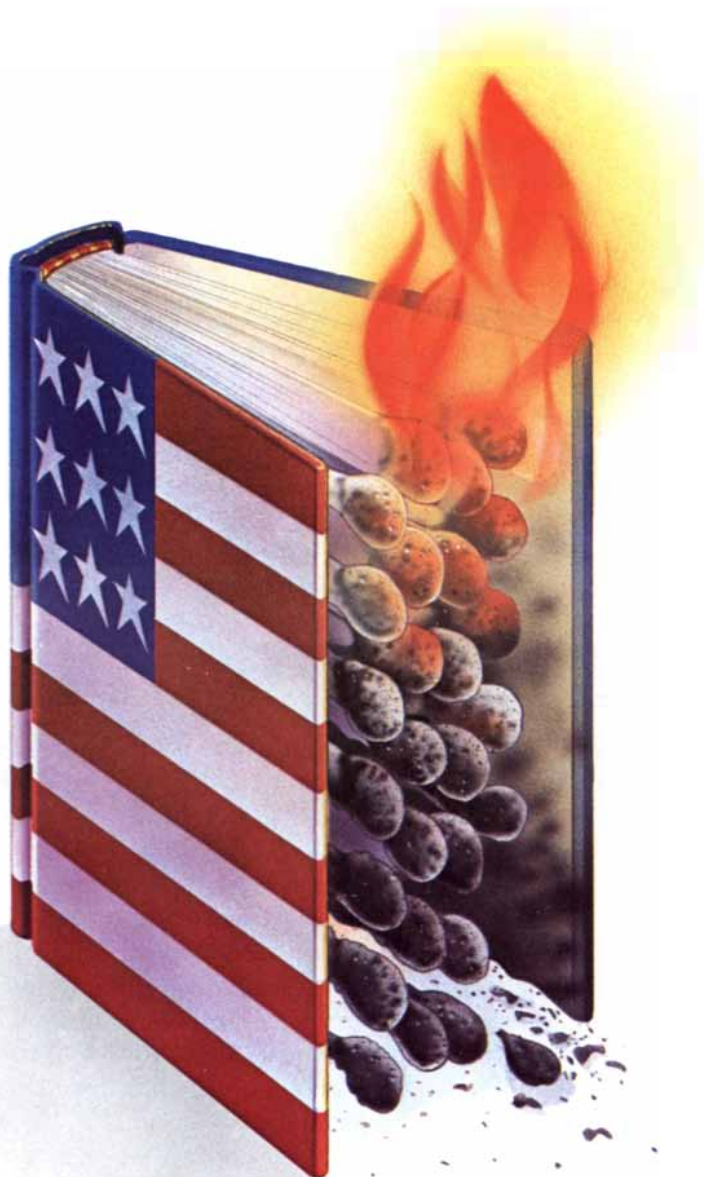


Figure 3. Illustration of dramatic improvements in response time (using comparable display terminals) of SNA/SDLC over older data link controls such as BSC. For long-propagation-delay circuits, such as in satellite technology, the improvements in response time can be better than an order of magnitude.



CENSORSHIP IN A FREE SOCIETY. IT'S A BAD MATCH.

Censorship is the greatest tragedy in American literature. It constricts the mind, teaches fear and leaves only ignorance and ashes.

Today, all over the country, books are being banned, burned and censored. Teachers, students, librarians, and book and magazine publishers are being harassed.

The attacks of these self-appointed censors are endorsed by our silence.

The freedom to read is one of our most precious rights. Do something to protect it.

Contact:

People For The American Way, P.O. Box 19900
Washington, D.C. 20036 or call 202/822-9450.



between the overlying sections of cord. It also depends on the diameter of the rod and cord, a geometry that affects the application of force by the top section to the bottom section. Bayman measured u and e for hitches of braided nylon string on a smooth steel rod; e was approximately 4, u about .2.

The clove hitch embodies both wraparounds and wrapovers. It is a common hitch whereby one can secure a cord to a fixture or to a heavier rope. From the free end the cord passes once around the rod and back over itself. Then it wraps once more around the rod before it emerges from under a second wrapover.

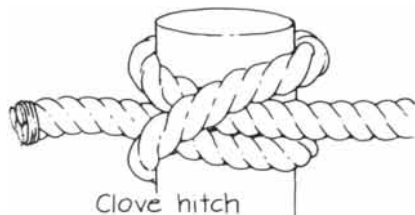
Sections of the cord are labeled in the bottom illustration at the left on the next page, beginning with 0 at the free end. Section 1 begins at the first wrapover and continues until the cord passes under the second wrapover. Section 2 emerges from the second wrapover and goes off to the load. (In application the hitch is pulled tighter than is shown.)

Is it a good hitch? Can the load on section 2 be arbitrarily large or will the load finally cause the free end to slip through the first wrapover and unravel the entire knot? Consider the bottom part of the first wrapover. On one side the free end has no tension. On the other side the cord has a tension (T_1) that acts to cause the free end to slide through the wrapover. The only opposing force is the friction generated by the top part of the wrapover where it presses the bottom part against the rod. If the hitch is to hold, T_1 cannot exceed the maximum possible value of this friction.

Next consider the two wraparounds of section 1. They do not slip if the tension in the part of the section near the second wrapover does not exceed $e^2 \times T_1$. Finally, consider the second wrapover. The cord will not slip under it if the tension (T_2) from the load does not overwhelm the opposing forces at the wrapover.

These relations of forces in the various parts of the hitch form a set of simultaneous inequalities that must be satisfied. The inequalities can be solved by a simple means of substitution, just as with simultaneous equations. The result is compact: all the conditions for avoiding slippage are satisfied if the value for u exceeds $1/e$. If this single condition is met, the load on the hitch can be arbitrarily large and the hitch will still hold. Additional load increases the pulls tending to loosen the knot but also increases the tensions that generate the opposing friction.

