

# SCIENTIFIC AMERICAN



TOPOLOGY OF MIRAGES

**\$2.50**

*June 1985*

# THE FIRST DIESEL SENSES INSTEAD



It has been snickered, over the years, that the average diesel runs so quietly, you can almost hear a piano drop.

Then there's the one about diesels having three speeds: slow, very slow and reverse.

In all fairness, though, the early diesels were created to help us through our gas

crises. Not to thrill us with their performance.

That is a mission reserved solely for the extraordinary machine you see here: the new BMW 524 turbo diesel.

An automobile whose technology has advanced the diesel beyond limits even some of the most forward-thinking engi-

neers thought unapproachable.

"At idle... the BMW doesn't sound as if it's going to drop its crankshaft," writes Car and Driver, "and the six cylinders turn over with very nearly the legendary smoothness of BMW's gasoline sixes."

No longer will starting your car in the

\*EPA-estimated figures are for comparison purposes only. Your actual mileage may vary, depending on speed, weather and trip length; actual highway mileage will most likely be lower. ©1985 BMW of North America, Inc. The BMW trademark and logo are registered. European Delivery can be arranged through your authorized U.S. BMW dealer.

# EL TO EXCITE THE OF OFFEND THEM.



morning become a neighborhood event.

"Compared to the initial snail-like acceleration of the Mercedes and Audi turbodiesels, the BMW virtually shoots off the line," continues Car and Driver magazine

The 524td is, in fact, the quickest diesel ever to grace our roads.

And "grace" is a key word here. For the 524td not only emulates the thrust and turbine-like sound of a BMW, but displays the same extraordinary agility as well.

In truth, it displays conventional diesel-like performance in only one area. It attains 24 mpg around town and 30 mpg on the

highway.\* In other words, as much as, or more than, many of its slower brethren.

The major difference being that, in the process, the BMW 524td will be quickening your pulse. As opposed to grating on your nerves.

**THE ULTIMATE DRIVING MACHINE.**





Thank Dad for believing you were very special  
every step of the way.



Send a gift of Johnnie Walker Black Label anywhere in the U.S.A. Call 1-800-243-3787. Void where prohibited.

12 YEAR OLD BLENDED SCOTCH WHISKY 86.8 PROOF. BOTTLED IN SCOTLAND. IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y., N.Y. © 1985.

## ARTICLES

- 37 THE CHOICE OF TECHNOLOGY, by Wassily Leontief**  
Managers have a powerful new analytical method for making investment decisions about technology.
- 46 THE IMMUNOLOGIC FUNCTION OF SKIN, by Richard L. Edelson and Joseph M. Fink**  
There are specialized cells in the epidermis that present foreign antigens to lymphocytes in the skin.
- 54 THE SEARCH FOR PROTON DECAY, by J. M. LoSecco, Frederick Reines and Daniel Sinclair** Is matter immortal? Theory says it is not, but so far no protons have been seen to decay.
- 78 GLOBULAR CLUSTERS, by Ivan R. King**  
Dense throngs of ancient stars, they tell much about stellar evolution and the history of the universe.
- 90 THE FIRST ORGANISMS, by A. G. Cairns-Smith**  
Not primordial soup but clay, it is argued, provided the fundamental materials from which life came.
- 102 THE SOCIAL ECOLOGY OF CHIMPANZEES, by Michael P. Ghiglieri**  
A uniquely flexible social order enables chimp society to adjust to changes in the abundance of food.
- 114 SIPHONS IN ROMAN AQUEDUCTS, by A. Trevor Hodge**  
The siphon formed a key element in the water-supply systems that made Roman urbanization possible.
- 120 THE TOPOLOGY OF MIRAGES, by Walter Tape**  
The distortions that create mirages can be analyzed topologically, without reference to atmospheric.

## DEPARTMENTS

- 6 LETTERS**
- 11 50 AND 100 YEARS AGO**
- 14 THE AUTHORS**
- 18 COMPUTER RECREATIONS**
- 30 BOOKS**
- 64 SCIENCE AND THE CITIZEN**
- 130 THE AMATEUR SCIENTIST**
- 136 BIBLIOGRAPHY**

PRESIDENT AND EDITOR	Jonathan Piel
BOARD OF EDITORS	Armand Schwab, Jr. (Associate Editor), Timothy Appenzeller, John M. Benditt, Peter G. Brown, Ari W. Epstein, Michael Feirtag, Robert Kunzig, Philip Morrison (Book Editor), James T. Rogers, 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 Eisler and Leo J. Petruzzi (Assistants to the Production Manager), Carol Hansen (Electronic Composition Manager), Carol Albert, Karen Friedman, Gary Pierce, William Sherman, 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
CHAIRMAN	Gerard Piel
EDITOR EMERITUS	Dennis Flanagan

# SCIENTIFIC AMERICAN

## CORRESPONDENCE

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

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

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

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

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

Name \_\_\_\_\_

New Address \_\_\_\_\_

Street \_\_\_\_\_

City \_\_\_\_\_

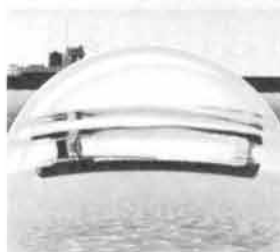
State and ZIP \_\_\_\_\_

Old Address \_\_\_\_\_

Street \_\_\_\_\_

City \_\_\_\_\_

State and ZIP \_\_\_\_\_



### THE COVER

The painting on the cover depicts one stage in visualizing the topological relation between an object's undistorted appearance and its mirage image. A mirage of a Great Lakes ore-carrying vessel is shown on the surface of an imaginary flexible "image sphere." The freighter's undistorted image lies in the background, displayed on a much larger "object sphere." The relation between the two is found in a topological operation called transfer mapping (see "The Topology of Mirages," by Walter Tape, page 120). In a transfer mapping the image sphere distorts, so that the image on its surface appears undistorted: the image sphere expands to meet the object sphere, and as it expands it wrinkles and folds. Eventually the image on its surface matches the undistorted image on the surface of the object sphere. In the example shown the image sphere would fold in several places: sections that bear images of the same part of the ship fold on top of one another. The scene on the cover could never actually be observed. The observer would not see both the image sphere and the object sphere; he would see only the mirage.

### THE ILLUSTRATIONS

Cover painting by Hank Iken

Page	Source	Page	Source
11	SCIENTIFIC AMERICAN	86	Stanislav Djorgovski, University of California at Berkeley
18-28	Andrew Christie		
31	<i>The Art of Describing.</i> © 1983, University of Chicago	88	Ivan R. King
		91	W. D. Keller, University of Missouri at Columbia ( <i>left</i> ); David W. Houseknecht, University of Missouri at Columbia ( <i>right</i> )
33	Jean Andrews Smith	92-97	George V. Kelvin
34	<i>The Astronomical Scrapbook.</i> © 1984, Sky Publishing Corporation and Cambridge University Press	98	W. J. McHardy, Macaulay Institute for Soil Research ( <i>top</i> ); W. D. Keller ( <i>bottom</i> )
38-45	Jerome Kuhl	99	W. D. Keller
47	Tung Tien-Sun, New York University Medical Center	100	Naganori Yoshinaga, Ehime University
48-50	Ilil Arbel	103	Constance S. Ghiglieri
51-52	Richard L. Edelson, Columbia University College of Physicians and Surgeons	104-112	Patricia J. Wynne
53	Ilil Arbel	115	A. Trevor Hodge, Carleton University, Ottawa
55	Frederick Reines, University of California at Irvine	116-118	Tom Prentiss
		119	A. Trevor Hodge
56-60	Gabor Kiss	120-121	Walter Tape, University of Alaska at Fairbanks
62	Edward Bell	122-123	Hank Iken, Walken Graphics
78	© 1983 California Institute of Technology ( <i>top</i> ); Ivan R. King, University of California at Berkeley ( <i>bottom</i> )	127	Walter Tape ( <i>top</i> ); Hank Iken, Walken Graphics ( <i>bottom</i> )
80-84	Ian Worpole	128	Walter Tape
		130-134	Michael Goodman

# Can you find the ITT computer?



You probably went straight to the ITT XTRA™ Personal Computer. The picture that *looks* like a computer.

But actually, there's a computer in each of these pictures.

The car, for example. It's equipped with our recently introduced anti-lock braking system. Which is controlled by an ITT computer.

Our System 12® telephone exchange and our 3100 business communications system are basically computers.

Even the "intelligent" ITT pump, which automatically regulates the temperature of the water that goes through it, couldn't function without a computer.

The point is, ITT computer technology exists in much that we do these days.

We've identified a select number of growing businesses that we're concentrating on. And many of them involve high technology.

You may not be able to see all the changes we've made in ITT yet.

But the results will be easy to spot.

Want to know more about ITT Corporation? Phone toll free 1-800-DIAL-ITT for continuously updated news.

## ITT

**It's a different world today.**

© 1985 ITT Corporation, 320 Park Avenue, New York, NY 10022

# LETTERS

To the Editors:

Dr. Whitehead vividly describes the breaching of whales ["Why Whales Leap," by Hal Whitehead; *SCIENTIFIC AMERICAN*, March] and concludes that it may be associated with social interaction among whales, perhaps in communication and play.

I interpret the purpose differently. The cover painting accompanying the article emphasizes the heavy accumulation of barnacles and whale lice on the belly of the whale. Perhaps a whale breaching has somewhat the same purpose as a dog scratching a flea—to get rid of an irritating parasite. Three pertinent facts presented by Dr. Whitehead, with my interpretation of each, are as follows:

1. *The humpbacks breach more in winter than in summer.* In winter they are in West Indies waters and in summer they are in the North Atlantic. Perhaps the warmer waters of the Caribbean induce a greater growth of parasites.

2. *Whales breach more as the wind picks up.* As the wind picks up, so do the surface water waves. Water waves induce an oscillating current, which, by hypothesis, moves the parasites back and forth on the whale's skin. The longer the parasite is, the more leverage it has to act on the skin. Such motion might irritate the whale.

3. *Rotund whales breach more frequently than streamlined whales.* Rotund whales have more surface area and swim more slowly, allowing parasites greater opportunity to flourish.

In a true breach the whale lands on its back, so that the parasite-infested belly is facing upward. The impact may generate a shock wave that travels through the whale and vibrates the parasites, dislodging some of the longer ones.

CYRIL GALVIN

Principal Coastal Engineer  
Springfield, Va.

To the Editors:

I (and other investigators) have dismissed parasite removal as a major cause of breaching for two principal reasons. First, as pointed out by the Soviet scientist A. G. Tomilin, even high-pressure hoses cannot dislodge most whale lice. Second, whales generally land on their backs, the least infested parts of their bodies.

Dr. Galvin raises some new and very interesting arguments, however. I do not know whether parasites grow fast-

er in the Tropics, although the idea seems reasonable. The idea of waves moving parasites, which then become irritating, is totally new to me and most interesting. Rotund whales do have more parasites than slim ones.

That shock waves created by the breach remove parasites from the whale's belly seems plausible. I believe Roger Payne has seen birds picking parasites from the water after a breach.

I do not think parasite removal can explain all breaches, however. Many breaches take place during stressful social circumstances, when whales would be concerned with things other than parasites.

HAL WHITEHEAD

Newfoundland Institute  
for Cold Ocean Science  
St. John's

To the Editors:

The article on Damascus steel ["Damascus Steels," by Oleg D. Sherby and Jeffrey Wadsworth; *SCIENTIFIC AMERICAN*, February] brings to mind two related applications.

First, the beauty of the Damascus-steel surface led to a kind of imitation: a patterned polishing of metal surfaces in a series of overlapping swirls. Perhaps the most conspicuous example of such polishing, called damascening, was on the aluminum engine cowling of Lindbergh's *Spirit of St. Louis*.

Second, early shotgun barrels were commonly fabricated by a Damascus-like process. A thin strip of steel was wrapped, spiral fashion, around a mandrel and forge-welded into a tube. After the mandrel was removed the rough barrel was finish-bored, turned down on the outside and then chemically treated to enhance the weld pattern. On inexpensive guns the pattern thus formed was a simple spiral, but on expensive guns very intricate patterns were produced by twisting or braiding multiple strips together before welding. Such barrels are very beautiful.

Toward the end of the 19th century techniques were developed for manufacturing a shotgun barrel inexpensively out of a single piece of steel, but Damascus barrels continued to be made. On some models of guns both Damascus and plain barrels were available, with Damascus an extra-cost option.

With the advent of smokeless powders in about 1900, the higher pressures they raised often ruptured Damascus barrels, usually along the weld lines. Barrels with the most intricate patterns appear to have been the most prone to failure. Probably gunmakers

did not follow all the forging procedures the article outlines, but it also appears that the lack of homogeneity of the steel concentrated tensile stresses in the discontinuities. By 1910 Damascus steel had lost all its prestige in shotgun barrels and went out of production.

JOHN S. HARRIS

Brigham Young University  
Provo, Utah

To the Editors:

I should like to compliment Oleg D. Sherby and Jeffrey Wadsworth for their fascinating article on Damascus steels. Unfortunately they seem to have made an important oversight, which may explain some of the inconsistencies in their test data.

The authors contend that theoretically the best Damascus-steel blades should show no damask pattern at all. Since it is well known that this is not the case, a conflict arises. When one considers the methods employed by Sherby and Wadsworth to produce their "Damascus" steel, however, it is understandable that they would draw this conclusion. Once we examine the way Damascus steel is actually made, it becomes clear exactly why the damask pattern is so highly valued.

A Damascus blade consists of thousands of layers of steel forming a lamination. These layers are welded together during the forging process. Essentially the blade maker starts with a billet of iron, which he and his assistants hammer flat and fold back on itself as many as 15 times. A weld is formed each time the billet is folded, thus doubling the number of layers. After five foldings there are 32 layers; after 10, more than 1,000; after 15, more than 30,000. This process is used today by several American and Japanese blade makers to produce Damascus-steel blades.

Because of its many layers, a Damascus blade is exceptionally flexible. Each lamination forms a slip plane with its neighbors, thereby allowing the blade to bend more than a conventional steel blade before breaking. The physical act of hammering and folding the billet helps to align the crystal structure of the steel, so that a high degree of uniformity throughout the length of the blade is attained. It is the heterogeneous composition of the steel, however, that causes it to be simultaneously hard and flexible.

This all boils down to a question of semantics. By definition Damascus steel is formed by physically laminating the steel in the way just described.



Hence the steel produced by Sherby and Wadsworth cannot legitimately be called Damascus steel. It would be a great mistake, however, to dismiss their work just because of an erroneous label. The technological applications for a mass-produced high-carbon are unlimited.

PAUL McNAMARA

Quincy, Mass.

To the Editors:

Mr. McNamara suggests the mechanism of Damascus-steel manufacture is quite different from the one we propose. Unfortunately, he is confusing two quite different art forms. The one we discussed was the manufacture of the *true* Damascus steel. The one Mr. McNamara describes is of a quite different origin, that of laminating steels of differing composition; the correct name for the product obtained by this approach is *welded* or *pattern-welded* (Damascus) steel. This confusion is understandable since, in some cases at least, pattern-welded steels were manufactured in an attempt to duplicate true Damascus steel (as, for example, in the gun barrels mentioned by Mr. Harris). Furthermore, essentially all the modern steel knives currently being manufactured and advertised in this country as Damascus steels are in fact of the pattern-welded variety.

The making of welded Damascus steel is a fascinating subject and has its own rich history, having been utilized in Indonesia and other Asian countries in ancient times and more recently in Europe and the U.S. On the other hand, the true Damascus steels are uniquely associated with Islamic countries (for example Persia, Turkey and India).

The major purpose of multiple folding procedures, as used by Japanese blade makers, was not to make a pattern-welded structure. It was, first, to homogenize the microstructure of the original 1.9 percent carbon casting (known as *tama-hagane*) and, second, to reduce the carbon content to about 1.3 percent carbon (an end product known as *uagane*). Fifteen folds, which represent about 30,000 layers, would lead to a layer thickness of .2 micrometer for a typical sword (a sword six millimeters thick). Such fine layers will not result in a visible damask.

OLEG D. SHERBY

JEFFREY WADSWORTH

Stanford University  
Stanford, Calif.

## A reminder of excellence...



**The Carlyle Hotel**

Luxurious guest rooms & apartments  
Peerless cuisine

Madison Avenue at 76th Street  
New York 10021  
Cable The Carlyle New York  
International Telex 620692  
Telephone 212-744-1600  
A member of the Sharp Group since 1967

## INVEST YOURSELF



**VITA** Putting Resources to Work for People

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

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.

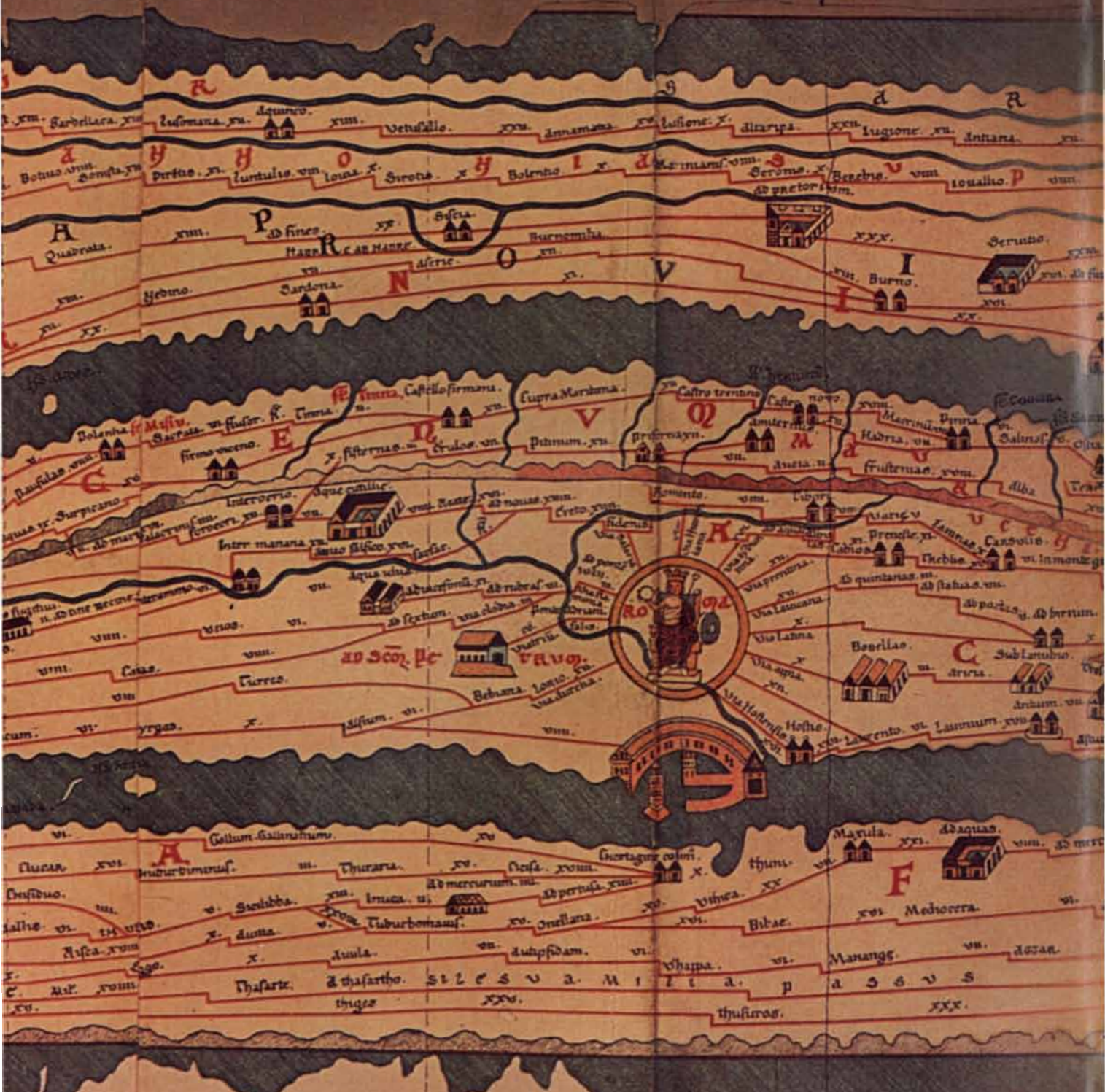
All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, 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.

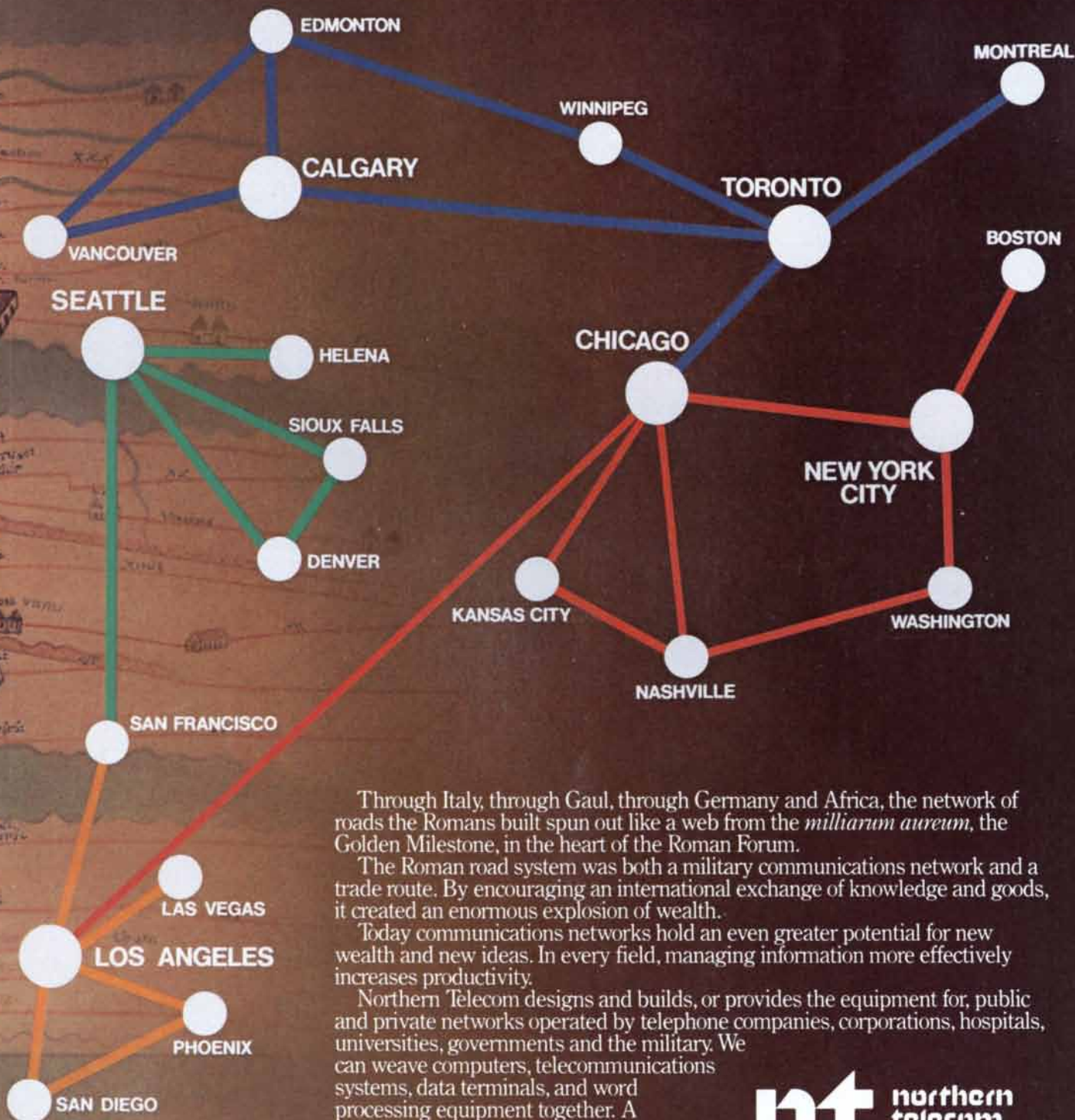
The road system of the Roman Empire—one of the world's earliest communications networks—took centuries to build. Messages travelled 40 miles in a day.