The first wrapover is the crucial element of the clove hitch. The force trying to pull the free end through the wrapover is equal to T_1 . The opposing force is at most equal to u times the tension in the cord passing over the top of the wrapover. That tension is easily related to T_1 because only one wraparound lies



Clove hitch



Constrictor knot



Two modified clove hitches



Modified constrictor knot



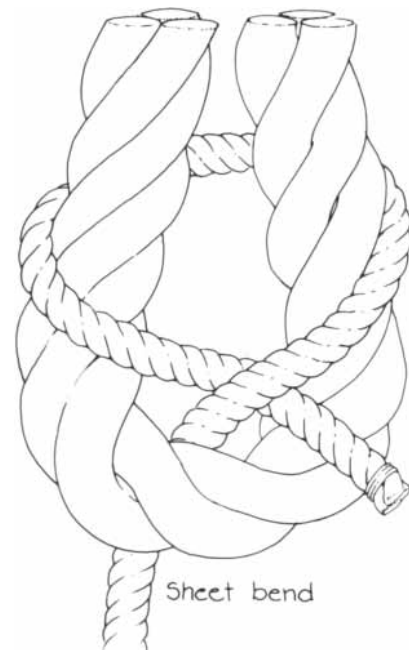
Double clove hitch



Fishhook knot



Groundline hitch



Sheet bend

between the two points in the cord. Since the tension in the top part is at most eT_1 , the friction at the wrapover is at most ueT_1 . If u exceeds $1/e$, the hitch holds any arbitrarily large load on the cord. An increase in the load merely provides more friction to lock the hitch in place.

Would additional wraparounds in the clove hitch make it hold better in the sense that it can hold an arbitrarily large load with a smaller value of u ? Two modified clove hitches are depicted in the top illustration on the opposite page. In the first hitch the cord is wrapped two more times around the rod before it crosses over the boundary between sections 0 and 1. In the second hitch the additional wraparounds are made just before the cord goes under its last wrapover and off to the load. These hitches are certainly more complex than the regular clove hitch, but are they stronger in the sense that they make less of a demand on the value of u ? Is one of the modified hitches better than the other?

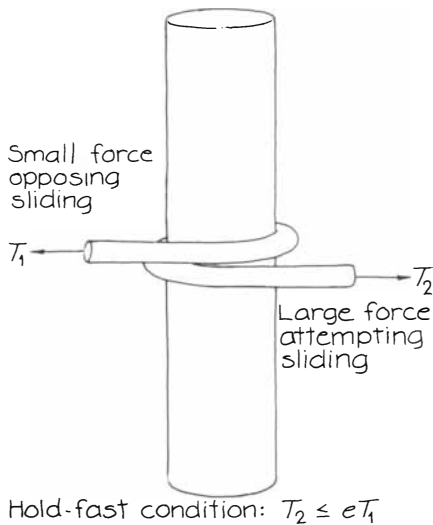
The key element in these modified clove hitches is the competition taking place near the free end. Consider the first knot. Section 1 is trying to make section 0 slide through the wrapover with tension T_1 . The top section of the wrapover is opposing the slide by pressing on the cord.

Does the cord slide? No, not if u is sufficiently large. Must it be as large as it is in the normal clove hitch? No, it can be considerably smaller because of the additional wraparounds. The tension in the top section of the first wrapover is now much larger than the force (T_1) tending to make the cord slide through the first wrapover. In the normal clove hitch the force from the wrapover can at most be ueT_1 . The (implicit) power of 1 on the e is from the single wraparound between the top part of the wrapover and the place where T_1 pulls on the bottom part. In the modified hitch the force from the wrapover can be as much as ue^3T_1 . Here the (explicit) power of 3 on

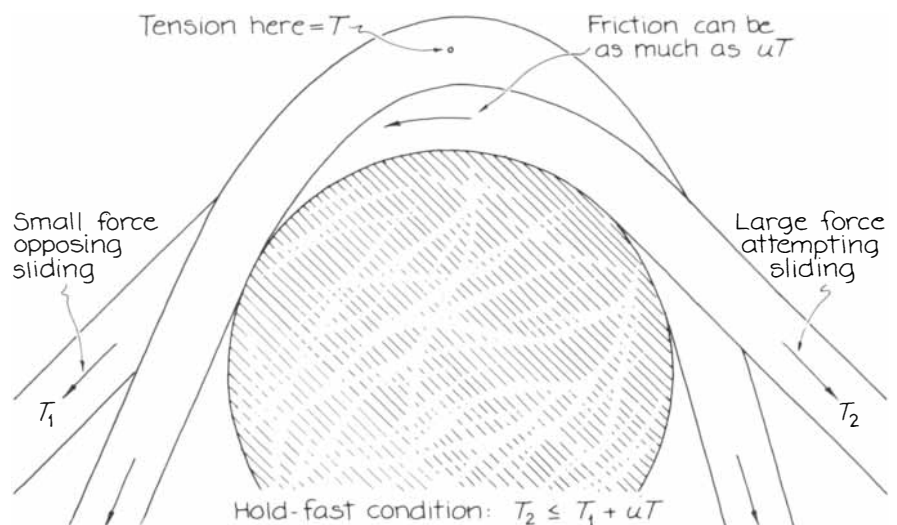
the e comes from the three wraparounds between the top and bottom parts of the wrapover. Hence the modified hitch stands a better chance of holding fast.

How about the second modification? It appears to be as complex as the first, but again the key element lies in securing the free end. This time the competition of pulls is precisely like that in a normal clove hitch. The extra wraparounds do not serve to secure the free end and so give no improvement over the normal clove hitch. They do give more friction on the cord, thereby reducing the value of T_1 when the cord is put under load, but the hitch still must have u in excess of $1/e$ (just as it does in the clove hitch) if it is to hold.