The Peutinger Table, a third century map of the Roman roads in Germany, Italy, and Africa.

Pomona	Ferme	Antrodice	Narni (Narnia)	Acquasana	Porto di Ferme	In Tiberino	Castellana	Castellana	Torre d'Amatone (79)	Tivoli	Palatino	Cervoli	Castellana	Castellana

At Northern Telecom we build networks that carry messages in milliseconds—and for longer distances.



Through Italy, through Gaul, through Germany and Africa, the network of roads the Romans built spun out like a web from the *milliarum aureum*, the Golden Milestone, in the heart of the Roman Forum.

The Roman road system was both a military communications network and a trade route. By encouraging an international exchange of knowledge and goods, it created an enormous explosion of wealth.

Today communications networks hold an even greater potential for new wealth and new ideas. In every field, managing information more effectively increases productivity.

Northern Telecom designs and builds, or provides the equipment for, public and private networks operated by telephone companies, corporations, hospitals, universities, governments and the military. We can weave computers, telecommunications systems, data terminals, and word processing equipment together. A world leader in digital telecommunications, we can build complete digital networks using only our own equipment.

We make information more valuable by making it more accessible.

**nt** northern  
telecom

For further information write:  
Northern Telecom Inc., Public Relations Dept.,  
259 Cumberland Bend, Nashville, TN 37228.  
Or call (615) 256-5900, Ext. 4264.

MORRISVILLE, NC • MORTON GROVE, IL • NASHVILLE, TN • RALEIGH, NC • SAN DIEGO, CA • SANTA CLARA, CA • WEST PALM BEACH, FL

© 1985 SCIENTIFIC AMERICAN, INC

# The Spirit of America



*Liberty by Ken Regan*

*Liberty. More than a symbol. A landmark to millions of soon-to-be Americans who passed her welcoming form and then themselves became part of the great tradition of freedom. A tradition toasted across the land with America's native whiskey: Kentucky Bourbon. Old Grand-Dad still makes that Bourbon as we did over 100 years ago. It's the spirit of America.*

*For a 19" by 26" print of Liberty, send a check or money order for \$6.95 to Spirit of America, P.O. Box 183L, Carle Place, New York 11514. For each print sold, we'll donate \$2.00 toward Liberty's restoration.*

## Old Grand-Dad



# 50 AND 100 YEARS AGO

## SCIENTIFIC AMERICAN

JUNE, 1935: "The discovery of the isotope of hydrogen, the so-called heavy hydrogen or deuterium, and its successful separation from the light variety, has excited great interest in scientific circles. One problem that challenges physicists is the structure of the atomic nucleus. We understand fairly well the structure of the electronic atmosphere of atoms, that is, the outside structure, but the structure of this minute central sun of the atom is quite unknown. In unraveling the structure of the nucleus, the deuterium nucleus, or deuteron, will certainly play an important part."

"Are some of the diffuse types of nebulae, which consist of highly rarefied gases, simply the later stages of new stars that explode, like the present new star in Hercules? At the Mount Wilson Observatory, Gustaf Strömberg thinks the Crab Nebula in Taurus represents a case of this kind and from a study of its spectrum concludes that the explosion took place about 900 years ago. Is it a coincidence that Chinese astronomers recorded a new star in the same spot 900 years ago?"

"It is now some 25 years since the first flint implements were found in a deposit of the Pliocene epoch in Suffolk, England. In those days it was generally believed, with an almost dogmatic intensity, that the earliest human beings had only appeared on the earth at a much later epoch. The announcement of the discovery at Suffolk was the signal for the waging of a fierce scientific battle, which has only recently come to an end. At long last the majority of competent investigators who have examined the specimens are agreed that they represent the work of man. With this acceptance the first stage in the fight for the recognition of the greater antiquity of man may be said to have terminated."

"Songs of wild birds in their native habitats are now being recorded by the American Museum-Cornell Ornithological Expedition. A parabolic reflector serves to gather the sound and concentrate it on a microphone, which in turn is connected to an apparatus for

recording on movie film. A telescopic sight on the reflector enables the operator to aim the reflector directly at the bird and thus secure greatest efficiency in recording the sound."



JUNE, 1885: "A few years since, some of the most eminent men of France conceived the idea of more firmly uniting the French and American people by the joint erection of some great work of art. The scheme was most favorably received in both countries. The Statue of Liberty, nearly completed, designed by the celebrated sculptor Auguste Frederic Bartholdi, was formally delivered to the United States Minister at Paris on the fourth of July, 1884. It was recently placed on board the *Iseré*, at Rouen; the vessel sailed for New York, May 20 last, and arrived early on the morning of June 17. The official welcome was on the 19th. The magnificent day, the enthusiastic crowds and the fine display of the tricolor and the stars and stripes made a pageant which will long be remembered."

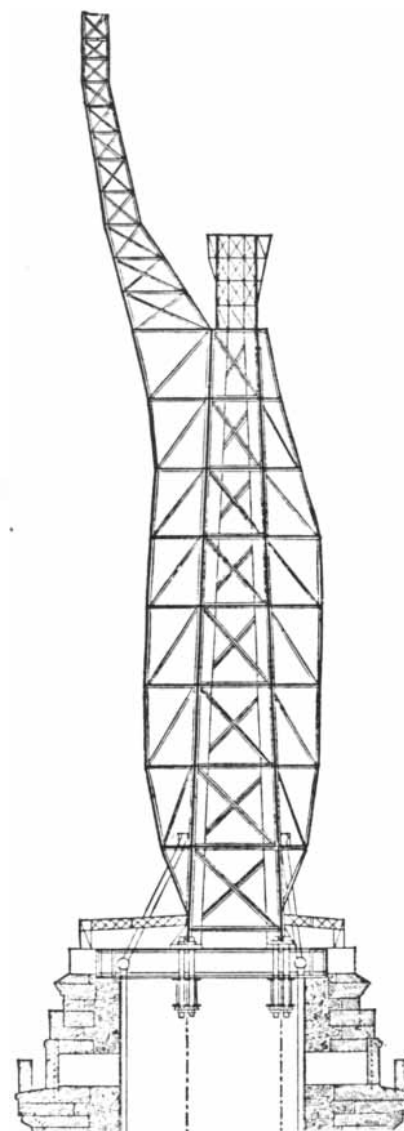
"There has recently been placed on exhibition, at the New York Museum of Natural History, an almost complete representation of the trees of the United States, between 400 and 500 trunk sections of the different species. These specimens are about 5 feet 8 inches long each, cut in such a manner as to display their barks and the transverse and longitudinal sections of the wood. This valuable collection includes examples of many curious and interesting species, of which probably the complete natural series could never have been viewed in their native soil by any single traveler."

"The industries of Japan possess a particular interest for foreigners, on account of the unique materials employed and the dignity which old age bestows. At a time when England as a nation did not exist, when the progressive peoples of modern Europe were to the polite world as barbarians and strangers, these ancient people were patiently at work, by slow degrees perfecting the details of their art, until now they produce wares without a rival in the markets of the world."

"MM. Brin, of Passy, are producing oxygen on rather a large scale by the barium oxide process. In two large retorts they calcine oxide of barium, passing over it a stream of air which has first passed through quicklime to

free it of carbonic acid. During this calcination the heat does not exceed about 500° C., at which temperature the barium oxide absorbs oxygen, becoming peroxidized. The nitrogen is drawn off and passed into gasholders, to be used for making ammonia, etc. When the barium oxide has absorbed as much oxygen as it can, the heat is raised to about 800° C., at which temperature the peroxide is decomposed, giving up again the absorbed oxygen, which is drawn off and pumped into a gasholder. MM. Brin make use of the oxygen so collected in many ways."

"The use of electric lights on athletic grounds has been tested for a few weeks at Williamsburg on Long Island, where the grounds of the Williamsburg Athletic Club are now lighted by electric lamps. By their light games were carried out in the evening."



Framework for the Statue of Liberty

# The uses

## Summary:

**Even the smoothest voice is discontinuous, especially in conversation. Data communications has bursts of message and periods of silence, too. Even TV has some "bursty" traits. GTE scientists are isolating silences and inserting other messages into them. This permits voice and data to coexist on the same channel at the same apparent time. The development stems from parallel research in microelectronics, silence detection, speech, voice compression and signal processing.**

Without basic change, or vast growth, telephone networks will be unable to cope with the anticipated traffic of the 1990's. The proliferation of personal computers and data terminals has already placed a strain

on switching and transmission facilities. It has also placed demands on networks that are much different from the original voice-communications concept, in which average time of connection was three minutes.

Today, far shorter and far longer connections abound, more subscriber lines are in demand, and there are growing needs for enhanced services and faster switching.

Out of research dating from 1979, GTE has developed a switching system that promises not only to triple present transmission capacity but also to process calls 20 times faster. The system is called Burst Switching.

## The nature of speech.

Our world is full of holes. Matter is mostly empty space. Conversation is mostly silence. But, even though speech is 2/3 silence interspersed with bursts of sound from 0.1 to 1.5 seconds long, if that speech goes over a telephone line, the line is locked up for the duration.

But, with Burst Switching, we can shoehorn other messages into the silences, automatically easing the pressure on transmission facilities. Theoretically, in fact, we triple transmission capacity.

## VHSIC.

Through Very High-Speed Integrated Circuits (in which we are currently researching devices with submicron feature size), we are able to make and break telephone connections at increasingly high speeds. Voice lines need be dedicated only for the very brief duration of voice bursts. At other times, channels are available for other voice messages, or for data streams which are also "bursty" in nature. In addition, video, because of its built-in redundancy, can be considered to have bursts, too.



# of silence.

## Message compression.

The capacity needed to transmit speech can be made even smaller if the information that must be sent to make it recognizable can be minimized. Our scientists have reduced the 64 kb/s signals to 16 kb/s while retaining high quality.

Thus, transmission-capacity requirement is reduced by a factor of four.

We are working, as well, on techniques for compressing video signals from 90 Mb/s to 64 kb/s. This will have special relevance for such activities as video conferencing.

So transmission capability grows and switching becomes faster—and we can now envision future telephone systems able to carry billions of simultaneous calls.

The box at the right lists some of the pertinent papers GTE people have published on Burst Switching and related subjects. For any of these, you are invited to write GTE Marketing Services Center, Department TPIIE, 70 Empire Drive, West Seneca, NY 14224.



Burst Switching experimental model.

## Pertinent Papers.

*Burst Switching—An Introduction*, IEEE Communications Magazine, November 1983.

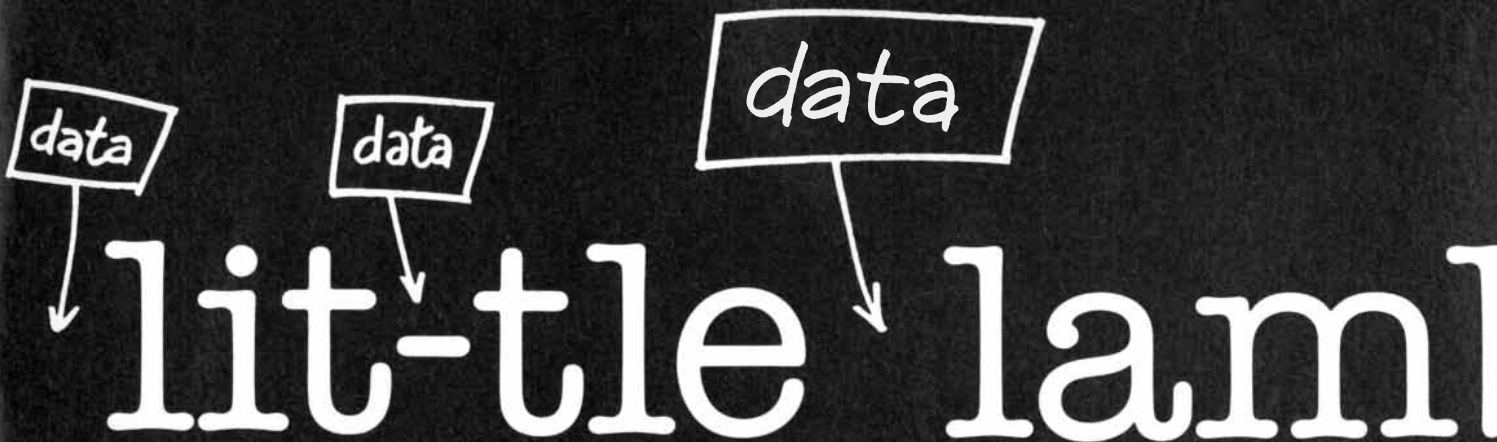
*New Switching Concept Integrates Voice and Data Bursts*, PROFILE, September 1983.

*A PCM Frame Switching Concept Leading to Burst Switching Network Architecture*, IEEE Communications Magazine, September 1983.

*Application of the Burst Switching Technology to the Defense Communications System*, Proceedings 1983 IEEE Military Communications Conference, MILCOM '83, Washington, D.C.

*Performance Evaluation of a Distributed Burst-Switched Communications System*, Proceedings Second Annual Phoenix Conference on Computers and Communications, March 1983.

*A Complementary Speech Detection Algorithm*, Proceedings of GLOBE-COM '83, November 1983.



In Burst Switching, the roughly 65% silence in speech can be filled with data streams and other messages, effectively tripling transmission capacity.

# THE AUTHORS

WASSILY LEONTIEF ("The Choice of Technology") is professor of economics at New York University and director of its Institute for Economic Analysis. Leontief earned a degree at the University of Leningrad in 1925 and went to the University of Berlin to get his doctorate. He then spent a year as adviser to the minister of railroads of the Chinese government in Nanking. In 1932, soon after moving to the U.S., he joined the faculty of Harvard University. There he developed the input-output method of economic analysis described in his work *Input-Output Economics*. He left Harvard in 1975 for New York University. Leontief was awarded the 1973 Nobel prize in economics; in 1984 the government of Japan conferred on him the Order of the Rising Sun for his part in formulating effective economic policies for that country.

RICHARD L. EDELSON and JOSEPH M. FINK ("The Immunologic Function of Skin") began collaborating in the study of the immunologic role of the skin in 1983. Edelson is professor and director of research in the department of dermatology at the Columbia University College of Physicians and Surgeons, where he is also associate director of the Columbia-Presbyterian Medical Center's General Clinical Research Center. A 1966 graduate of Hamilton College, he became interested in immunology while he was at the Yale University School of Medicine. After getting his M.D. in 1970 he did an internship at the University of Chicago Pritzker School of Medicine and a residency in dermatology at the Massachusetts General Hospital. Edelson joined the Columbia faculty in 1975. Fink earned an undergraduate degree at Brooklyn College of the City University of New York in 1979 and a D.D.S. at the Columbia University School of Dental and Oral Surgery in 1983. There he developed an interest in the immunologic mechanisms involved in periodontal disease. The similarity of the oral epithelium to skin led him to study the immunologic function of skin during a research postdoctoral fellowship in Edelson's laboratory. Fink is now an associate research scientist in the department of periodontics at Columbia.

J. M. LOSECCO, FREDERICK REINES and DANIEL SINCLAIR ("The Search for Proton Decay") are experimental physicists. LoSecco is assistant professor of physics at the Cal-

ifornia Institute of Technology. He is a graduate of the Cooper Union School of Engineering and Science and of Harvard University, where he got his Ph.D. in 1976. He did research at Harvard and at the University of Michigan, concentrating on the behavior and properties of the neutrino as well as on the decay of the proton, before moving to Caltech in 1981. Reines is professor of physics at the University of California at Irvine. He is a graduate of the Stevens Institute of Technology and New York University, where he got his doctorate in 1944. In 1959, after 15 years on the scientific staff of the Los Alamos National Laboratory, he joined the faculty of the Case Institute of Technology. In 1966 he moved to Irvine, where he was the founding dean of the school of physical sciences. For his work, which includes studies of muons, neutrino scattering and proton stability, Reines received the J. Robert Oppenheimer Memorial Prize in 1981. Sinclair is professor of physics at the University of Michigan. He studied at the University of Glasgow, which awarded him a Ph.D. in 1957 for research on the properties of pions. He then taught at the University of Michigan, first as an instructor in the department of physics and eventually as professor.

IVAN R. KING ("Globular Clusters") is professor of astronomy at the University of California at Berkeley. He got his A.B. at Hamilton College in 1947 and his Ph.D. from Harvard University in 1952. After serving in the U.S. Navy and working for the Department of Defense he took a position in the astronomy department at the University of Illinois at Urbana-Champaign. He remained there for eight years and then moved to Berkeley in 1964. For the past several years, as a U.S. member of the European Faint Object Camera team, King has been active in preparations for observations to be made with the Space Telescope after its launching in 1986.

A. G. CAIRNS-SMITH ("The First Organisms") is reader in chemistry at the University of Glasgow. He holds a bachelor's degree (1954) and a Ph.D. in organic chemistry (1957) from the University of Edinburgh. After receiving his doctorate he moved to Glasgow, where he teaches classes in organic chemistry and the history of chemistry. His investigations into clays and the first organic molecules resulted in a book, *Genetic Takeover and the Min-*

*eral Origins of Life*. Recently Cairns-Smith has become interested in the banded iron formations of ancient sedimentary rock and what they reveal about photochemical processes in Precambrian seas.

MICHAEL P. GHIGLIERI ("The Social Ecology of Chimpanzees") is a wildlife biologist and specialist in primate behavior. He is a graduate of California State University at Hayward, where he got a B.S. in 1971 and an M.A. in 1973; for his master's thesis he did a study of the behavior of captive lowland gorillas. He obtained his doctorate in 1979 from the University of California at Davis for the research described in this issue of SCIENTIFIC AMERICAN. During and after graduate school he conducted wildlife studies and led wilderness river tours and treks in Turkey, Ethiopia, Kenya, Rwanda, Tanzania, Papua New Guinea, Indonesia and the U.S. In 1984 Ghiglieri published *The Chimpanzees of Kibale Forest*, a book based on his doctoral research.

A. TREVOR HODGE ("Siphons in Roman Aqueducts") is professor of classics at Carleton University in Ottawa. He was born in Belfast and educated at the University of Cambridge, where he received a B.A., an M.A. and a doctorate in classical archaeology. His doctoral thesis was published in book form in 1960 by the Cambridge University Press as *The Woodwork of Greek Roofs*. Following his graduation he taught at Stanford University, Cornell University and the University of Pennsylvania before he joined the faculty at Carleton in 1960. His chief current interests are Roman aqueducts and the Greek colonization of southern France. Hodge has devised several programs for instructional television, and he is a regular commentator on the Canadian Broadcasting Corporation's radio network.

WALTER TAPE ("The Topology of Mirages") is associate professor of mathematics at the University of Alaska at Fairbanks. He earned his bachelor's degree at Princeton University and his doctorate from the University of Michigan. His main interest, recreational as well as professional, is meteorological optics, which includes phenomena such as rainbows and mirages. Such phenomena, Tape writes, "appeal to me both for their natural beauty and for the mathematical ideas that are related to them. I enjoy photographing them, partly for scientific reasons and partly for the excitement of the chase: one always hopes to capture a rare feature."



# You may never drop your Leica from an airplane.

But isn't it nice to know you could if you wanted to?

For 60 years, professionals and serious amateurs have considered Leica® the world's finest camera. A camera in a class by itself.

One reason is the unquestioned superiority of Leitz® lenses. Another is the almost fanatical dedication to perfection that governs every step of our manufacturing process.

## A lifetime camera.

Leica cameras are built to last a lifetime. Or even longer.

Of all the Leicas sold in the United States since 1929, an amazingly high number are still in use. Which is no accident.

Leica builds cameras of metal, not plastic like other manufacturers. Using plastic would not be Leica's way. Instead, we use the highest-quality brass, zinc, aluminum, magnesium and titanium. Materials used in spacecraft.

Leica bodies are made of rugged die-cast aluminum. Bottom plates are 0.8mm of solid brass.

## Copper, zinc and nickel.

Leica top plates, which protect the prism, electronics and vital controls, are made of fatigue-resistant zinc. Die cast in a patented

Leitz process, not stamped out like top plates from other manufacturers. Which could result in serious weaknesses.

Once cast, top plates are smoothed with jeweler's tools. Coated with layer upon layer of copper and nickel. Chrome plated with a black or silver finish.

No wonder our cameras are legendary for ruggedness.

## An extraordinary difference.

But ruggedness isn't the only measure of Leica's manufacturing superiority.

Another is precision. Take our bayonet lens mounting, for example. Made of a special brass alloy with three different coatings. So tough it can never wear down, no matter how often you change lenses. Maintaining a lens-to-film distance accurate to .01 mm. Or our film channel. Machined to an incredibly precise tolerance. To hold the film flat for an exceptionally sharp photograph.

## An extraordinary warranty.

Ruggedness and precision. Only two of the ways in which Leica builds better cameras.

We could write about hundreds of them.

But perhaps the greatest proof of our superior construction is our superior warranty.

Every U.S. warranted Leica comes with The Passport Protection Plan. An extraordinary warranty, protecting you against any and all damage to your lens or camera. Because for two full years, Leica will repair your damaged camera or replace it. No matter how extensive the damage or how it occurs.

Even if your camera gets dropped from an airplane. Which we don't really recommend.

Although we know at least one Leica that survived it.

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

Leitz means precision. Worldwide.





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

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

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

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

## TURBODIESEL POWER, DIESEL DURABILITY

Yet consider the bottom line. The Turbodiesel you will be living with and maintaining and paying the bills for, year in and year out, is a true-blue diesel. No complex

electrical system. No conventional tune-ups. A durability factor that has become part of automotive folklore.

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

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

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

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

vivid level of performance.

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

## SEDAN, STATION WAGON AND COUPE

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

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

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

## EXOTIC, YET PRACTICAL

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



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

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

Sedan or Station Wagon or

Coupe, Mercedes-Benz Turbo-diesel power is harmonized with high standards of performance in every sense of the word.

From suspension to steering to brakes, every Turbodiesel is engineered to be a precision driving instrument. "There's a cornucopia of driving delights at your disposal,"

concludes *Car and Driver*—suggesting that in driving precision there is driving pleasure.