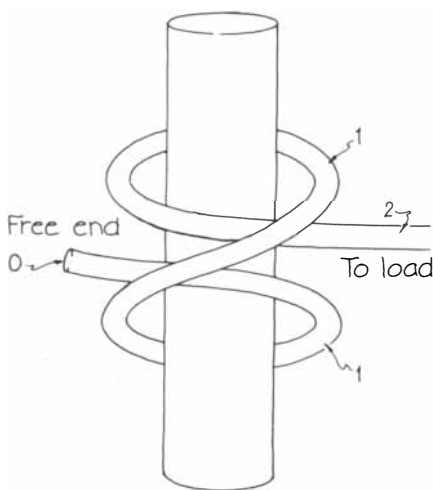
Something additional can be concluded from the first modified clove hitch. If it is to hold, the value of u must exceed $1/(e + e^2 + e^3)$. If u is larger than $1/(e + e^2)$, however, the first wraparound (section 1) serves no purpose and can be eliminated without jeopard-



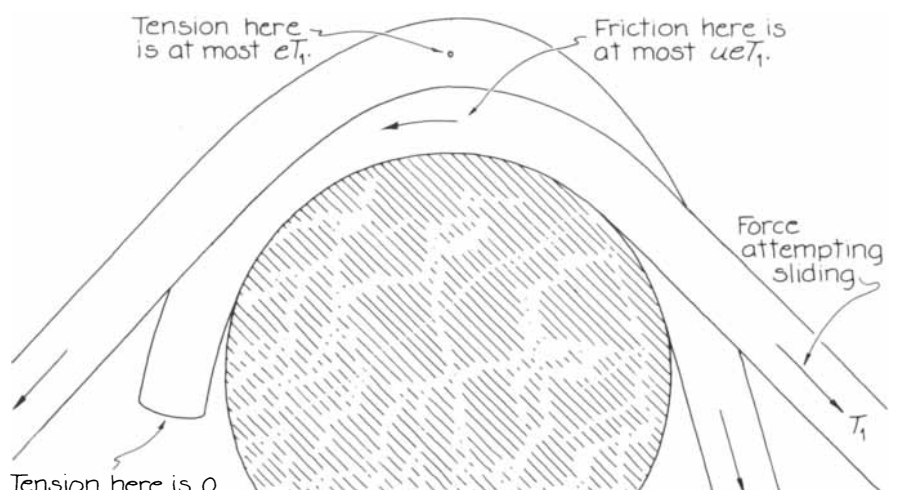
The forces in a wraparound



The forces in a wrapover



A clove hitch



The first wrapover of a clove hitch

dizing the integrity of the hitch. Section 0 might as well pass through a wrapover and become section 1. Furthermore, if u is larger than $1/e$, section 2 serves no purpose and can also be eliminated. Thus if u is larger than $1/e$, there is no reason to put extra wraparounds into a clove hitch because the hitch is already certain to hold fast.

Another possible modification is the double clove hitch depicted in the middle illustration at the left on this page. This hitch differs from the normal clove hitch because after the cord passes under the second wrapover it does not go directly to the load. Instead it makes another wrapover and another wrap-around and then passes under itself to go to the load.

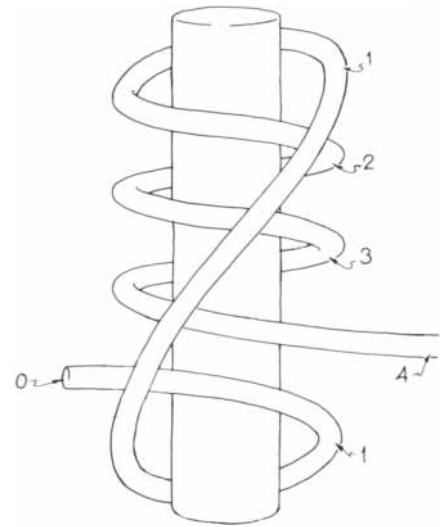
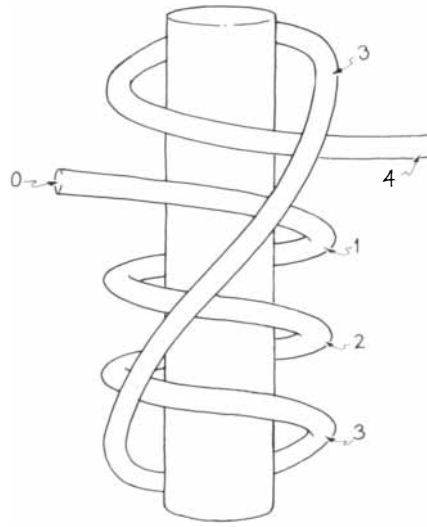
Does the extra hitching reduce the requirement on the value of u ? No, at least not with the assumptions I have made about the primary sources of friction in knots. If the double clove hitch is to hold, u must still exceed $1/e$. The extra hitching serves primarily as a safety feature. If section 0 slips through its wrapover, you still have a clove hitch that must be untied before the cord loosens completely from the rod.

A knot similar to the clove hitch is the groundline hitch shown in the middle illustration at the right on this page. It differs from the clove hitch in the way the early stages of the knot are wrapped. Section 0 passes under a wrapover made by a late section in the knot, that is, a section near the one under full load. Section 1 wraps once around the rod and then passes under itself to begin section 2, which wraps once around the rod before it passes under the last wrapover and begins section 3.

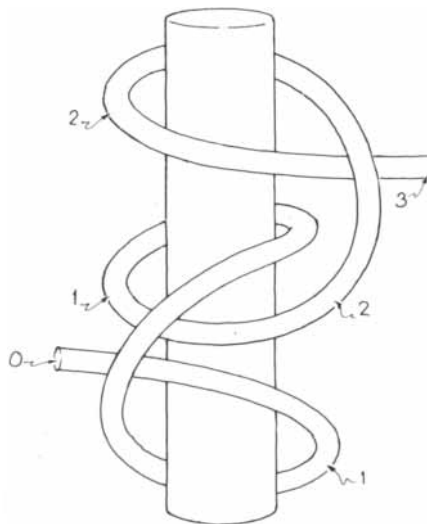
An analysis of the simultaneous inequalities indicates that this hitch holds fast if u and e meet the condition that $ue(u + e)$ exceed 1. Since e is greater than 1, the condition is met with a small value for u , smaller than the value of $1/e$ demanded by the clove hitch. Hence there is a range of values for u in which the groundline hitch holds but the clove hitch does not. If u is larger than $1/e$, both hitches serve equally well.