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

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

**MORE THAN POWER**

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



**Engineered like no other car in the world**

*SEE YOUR AUTHORIZED MERCEDES-BENZ DEALER*



# COMPUTER RECREATIONS

*Analog gadgets that solve a diversity of problems and raise an array of questions*

by A. K. Dewdney

Exactly one year ago a collection of analog gadgets in these pages set off an avalanche of similar devices from inspired readers. I am still extricating myself from a vast heap of wood boards, rubber bands, strings, balls of polystyrene, fish tanks, lead weights, canisters, tubing and stopcocks. In the process I have, I think, identified the best of these gadgets and have arranged them into a kind of gallery through which the reader is invited to wander.

Analog gadgets are mechanical devices that solve specific problems by virtue of the fact that their construction or behavior is analogous to the elements of the problem. For example, the June 1984 column described SAG, the Spaghetti Analog Gadget. Lengths

of uncooked spaghetti were used as analogs for numbers. To sort the numbers in decreasing order, gather the spaghetti into a bundle held vertically and bring it down rather sharply on a tabletop; select the longest rod. Continued selection of the longest remaining rod produces the desired decreasing sequence of numbers. In addition to SAG, I presented gadgets for finding shortest paths, convex hulls and minimum trees. I even displayed a gadget for factoring numbers that consisted of mirrors and a laser beam.

The latest collection includes several ingenious new gadgets for solving problems in statistics, network theory, algebra and arithmetic. On leaving the gallery I shall reexamine some important issues that arise from the prospect

of analog computing: How accurate are analog computers and how much time do they really take to compute? Are there some analog computers that outperform digital machines?

The first of the new gadgets solves a certain problem in statistics by means of a wood board, rubber bands, nails and a smooth, rigid rod. A set of data points plotted on a sheet of graph paper may present a linear trend to the eye. If a linear relation really governs the points, what straight line best displays the relation? The gadget suggested by Marc Hawley of Mount Vernon, Ind., supplies one possible answer:

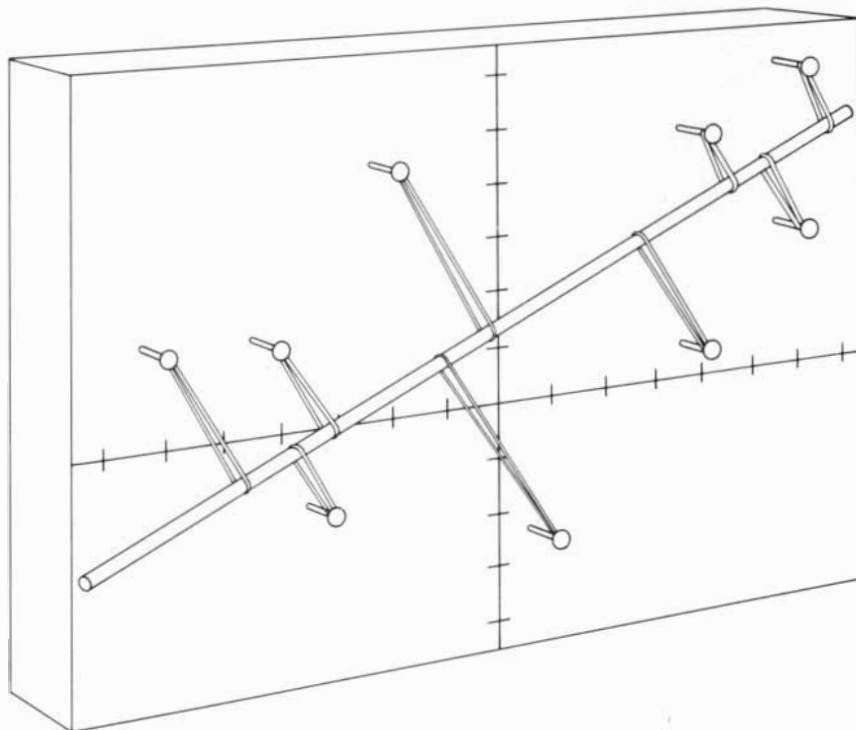
Plot the data points on a wood surface and drive a nail partway into the wood at each point. Next, slip a number of uniform rubber bands onto the rod, one band for each nail. Fit the rod approximately into place and pull each band over one of the nails. When the rod is released, it wiggles and shivers quickly into an equilibrium position [see illustration on this page].

The equilibrium position minimizes the total energy of the system; therefore the sum of the distances from the nails to the rod has also been minimized. In terms of such distances, the rod's final position indicates the straight line that best fits the data. It is not such distances but their squares that appear in the formulas for linear regression used by statisticians. Hawley's gadget computes something at least as complicated.

A charming string gadget was suggested by Jos Wennmacker of Nijmegen, Holland. Suppose we wish to know the longest path any message might travel in a communications network shaped like a tree. This path will be what combinatorial mathematicians call the diameter of the tree. In order to find the diameter Wennmacker reconstructs the tree by knotting together an analogy out of pieces of string. Each string is scaled to a specific communications line of the network. Two simple steps complete the computation. Pick up the string tree at any node and allow it to dangle freely. Now pick up the tree anew at the lowest node and dangle it once more. The longest path in the tree runs from the top node to the bottom one [see illustration on page 22].

When I initially encountered Wennmacker's gadget, my immediate reaction was, "It can't be that simple. Surely I must continue to select the bottom node and dangle the tree several more times." But one does not need to do that. Kudos in this column for the most elegant argument.

In last June's column I stated that the problem of finding the longest path in an arbitrary network was what theo-



*A gadget for finding the line that best fits a series of data points*



# When it comes to Dad, the sky's the limit.

Wild Turkey. It's not the best because it's expensive.  
It's expensive because it's the best.



Now you can send a gift of Wild Turkey® 811 Proof anywhere\* by phone. Call Toll Free 1-800-CHEER-UP (Arizona 602-957-4923).  
\*Except where prohibited. Major credit cards accepted. Austin, Nichols Distilling Co., Lawrenceburg, KY © 1985.

© 1985 SCIENTIFIC AMERICAN, INC



## For everyone who ever tried doing five things at once

**The perfect computer program  
for someone as busy as you.  
It lets you keep several other  
programs working at once.**

Do you ever go in so many directions  
so fast not even a computer can keep up  
with you?

Well, now an IBM Personal Comput-  
er can — thanks to IBM TopView.

TopView is a new kind of software  
that lets you switch between other pro-  
grams as quickly as you can change your  
mind, even run several programs at the  
same time.

Once you load TopView into your  
computer, you load the other programs  
you use most—as many as your com-  
puter's memory will permit.

After that, the greatest distance  
between two programs is just a couple of

keystrokes, or (optional) mouse moves.

There's no waiting and a lot less  
diskette swapping.

But when you're *really* busy is when  
TopView really shines, letting you do  
many jobs simultaneously.

For example, you can print a letter,  
while you search a file, while you analyze  
a spreadsheet, while your clock/calen-  
dar reminds you that your automatic  
dialer is about to place a call for you.



## ...IBM presents TopView.

And you can see everything through on-screen "windows" and control it all with easy-to-use pop-up menus.

You can even make unrelated programs work together: say a "Brand Y" spreadsheet with a "Brand Z" word processor:

But simplest of all is a certain "Brand IBM", namely the IBM Assistant Series—for filing, writing, planning, reporting and graphing.

Many other popular programs also work with TopView, and the number is growing.

Naturally, the more computer memory you have, the more TopView can help you. At least 512K is recommended.

And the price is only \$149\*.

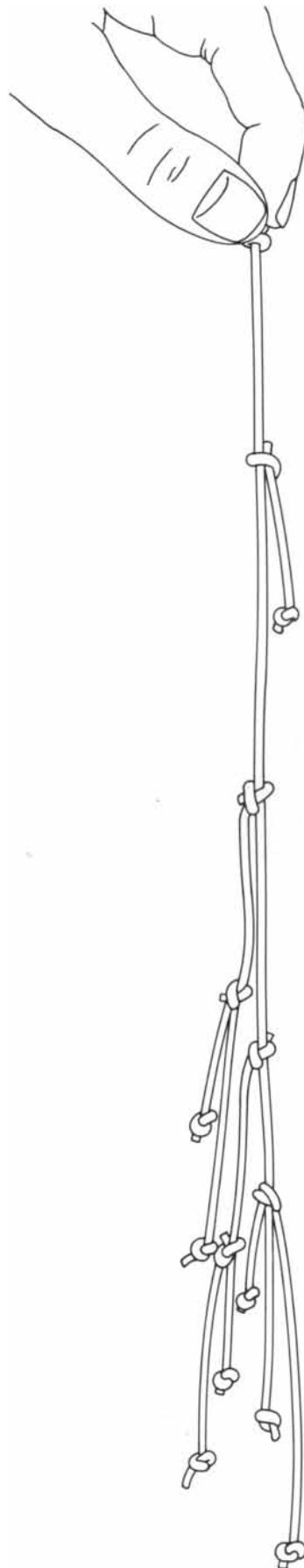
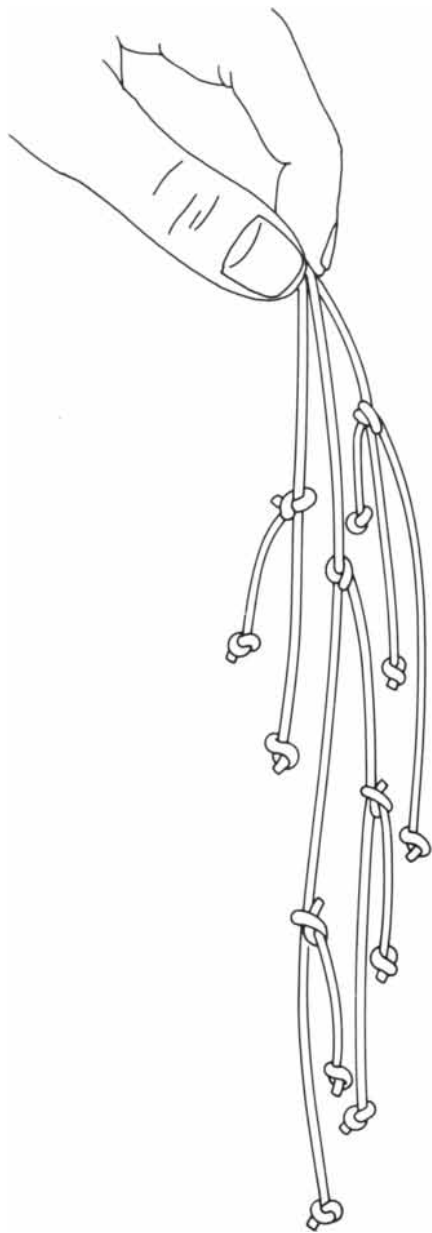
Beyond that, all you need is to be the kind of person who never does a single thing all day, but who wants to do everything, at once.

To learn more, call an IBM marketing representative, or visit an IBM Product Center or Authorized IBM PC or Software Dealer:

For the store nearest you, and a free brochure, call 800-447-4700. (In Alaska and Hawaii, 800-447-0890.)

Personal Computer Software

\*IBM Product Center price.



*A string gadget that finds the longest path in a tree-like network*

reticians call *NP*-complete. This fact means that as a practical matter the problem cannot be solved in a reasonable amount of time on a digital computer. Have we at last demonstrated the superiority of an analog device to its digital cousins? Not quite. It turns out that if the network is a tree, there is a digital algorithm for finding a longest path rapidly.

The next room in the gadget gallery contains a deceptively simple device suggested by M. Laso of the Swiss Federal Institute of Technology in Zurich. Here we find clamped into a vise a beam of aluminum that sports a needle at one end [see illustration on page 24]. The needle is superposed on a finely calibrated scale.

Prior to our entry into the room someone clamped the beam into the vise so that exactly three meters protrude. At the vise the beam is level. At the end from which the needle protrudes the beam droops slightly. Approaching the scale, we note the needle gives a reading of 81. This is the fourth power of 3.

According to the theory of elasticity, the deflection of a beam supporting only its own weight is proportional to the fourth power of its length. Laso points out that the same gadget could be used to compute the fourth root of a number by sliding the beam through the vise until the needle points to that number on the scale. It is even possible to compute third powers and roots by using a new scale. With the beam clamped in a fixed position, the needle points to zero. Next a weight proportional to a given number is placed on the end of the beam. In this case the theory of elasticity predicts that the beam's deflection is proportional to the third power of the weight.

In the same room is a slab of wood bearing a map, through which three holes have been drilled. Three strings pass through the holes. Below, weights are attached to the strings; above, the strings are attached to a small brass ring [see illustration on page 25]. J. H. Lueth of the United States Metals Refining Company in Carteret, N.J., calls this device *SLAG*, for Smelter Location Analog Gadget. The problem solved by *SLAG* is to find the optimal location of a refinery so that the cost of transporting three major ingredients is minimized. If ore, coal and limestone cost  $A$ ,  $B$  and  $C$  dollars per mile per ton to transport and if the distances from the refinery to these sources is  $a$ ,  $b$  and  $c$  miles, then the total cost of delivery is  $aA + bB + cC$  dollars. The three holes are drilled through the board at places corresponding to the geographic locations of the three sources. Once the strings have been passed through the





## NOW TWA'S FREQUENT FLIGHT BONUS<sup>SM</sup> PROGRAM COVERS THE WORLD LIKE NO ONE ELSE.

Only TWA can give you free travel to  
so many exciting award destinations...so fast!

When it comes to free travel around the world, no other frequent flyer program delivers like TWA's. We count every mile you fly on TWA... *plus* the miles you fly on Eastern, Qantas and PSA. So in TWA's FFB<sup>SM</sup> program, you can really earn miles fast.

And when you're ready to cash in those miles for free travel, you'll find that TWA has the world covered, too. With exciting awards to Europe... the Caribbean...

Australia... the Far East... and over 170 places in between. Even an unforgettable free trip for two around the world... First Class all the way. No other U.S. airline can give you that!



**TWA overshadows the competition.**

And when it comes to service, TWA takes care of you like no one else. Fly often, and you'll enjoy complimentary upgrades to our roomy business class on widebody

flights... or to our incredibly comfortable First Class, with its luxurious Sleeper-Seats<sup>SM</sup>. Without cashing in a single mile!

So why look anywhere else? Join the one frequent flyer program that really delivers the world... TWA.

Simply mail the attached coupon, or call toll-free:

1 800 325-4815, Agent 114



**Enroll in TWA's Frequent Flight Bonus program today.**

We'll give you a fast 3,000 bonus miles just for joining. So we can give you the world even faster.

PLEASE PRINT CLEARLY

Name \_\_\_\_\_

Address (U.S. only) \_\_\_\_\_

City/State/Zip \_\_\_\_\_

Telephone (\_\_\_\_\_) \_\_\_\_\_

If you are a member of TWA's Ambassadors Club\*, please write your membership number: \_\_\_\_\_

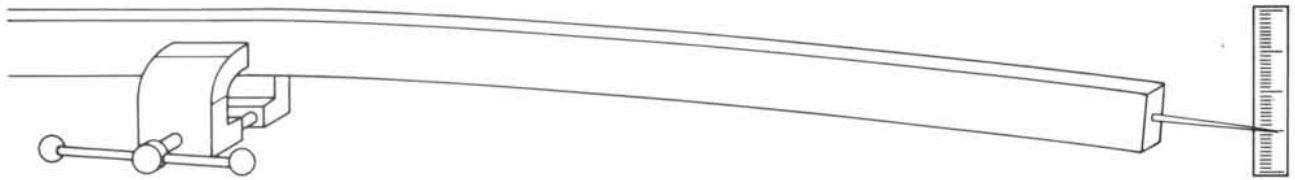
Mail to: TWA's Frequent Flight Bonus Program  
P.O. Box 767, Murray Hill Station, New York, NY 10156

Please allow 3-4 weeks for delivery of your TWA Enrollment Portfolio, which contains all the rules and conditions for participation in the program.

SA 114

LEADING THE WAY<sup>SM</sup> TWA.

© 1985 SCIENTIFIC AMERICAN, INC



*A beam-deflection gadget computes the fourth power of its length*

holes and attached to the ring, weights proportional to  $A$ ,  $B$  and  $C$  are attached to the appropriate strings. Released, the brass ring quickly slips into position, thereby revealing the optimal smelter location on the map.

The next room of the gallery is filled with a multitude of glass and brass gadgets; the air is heavy with the smell of soapy water.

A number of these aqueous computers were brought to our attention by Dale T. Hoffman of Bellevue Community College in Bellevue, Wash. A soap film between glass and a stepped surface provides an analog for a light ray passing from one transparent medium to another: the film is straight over both levels, but at the step where it drops it also bends abruptly [see *illustration on page 28*]. If the step size between the levels is changed, the soap film's angle changes as though the simulated media had changed their indexes of refraction.

This simple gadget in more complex form can also solve a pipeline problem. Imagine a terrain stretching between two cities that is subdivided into distinct regions, each of which corresponds to a specific pipeline construction cost. Here is a swamp, over there is high ground and off to one side is forest. In order to model this problem Hoffman proposes that a region be represented by a flat surface cut to conform to the outline of the area. Each surface is placed at a height proportional to the associated cost of construction. A peg is then inserted at each city, a glass cover is put over the gadget and the entire assembly is immersed in a soap solution. A film forms that bends at two points. It indicates the course of a pipeline whose cost is a minimum over all possible routes. It would be enjoyable to explore more of Hoffman's gadgets [see "Bibliography," page 136], but we are drawn to some exotic glassware.

Five graduated glass cylinders are connected at their bases by tubes. A stopcock interrupts the free passage of water from one cylinder to the next. Thus when the cylinders are filled to different levels representing five distinct numbers, nothing happens. But when the stopcocks are all opened, the water is free to seek its own level, which in this case is at the average of

the five numbers input to the gadget. The device comes from Sartore Marco of San Remo, Italy.

The final gadget I shall describe in detail was suggested by Peter F. Ash of St. Joseph's University in Philadelphia, Pa. It solves cubic equations and can be elaborated into versions capable of solving much higher polynomials. A cubic equation has four terms,  $ax^3$ ,  $bx^2$ ,  $cx$  and  $d$ . These are added together and set equal to zero. To solve such an equation one must find a value of  $x$  that yields a sum of 0.

The gadget now before us [see *illustration on page 28*] solves a specific cubic equation. It consists of a large water tank, a balance beam, two scalepans and a variety of solids to represent the terms of the equation. The solids have rounded surfaces, as though turned on a lathe. The  $x^3$  term is represented by a paraboloid. Hung apex down and immersed to a depth of  $x$  centimeters, the paraboloid displaces a volume of  $x^3$  cubic centimeters. The  $x^2$  term is represented by a cone that displaces  $x^2$  cubic centimeters. A cylinder represents  $cx$  and a sphere represents  $d$ . The sphere is always immersed.

The four solids are hung from a balance beam that has scalepans at each end so that its fulcrum coincides with its middle. The beam is suspended over a water tank. The paraboloid is hung  $a$  centimeters to the left of the fulcrum if  $a$  is negative. Otherwise it is hung  $a$  centimeters to the right. The same rule applies to the other three solids.

To solve the cubic equation, one holds the balance beam level and fills the tank with water until three of the solids just touch the surface. The fourth solid, the sphere, is already immersed. If the balance beam is now released, one arm will probably be heavier than the other. Consequently weights are added to one of the pans so that the beam is balanced and does not have to be held. The current water level is marked as zero.

So far I have been describing the preprocessing phases of this gadget. Now comes the analog solution. As the tank is slowly filled with water the beam becomes unbalanced, but a little later it balances again. The water is immediately turned off and its new level is recorded. The difference between the old level and the new level is the

value of  $x$  that satisfies the equation. Of course, the equation might not have a solution. In such a case we could fill the tank to the brim and still the beam would remain unbalanced.

How does this remarkable gadget work? According to the principle of Archimedes, an immersed solid suffers an apparent loss of weight that is proportional to the volume submerged. The effect of immersion on the balance beam is multiplied by the distance between each solid and the fulcrum. Thus the new equilibrium is achieved only when the equation is satisfied, in effect, by the new water level.

It would be wonderful if I could double the length of this column to include full descriptions of all the gadgets in the gallery. It is possible to mention only a few more.

One of them is an engaging if highly impractical alternative to SAG. Suggested by Michael Gardner of Findlay, Ohio, it is called GAS, for Gasoline-Activated Sorter. GAS requires the use of one Volkswagen for each number to be sorted. Each Volkswagen is filled with an amount of gasoline reflecting the number. Drivers of equal weight enter their respective vehicles and the convoy sets off down some little-used road. An additional Volkswagen (with a full tank of gas) follows the convoy and its driver notes the order in which the other cars have run out of gas as it passes them. In this way the numbers are sorted.

The string gadget for finding shortest paths in a network described in the June 1984 column can itself be simulated by a simple electrical network, according to Stephen Fortescue of Canoga Park, Calif. Each edge of the network is replaced with a parallel circuit that has two branches. Each branch consists of a Zener diode connected serially to a light-emitting diode (LED). Diodes will normally pass current in only one direction, and strange as it may seem, the Zener and the LED on both branches are oriented in opposite directions. Furthermore, with respect to the order of components one branch is the reverse of the other. Zener diodes can, however, be made to conduct current in the abnormal direction if the applied voltage is high enough. One selects Zener diodes with breakdown voltages directly proportional to

the length of edge they replace. To find the shortest path between two nodes apply the voltage between them. As the voltage gradually increases, the shortest path between the nodes suddenly lights up.

No one has found an alternative that bests the rubber-band gadget in finding the convex hull of a set of points in a plane. Nevertheless, Elio Lanzoni of Modena, Italy, has invented a device that finds the circumscribed circle, in other words the smallest circle that contains all the points. Lanzoni drills holes through a board to represent the points. He prepares several strings of the same length, tying them together at one end and passing the other ends through the board, one end per hole. The lower ends are then attached at corresponding points to a large, flat piece of plywood. When the plywood is suspended below the board, the knotted upper end of the strings hovers over the center of the circle sought. At least three of the strings are tight. These are radii of the circle.

My earlier promise to address the serious issues raised by analog computers sprang from a certain anxiety about the gadgets in particular and this approach to computing in general.