Why can the groundline hitch hold even when u is too small for the clove hitch to hold? The reason is found in the way the free end is tucked. In the clove hitch the free end is under the cord of section 1 after it has made one wrap-around. In the groundline hitch the free end is tucked under a section of the cord closer to the region of full load. Therefore the tension in the top part of the wrapover that fixes the free end is larger than it is in the clove hitch. In general a hitch is stronger (in the sense that less is demanded of u) when the free end is tucked in this way.

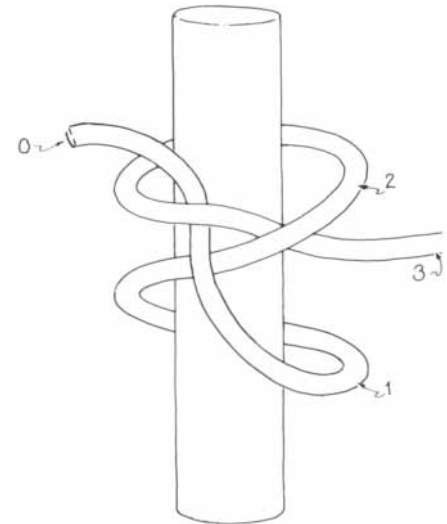
Bayman presents another example of how to tuck the free end beneficially. In the constrictor knot the free end of the



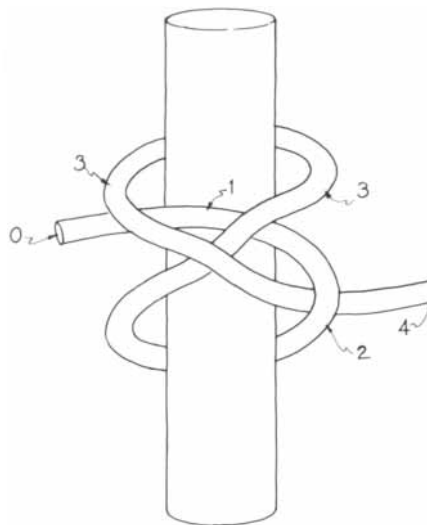
Modified clove hitches



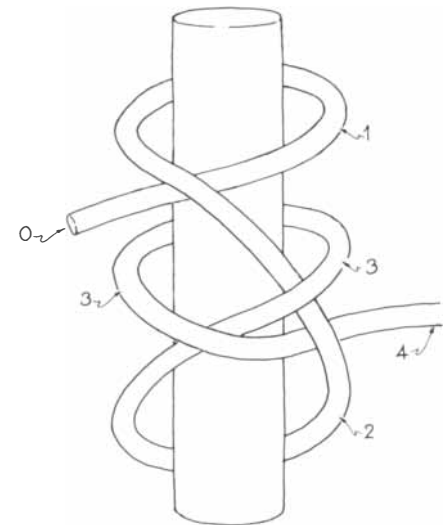
A double clove hitch



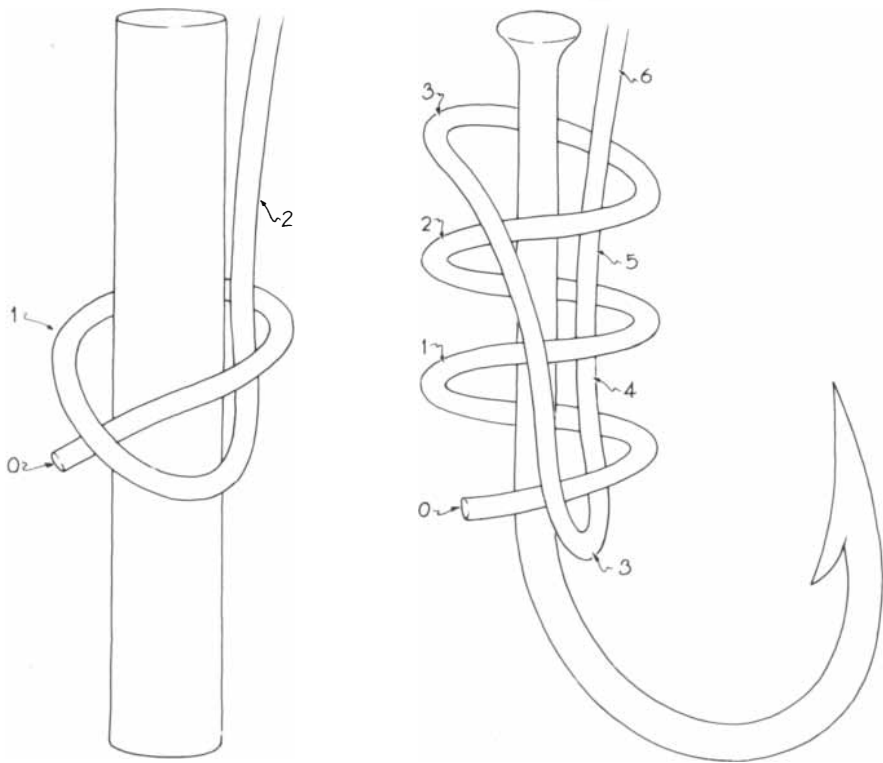
A groundline hitch



A constrictor knot



Benjamin F. Bayman's modified constrictor



Knots for a fishhook

cord runs under two wrapovers, each of which is under large tension. The solution of the simultaneous inequalities yields the lower limit of u if the hitch is to hold fast: it must be larger than $1/(e(2+e))$. Since this requirement is milder than the one for the clove hitch, the constrictor hitch is a better knot when you think the value for u may be low. If u is greater than $1/(2e)$, the friction in the wrapover marking the boundary between sections 1 and 2 is great enough so that the tension in section 1 can be zero and still the cord will not slip. If u is greater than $1/e$, section 2 is

no longer required and can even have zero tension without jeopardizing the hitch. Hence the full constrictor hitch serves well if u is small compared with $1/e$, but the full hitch is not needed for larger values of u .