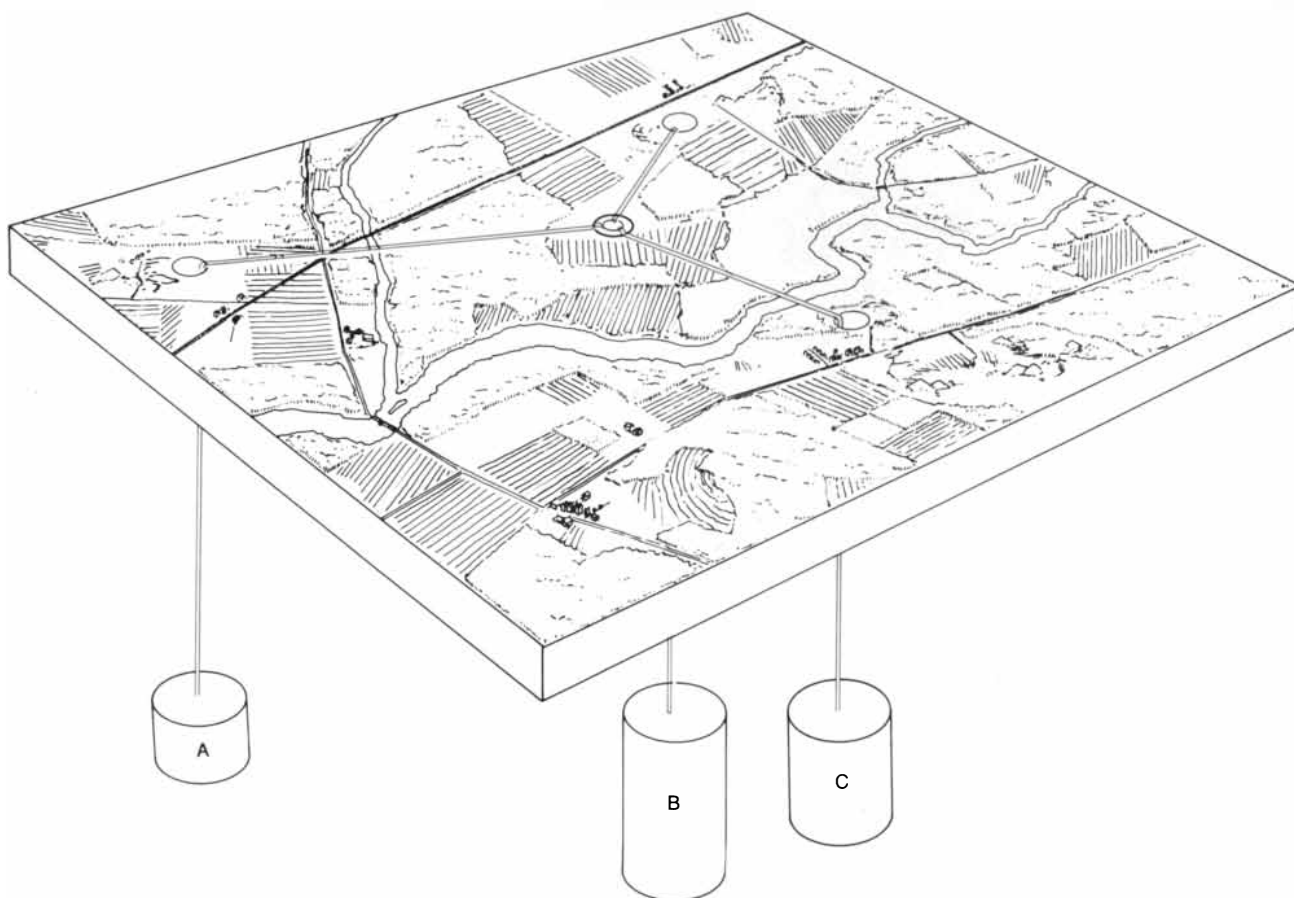
In response to last June's column a few readers wrote critical notes about my calculation of the speed at which gadgets compute. Some others questioned the accuracy of gadgets and even challenged me to define exactly what I meant by an analog computer. These challenges lead us to the heart of a question involving the ultimate relation between matter and information.

In the first column on this subject I analyzed the pre- and postprocessing phases of analog computation. In the SAG machine one first cut spaghetti rods to lengths corresponding to the numbers to be sorted. Later one read them off by measuring the rods. In almost all cases this was the most time-consuming operation. The essential computation is embodied in a slam, a tug or a snap signifying the arrival of an equilibrium state—the solution. In a sense, however, the processing phases are foreign to analog computing. Our use of numbers as a kind of mental currency leads us to demand input and output in this form. But if by an analog computer we mean a physical process or an abstract model of a physical process, the inputs and outputs must themselves be physical. Digital computing has a reciprocal debility. What if I had to evaluate a digital computing

scheme by first deciding how quickly I could convert a number of physical variables into numeric ones?

It is perhaps better to pursue the meaning of analog computation without worrying, for the moment, about how to get back and forth between the digital and analog realms. What, then, is the analog realm? It may be reasonable to idealize this realm somewhat in the manner of an alien being who visited our planet quite a few years ago. The incident was mentioned in a book by a predecessor of mine in this department. The alien (let us call him Martian Gardner) landed his spaceship and proceeded to convert all the books on the earth into a single, enormous number. The process was simple in principle: considered as a very long string of words, the books could equally well be regarded as a very long string of digits. The alien then merely inserted a decimal point in front of this number and made a tiny mark on a beautiful duron rod kept aboard the spaceship. The nick divided the rod precisely in the ratio indicated by the enormous decimal number. Thus was the written tradition of humankind reduced to a nick on a rod.

Imagine, then, that the analog realm consists of such ideal matter obeying



*The smelter location analog gadget (SLAG)*


# We keep raisin

First, Honda created the Accord 4-Door Sedan. Elegantly styled. Superbly engineered.

It quickly became the number one selling small car in America\* and set the standard by which other automobiles in its class would be judged. But no other car could be an Accord. No other that is until Honda introduced the

luxurious Accord LX 4-Door Sedan.

To the proven styling and engineering of the Accord 4-Door, Honda added even more standard features. Like power windows and door locks. A four-speaker AM/FM electronic tuning stereo with autoreverse cassette. Michelin radials. And air conditioning.

 © 1985 American Honda Motor Co., Inc.



# g the standard.

Ahhh, but then there were three. Because soon Honda designed another very special Accord 4-Door Sedan replete with leather seats, a power Moonroof, fuel injection, dual power mirrors, alloy wheels, all standard. And so was created the Honda Accord SE-i.

The Accord 4-Door Sedan. The Accord

LX4-Door. The Accord SE-i. At Honda, after we set the standards, we keep raising them.

\*Based on 1984 calendar year *Ward's Automotive Reports* and EPA Interior Volume Index for subcompacts.

**HONDA**

The Accord 4-Door Sedans



ideal laws in an ideal space. This realm is inhabited by differential equations, some of which describe quite extraordinary things. For example, there are differential equations that define a continuous form of Turing machine. The continuous machine can do everything its discrete counterpart is capable of and probably much more. I would only produce such equations for those who wonder what I mean by an analog computer in an ideal sense. To build such a machine from electronic components would probably result in a critical loss of accuracy and speed.

Anastasios Vergis of the University of Minnesota and Kenneth Steiglitz and Bradley Dickinson of Princeton University [see "Bibliography," page 136] have created a device seemingly closer than the analog Turing machine to reality. Consisting only of shafts, gears and cams, it solves a certain problem in logic known as three-satis-

fiability: a logical expression is formed of three-clauses, each a sum of three literals. Each literal is a logical variable (such as  $x$ ) or its negation (such as  $\bar{x}$ ) and the clauses are multiplied in the manner of the following example:

$$(x + y + z) \cdot (\bar{x} + \bar{y} + z)$$

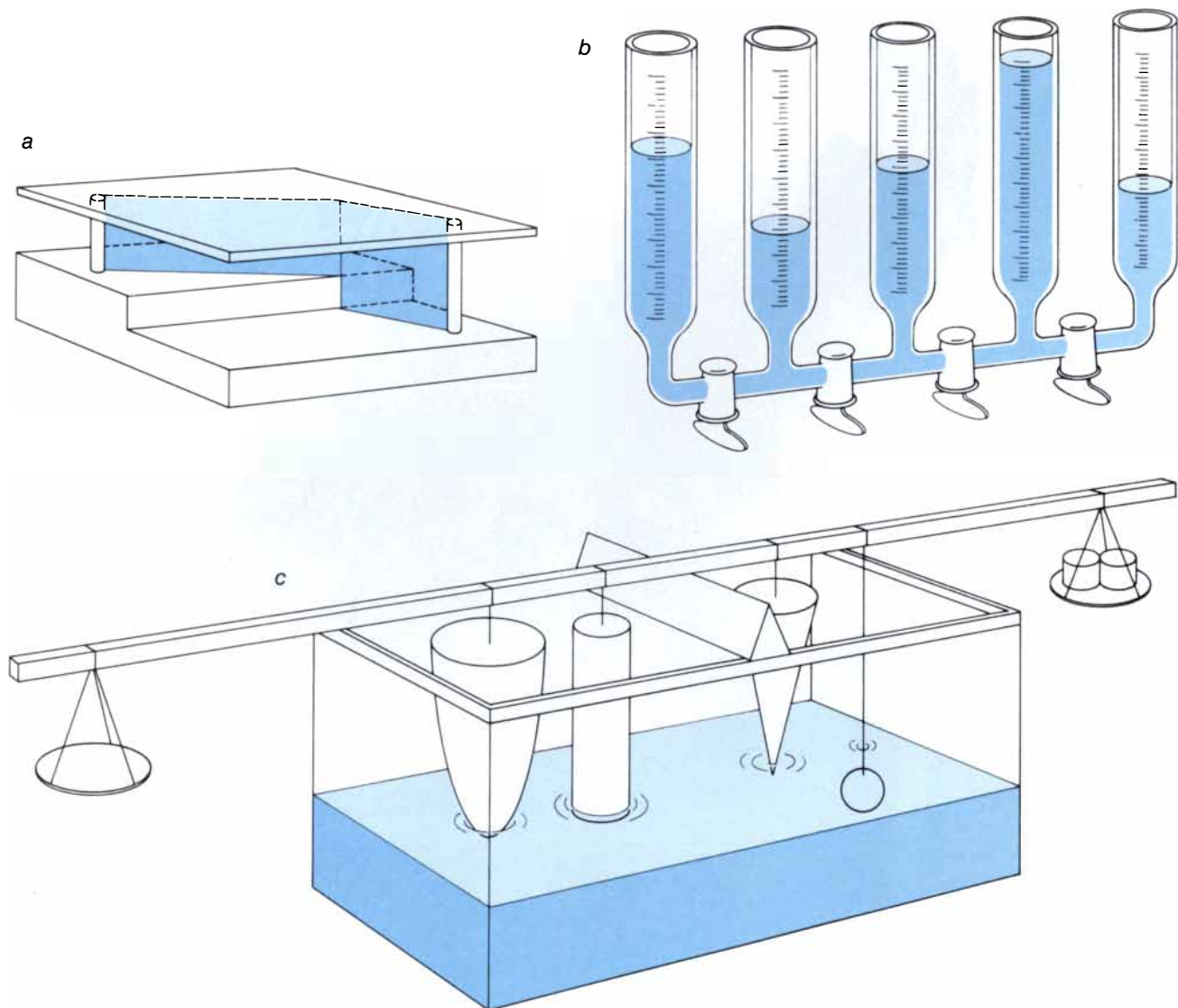
Is there a way to assign values of true and false to  $x$ ,  $y$  and  $z$  so that the entire expression is true? If the expression consisted only of the two clauses shown, the answer is yes. Let  $x$  be true,  $y$  be false and  $z$  be true. It is obvious at a glance that both clauses contain at least one true literal and so the entire expression becomes true. Although the problem appears simple presented in this way, it is really extremely difficult to solve on any level of generality. Actually it has a property dreaded by all computer scientists. It is *NP*-complete: no algorithm known (or ex-

pected) solves such a problem in less than exponential time. To satisfy  $n$  clauses requires  $2^n$  steps.

Yet the three theorists have found a configuration of gears, shafts and cams that can be set up to embody any instance of the three-satisfiability problem. The instance is satisfiable if and only if a particular shaft can be turned.

At the same time, because it is *NP*-complete, the three-satisfiability problem has a special feature: solve it quickly and one can solve any *NP*-complete problem quickly. But this is generally supposed by theorists to be impossible for a digital computer. Have the authors discovered an analog device that outperforms any digital one? They continue to search for some flaw in their machine's design.

If the machine actually performs according to the abstract description, the question remains of whether it would perform if it were actually construct-



Aqueous gadgets for illustrating refractions (a), averaging numbers (b) and solving a cubic equation (c)

ed. Something fundamental in the material world may conspire against this. It may be a basic law of nature that all physical systems can be described and simulated as quickly in digital terms as in any other. Indeed, the idea has been with us for some time that the universe is essentially digital in all respects. This would mean that any analog computer threatening to outperform its digital rivals would immediately become subject to flaws in accuracy or speed. The only possible advantage would lie in a high degree of parallelism. Water seeks its own level.

**R**esponse to the call for Core War network volunteers was encouraging. The director will be Mark Clarkson, who lives at 8619 Wassall Street, Wichita, Kan. 67210. Readers who would like to be on the network mailing list or who would like to volunteer for special functions should get in touch with Clarkson. In a future column, when planning is complete, I shall describe network activities and available programs.

The possibility of a DOS DOCTOR program on disk that interdicts viral infections of personal computers may be more remote than I thought. Norman Ramsey of Ithaca, N.Y., concludes that DOS DOCTOR can function diagnostically but not therapeutically. Ramsey designed a program that relied on "a second viral dos, which writes itself on diskettes exactly as does the diseased dos but does not cause disease." The benign virus thwarts infection by preemptively occupying memory space. His program, Ramsey adds, had some amusing bells and whistles. "It could even display a message at boot time, 'Fear not—DOS DOCTOR is with you,' to indicate to the user that his DOS is safe." Unfortunately Ramsey went on to invent a virus smart enough to evade its remedial analog.

The doctor-cum-virus idea also occurred to Joe Dellinger, a biophysicist at Stanford University. Strangely enough, his intention was to create not a doctor but merely an innocuous virus that could pass unnoticed from disk to disk. Dellinger was inspired by the analogy between programs and living creatures. Both might have parasites. In particular, a disk operating system could carry a virus whose continued reproductive success would hinge in part on its never being noticed. A key feature of Dellinger's virus is to infect only the DOS on 48K slave disks: master disks thus always contain a clean DOS. Taking up almost "no room" on infected disks, this undetectable virus checks the host DOS each time it is copied onto a new disk. Its vested interest in the continued health

of its host causes it to make certain that the copying is accurate. Dellinger reports that he shared the program for the benign virus with friends. Two of them, he fears, were careless, so that the strain may have spread beyond California. How can you tell whether your DOS has Dellinger's virus? He says there may be a half-second swishing noise when the virus copies itself.

Software viruses may not be as virulent as I implied in the March "Computer Recreations." Kenneth L. Kashmarek of Eldridge, Iowa, points out that a virus cannot spread to a disk carrying an operating system different from the host disk's. Kashmarek write-protects his disks, a precaution that will certainly halt the spread of any virus—unless, of course, the write-protect software has already been circumvented. He also asks whether it is right to discuss in public media viruses and other computer diseases. I do not doubt that compromised software is a serious subject. I think that computer epidemics of the kind I have been describing loom in the near future. In my opinion, "forewarned is forearmed." It is my hope that public discussion will spur research into antidotes.

Meanwhile the source of the rumor that sparked the invention of Core War has finally been revealed. In the May 1984 column I told the story of CREEPER and REAPER. In the March 1984 column I described ANIMAL, a game program similar to the Core War software that replicated itself in the computers of all players who used the same time-sharing system. The author of this program, John Walker of Sausalito, Calif., has stepped forward to claim responsibility for the most successful version. Actually ANIMAL is only part of the story. Inside the game program was another piece of software called PERVADE that was responsible for reproducing the program. Written in January, 1975, PERVADE was a subroutine able, when called, to create "an independent process that, while the host program was going about its business, would examine the directories accessible to its caller. If a directory did not contain a copy of the program, or contained an older version, PERVADE would copy the version being executed into that directory. PERVADE was very fastidious and took great care not to destroy, for example, an unrelated user program with the same name."

Enthusiasm for Wa-Tor (see last December's column) continues. Milt Boyd at Pinetree, P.O. Box 267, Amherst, N.H. 03031, will organize a user group for this and related software. Boyd is ready to collect versions of WATOR created by readers.

## "A BRILLIANT, ORIGINAL LEAP INTO THE FUTURE

rather than a restatement of the challenges of the past ... *Star Wave*, like quantum physics itself, is fascinating, otherworldly, counterintuitive."—RICHARD M. RESTAK, M.D., author of *The Brain*, writing in the *New York Times Book Review*

"*Star Wave* is full of thought-provoking material... Dr. Wolf's lucid book constitutes a great leap towards our understanding of the brain."—E. RAMON-MOLINER, M.D., Ph.D., Professor of Neuroanatomy, School of Medicine, Sherbrooke University, Quebec

"It gives one much to think about, especially someone like me who out of necessity has to adopt a more pedestrian approach to the many issues dealt with in this book."—JOHN A. WHEELER, Ph.D., Professor of Physics and Director, Center for Theoretical Physics, University of Texas at Austin

Ψ\*Ψ  
**STAR WAVE**  
Mind, Consciousness, and Quantum Physics  
An original interpretation of what quantum physics tells us about the human mind  
**Fred Alan Wolf**  
From the author of *Taking the Quantum Leap*, winner of the American Book Award

\$19.95  
at all bookstores

**MACMILLAN PUBLISHING COMPANY**  
866 Third Avenue, New York, NY 10022

# BOOKS

## *Seas of iron, patrician burghers, Vermeer's women, capsaicin, Tycho's nose, neutrinos*

by Philip Morrison

**T**HE CHEMICAL EVOLUTION OF THE ATMOSPHERE AND OCEANS, by Heinrich D. Holland. Princeton University Press (\$75; paperbound, \$24.50). The earth entered its Iron Age long, long ago. Iron-rich sediments are varied and common over geologic space and time, to be sure. But quantitatively most of the iron ore we know is found in half a dozen enormous deposits of a single kind, from Lake Superior to Western Australia. All were laid down during only a tenth of the planet's history, the half-billion-year epoch that closed about two billion years ago. These rocks are the banded iron formations, millimeter layers of iron oxides separated by layers of quartz chert. They form a grand pattern that maintains its coherence cross-country for hundreds of miles. The formations pile up half a mile thick overall, a third of the dark rock consisting of iron. The delicate layers are probably chronological markers as complex in their origin as tree rings, although we are not quite certain of that interpretation.

It is the "sheer magnitude" of the processes, which marshaled so much iron and precipitated it tranquilly from solution, that constrains local theories. Slowly filtering groundwaters, great rivers running down to shallow estuaries, volcanoes, midocean ridges—all fail to supply enough dissolved iron. Such sources can be excluded for another reason: no load of sedimentary debris dilutes the uniform darkly banded rock. Only the ocean to which all rivers flow is an iron source plentiful enough.

Yet seawater today holds a few parts per billion of dissolved iron; the soup is a thousandfold too thin. Silica for the chert must also be supplied at a concentration much above that of today's seas. Modern clues from ocean-bottom cores and analyses of the Black Sea support in detail a tricky biochemical scenario. The plot turns on the amount of oxidation present. Too much oxygen, and no iron will dissolve; too little, and manganese will

precipitate instead. The banded irons have no manganese, although such deposits are widespread at later epochs.

All that incompletely oxidized iron speaks eloquently for some oxygen, yet not too much. From the simple balances struck, the nature of ocean and air in that time emerges, a tale of meager organic reduction in the deeps, followed by an intense upwelling of iron-rich currents offshore, whose nutrients supported vigorous microbial life during tens of millions of years. Minor iron-bearing layers of three other mineralogical types are present in the formation as well; they all find a natural place in the model.

The seagoing narrative is supported by distinct evidence from land environments. Some deep old ores, like those of the Witwatersrand, are rich in anciently weathered grains of uraninite, a mineral too vulnerable to oxidation to accumulate under the more corrosive oxygenated atmosphere of latter days. Oxygen and carbon dioxide pressures can be monitored from those times by the survival of the grains.

Photosynthesizing life was abundant in the shallows where the banded iron formed but sparse in the open seas. The molecular oxygen that multicellular forms would one day demand was surely present, but no candle would have stayed alight in that atmosphere, oxygenated to about one part in 50 of today's level. Temperatures were modern; the tiny unicellular and filamentary forms we see in those ancient rocks lived about as their counterparts live today. They were cells without nuclei (of course), biochemically adept but thriving only close to pond shore or tidewater, not on dry land or in the open sea. The organisms that hungrily scavenge silica from today's seawater to deposit it on the bottom as ooze made of their intricate opaline skeletons had not yet evolved.

The oldest geochemical samples we have, 3.8 billion years old, come from a thoroughly mapped sequence of

much-disturbed sedimentary rocks in Greenland off Baffin Bay. Before that date, back to the time the earth was formed .7 billion years earlier, all evidence is celestial: it consists of meteorites and moon rocks as well as newly formed stars drafted as stand-ins for the early sun. The first fifth of this up-to-date monograph treats of that time, with distinct uncertainty.

The bulk of the volume is devoted to the long interval between the oldest samples we have found on the earth and the coming of animal life about .6 billion years ago to the shallow sea floor. The last fifth surveys air and sea during classical geologic time, signaled by fossils big enough to display.

Earth history can be seen as relatively dull, since life can hardly survive truly interesting times. To Heinrich Holland, life prospers only as it adapts to the geochemical circumstances that it helps to bring about. This is less of a wonder than the "intriguing and charming" hypothesis of Gaia, the idea that overall geochemical control is actively set by the biota at some desirable optimum. Ours is the best of all possible worlds only for the well-adapted; pervasive change destroys the rest.

The expected sequence of compounds that forms as seawater slowly evaporates appears to be captured in the salty record. There cannot have been much variation in seawater since the animals came; a factor of two or so is not excluded for certain components. Oxygen went early on the increase, or the calcareous organisms could not have developed. Whether the shift to today's heady air began soon or late after the time of the banded iron is not chemically clear; there are biological hints of early change.

The great extinction 65 million years ago is to be ascribed to the cosmic collisions that brought in the iridium tracer accompanying the event worldwide, but results of the inquest are not yet final. (Professor Holland's opinion is of weight; he is no man to be easily swayed by trendy inferences from uncertain data. As it is, this book was finished before those repeated comet collisions were suggested.)

Important facets of ocean chemistry have recently changed. The sulfur-isotope ratios in the salt beds clearly do not remain constant over long times. The strontium-calcium ratio in the shells of marine organisms has changed even within the past 100 million years; these two effects, and a few others less well established, may reflect variation in the amount of volatiles recycled through the hot basalts of the midocean ridges. The seawater that emerges from such encoun-



ters certainly contains dramatically increased amounts of calcium.

The text itself begins at the sun's beginning. But there is virtue in working back from harder evidence. The oldest rocks make plain that at the time of their deposition geologic processes were more or less the present ones. The crust was already continental, thick and folded; volcanoes were at work, and the seas were neither all-covering nor everywhere shallow.

Before then most conclusions are tentative; they march in accordance to still uncertain models for the origins of the solar system. The early atmosphere seems more likely to have been the mildly reducing mixture of carbon dioxide and nitrogen (no free oxygen) than the ammonia and methane favored two decades ago; both gases exposed to the bright solar ultraviolet unfiltered by ozone would have lacked stability. Sizable bodies of water were early to appear, as soon as high-pressure steam could condense; take-up of water by even red-hot rock seems slower than the early outgassing.

If the astronomers are right in associating the early sun with the strong ultraviolet flux seen in newly formed T Tauri stars, its high-energy photochemical effects on the atmosphere were dramatic. The famous synthesis experiment by Stanley Miller and Harold Urey, and many variations on its theme, make plausible a rain of organic compounds into the ocean, able to take part in still more complex reactions wherever local circumstances, from lava flows to cometary impact, had provided concentrated free energy and scarcer ingredients, such as phosphate. Chemical equilibrium is simple, but it only hints at reality. Life arose in some specialized geologic context, it would appear, during that first unseen but active tenth of earth history.

Technical but unusually clear, this comprehensive treatise is readable without any special knowledge of geochemistry by anyone sketchily familiar with physical inorganic chemistry. There are a few compilations of data too detailed to be of use except for research reference. The reader misses an account (even if only in refutation) of the recent daring proposal that the carbon cycle needs radical revision, to include protracted large-scale outgassing of carbon from the depths in the form of hydrocarbons.

**T**HE LAND OF STEVIN AND HUYGENS: A SKETCH OF SCIENCE AND TECHNOLOGY IN THE DUTCH REPUBLIC DURING THE GOLDEN CENTURY, by Dirk J. Struik. D. Reidel Publishing Company (\$39.50; paperbound, \$19.50). THE ART OF DESCRIBING: DUTCH ART IN

THE SEVENTEENTH CENTURY, by Svetlana Alpers. The University of Chicago Press (\$37.50; paperbound, \$17.50). Every scientific reader in English is more or less at home with the illustrious sequence: Gilbert, Hooke, Newton. Fewer of us know much of their contemporaries nearby: Stevin, Huygens, van Leeuwenhoek. Newton shines with unmatched brilliance, but on two sides of the North Sea the other luminaries are all of a similar magnitude. This cheerful account by a well-known senior American mathematician and historian of science, first published in Dutch quite a while ago and carefully revised, helps to redress our provincialism.

Struik has written a modest history, devoting a chapter each to the main names, in which their work is clearly set in the context of lowlands history. Maritime excellence went right along with the spirit of enterprise and expansion that arose there under institutions dominated in the absence of a powerful class of landed gentry by wealthy patrician burghers, seekers of profitable investment at home and abroad, who shared rule with princes of remarkable personal ability and tolerance. Simon Stevin wrote by choice in plain Dutch, whereas Gilbert spoke only in Latin to the learned about that great magnet the earth. Engineer and prolific expositor of practical calculation, Stevin was enough of a cosmic scientist to be a premature Copernican. He also tested Aristotle on the fall of light and heavy balls of lead, listening for the two thumps on a board. He and his partner Jan Cornets de Groot, burgomaster of Delft, reported the

gravitational dead heat 20 years before Galileo even went to teach at Pisa.

A brief story of these big names is filled out by an account of 100 memorable careers within Dutch and Belgian science and technology of the time, in addition to sketches of some precursors such as Mercator and Vesalius. Pioneers in making and publishing maps, in hydraulic engineering, in navigation and computing, collectors and illustrators of living forms in the far-flung new colonies, these lively people fill the pages with a sense of the boundless energy released at the end of foreign rule in about 1590. It has long been the fashion to note a parallel between the winning of American independence and the Dutch; our Founding Fathers might match Franklin and Jefferson against Simon Stevin and Constantijn Huygens, towering statesman and gifted poet, still better known in his homeland than his physicist son Christiaan. But the American Enlightenment does not rival that scientific flowering of the Netherlands, far richer, more populous and more cultivated than the raw lands of the 13 states.

Professor Svetlana Alpers is a brilliant analyst of ideas, working at Berkeley within the modern discipline of the history of art, who takes her data from the dazzling visual (and textual) legacy of the celebrated artists who might have jostled any of the engineers, craftsmen and scholars in the bustling streets of 17th-century Amsterdam and Delft. She offers a riveting new critique, painting by wonderful painting, to support a rather audacious hypothesis, not wholly new.

A brief review in the vein of a gener-



Jan Christaensz. Micker's View of Amsterdam

al scientific reader can hardly present the depth and fullness of thought here, the pages studded with the striking detail and power of the pictures—still life, landscape and domestic colloquy—so illustrative of the period. A simple outline is believable enough. The painters of the Italian Renaissance sought to tell a tale in a painting, drawing on symbol and myth, demanding prior knowledge of the viewer and imposing an iconographic search on their historians. You are called on to recall the foam-born Venus, however splendidly painted are the scallop shell and the Florentine woman.

Contrast the painting with the *View of Delft* by Jan Vermeer. There is the city as it appears across the water. The painting lies within a printmakers' tradition of city views, but it is marvelously converted into color, devoted in detail and striking in the contrast between light and shadow. There is some evidence (although a plethora of claims) that the artist's work was aided by the real image of a lens. The perspective is not that of a human viewer standing at the shore but is taken from high above. For these painters the picture surface was a record of nature, a nature mediated by optical means and by consummate skill with his medium. South of the Alps optical instruments were distrusted as being bearers of mere illusion; here they were welcome. The picture was the surrogate of the retinal image, itself (the northerner Kepler notes) optically generated by a lens not dissimilar to the products of the polishers of Middelburg.

This is the art of description, an art of surfaces and textures, often internal, the lemon opened as if anatomized. When texts occur, the words are usually made plain to read. Everywhere images are reflected in curved surfaces, perspective is not that of a viewing human spectator but that of an isolated eye, its projections taken to the working surface rather than to the invisible window of Alberti. The mapping impulse is strong; the Dutch in fact decorated their walls with many fine maps, and Vermeer himself, in his masterful *Art of Painting*, celebrates a map of the Netherlands, a real map, its varnished surface torn and bearing insets that stemmed from his own representation of Delft.

From this standpoint it makes sense to join Dutch art and science; they were both descriptions of the world. It would be shallow to see them as mere craft tricks carried to perfection. Hooke speaks of requiring more a "sincere Hand, and a faithful Eye," than any "depth of Contemplation," but he goes on to discuss the need for comparison of modes of lighting in order to

come to the nature of the microscopic view. Meaning is read out of the texts implied by an Italian painting; it is found visually in the Dutch by the use of eye and mind. Such a find is more of the world than it is of the culture, less imposed, although of course not at all simply given. Willebrord Snel van Royen, the pioneer of the law of refraction used so well by Descartes, spent much of his short life in measuring a net of triangles across the country, on a baseline that he fixed near Leiden when the flooded countryside froze to level ice. Maps arose from the travelers' skill and care; they became applied science in these years, for science is insightful description.

Of course, even Vermeer goes beyond description in his portraits of event and feeling. One of the most interesting of Alpers' penetrating arguments treats the role of women in Vermeer's work; for him women are "self-possessed," not mere attributes of men and families. Such psychological arguments are in the end quasi-statistical, explanations of trends and tendencies. There are exceptions: Galileo is one great southerner who trusted optics, and unmatched Rembrandt is a Dutch painter whose strong allegories in biblical costume are "Italian painting with a difference." Some exceptions!

Alpers' conclusion concerning Rembrandt is persuasive, but Galileo did not remain alone in Italy. Marcello Malpighi is only the greatest investigator of that place and century to depend on the microscope. Not easy to read because of its subtlety and learning, this bargain of a book is a pleasure to the eye and mind of serious scientific readers; it illuminates the nature of modern science during its first growth by a clear light cast from the side.

**PEPPERS: THE DOMESTICATED CAPSICUMS**, written and illustrated by Jean Andrews. University of Texas Press (\$35). A remarkably simple molecule is responsible for the fragrance of vanilla. It is a single six-member carbon ring trimmed with three small atomic tabs. Replace one of those tabs—the aldehyde CHO—by the appropriate, somewhat lumpy side chain (an amide of a methyl nonenic acid) built of half a dozen carbon atoms and an essential nitrogen link, and you have one of the five or more capsaicins, specific vanillyl amides, that constitute the "powerful pungent principle" of this all-American genus of plants. They are now the most important of all spice crops worldwide.

This striking monograph proceeds from a decade of work by a gifted Texas gardener, scholar and artist. "Peppers just get to you," Dr. Jean Andrews

admits. Attracted first as artist by their visual appeal, she became caught up, seriously and lightheartedly at once, in the scholarly lore that has grown around so prized and ancient a domesticate; the species has spread since Columbus' time from its New World home to the shores of Lake Balaton, the Niger delta, the Malabar coast and the headwaters of the Yangtze. In our brave new chinalated world there is reportedly even a Taco Bell in Paris.

The central element of the book is a bright display of paintings of 32 varieties representing the five currently recognized species of *Capsicum*. Drawn from the living plants—at one time Dr. Andrews had almost 90 cultivars in the garden, to allow choice among her fast-changing models—they are rendered with persuasive freshness, the varied scarlet, green and golden fruits as decorative as they are precise. "I can vouch for the almost clocklike behavior of the pepper flower," she writes: most of them open within two hours of sunrise for just a few days.

The pepper plant usually has a very leafy habit, but in these artful representations the leaves are suppressed into careful outlines, the better to display the stems set with bud, blossom and fruit, composites over time. No such careful pictorial survey of the genus has been published since the "beautifully illustrated" Latin monograph of K. A. Fingerhuth, published in 1832. Andrews uses the latest taxonomic framework, a consensus sponsored by the Food and Agriculture Organization, not yet published in full.

The 32 sample forms each are treated in a page or two of descriptive and historical comment. The first variety displayed is one known to every supermarket chile shopper: the Long Green/Red Chile, growing today by the square mile in the flat fertile lands around Oxnard, 30 coastal miles beyond Malibu into reality. It is the leader for tonnage and total value among all hot peppers grown in the U.S. The ubiquitous tins of roasted peeled green chiles still bear the name of Emilio Ortega, who brought the strain back to California from New Mexico to supply his cannery in Anaheim at the turn of the century. The last variety presented in the plates is an Andean species, which has violet blooms and dark seeds. Dominant in the highlands, the form is now grown mostly in home gardens as far north as Michoacán. The diagnostic dark seeds are worth noting; the long-day crops of the Andes are murderously pungent. (Peruvians say admiringly that this variety, *rocoto*, is able to raise the dead, or even kill a gringo!)

Capsaicin acts specifically on nerve

endings for pain and heat presented by the skin and by some mucous membranes. It is without color or flavor, but one part per million is perceptibly warm to the tongue and a quantity 10 times greater causes persistent burning; really hot fruits may contain 1,000 times as much as that. Prolonged use raises the chemical threshold of the nerve endings affected, although not their response to physical stimuli; people become accustomed to the daily heat of their chili peppers (tourists, beware). The sensation does not extend to the usual effects of a burn; the stuff does not redden the skin, dilate the capillaries or blister.

The glands that secrete the capsaicins are situated where the seed-bearing tissue joins the outer pod wall; the cross-wall partition contains most of the compound. Only a tenth or less is found in the seeds or the outer wall. Presence of the material is controlled by a single dominant genetic element, but the amount is variable even from fruit to fruit on the same plant. Maturity, rather than climate or cultivar, is what counts. Of course, a particular genetic variety, like the sweet, or bell, pepper, may lack capsaicin entirely.

Kids have long been punished or tested by the burning effects of the peppers. The usage goes on today; an old codex reproduced here bears witness to the pre-Columbian practice. A paper by two psychologists raises a burning question: "How do tens of millions of little chili haters become chili lovers each year?" The answer is far from complete; the authors relate the enjoyment of a basic adverse reaction to the sense of "constrained risk." They argue that "benignly masochistic activities are uniquely human," from horror movies to skydiving and tonic waters. There is surely some link with the more dangerous habituations at higher levels of the nervous system.

That bell peppers are a larger crop in the U.S. than all hot ones combined, in both tonnage and value, is demonstration that the fruits offer more than the delicious burn. In recent years the pigments of the plants have become economically important. Now that natural food colors have replaced the prohibited dyestuffs, the carotenoids of this genus are widely used. One extractable red pigment, capsanthin, constitutes about a third of the colorants in ripe red peppers. A great many processed foods are now tinted with pepper pigments; other foods, such as egg yolks and the skin of chickens, gain color from pigment given in feed. The feathers of cage birds are similarly touched up. The pepper fruits evidently brighten, to both eye and palate, the bland starchy diets of much

of the world (and also supply welcome vitamins).

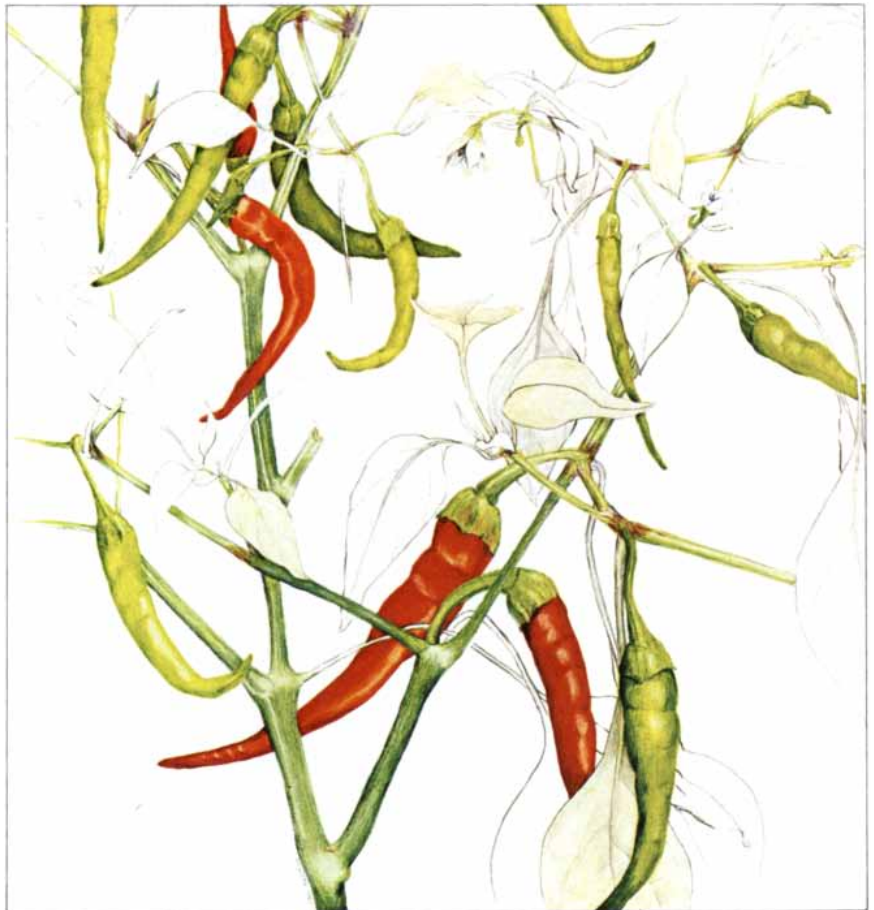
The discovery of vitamin C was made possible by the peppers. Albert Szent-Györgyi was on the track, but a year's work in Minnesota extracting the substance from beef adrenals had yielded him just 15 grams, not enough for anything. One evening, back home in Szeged, global hearth of paprika, he took a dinner dish he had not liked to the lab. "By midnight... I knew I had found a treasure trove." Within a month he had prepared three pounds of ascorbic acid from paprika. The synthesis and the Nobel prize were soon his. Peppers are rich in vitamins A and C, although their contribution to total intake is often unimportant.

The diverse feast for eye and mind that first-rate economic botany usually provides is here delightfully set out anew, from prehistory through Pickle Packers International at One Pickle and Pepper Plaza in St. Charles, Ill., to the scholarly biennial National Pepper Conference. Dr. Andrews has herself discovered an antidote to chili-burned hands: dip them in a solution of one part common household bleach, sodium hypochlorite, in five parts water and rinse. There is a chemical explana-

tion too. This book with its lasting set of vivid botanical portraits is dense with such gains and pleasures.

**T**HE ASTRONOMICAL SCRAPBOOK: SKYWATCHERS, PIONEERS, AND SEEKERS IN ASTRONOMY, by Joseph Ashbrook. Edited by Leif J. Robinson. Sky Publishing Corporation & Cambridge University Press (\$19.95). ASTROPHYSICAL TECHNIQUES, by C. R. Kitchin. Adam Hilger Ltd., distributed in the U.S. by IPS Inc., P.O. Box 230, Accord, Mass. 02018 (\$65; paperback, \$25). The late Joseph Ashbrook was for 25 years the learned and unruffled editor of the best-known of all American amateur-astronomy journals. He was not a trained historian but a professional astronomer of lifelong devotion and prodigious memory, who took delight in the half-forgotten volumes that crowd the stacks of the library at the Harvard College Observatory. Six times a year he used a few pages to tell some story he had fished out of those overflowing annals, with "a particular fondness for misfits, ill-conceived projects, and far-away places."

This volume is an illustrated anthology of 83 of his best pieces, selected



*Cayenne pepper, a prized and ancient domestic*

by his friend and successor as editor. They begin with the silver nose prosthesis of Tycho Brahe (which turned out to have a high copper content) and end with a famous enigmatic woodcut, one widely misused. Notes cite precise references, obscure and diverse, themselves an education in scholarship.

One of the most influential of the Ashbrook pieces reviews the sport of observing very thin crescent moons, seriously examined as part of chronology, since many calendars old and new begin with new-moon sighting. A moon 20 hours old is only a glowing wire of fine gold, yet there are certified reports of 15-hour finds, hard to accept by one who has tried the art. Another note summarizes the appearance—this year we mark its centennial—of the supernova near the nucleus of the Andromeda galaxy, first seen in August of 1885. A contemporary drawing is reproduced, along with a modern photograph of the starlike nucleus, which lies not far from the supernova. That event, vivid and puzzling, is forever entangled with the slow recognition of the distance of external galaxies.

The collection ends with that much reproduced cut of the medieval pilgrim thrusting his head out of the sphere of the fixed stars to examine the wheels beyond. Dr. Ashbrook documents how within the past decades the scholars came tardily to realize that the picture was not old at all but a clever 19th-century pastiche done by a gifted popularizer of astronomy. Camille Flammarion, first-rate artist and engraver, presented it along with a suitable caption in a book he published in 1888. But brilliant imitation, like

irony, is hazardous. Detached from its origin by a single error, then probably cast to the winds by some commercial picture library, the image has since sprouted in a wide variety of works by authors who captioned it more or less at face value, all unknowing of the source. It adorned the dust jacket of a widely read popular chronicle of science only last year.

Gaudy images and excited reports have kept us all aware that we live in a time of blossoming new astronomies. It is novel instrumentation, based on the powerful physics and engineering of the day, that is at the foundation of what is new; the theorists are glad enough to apply old Newton or Rayleigh or Einstein. This unusual textbook by an astronomer at Hatfield Polytechnic near London is a pioneer effort to "provide a coherent state-of-the-art account" of the wide range of techniques thriving these days in observatories from mountaintop to mine shaft, with a little attention to orbit. Its level is that of an undergraduate student of science; it centers on function and principle, with the intent of orienting a prospective user rather than preparing a reader to engage in detailed design.

The text considers detectors over the entire gamut, from radio to gamma rays, and on to neutrinos and gravitational radiation as well. Its next division deals with imaging, from the familiar photographic plate to scanning interferometry and even radar. Photometry, spectroscopy old and new, polarimetry and instrumentation for studies of bright sunlight close the volume. Problems are proposed for each

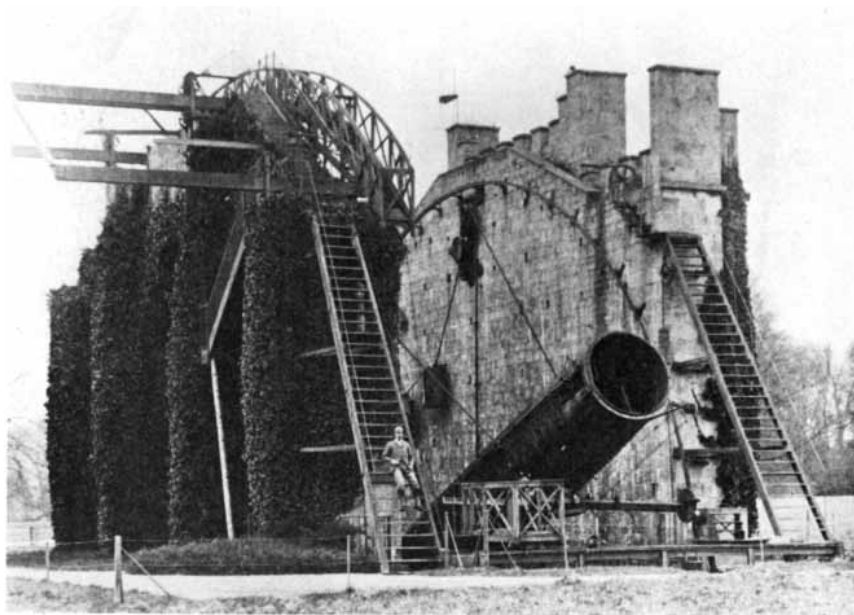
main section. They are rather easier than one might expect.

The strength of such a book is its comprehensive quantity; the weakness has the same roots, the sketchy nature of accounts given over so wide a range. Dr. C. R. Kitchin is at his best when he is near physical optics. He brings a helpful unity of insight into the nature of telescope resolution, radio interferometry and the most-used new spectroscopes, the spectroscopic adaptations of interferometric schemes. Criteria such as signal-to-noise ratio, quantum noise, bandwidth and dynamic range, which open another road to a deep unity, are not ignored, but they are not fully exploited.

Most missed is some systematic tabulation of performance for the many systems treated. The eye will detect a signal at the level of  $10^{-15}$  watt; the big optical telescopes do 100 million times better; gamma-ray detectors, whose designers will settle for a few photons of the right kind every day, are only as power-sensitive as is the unaided eye in the optical band.

Ingenuities are unfolded here that range from vidicon cameras to vibrating gravitational-wave bars. One neat diagram shows the grazing-incidence X-ray imaging reflectors with metal mirrors, famous from the results of the Einstein satellite. Their resolution is limited by surface irregularities at the scale of X-ray wavelengths, about 10 angstroms, rather than by diffraction effects. In the world of optical light, consider the Treanor direct-vision spectroscope. Three prisms are made up into a glass block with parallel entry and exit faces. It spreads the spectrum without deviating the direction of the light beam, by combining optical glasses that have matching refractive indexes at the undeviated wavelength but different rates of index change with color. Because the telescope beam is larger than the prism placed in its path, every star in the field will appear as an undeviated image at the reference wavelength, surrounded by a short ribbon spectrum. The device is used for radial-velocity studies of star fields.

There are dozens of such schemes here, described rather in general, although often with the key formulas that govern them. Specific instruments rarely appear, but available photographic emulsions and the optical filter schemes for standardized color bands are treated in concrete detail. The intricate master art of digital data collection and treatment is hardly touched on. For readers with some background in physics the book is an all but unique public door to the swift-growing architecture of the new astronomies.



*Telescope with which the third Earl of Rosse discovered spiral nebulae*



His tour of duty was over. This was his final good-bye. He remembered all the good times, the joking and that special closeness that comes from sharing not only victory, but defeat.

As he shook hands with Willi, Rolf, Dieter and the others, he realized they had become brothers.

And that was something he'd never forget.

**Call West Germany. Ten minutes can average just 81¢ a minute.\***

Saying good-bye is never easy—but saying hello is, with AT&T. A ten-minute phone call to West Germany on AT&T can average as little as 81¢ a minute.

Just dial the call yourself any night from 6 pm until 7 am.

If you don't have International Dialing in your area, you'll still get the same low rate as long as special operator assistance is not required.