Bayman considers a modification of the constrictor hitch in which section 1 is wrapped around the rod once before it tucks under itself. Compare this knot with the full constrictor hitch. They differ only in the tying of section 1. In the regular hitch section 1 tucks under a section that is subject to a large tension, and so it is secure. In the modification

section 1 gains some friction from the wraparound but then merely tucks under itself. Less tension presses down on the cord at the wrapover dividing sections 1 and 0.

This modified knot is certainly not as strong as the usual constrictor hitch. Is it better than a hitch in which section 1 is not tucked away at all? No, tucking section 1 under itself serves no purpose. In order for section 1 not to slip through the wrapover dividing it from section 2 the value of u must exceed $1/2e$.

Suppose it does not. Does the extra tie of section 1 save the hitch? No, because its wrapover on itself requires u to be larger than $1/e$ if the knot is not to slip. Therefore if the rest of the hitch begins to fail, the extra wraparound and wrapover of section 1 do not prevent the failure. You might as well leave them out, or better yet tie the full constrictor hitch.

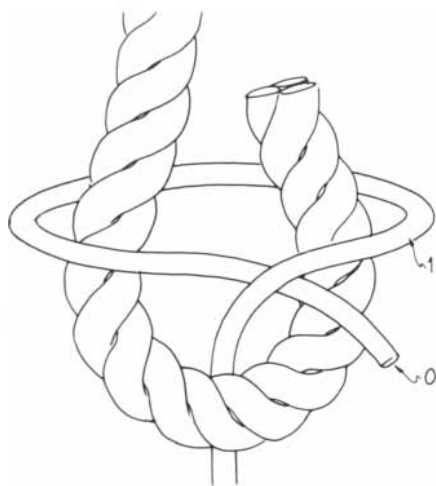
The top illustration on this page depicts a knot in which the cord leaves the knot parallel to the rod. This knot may be better than the clove hitch in some applications. I have not encountered a formal name for it; because it is similar to a knot for securing a fishing line to a straight hook (one without an eye the line can be tied to), I call it a fish-hook knot.

When I solve the simultaneous inequalities for this knot, I find that (within the assumptions I have been applying) the hook knot demands the same kind of value for u as the clove hitch. In both hitches section 1 tries to pull the free end through a wrapover. Also in both knots the tension in the top section of the wrapover is T_1 multiplied by the factor due to one wraparound and thus is at most eT_1 . If the cord is to remain in place, the value of u must exceed $1/e$. Although the two knots look different and may serve different functions in application, they are identical in their ability to withstand an arbitrarily large load.

If a cord or a fishing line is to be tied to a straight hook, the cord ought to be wrapped around the shaft of the hook several times before it is wormed through several wrapovers. How effective are the extra wraparounds? Does adding one more wraparound make a significant change in the demand on the value for u ?

When I solved the simultaneous inequalities for this knot, I found that the minimum value of u is driven downward sharply with each additional wraparound. For the knot shown the value for u must exceed $1/(e+e^2+e^3)$. If you eliminate one of the wraparounds, the term e^3 drops out of the formula. Since e is larger than 1, the loss of one wraparound can have a significant effect.

The bottom illustration at the left on this page shows a sheet-bend hitch designed for joining a cord of small diameter to a larger rope. The rope forms a U . From the free end the cord wraps



A sheet-bend hitch

	$j=1$	$j=2$	$j=3$
$i=1$	1	0	$-ue$
$i=2$	-1	1	$-u$
$i=3$	0	$-e$	$1-ue$

Determinant = $1-ue(2+e)=0$

$u = \frac{1}{e(2+e)}$ for critical case

Constrictor-knot determinant

around the full U once and then back over itself before it passes through the bottom of the U and off to the load.

The "lock" on this knot is in the wrapover. In the top part the cord has a large tension, almost as large as in the section under full load. The bottom part of the wrapover is therefore pressed hard against the rope. If the wrapover fails, the entire sheet bend fails.

To analyze the sheet bend's friction requirement I make several assumptions. First I assume the rope is fairly stiff so that in the wrapover the bottom part can be squeezed between the rope and the top part. I also assume the wrap-around on the U is identical with one around a rod of large diameter. Within the limits of these assumptions the requirement on u for the cord to hold is that it must exceed $1/e$, just as with a clove hitch. The sheet bend is in the same class as a clove hitch.

The sheet bend can be improved by wrapping the cord around the U more times before making a wrapover and passing the cord through the bottom of the U . Each added wraparound increases the effectiveness of the wrapover by widening the difference in tension between the top and bottom parts of the wrapover.

All the knots I have discussed can be investigated by solving simultaneous inequalities. Bayman has devised a quicker solution in which a determinant is constructed from the coefficients in the inequality statements. If the quantities of e and u lie at the crucial values where the cord is on the verge of slipping, the simultaneous statements are equations and not inequalities. Hence at the crucial values the determinant is equal to zero. Establish it at zero and then expand it in the usual way. The resulting equation is solved for u . This value of u must be exceeded if the knot is to hold under load.

The first step in constructing a determinant is to section the hitch for analysis. The free end is labeled 0 as before. Each of the succeeding sections is defined as a length of cord that begins at the bottom of a wrapover and extends until it must pass under another one. The determinant is laid out on a chart labeled with i vertically and j horizontally. These letters represent the various sections in the hitch. Neither section 0 nor the last section (the one under the full load) is included.

The determinant is filled in by horizontal rows. Within a row the spaces under the values of j are determined through three rules: (1) If j matches i , the space has a term of 1; (2) if j is the section where i begins, the space has a term of $-ue^n$, where n is the number of wraparounds j makes before it passes over the beginning of i ; (3) if j is the section just before i in the hitch, the space has a term of $-ue^n$, where n represents the

total number of wraparounds the previous section has.

The determinant for the full constrictor hitch is worked out in the bottom illustration at the right on the opposite page. The top row is for section 1, the middle row for section 2 and the bottom row for section 3. In the top row j and i match in the first space, and so a 1 is inserted. The other rules do not contribute any further terms to the space. They also do not contribute anything to the space below $j = 2$, which is left at zero.

The space below $j = 3$ does have a contribution because section 1 begins under section 3. Since section 3 has been through one full wraparound when it crosses over the beginning of section 1, the value of n in the term contributed by the second rule is 1.

The second row pertains to section 2 of the hitch. The space below $j = 2$ has a value of 1 because of the match of i and j . The space below $j = 1$ deals with the section of the hitch just ahead of section 2. Section 1 had no full wraparounds, and so the term from rule 3 that is to be filled in is $-e^0$, or -1 .

The space below $j = 3$ relates to the beginning of section 2. Since it begins under section 3 before that section has made a wraparound, the contribution to the determinant is $-ue^0$, a term that reduces to $-u$.

The bottom row in the determinant pertains to section 3 of the hitch. The space below $j = 1$ has no contribution. The space below $j = 2$ has $-e$ because section 3 is preceded by section 2, which has one wraparound. The space below $j = 3$ has two contributions based on the three rules. The match of j and i contributes a term of 1. A second term is contributed because section 3 begins under itself after it has made one wraparound. This second term is $-ue^1$, which simplifies to $-ue$.

Once the determinant is constructed it is set equal to zero and solved by multiplying terms along diagonals or by reducing the determinant to smaller determinants. The resulting equation involves both u and e . The solution for u in the equation is the critical value for u . If u is smaller than this value, the cord slips because the wrapovers will not hold under tension. If u is larger than the critical value, the cord can be expected to hold under load.

A great deal more can be done on the analysis of hitches and other knots. Besides investigating new knots or improving old ones you might like to strengthen the theoretical analysis. For example, Bayman points out that not all wrapovers are identical in their ability to squeeze an underlying section of cord, particularly when one section of cord forms the top part of several adjacent wrapovers. If you do strengthen the theory or if you find interesting new knots, I would enjoy hearing from you.



**Better Than
Jogging,
Swimming
or Cycling**

NordicTrack
**Jarless Total Body
Cardiovascular Exerciser
Duplicates X-C Skiing For The
Best Way To Fitness**

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for **FREE BROCHURE**
Toll Free 1-800-328-5888 MN 612-448-6987
PSI 124 F Columbia Crt., Chaska, Minn. 55318

SCIENTIFIC AMERICAN

is now available
to the blind and
physically handi-
capped on cassette
tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

MATHEMATICAL GAMES

- ON WELL-QUASI-ORDERING FINITE TREES. C. St. J. A. Nash-Williams in *Proceedings of the Cambridge Philosophical Society*, Vol. 59, Part 4, pages 833-835; October, 1963.
- NUMBER THEORY: THE THEORY OF PARTITIONS. George E. Andrews. Addison-Wesley Publishing Co., 1976.
- TREES AND BALL GAMES. Raymond M. Smullyan in *Annals of the New York Academy of Sciences*, Vol. 321, pages 86-90; 1979.
- PROVING TERMINATION WITH MULTISSET ORDERINGS. Nachum Dershowitz and Zohar Manna in *Communications of the ACM*, Vol. 22, No. 8, pages 465-476; August, 1979.
- ACCESSIBLE INDEPENDENCE RESULTS FOR PEANO ARITHMETIC. Laurie Kirby and Jeff Paris in *The Bulletin of the London Mathematical Society*, Vol. 14, No. 49, Part 4, pages 285-293; July, 1983.

TRAUMA

- THE HEALING HAND: MAN AND WOUND IN THE ANCIENT WORLD. Guido Majno. Harvard University Press, 1975.
- THE EPIDEMIOLOGY AND PREVENTION OF INJURIES. Susan P. Baker and P. E. Dietz in *The Management of Trauma*, edited by George D. Zuidema, Robert B. Rutherford and Walter F. Ballinger II. W. B. Saunders Company, 1979.
- INJURY CONTROL. W. Haddon and S. P. Baker in *Preventive and Community Medicine*. Duncan W. Clark and Brian MacMahon. Little, Brown & Co., 1981.
- TRAUMA MANAGEMENT, VOL. 1: ABDOMINAL TRAUMA. Edited by F. William Blaisdell and Donald D. Trunkey. Thieme-Stratton, Inc., New York, 1982.

THE PURIFICATION AND MANUFACTURE OF HUMAN INTERFERONS

- VIRUS INTERFERENCE, 1: THE INTERFERON. A. Isaacs and J. Lindenmann in *Proceedings of the Royal Society of London, Series B*, Vol. 147, No. 927, pages 258-267; September 12, 1957.
- THE INTERFERON SYSTEM. Edited by W. E. Stewart. Springer-Verlag, 1979.
- INTERFERONS. Edited by S. Pestka in *Methods in Enzymology*, Part A, Vol. 78; Part B, Vol. 79; 1981.

- AMINO ACID SEQUENCE OF A HUMAN LEUKOCYTE INTERFERON. Warren P. Levy, Menachem Rubinstein, John Shively, Ursino Del Valle, Chen-Yen Lai, John Moschera, Larry Brink, Louise Gerber, Stanley Stein and Sidney Pestka in *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 78, No. 10, pages 6186-6190; October, 1981.
- THE INTERFERON SYSTEM: A REVIEW TO 1982. Samuel Baron, Ferdinando Dianzani and G. John Stanton in *Texas Reports on Biology and Medicine*, Vol. 41, Part 1, Part 2; 1982.
- THE HUMAN INTERFERONS—FROM PROTEIN PURIFICATION AND SEQUENCE TO CLONING AND EXPRESSION IN BACTERIA: BEFORE, BETWEEN, AND BEYOND. Sidney Pestka in *Archives of Biochemistry and Biophysics*, Vol. 221, No. 1, pages 1-37; February 15, 1983.

MAGNETIC FIELDS IN THE COSMOS

- THE SOLAR WIND. E. N. Parker in *Scientific American*, Vol. 210, No. 4, pages 66-76; April, 1964.
- COSMICAL MAGNETIC FIELDS: THEIR ORIGIN AND THEIR ACTIVITY. E. N. Parker. Oxford University Press, 1979.
- MAGNETIC STARS. E. F. Borra, J. D. Landstreet and L. Mestel in *Annual Review of Astronomy and Astrophysics*, Vol. 20, pages 191-220; 1982.