**AT&T International Long Distance Service.**

# On taking leave of Germany... and his teammates.

### Germany

Rate Level	Average Cost Per Minute For a 10-Minute Call**	Hours
Economy	.81	6pm-7am
Discount	\$1.02	1pm-6pm
Standard	\$13.5	7am-1pm

\*\*Average cost per minute varies depending on length of call. First minute costs more; additional minutes cost less. All rates are for calls dialed direct from continental U.S. during hours listed. Add 3% Federal Excise Tax. For further information, call our International Information Service, toll free 1 800 874-4000.

\*During Economy time periods

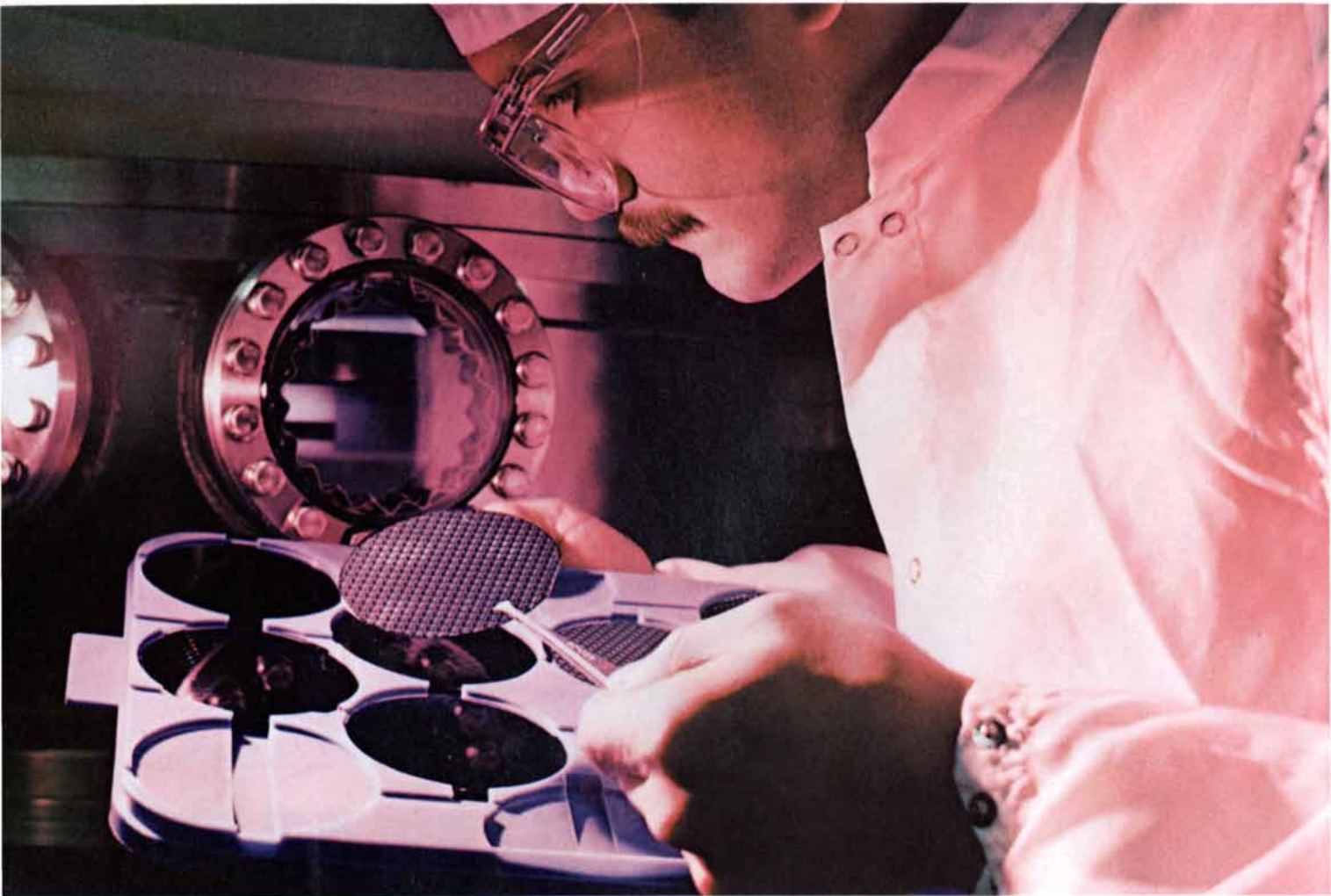
© 1985 AT&T Communications



**AT&T**

The right choice.

# A proving ground for computers.



Nobody makes more automotive computers than General Motors.

No other automotive company subjects their computers to more torture tests. 40 degrees below zero to 185 above. Electron microscope scanning. Moisture. Vibration. Testing that equals 10 years of typical driving.

Our computerization gives you improved driveability. Smooth performance regardless of speed. An indoor climate control system. And greater fuel economy.

And the reliability of our computers makes the operation of your new GM car or truck safer and more efficient.

We believe in taking the extra time, giving the extra effort and paying attention to every detail. That's what it takes to provide the quality that leads more people to buy GM cars and trucks than any other kind.

That's the GM commitment to excellence.



## Nobody sweats the details like GM.

Chevrolet • Pontiac • Oldsmobile • Buick • Cadillac • GMC Truck

# The Choice of Technology

*The ripple of a new technology throughout the economy leads to effects that are not predictable by examining each industry in isolation. Every decision to introduce technology could be based on data available to all*

by Wassily Leontief

The manager of a steel mill deliberates the purchase of an electric furnace that will modernize the mill's operations. The new furnace will reduce the costs of labor, but it will also consume prodigious amounts of electric power. The manager assumes there will be no change either in the costs of scrap and other raw materials or in the price of the steel product. Given such constraints a cost comparison shows that the high price of the new furnace and its auxiliary equipment would reduce rather than increase the overall rate of return on the mill's total capital investment. The plans for the furnace are scrubbed.

If the manager's analysis had taken account of technological changes in other sectors of the economy, the decision to buy the furnace might have been different. Imagine that as new electric furnaces are introduced by the steel industry, public utilities adopt more efficient methods of generating electric power. At the same time, imagine that the automobile industry shifts to new models of cars that increase the demand for high-strength steel. Such changes could readily lead to shifts in the unit cost of electricity and in the price of high-strength steel. The investment in the electric furnace might well have paid off.

It is unfortunate that many real managers must survey their purchasing alternatives from the same limited vantage as the one assumed by my fictional manager. Real managers obviously have a thorough knowledge of their current technology, and it is likely they have fairly reliable information about the technology that could replace the current one. Furthermore,

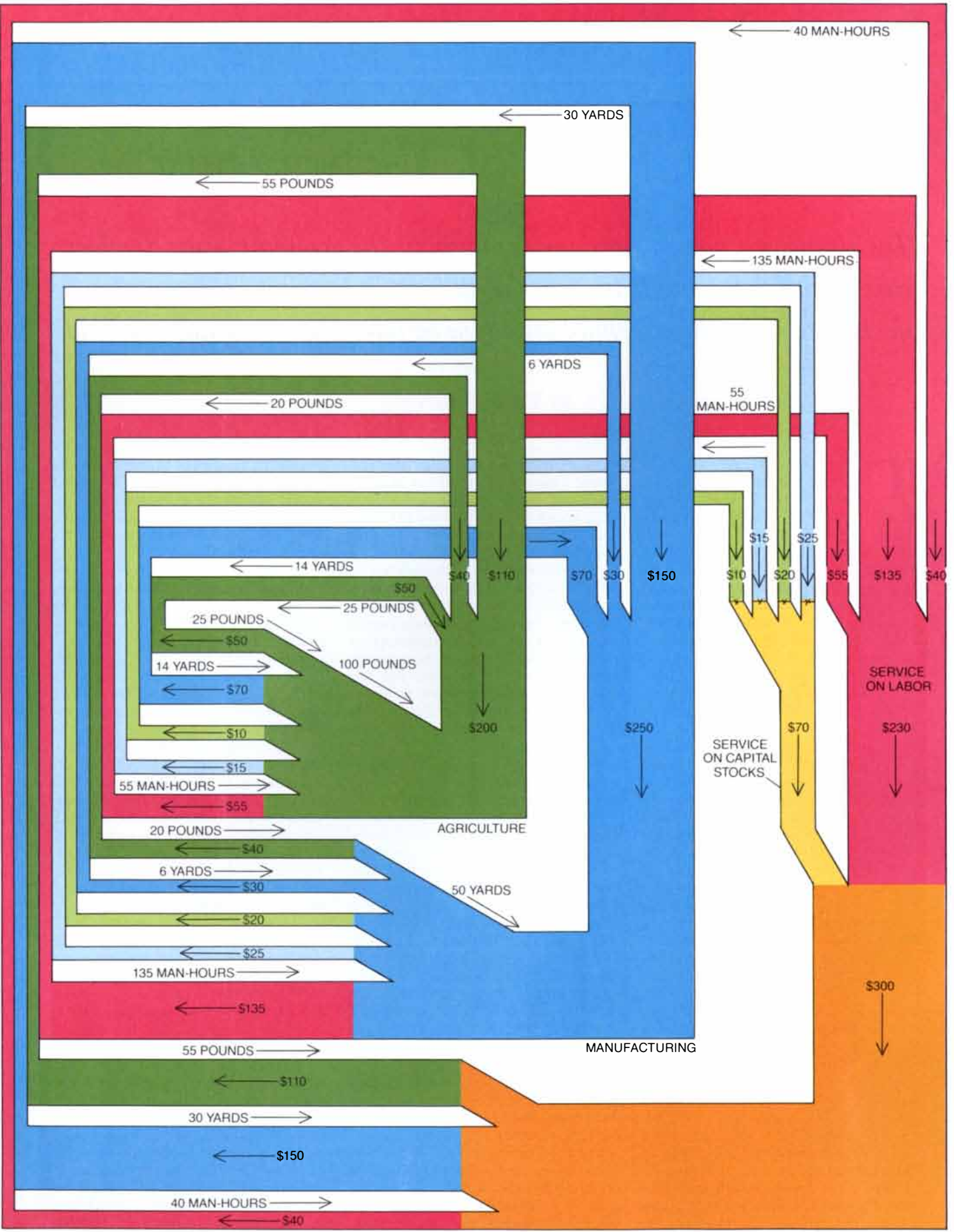
most of them acknowledge the economic truism that the effects of every economic sector propagate to every other sector by means of the price system. Yet managers know very little about the new technologies whose introduction may be contemplated by other industries. The estimated costs and benefits of the new technology in the home industry are consequently based on designs and technologies operating in other sectors of the economy at the time of the estimate. The fact that the introduction of new technology in other industries could alter the investment climate for the new technology in the home industry is to a large extent ignored. In the absence of adequate data or a sound analytic technique there is little else one can do.

Fifty years ago I developed a technique called input-output analysis that could be adopted by each economic sector to make a more informed decision about the introduction of new technology. Two years ago Faye Duchin and I, together with seven of our colleagues at the Institute for Economic Analysis of New York University, assembled the data needed in order to apply input-output analysis to the current prospects for technological change. The data describe several possibilities for the workings of the American economy in the year 2000. They are based on the input needs of technology that can be expected to replace the present methods of production in the next 15 years. Our method did not require us to make any projections about unknown, future technology. On the contrary, the technology we considered is already well understood, but it has not yet been widely intro-

duced. Its readiness to begin playing a role in the economy is based on expert judgment, primarily by engineers, and sometimes on experience with newly constructed plants in the relevant industries.

An input-output analysis of any economy based on such data can lead to two major benefits. First, it can show each industrial sector how to realize its maximum possible rate of return on invested capital. Second, because the choices by each sector needed to realize this potential are predictable, the analysis might eventually enable the strategic planner to develop a picture of the future economy that is based on empirical data and on the self-interest of the various economic sectors. For example, the Japanese government, in cooperation with private industry in Japan, has recently undertaken such an analysis of the entire Japanese economy. That is precisely the kind of task for which input-output analysis was originally developed.

Any analysis of an economic system that purports to describe the future ebb and flow of real goods and services must rely as heavily as possible on current empirical data. There is a wide consensus that the ongoing introduction of computers, computer-numerically-controlled (CNC) machines, robots and computerized telecommunications will profoundly affect the structure of the U.S. economy in the next two decades. At the General Motors Corporation, for example, managers predict that in two years 90 percent of all new capital investments will be in CNC machines, which can be programmed to follow a predeter-





mined sequence of steps in the cutting and forming of metal. The market for industrial robots is expected to grow at an annual rate of between 30 and 40 percent in the next five years. In the office, particularly in industries such as banking, insurance, legal services and government, computers are expected to assume most of the routine, clerical work, thereby displacing a large number of white-collar clerical workers.

Accordingly our study focused primarily on technological changes related to the introduction of computer-based automation. We made no effort to assess the economic effects of the technological changes that can be expected in agriculture from the genetic engineering of crops, in mining from improved methods of prospecting and mineral recovery or in various industries from the substitution of materials such as plastics, ceramics and fiberglass for metals [see "Industrial Microbiology," SCIENTIFIC AMERICAN, September, 1981, and "The Mechanization of Work," SCIENTIFIC AMERICAN, September, 1982].

We considered four separate scenarios, or sets of assumptions about the rate at which computerized automa-

tion is introduced into the economy. In the first scenario we assumed that no automation or any other technological changes are made after 1980. On the other hand, the first scenario does not presuppose a static economy: the final demand for goods and services is projected to continue growing as it has in recent years through the year 2000. Obviously the first assumption made in this scenario is already counter to the facts, but it serves as a baseline against which one can compare the other three scenarios.

In the second and third scenarios we assumed the investment in computer-based automation would grow at increasingly rapid rates; in the fourth scenario we assumed the rate of increase is the same as the rate in the third scenario, but we added the assumption that the final demand for computers and computer-based products as well as the intermediate demands for such products would grow at a faster rate than it would in the third scenario. In this article I shall discuss only the results of comparisons made between the first scenario, which I shall call the old technology, and the third scenario, for which the investment in computer-based automation is most rapid. I shall

refer collectively to the technologies introduced under the third scenario as the new technology. To give a rough sense of the differences in our assumptions, the total investment in computers and robots is about 15 percent higher under the new technology than it is under the old technology in the 1980's; in the 1990's the corresponding total investment is about 30 percent higher.

It is worth mentioning a few of the technological changes that would be brought about under our third scenario. One of the most striking possibilities is that the labor force required under the new technology would be significantly smaller than the labor force required under the old technology to produce the same bill of goods. Although we project employment increases for both technologies, the new technology would require roughly 11 million fewer workers than the old technology would by 1990. By the year 2000 the corresponding difference is 20 million workers.

The composition of the labor force is also different under the two technologies. Professionals would make up almost 20 percent of the labor force in the year 2000 if the new technology is adopted, whereas they would make up only 14.5 percent under the old technology. In 1978 professionals made up 15.6 percent of the labor force. The proportion of service workers would be about 2 percent higher under the new technology in the year 2000 than it would be under the old technology, although in both scenarios the proportion would be slightly greater than it was in 1978.

For managers and clerical workers the trend is in the opposite direction. Under the old technology managers would make up nearly 11 percent of the labor force in the year 2000, and clerical workers would make up more than 18 percent. Under the new technology the proportion of managers would be only 7.2 percent, and the proportion of clerical workers would be only 11.4 percent. In 1978 9.5 percent of all workers were managers and 17.8 percent were clerical workers.

I can only suggest here a few of the other changes that would take place under the new technology. One important change is a reduced requirement for iron and various alloys of iron. The reduction follows in part because CNC machines generate less scrap metal than workers do. There is also a significant increase projected in the demand for nonferrous metals and a decrease projected in the demand for paint. It is estimated that robots employed for industrial painting can save between 10

	AGRICULTURE	MANUFACTURING	HOUSEHOLDS	TOTAL
AGRICULTURE	25 POUNDS	20 POUNDS	55 POUNDS	100 POUNDS
MANUFACTURING	14 YARDS	6 YARDS	30 YARDS	50 YARDS
CAPITAL STOCKS FROM AGRICULTURAL PRODUCTS	50 POUNDS	100 POUNDS		
CAPITAL STOCKS FROM MANUFACTURING PRODUCTS	30 YARDS	50 YARDS		
LABOR	55 MAN-HOURS	135 MAN-HOURS	40 MAN-HOURS	230 MAN-HOURS

**INPUT-OUTPUT MODEL** of an economy traces the flow of goods and services among its various sectors. The simple economy represented in the diagram is divided into three sectors: agriculture (*dark green*), whose output is measured in pounds of crop; manufacturing (*dark blue*), whose output is measured in yards of cloth, and households (*orange*), which consume the end products of the economy, own the capital stocks of agriculture and manufacturing and provide labor measured in man-hours (*red*). The outputs of agriculture, manufacturing and labor are given by the rows to which they correspond in the table above; the total output is given by the rightmost number in each row. The input and the capital stocks needed to produce the total output of a sector is given by the numbers in the column to which the sector belongs. For example, to produce its total output of 50 yards of cloth, manufacturing requires 20 pounds of crop, six yards of its own output, 100 pounds of capital stocks made up of agricultural products, 50 yards of capital stocks made up of manufactured products and 135 man-hours of labor. In the diagram on the opposite page labor and material inputs flow into each sector from left to right and flow out of each sector from bottom to top. Money circulates through the sectors in the opposite, clockwise sense. Money paid to investors in the household sector by the other two sectors for the services of capital stocks from agriculture is shown in light green; money paid by the two sectors for the services of capital stocks from manufacturing is shown in light blue. The total annual returns on all these investments are shown in yellow. The steady state of flows is achieved by successive approximations of the price ratios among unit outputs of crop, cloth and labor. If the cost of labor in the simple economy is \$1 per hour and the rate of return on capital is 10 percent, the crop is worth \$2 per pound and the cloth is worth \$5 per yard. The width of each pipe in the diagram is proportional to the number of dollars in the pipe when the prices have reached this equilibrium. More detailed analyses can be done if more sectors are listed. Recent input-output tables have divided the U.S. economy into more than 600 sectors.

and 30 percent of the paint ordinarily applied by painters.

Industries with large information-processing needs whose operations are in small establishments will add computers rapidly to their capital stock. Such industries include retail trade, real estate, hotels, amusements and educational services. Desktop computers and electronic cash registers will make up the bulk of these investments.

The most sophisticated automation we envision in the factory is the so-called flexible manufacturing system. In such an installation several CNC machines are assembled into a manufacturing unit that carries out several op-

erations automatically. For example, one such unit could bore a hole in a block of metal, turn the block on a lathe, mill the turned workpiece into a specified shape and then finish the piece in a wash. All the machines in the system are controlled by a hierarchy of computers, and they are linked by a conveyor. Among the most likely candidate industries for the early introduction of flexible manufacturing systems are the makers of machine screws and stampings, metalworking machinery and aircraft.

Capital goods would make up a larger proportion of the total national output under the new technology than

they would under the old. The output of goods for use by other industries would be almost 9 percent higher under the new technology than it is under the old, and investment would be more than 42 percent higher. The increased production of capital goods would slow the current transfer of employment from manufacturing sectors to service sectors over the next 15 years.

In order to compare the overall productivity of the economy under the old and new technologies it is essential to describe both technologies within each economic sector in some detail. A concise description of each technology can be understood as a kind of cooking recipe. The inputs of the industrial recipe, like the ingredients of an ordinary recipe, are specified in the amounts needed to produce one unit of output. Both labor and the goods and services provided by the various sectors of the economy are considered inputs.

An industrial recipe, like the one used in the kitchen, must also specify the number of pots and pans, baking ovens, blast furnaces, metalworking machines, industrial buildings and so on necessary to process the inputs. In other words, the recipe must list the capital stocks needed by an industry, measured, say, in the number of units of each stock required per unit of the annual industrial output. The dollar value of these stocks represents the annual capital investment of the industry per unit output, and the owners of capital stocks in each industry seek to maximize the rate of return on this investment. The rate of return achieved under different technologies, stated as a percentage of the annual capital investment, determines which technology is chosen.

A systematic tabulation of the industrial recipes employed by all the sectors into which the economy is divided can provide a concise and detailed description of the technological structure of the economy at a given time. The structure determines the inputs each sector must receive from other sectors, and it specifies how much of the output of each sector must be delivered to each of the other sectors. Moreover, the technological structure determines how many workers must be employed by each sector in each kind of job and how many machines and other capital goods of various kinds each sector must maintain.