INTERSTELLAR MATTER IN METEORITES

- ISOTOPIC ANOMALIES IN METEORITES. F. Beegemann in *Reports on Progress in Physics*, Vol. 43, Part 4, pages 1309-1356; 1980.
- COMPOUNDS IN METEORITES AND THEIR ORIGINS. R. Hayatsu and E. Anders in *Topics in Current Chemistry*, Vol. 99, pages 1-34; 1981.
- NOBLE GASES IN METEORITES: EVIDENCE FOR PRESOLAR MATTER AND SUPERHEAVY ELEMENTS. E. Anders in *Proceedings of the Royal Society of London, Series A*, Vol. 374, No. 1757, pages 207-238; February 4, 1981.
- INTERSTELLAR CARBON IN METEORITES. P. Swart, M. Grady, C. Pillinger, R. Lewis and E. Anders in *Science*, in press.

THE CHEMICAL DEFENSES OF TERMITES

- CHEMICAL DEFENSE BY TERMITE SOLDIERS. Glenn D. Prestwich in *Journal of Chemical Ecology*, Vol. 5, pages 459-480; 1979.
- CHEMISTRY, DEFENSE AND SURVIVAL: CASE STUDIES AND SELECTED TOPICS.

Thomas Eisner in *Insect Biology in the Future*, edited by Michael Locke and David S. Smith. Academic Press, 1980.

CHEMICAL SELF-DEFENSE BY TERMITE WORKERS: PREVENTION OF AUTOTOXICATION IN TWO RHINOTERMINIDS. Stephen G. Spanton and Glenn D. Prestwich in *Science*, in press.

RATIONAL COLLECTIVE CHOICE

- SOCIAL CHOICE AND INDIVIDUAL VALUES. Kenneth J. Arrow. John Wiley & Sons, Inc., 1963.
- AGGREGATION OF PREFERENCES. Donald J. Brown in *Quarterly Journal of Economics*, Vol. 89, No. 3, pages 456-469; August, 1975.
- CHOICE, WELFARE AND MEASUREMENT. Amartya Sen. The MIT Press, 1982.
- ACYCLIC COLLECTIVE CHOICE RULES. Douglas H. Blair and Robert A. Pollak in *Econometrica*, Vol. 50, No. 4, pages 931-943; July, 1982.

THE STAVE CHURCHES OF NORWAY

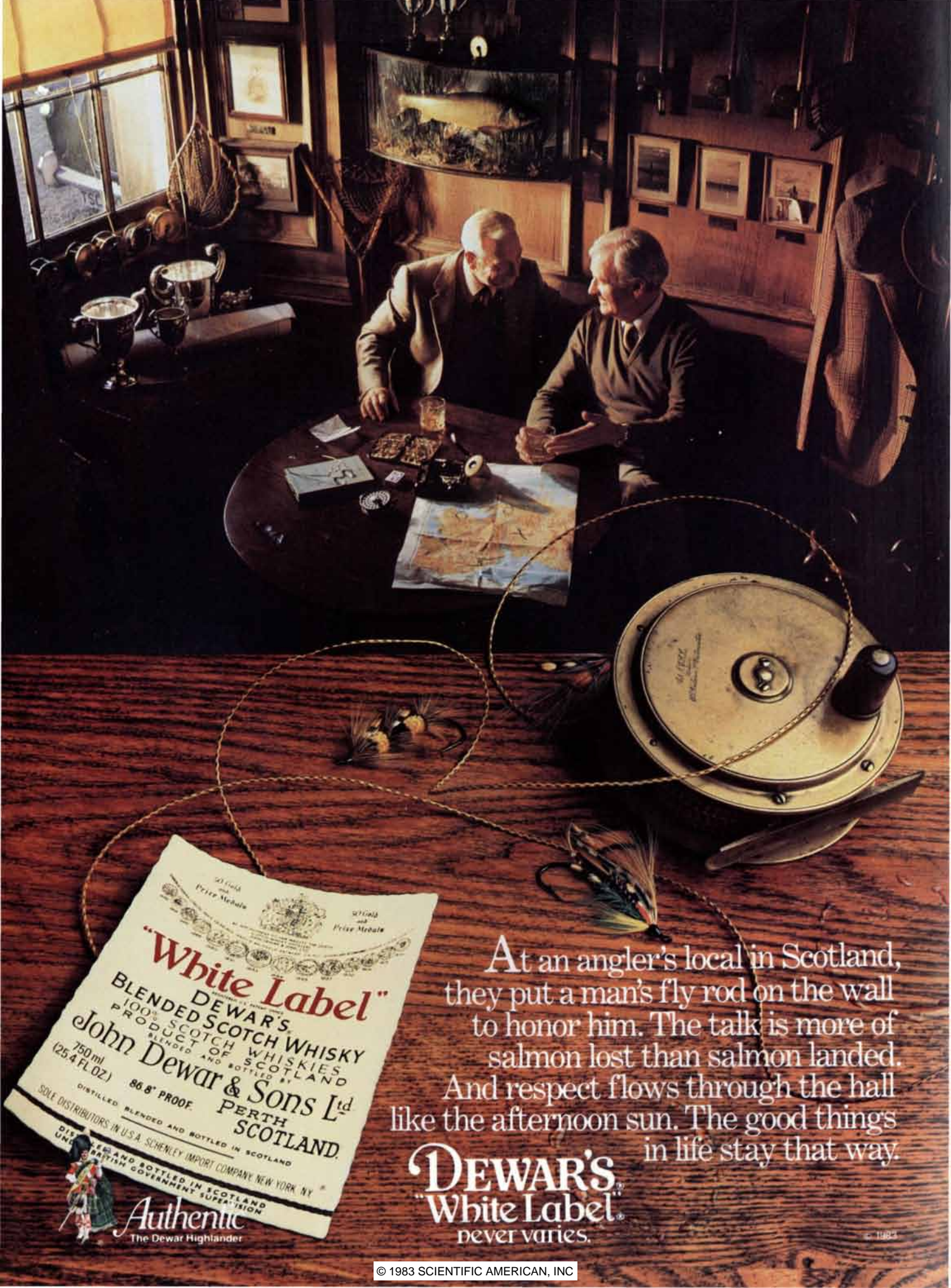
- STAV OG LAFT I NORGE/EARLY WOODEN ARCHITECTURE IN NORWAY. Gunnar Bugge and Christian Norborg-Schulz. Byggekunst, Norske Arkitekters Landsforbund, Oslo, 1969.
- NORWEGIAN STAVE CHURCHES. Roar Hauglid. English text by R. I. Christophersen. Dreyer, Oslo, 1970.
- THE STAVE CHURCHES OF KAUPANGER. Kristian Bjerkens and Hans-Emil Lidén. Fabritius Forlag, Oslo, 1975.
- NORSKE STAVKIRKER: BYGNINGSHISTORISK BAKGRUNN OG UTVIKLING. Roar Hauglid. Dreyer, Oslo, 1976.