Given such technological information one can set up a system of equations from which the prices of all inputs and outputs within the economy can be calculated. The equations show that the price of any good or service depends not only on the wages of

	AGRICULTURE	MANUFACTURING
AGRICULTURE	.25 POUND	.4 POUND
MANUFACTURING	.14 YARD	.12 YARD
CAPITAL STOCKS FROM AGRICULTURAL PRODUCTS	.5 POUND	2 POUNDS
CAPITAL STOCKS FROM MANUFACTURING PRODUCTS	.3 YARD	1 YARD
LABOR	.55 MAN-HOUR	2.7 MAN-HOURS

a UNIT PRICE OF CROP =

$$\begin{aligned}
 & (.25 \text{ POUND INPUT PER POUND PRODUCED}) \times (\text{UNIT PRICE OF CROP}) \\
 & + (.14 \text{ YARD INPUT PER POUND PRODUCED}) \times (\text{UNIT PRICE OF CLOTH}) \\
 & + (.5 \text{ POUND CAPITAL STOCK PER POUND PRODUCED}) \times (\text{UNIT PRICE OF CROP}) \times (10\%) \\
 & + (.3 \text{ YARD CAPITAL STOCK PER POUND PRODUCED}) \times (\text{UNIT PRICE OF CLOTH}) \times (10\%) \\
 & + (.55 \text{ MAN-HOUR INPUT PER POUND PRODUCED}) \times (\text{UNIT PRICE OF FARM LABOR})
 \end{aligned}$$

b UNIT PRICE OF CLOTH =

$$\begin{aligned}
 & (.4 \text{ POUND INPUT PER YARD PRODUCED}) \times (\text{UNIT PRICE OF CROP}) \\
 & + (.12 \text{ YARD INPUT PER YARD PRODUCED}) \times (\text{UNIT PRICE OF CLOTH}) \\
 & + (2 \text{ POUNDS CAPITAL STOCK PER YARD PRODUCED}) \times (\text{UNIT PRICE OF CROP}) \times (10\%) \\
 & + (1 \text{ YARD CAPITAL STOCK PER YARD PRODUCED}) \times (\text{UNIT PRICE OF CLOTH}) \times (10\%) \\
 & + (2.7 \text{ MAN-HOURS INPUT PER YARD PRODUCED}) \times (\text{UNIT PRICE OF MANUFACTURING LABOR})
 \end{aligned}$$

c UNIT PRICE OF CROP =

$$\begin{aligned}
 & 1.757 \times (\text{PRICE OF FARM LABOR PER POUND OF CROP}) \\
 & + .383 \times (\text{PRICE OF MANUFACTURING LABOR PER YARD OF CLOTH}) = \$2 \text{ PER POUND}
 \end{aligned}$$

d UNIT PRICE OF CLOTH =

$$\begin{aligned}
 & 1.351 \times (\text{PRICE OF FARM LABOR PER POUND OF CROP}) \\
 & + 1.577 \times (\text{PRICE OF MANUFACTURING LABOR PER YARD OF CLOTH}) = \$5 \text{ PER YARD}
 \end{aligned}$$

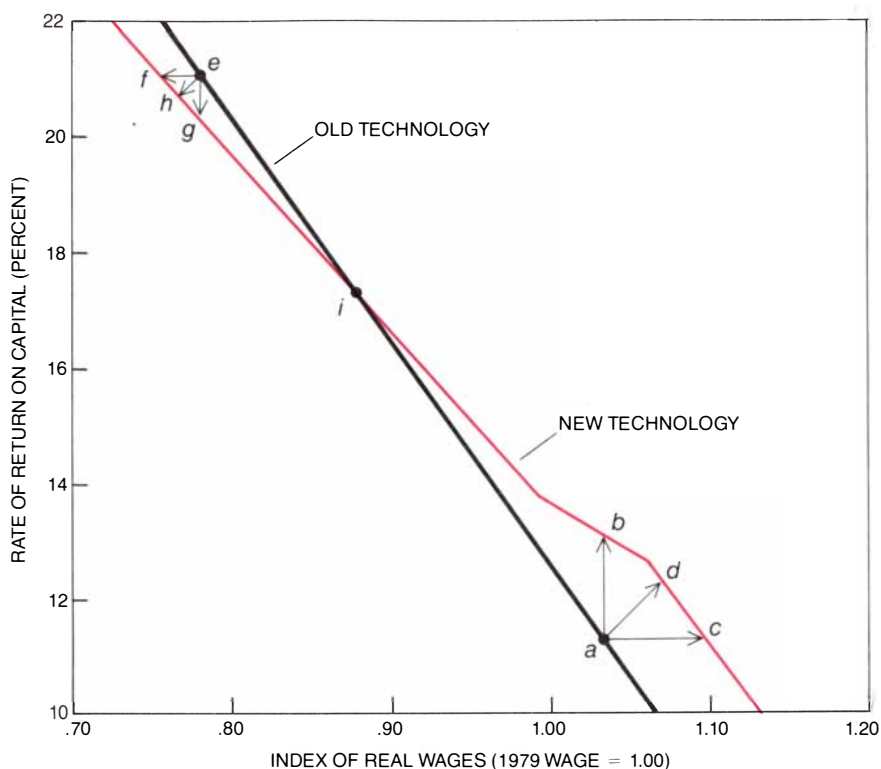
**CALCULATION OF PRICES** for goods and services is illustrated for a three-sector economy. The annual rate of return on invested capital is assumed to be 10 percent, and the entries for material inputs and capital stocks in the input-output table at the top of the illustration are given in units of input and stock necessary for each unit of output to maintain production at a certain annual level. For example, the entry in the first row of the second column shows that .4 pound of crop from the agricultural sector is needed for every yard of cloth produced by the manufacturing sector. The number is derived from the empirical ratio of the 20 pounds of crop needed to produce the 50 yards of cloth in the economy shown in the table on the preceding page. The calculation of unit prices for crops and cloth begins by equating each price with the sum of the costs of its inputs. The procedure leads to two equations (a, b) each of which includes the two unknown quantities, namely the unit price of crop and the unit price of cloth. They are solved in the lower part of the illustration (c, d) in terms of the costs of farm and manufacturing labor needed for each unit of output. For the illustration the wages in both sectors are each assumed to be \$1 per hour.

workers and the rate of return on capital prevailing in the industry producing that good or service but also on the wages paid and the rates of return earned in all other industries as well.

Once the prices have been determined for a given combination of wages and rate of return on capital, the measurement of the wages can be adjusted for changes in the average price level of consumer goods. One assumes that each consumer buys a given combination of products each year, and the total cost of that combination is computed for all years of interest to the economist. The ratio of this cost in any given year to the cost in some base year is the cost-of-living index. The money wages, or the actual dollar payments for labor in each year, can then be converted into real wages, or the effective wages under any new combination of money wages and rate of return. The real wages are equal to the money wages divided by the cost-of-living index.

As long as the technological base of an economy remains unchanged, or in other words as long as the industrial recipes remain the same, there is a definite, one-to-one relation between the rate of return on capital and the level of real wages. Under a fixed technology there is only one corresponding level of real wages for any given rate of return on productively invested capital. To put the point another way, to a given level of real wages there is only one corresponding rate of return on capital. Not surprisingly, the relation between the rate of return and real wages is an inverse one: within the framework of any fixed technology the interests of investors and wage earners collide. An increase in the rate of return on capital implies a reduction in the real wages and vice versa: an increase in the real wages implies a reduction in the rate of return. The reciprocal relation between real wages and rate of return for a given technology can be interpreted as a measure of the productivity of the economy based on that technology.

A change in the technological base of the economy can be described as a replacement of old industrial recipes by new recipes in some or all of its sectors. Such a change is bound to bring about a shift in the one-to-one relation between real wages and the rate of return on capital. The new relation can be determined by setting up a new system of equations based on the new industrial recipes whose introduction is being studied. Again the equations can be solved for the prices of goods and services, given the level of money wages and the rate of return on capital. The level of real wages corre-



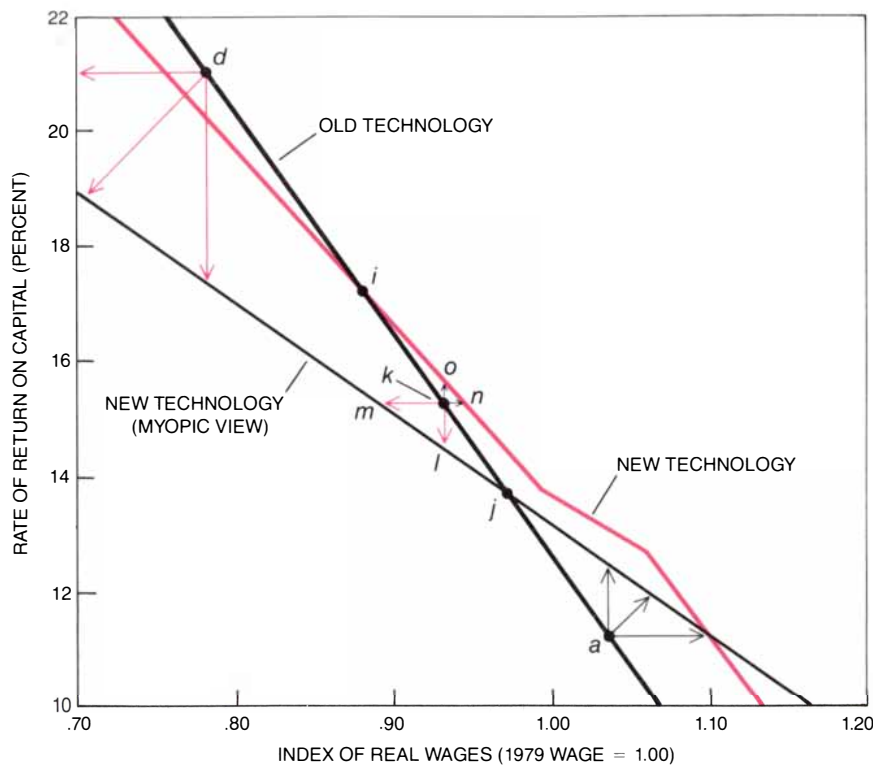
**RELATIONS** between real wages and the rate of return on capital are shown for the U.S. economy operating under two distinct sets of technological conditions. The line in black shows the relation that can be achieved under the combination of industrial recipes employed by various sectors of the economy at the end of the 1970's. This combination, the so-called old technology, is reflected in an input-output table describing the flows of goods and services among 89 sectors of the economy. For any given money wages and for any rate of return on capital the unit price of each output can be calculated just as it was for the three-sector economy shown in the illustration on the opposite page. The level of real wages corresponding to the assumed money wages and rate of return can then be calculated from a standard cost-of-living index. The real wage is set equal to \$1 per hour for the 12.5 percent rate of return on capital that was paid in 1979. Note that if the rate of return were to increase, the real wages would fall; if the real wages were to increase, the rate of return would fall. The line in color shows the results of the same calculations carried out for a second combination of industrial recipes. The second combination of recipes summarizes the so-called new technology, or the economic recipes that could prevail by the year 2000 as a result of the introduction of computer-based automation. The superposition of the two lines shows how the choice between the two technologies depends on the current rate of return on capital (or on the current real wages). For example, if the rate of return were 11 percent under the old technology (a), the introduction of the new technology would enable it to increase to about 13 percent (b). If one were content with the 11 percent rate of return on capital, real wages would increase from \$1.03 to almost \$1.10 per hour under the new technology (c). Both owners and wage earners could also share in the benefits of the new technology (d). On the other hand, if the rate of return were as high as 21 percent under the old technology, real wages would be about 78 cents per hour (e) and the costs of a changeover to the new technology would lead to a drop in the real wages (f), a drop in the rate of return (g) or both (h). There is neither incentive nor disincentive to switch from the old to the new technology if the rate of return at the time of the switch were about 17.5 percent (i).

sponding to each rate of return can then be calculated from the cost-of-living index.

The shift to the new technology in some or all of the economic sectors could lead to a rise in the overall productivity of the economy. For example, it might allow the rate of return on capital to rise without any consequent reduction in the level of real wages. Conversely, it might allow the wage rates to increase without a reduction in the rate of return on capital. Indeed, although the interests of owners and wage earners collide within any single

technology, a change in the technology can be in the interests of both.

The input-output analysis of prices within a given technology can best be understood by considering a simple numerical example. Imagine a simple economy that is made up of only three sectors: an agricultural sector, a manufacturing sector and a household sector that consumes all the end products and provides a labor force. The agricultural sector produces only crops, which are measured in pounds, and the manufacturing sector produces only



**MYOPIC ANALYSIS** of the effects of a new technology gives rise to an inaccurate assessment of the costs and benefits of the technological change. The thin black line shows the relation between real wages and the rate of return on capital that would be expected if the only sector of the economy to adopt the new computer-based technology were the steel industry. The thick black line and the colored line show the relations under the old technology and the new technology respectively. The graphs show that the myopic analysis would correctly lead to the choice of the new technology if the current economic climate were represented by point *a*, and it would correctly advise operating under the old technology if the economic climate were represented by point *d*. If, however, the rate of return were represented by any point *k* between 14 percent (*j*) and 17.5 percent (*i*), the manager would inadvisedly reject the new technology on the grounds that losses in the rate of return (*l*) or in real wages (*m*) are predicted by the myopic analysis. A more penetrating analysis, in which the change to the new technology by the steel industry is accompanied by similar changes in other sectors of the economy, shows that even at these high rates of return the new technology could bring benefits to wage earners (*n*) or to stockholders (*o*). Unfortunately, because of the lack of data about the entire economy, the myopic analysis is often the only scenario that can be rigorously considered by managers contemplating technological change.

cloth, which is measured in yards.

Not all the output of the agricultural sector is bought and consumed by the household sector. Some of it, such as cotton fiber, is used as a raw material by the manufacturing sector to make cloth, and some of it, such as the alfalfa crop, is plowed back into the fields to renew the supply of nitrogen in the soil. Similarly, not all the output of the manufacturing sector goes into the household sector as end product. Part of the cloth output of the manufacturing sector is bought by the agricultural sector to make up bags of grain, and part of it is used by the manufacturing sector itself to make special work clothing.

For concreteness, suppose the recipe for one pound of agricultural crop calls for the input of .25 pound of crop, .14 yard of cloth and .55 hour of labor. (Rain, soil and sunlight are assumed to be free, or in other words

noneconomic, commodities.) The recipe for producing one yard of cloth calls for .4 pound of crop, .12 yard of cloth and 2.7 hours of labor. These quantities are known as input coefficients. If the actual output of the agricultural sector is 100 pounds of crop, the input coefficients for agriculture are simply multiplied by 100: one needs 25 pounds of crop, 14 yards of cloth and 55 hours of labor. If the actual output of the manufacturing sector is 50 yards of cloth, one needs 20 pounds of crop, six yards of cloth and 135 hours of labor.

An input-output table displays all current inputs absorbed by each economic sector as a column of numbers. Because every economic input is also an output, each number in the column is also designated by the name of the row of the table in which it is found. Hence the column labeled "Agriculture" in my previous example includes

the quantities 25 pounds, 14 yards and 55 hours of labor. The column labeled "Manufacturing" includes the quantities 20 pounds, six yards and 135 hours of labor.

Once the input recipes are written down in the table the distribution of the output of each economic sector to itself and to every other sector can be read from left to right across a row. Thus of the 100 pounds of crop grown by the agricultural sector, the first row of the table shows that 25 pounds must be returned to that sector, 20 pounds are sold to the manufacturing sector and 55 pounds remain for consumption by households. Similarly, of the 50 yards of cloth produced by manufacturing, the second row of the table shows that 14 must go to agriculture, six must go back into manufacturing and 30 remain for consumption by the households [see illustration on page 39].

The material and labor inputs alone are not sufficient to produce crops or cloth. In each sector of the small economy there are capital stocks as well. In agriculture these include barns built with wood from the agricultural woodlots and drying sheds that can be closed in a storm with canvas cloth from the manufacturing sector. In manufacturing the capital stocks include buildings made of wood from the agricultural sector and belts for transmitting power made of woven fabric from the manufacturing sector.

I have already pointed out that certain quantities of capital stocks must be in place before an industry can transform its inputs into outputs. Imagine that for each pound of crop one needs a capital stock of .5 pound of crop and .3 yard of cloth. Furthermore, for each yard of cloth one needs a capital stock of two pounds of crop and one yard of cloth. These numbers are called the capital coefficients.

For a given rate of return on the capital per unit output of an industry, the unit price of each product in the economy can now be calculated. Each unit price must be equal to the cost of the material inputs absorbed plus the cost of the capital stocks employed per unit of output plus the cost of the labor inputs per unit output. For example, the cost of the material inputs to agriculture per pound of agricultural production is .25 times the cost per pound plus .14 times the cost per yard of cloth. If the annual rate of return on capital is 10 percent, the cost of the capital needed per pound of agricultural production is 10 percent of .5 times the cost per pound plus 10 percent of .3 times the cost per yard of cloth. The cost of labor is simply the

product of the labor input coefficient .55 and a given hourly wage [see illustration on page 40].

Note that once the hourly wage and the annual rate of return on capital are fixed, the statement or equation that gives the price of a pound of crop includes only two unknown quantities: namely the price of a pound of crop and the price of a yard of cloth. A second, similar equation can be constructed from the input coefficients and the capital coefficients for the price of a yard of cloth. In short, there are two equations in which the same two unknown quantities appear. Since the equations are derived independently of each other, both unknown quantities can be determined. In this example it turns out that if the wage in both sectors of the simple economy is \$1 per hour, a pound of crop is worth \$2 and a yard of cloth is worth \$5.

In general, for an economy having  $n$  sectors that produce one output each, there are  $n$  unit prices that must be determined, and the recipes for the  $n$  outputs of the economy give rise to  $n$  equations, each including at most  $n$  unknown quantities. Straightforward mathematical methods then lead to the solution of each of the  $n$  unit prices. When the industrial recipes are based on the standard format of the input-output table, the techniques of linear algebra and matrix manipulation, which are readily available in computer-software packages and even in some hand-held calculators, make the calculations fast and convenient.

The calculation of the prices can be carried out in this way for any combination of wage level and rate of return. The wages in each case are the money wages, and for simplicity we assumed they are always paid at a fixed rate in constant 1979 dollars. The prices of goods can then be calculated from the equations as we vary the assumed rate of return. It can be shown mathematically that the assumption of constant money wages imposes no loss of generality on our final conclusions. Both the ratios of the prices among the different goods and the level of real wages depend only on the relative levels of money wage rates, not on their absolute levels.

To determine the set of possible values for real wages and rates of return on capital for the U.S. economy under the old technology we divided the economy into 89 input-output sectors. Our picture of the old technology is derived from data on input and capital coefficients compiled by the Department of Commerce for the year 1978. In 1979, the base year for the study, the rate of return on capital in-

vested was about 12.5 percent, and the corresponding real wages are set equal to \$1 an hour. If the rate of return is assumed to be 20 percent, the real wages fall to about 80 percent of their 1979 value. On the other hand, if the rate of return is assumed to be only 5 percent, the real wages increase by more than 20 percent of their 1979 value.

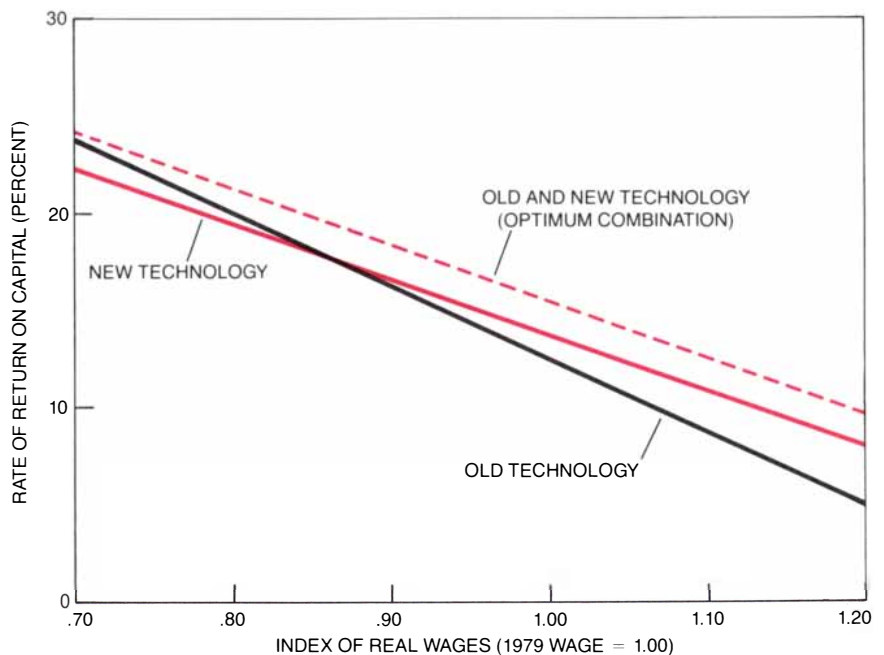
Each solution of the price system for the 89 sectors of the economy therefore represents a distinct relation between wages and rate of return on capital. Our next task was to explore the effects these relations would have on the relative prices of various goods. The relative price of a product is its money price divided by the cost-of-living index. For the purposes of our analysis we adjusted the relative prices of all goods to \$1 for the relation between real wages and rates of return that prevailed in 1979.

The results show that a rise in the rate of return, which is accompanied by a corresponding fall in real wages, leads to a rise in the relative prices of goods from such sectors of the economy as the utilities industry, the mining of ferrous ores and the livestock industry. In the utilities industry, for example, an increase in the rate of return on capital from its 1979 base value of 12.5 percent to 20 percent would lead to a 25 percent increase in the cost of utilities. On the other hand, with the same rise in the rate of return the relative prices of wood containers and

retail-trade services tend to fall. For example, the price of wood containers would decrease to approximately 92 percent of its relative price in the 1979 economy.

The reason for the difference is the varying proportion of economic value added by capital and by labor to the input materials that make up the final product. Such industries as the utilities industry require relatively large amounts of capital. The prices of such products are therefore quite sensitive to a rise in the rate of return on capital. On the other hand, the manufacture of such goods as wood containers is relatively labor-intensive, and so the decline in real wages that accompanies the rise in the rate of return on capital more than offsets the rise in the cost of capital.

To compare the reciprocal relations between wages and rate of return under the new technology with those necessary under the old technology, we had to solve another system of 89 equations for the 89 unknown unit prices of the outputs of each of the 89 economic sectors. The equations were built up just as they were for the old technology. The input coefficients and the capital coefficients from our third scenario about the introduction of computer-based automation were appropriately introduced in each of the 89 equations. A computer then solved the equations for each of several as-



**OPTIMUM COMBINATION** of old technology in some sectors of the economy and new technology in others can be determined for any given rate of return on capital by solving a problem in linear programming. The resulting relation between rate of return and real wages is superimposed on the relations attainable if all sectors adopt only the old technology or only the new. The optimum combination can improve real wages and rate of return.

sumed values of the rate of return.

The reciprocal relations between wages and rate of return under both the old and the new technologies can be plotted on a graph. The horizontal axis represents real wages and the vertical axis represents the rate of return [see illustration on page 41]. Under the new technology the rate of return of 12.5 percent that prevailed in 1979 would have given workers real wages more than 6 percent higher than the wages they had under the old technology. Had workers instead been content with the same real wages they had in that year, the new technology would have made it possible to realize an annual return of nearly 14 percent. Hence given the 1979 relation between wages and rate of return both workers and owners could have been better off economically under the new technology than they were under the old.

The levels of real wages for given rates of return under the new technology remain superior to those obtainable under the old technology as long as the rate of return is less than 17.5 percent. At that level the real-wage rate is 88 percent of its base value for both the old technology and the new. If the rate of return were to rise higher than 17.5 percent, neither workers nor owners could benefit from a changeover to the new technology. The cost of capital improvements and expansion would

then be high enough to wipe out the benefits the new technology could otherwise bring about. This analysis may explain, incidentally, why entrepreneurs in Japan often seem more eager to adopt new technology than their American counterparts. Because they are often satisfied with a lower rate of return on invested capital, that rate is more likely to be improved on by the introduction of the new technology.

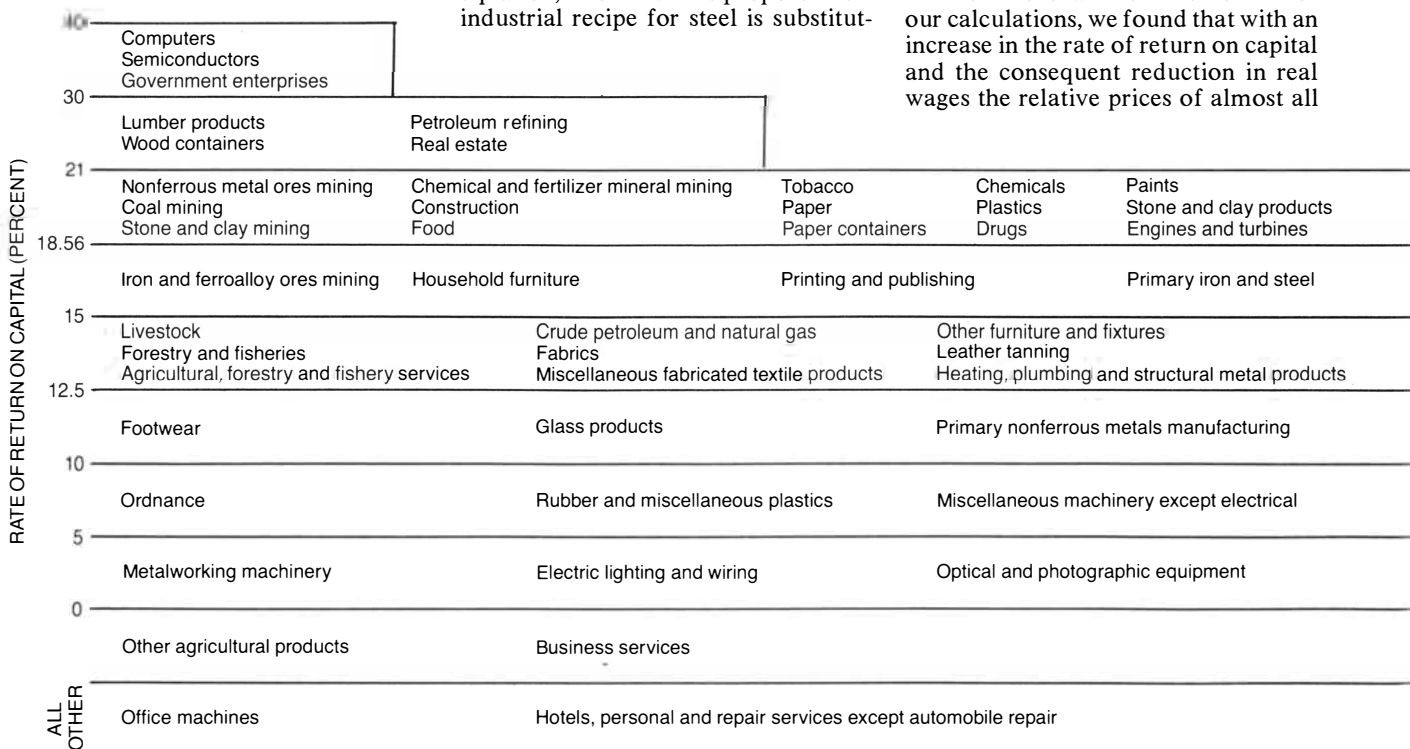
With these relations among wages, rates of return and technologies in mind, one can begin to give a quantitative explanation for the failure of certain industries to introduce new technology. Imagine once again the predicament faced by the manager of a steel mill. Given the supply of capital reflected in the rate of return on capital, the manager must decide whether or not to modernize the mill. As I have noted, however, cost comparisons between the old technology and the new are generally not based on the systematic use of a detailed, empirical model of the national economy. Instead they are typically based only on the prices observed in the economy at the time the choice is being made. These prices reflect the dominance of the old technology not only in the steel industry but also in all the other industries.