DIGITAL TYPOGRAPHY

- AN ATLAS OF TYPEFORMS. James Sutton and Alan Bartram. Hastings House, 1968.
- TRANSMISSION AND DISPLAY OF PICTORIAL INFORMATION. D. E. Pearson. Pentech Press, Ltd., 1975.
- ANCIENT WRITING AND ITS INFLUENCE. Berthold Louis Ullman. University of Toronto Press, 1980.
- CHARACTER GENERATION BY COMPUTER. Ph. Coueignoux in *Computer Graphics and Image Processing*, Vol. 16, pages 240-269; 1981.
- THE CONCEPT OF A META-FONT. Donald E. Knuth in *Visible Language*, Vol. 16, No. 1, pages 3-27; Winter, 1982.
- DIGITAL TYPEFACE DESIGNS. Charles Bigelow. Seybold Publications, 1983.

THE AMATEUR SCIENTIST

- THEORY OF HITCHES. Benjamin F. Bayman in *American Journal of Physics*, Vol. 45, No. 2, pages 185-190; February, 1977.



50 Gold and
 Prize Medals
 50 Gold and
 Prize Medals
"White Label"
 DEWAR'S
 BLENDED SCOTCH WHISKY
 100% SCOTCH WHISKIES
 PRODUCT OF SCOTLAND
 BLENDED AND BOTTLED BY
John Dewar & Sons Ltd
 PERTH
 SCOTLAND.
 750 ml
 (25.4 FL OZ)
 86.8° PROOF
 DISTILLED, BLENDED AND BOTTLED IN SCOTLAND
 SOLE DISTRIBUTORS IN U.S.A. SCHENLEY IMPORT COMPANY, NEW YORK, N.Y.
 DISTILLED AND BOTTLED IN SCOTLAND
 UNDER SUPERVISION
 OF THE
 BRITISH GOVERNMENT
Authentic
 The Dewar Highlander

At an angler's local in Scotland,
 they put a man's fly rod on the wall
 to honor him. The talk is more of
 salmon lost than salmon landed.
 And respect flows through the hall
 like the afternoon sun. The good things
 in life stay that way.

DEWAR'S
"White Label"
 never varies.

© 1983

Tomorrow's Window



on the Universe

Color displays are coming of age

The cathode ray tube (CRT) is no longer just a vehicle for television broadcasts. It's an indispensable tool for highlighting scientific and business data, providing colorful, three-dimensional graphics for industry and home entertainment. Such sophisticated applications require screens with much higher resolution. Moreover, future color displays will have to be thinner, to conserve space and suit modern designs, and smaller, to enable the development of truly portable, full-function computers and a myriad of compact visual devices.

Hitachi probes the limits of resolution

To meet the need for high-resolution color display, Hitachi engineers have come up with a patented dry process for applying red, green and blue color phosphors to the CRT screen. This innovative process produces highly accurate patterns of phosphor dots, each one just 95 microns in diameter, as opposed to 120 microns in the conventional slurry process. In fact, the color images it yields are so sharp that graphically designed dot sequences are almost indistinguishable from continuous lines and curves — an especially important feature in computer-aided design.

Flattening the screen of the future

Concurrently, Hitachi has begun working on new technologies to allow production of televisions so thin you can hang them on the wall like pictures and portable color monitors you can easily carry from room to room. One of these technologies involves a color plasma display using gas discharged in a tiny cell to illuminate phosphors, instead of using the usual electron gun. It could be the key to both small and large screens just 5 cm thin. Other display processes under development include bichromatic liquid crystal displays (LCDs), electroluminescents, light-emitting diodes (LEDs), and fluorescent display tubes — all aimed at turning the color display unit into a real window on the universe.

Technical excellence is embodied in all Hitachi products

The development of improved color displays is just one case demonstrating Hitachi's technological strength. You'll find other examples in all of Hitachi's products, from complete CRT units and dot matrix LCDs to a whole host of high-quality electronic appliances. Our comprehensive technological expertise is your guarantee of convenience, easy operation and high reliability in every product that bears the Hitachi brand.



A World Leader in Technology

Inquiries to: Hitachi America, Ltd., Chicago Office, Electron Tube Sales & Service Div.
500 Park Boulevard, Suite 805, Itasca, Ill. 60143 Tel: (312) 773-0700



OBVIOUS LUXURY...SURPRISING PERFORMANCE

Introducing the luxury sedan that was born to perform. The magnificent Datsun Maxima boasts a fuel-injected 2.4 liter six cylinder engine that merges surprising power with unsurpassed elegance.

Inside, virtually every conceivable creature comfort: cruise control on both the 5-speed and automatic, eight speaker stereo including cassette

deck with Dolby, air conditioning, power windows, mirrors and door locks. A voice will tell you when your "Right door is open"—one of six vocal reminders. All this is standard.

The optional sunroof is power driven. The optional leather package includes a new digital control panel that monitors the engine's functions electronically. The Datsun Maxima is available in

four-door sedan or wagon, gas or diesel. Compare it feature by feature with any other car in its luxury class. Compare the performance. Then compare the price. Most importantly of all, remember it's built by the Nissan Motor Company, Ltd.—whose name stands for quality world-wide.

NISSAN
WE ARE DRIVEN
DATSUN