Costing out a new technology in this way calls for solving only one equation, into which the proposed new industrial recipe for steel is substitut-

ed. The prices of the input materials and the capital goods specified in the equation are not unknown quantities, as they are in the input-output analysis. Instead they are merely constants that have already been determined by the economy under the old technology. The new price for a unit of steel output then determines the real wages for a given rate of return on capital in the way I described above.

This method of determining the costs of a new technology is bound to be biased, except in the unlikely case that even after the steel industry adopts the new technology all the other industries continue to stick with the old. We calculated the relation between wages and rates of return under this new set of costs and then compared it with the relation that prevailed under the old technology. We found that the manager of the steel mill must rule against the introduction of new technology whenever the rate of return is greater than 14 percent. Thus if the rate of return happens to lie between 14 and 17.5 percent, the manager will make the wrong decision for the company [see illustration on page 42]. In that case the manager would choose not to switch to the new technology, even though a more comprehensive and correct assessment shows that both employees and owners of the mill would benefit from such a change.

When we examined the results of our calculations, we found that with an increase in the rate of return on capital and the consequent reduction in real wages the relative prices of almost all



**ECONOMIC INCENTIVE** to introduce computer-based automation in a given sector of the economy depends on the prevailing rate of return on capital and on the real wages to which each rate corresponds. For each rate of return shown at the left in the chart the

combination of old and new technologies that maximizes the real wages was determined. Each calculation yielded a list of sectors that should choose the new technology. The chart presents each list and shows the relations among them. For example, the three sectors list-

goods and services would decline. The exceptions would be the relative prices of hospital, health care and educational services, which would rise more than the relative prices of most other goods. This finding prompted us to explore what would happen if new technology were introduced everywhere except in these three sectors.

Our results show that the introduction of the new technology in health and education hampers rather than advances the income-generating capacity of the economy. Indeed, under the present institutional setup both sectors are being heavily subsidized by the Government transfer of income earned in other sectors. A sharper theoretical formulation, which explicitly takes account of taxes and subsidies, should make it possible to trace the underlying input-output relations in greater detail.

In the meantime it suffices to observe that in both education and health care new technology can be expected to improve the quality of the final product and with it the benefits to consumers. To the extent to which these benefits derive from entirely new goods, they cannot legitimately be included either in the cost-of-living index or in the measure of real wages employed in the present formulation of the input-output model. In the current state of our empirical understanding a decision to introduce new technology in these sectors must essentially be a political one; a decision in the opposite direction, or in other words toward fiscal retrenchment, must be

understood as being political as well.

I have treated the choice of technology for the most part as an elimination contest between two teams: the old technology and the new. Actually, of course, neither team is an indivisible set; individual players, or in other words individual technologies, can be put into the lineup or returned to the bench. At the risk of placing too great a burden on a fragile data base, my colleagues and I therefore tried to answer one final question: What combination of technologies, old in some sectors and new in others, would yield the highest real wages for given rates of return on capital?

The problem of selecting such an optimum combination can be treated mathematically as a problem in linear programming. The result is a list of economic sectors that, for each possible level of the rate of return, would benefit from the introduction of new technology [see illustration on these two pages]. It turns out that the list is different for each rate of return, and in a sense it is cumulative: every sector that should optimally choose the new technology at a high rate of return on capital should also choose the new technology at every lower rate. For example, if the rate of return were between 30 and 40 percent, only the computer industry, the semiconductor industry and government enterprises should adopt the new technology. If the rate were between 21 and 30 percent, these three sectors should be joined by the makers of lumber products and wood containers, the petroleum-refining in-

dustry and the real-estate industry. As the rate of return continues to fall, an increasing number of industries should choose the new technology.

It is interesting to speculate on the combination of old and new technologies that would result if every industry pursued the myopic economic analysis I illustrated above for the steel industry. It turns out that the choices within the industries would be quite different from what they should be according to the linear-programming solution. On the other hand, the real wages that can be attained for given rates of return are only slightly inferior to their optimum values.

There are many other directions in which our current findings and procedures could be pursued. The analysis I have outlined could be applied to input-output models much larger and more sophisticated than the ones we employed. Costs derived from our model could also be applied in the analysis of multiregional input-output models, such as the large United Nations model of the world economy that I published in 1977 with Anne P. Carter and Peter A. Petri of Brandeis University. No significant advance in either direction, however, can be made without the creation of a new and substantially expanded data base. "Money, money and more money" was and still is the prescription for conducting a successful war; data, data and more data is the prescription for a successful explanation of how the economic system actually works.

Construction machinery Materials-handling equipment Service-industry machinery	Other transportation equipment Transportation and warehousing Radio and television broadcasting	Wholesale trade Finance Insurance	Miscellaneous textile goods Apparel	
Metal containers	General industrial equipment	Electronic components	Miscellaneous manufacturing	
Screw-machine products Other fabricated metal products Farm and garden machinery	Special industrial equipment Electric industrial equipment Radio, television and communications equipment	Electron tubes Miscellaneous electrical machinery Scientific and controlling instruments	Communications, except radio and television	
Household appliances	Aircraft and parts	Retail trade		
Motor vehicles	Electric, gas, water and sanitary services	Nonprofit organizations		
Automobile repair services				
Eating and drinking places	Amusements	Hospitals	Health services except hospitals	Educational services (private)

ed just under the top step should adopt the new technology for all rates of return below 40 percent. For rates above 30 percent all other sectors should adopt the old technology. If the rate were 30 percent or less, the sectors listed just under the second step should also

shift to the new technology. Sectors such as health and educational services would continue operating under the old technology no matter how low the cost of capital if their decisions about technology were based solely on wage rates and the rates of return on capital.

# The Immunologic Function of Skin

*The body's largest organ is more than a passive protective covering: it is also an active element of the immune system. Specialized cells in the skin have interacting roles in the response to foreign invaders*

by Richard L. Edelson and Joseph M. Fink

The elegant simplicity of human skin camouflages rich complexity and multiple functions. The skin is only a few millimeters thick, but it is the body's largest organ and within it a variety of highly specialized cells are organized into intricate structures and subsystems. One of the skin's most remarkable roles has been recognized only recently: it is an integral and active element of the immune system.

In retrospect one can see that the discovery of this active role in the immune response should not have come as a surprise. The skin is the body's interface with the external environment. The skin of human beings is particularly vulnerable because it is covered only sparsely with hair. It is reasonable to suppose mankind could not have survived the infections arising from multiple skin wounds if the body's outer covering had not been able to mobilize a potent and sophisticated immune response.

Obscure but important normal functions of this kind are often brought to light by the study of malfunction in disease. The fact that the skin has a more than passive role in certain diseases affecting it was first suggested by new understanding of malignancies of lymphocytes, the white blood cells that control the functions of the immune system.

By 1970 it was known from studies in mice that lymphocytes can be divided into two functionally distinct major populations: *B* cells (which mature largely in the bone marrow) and *T* cells (which mature in the thymus gland). *B* lymphocytes are implicated in humoral immunity. They synthesize antibodies, which react with specific antigens such as molecules on the surface of infectious organisms or malignant cells. *T* lymphocytes mediate cellular immunity. For example, they destroy cells infected by viruses, initiate the cellular response to bacterial invasion and react against incompatible

transplanted tissue. Subgroups of *T* cells modulate the immune response. "Helper" *T* cells enhance the maturation of *B* cells into antibody-secreting cells and the expansion of specific populations of *T* cells; "suppressor" *T* cells diminish *B*-cell function and limit the size of *T*-cell populations.

In 1972 it first became possible to classify human lymphocytes as either *B* or *T* cells, and a number of investigators undertook to reclassify the lymphocytic malignancies as proliferations specifically of one or the other cell type. These cancers include leukemias, in which large numbers of malignant lymphocytes circulate in the blood, and lymphomas, in which the malignant cells accumulate in the lymph nodes and organs. The result of the reclassification was a surprise. Even though there are normally at least three times as many *T* cells in the blood as there are *B* cells, it turned out that the very large majority of adult leukemias and lymphomas are malignancies of *B* lymphocytes.

At the National Institutes of Health, Ira Green, Philip S. Schein, Charles H. Kirkpatrick, Ethan M. Shevach, Marvin A. Lutzner and one of us (Edelson) found, however, that one broad group could regularly be identified as malignancies of *T* cells: certain leukemias and lymphomas in which the skin is significantly infiltrated by malignant cells. The dichotomy was striking. Virtually all lymphocytic malignancies in which there is widespread infiltration of the skin were shown to be of *T*-cell origin; most of those in which the skin is spared were shown to be of *B*-cell origin. In other words, malignant *T* cells seemed to have a particular affinity for the skin.

One implication of this finding was that all these malignancies of *T* cells having affinity for skin might be different representations of a single disease. Moreover, the affinity of the

malignant *T* cells for skin was likely to reflect a similar characteristic in the normal *T* cells from which they had arisen; a significant population of *T* cells must be resident in the skin. Scattered "passenger" lymphocytes had often been observed in normal human skin, but their presence had previously not been considered significant. What we were witnessing appeared to be a malignant amplification of a normally inapparent interaction between skin and certain resident *T* cells.

We found three additional common denominators in patients showing the various clinical manifestations of *T*-cell malignancies involving the skin. First, there is a natural evolution in individual patients. The earliest lesions observed in a patient tend to be "epidermotropic" ones: they are characterized by the migration of malignant cells into the epidermis, the outer layer of the skin. These early lesions are progressively replaced by "nonepidermotropic" lesions, in which the epidermis is spared while malignant cells accumulate deeper in the skin and subcutaneous tissues. If only epidermotropic lesions are present, there is often no evidence of disease in organs other than the skin. The development of nonepidermotropic lesions, on the other hand, is correlated with dissemination of malignant cells to internal organs and with a substantially worse prognosis.

In other words, what had been considered to be different types of *T*-cell malignancies involving the skin are instead clinical manifestations of different time frames in the evolution of a single type of malignancy. Because the most distinctive biologic feature of these malignant *T* cells is their initial affinity for epidermis, we named this single disease cutaneous *T*-cell lymphoma (*CTCL*). Once it was recognized as being a single entity, *CTCL* was seen to have a cumulative incidence exceeding that of Hodgkin's dis-



ease, making it the commonest adult lymphoma.

A second common denominator of patients with *CTCL* is that the first clinical appearance of the disease is a skin rash, even though the malignant cell is a type of white blood cell rather than a cell known to be native to the epidermis. This finding was supported by Henry Rappaport of the University of Chicago Pritzker School of Medicine and Louis B. Thomas of the National Cancer Institute, who examined autopsy samples from 45 patients with *CTCL*. They found several who had only skin disease and no apparent internal involvement.

Finally, we established that *CTCL* is regularly a malignancy of a specific class of *T* cells: the helper *T* cells. This observation was made by exploiting

an experimental system developed by Samuel Broder and Thomas A. Waldmann of the NCI. Purified normal *B* cells grown in the laboratory, even in the presence of the *B*-cell stimulant pokeweed, failed to produce antibodies. Addition of either normal *T* cells or *CTCL* cells induced the *B* cells to synthesize large amounts of antibody. Clearly the malignant *T* cells often retained the characteristic capability of normal helper *T* cells.

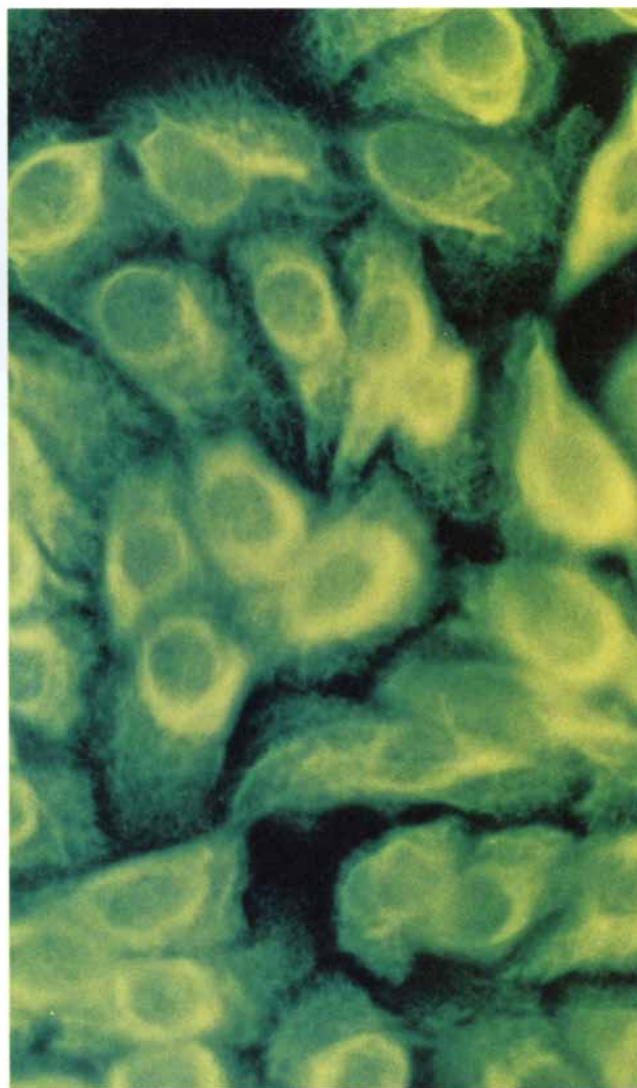
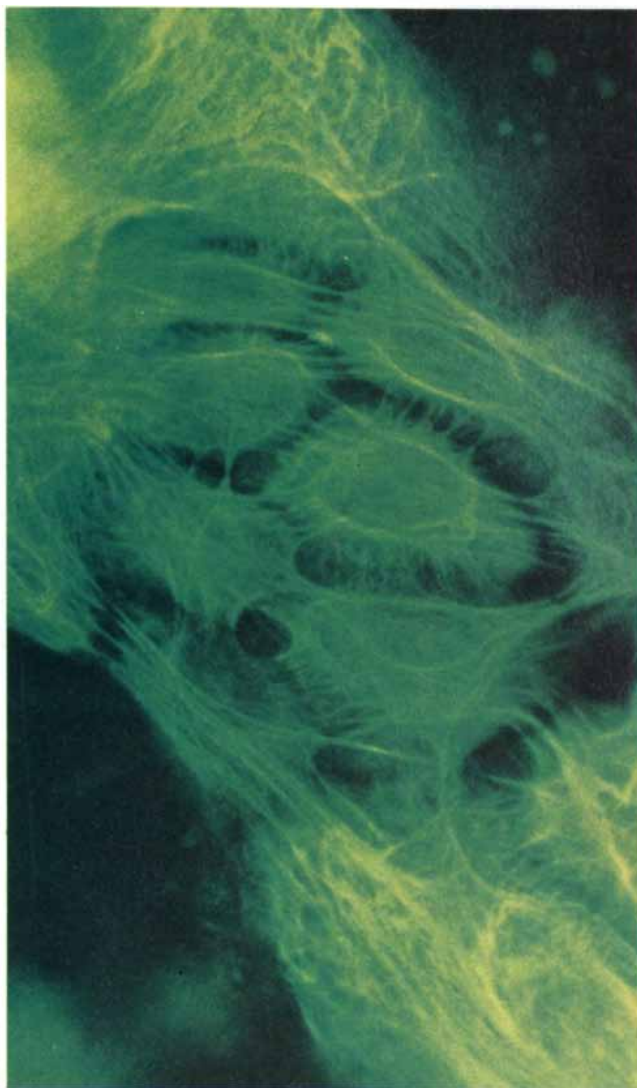
An extraordinary story was unfolding. *CTCL* was identified as a malignancy of helper *T* cells with a pronounced preference for the environment of the epidermis. In 1974 we were able to reach a striking conclusion: there must normally be a population of *T* cells that interact with epider-

mal cells in a dynamic way, and *CTCL* must be a malignancy of those cells.

At this juncture one of us (Edelson) became intrigued by an exciting possibility. Might the skin, like the thymus, be a site where at least certain types of *T* cells undergo maturation?

The thymus gland, situated in the upper chest, is composed of several classes of cells. The large, immobile epithelial cells belong to a broad category of cells that includes such diverse types as those lining the intestinal tract, bronchial tree and blood vessels and serving as secretory cells in many glands. Interspersed among the epithelial cells are spidery dendritic cells distinguished by long, thin cytoplasmic extensions.

The immature lymphocytes that migrate to the thymus from the bone



**IMMUNOLOGIC FUNCTION** of human skin is implied by its structural relation to the thymus gland, an immune-system organ in which *T* lymphocytes undergo maturation. Photomicrographs made by Tung Tien-Sun of the New York University Medical Center demonstrate a striking similarity between keratinocytes, the ma-

ior cells of the epidermis (left), and thymic epithelial cells (right). Cells were incubated with an antibody to keratin, a protein made by keratinocytes; bound antibody was indirectly stained with fluorescein, a yellow-green fluorescent dye. A wavy network of keratin fibrils is revealed in the thymic cells as well as in the skin cells.

marrow are called thymocytes while they are in the gland. As they mature in the thymus to become cells that are recognizable as *T* lymphocytes they encounter the cellular components of the thymus as well as a series of thymic hormones. The maturation of the cells takes place sequentially as they move from the outer part (the cortex) of the gland to the inner part (the medulla) before being exported to other parts of the body. Significantly, it was known that even after completing their "education" in the thymus, many *T* cells need to undergo further maturation before they finally become fully functional. The site of this obligatory post-thymic maturation was not known.

Our discovery of a major *T*-cell population in human skin suggested that the skin might have functions analogous to those of the thymus. A strong indication that there indeed is a close relation between the thymus and the

skin came from an experiment of nature. For the past two decades biologists have exploited "nude" mice for immunological manipulations. These mutant mice are named for their lack of mature hair, the major epidermal appendage. They are of interest to immunologists because they also congenitally lack a thymus gland and therefore fail to develop a normal contingent of functional *T* cells. As a result they do not reject tissue transplants. Moreover, their immune system can be partially reconstituted by the transfusion of one or another type of *T* cell. They therefore offer a host of experimental opportunities.

Nude mice have been crossbred with many other strains of mice. The "congenic" offspring have the nude-mouse genes coding for the absence of the thymus as well as various other genes derived from the other mice. Given all the genetic rearrangements the nude-

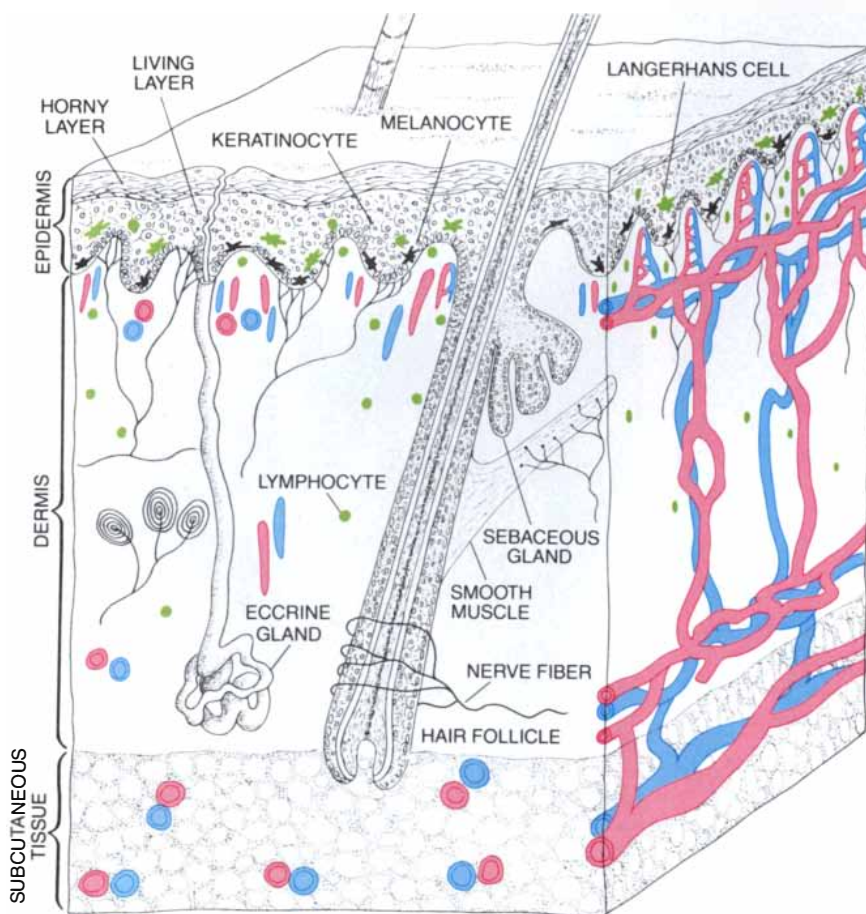
mouse chromosomes have undergone over the past 20 years, it is of major interest that it has never been possible to separate the genes responsible for the absence of the thymus from the genes responsible for the absence of mature body hair. Either the genes coding for the development of the thymus are identical with those coding for the production of normal hair or the two sets of genes are closely linked on the same chromosome.

There was another clue linking the skin with the thymus. Electron microscopy had revealed that some thymic epithelial cells carry distinctive granular structures. They seemed to be identical with the "keratohyalin" granules seen in keratinocytes, the major cells of the epidermis. Keratinocytes synthesize keratin, the primary structural protein of the skin's horny outer coat and of hair.

The mammalian thymus gland and epidermis, then, were known to have genetic and structural similarities. Marian R. Rubenfeld in our laboratory at the Columbia University College of Physicians and Surgeons collaborated with Allen E. Silverstone of the Memorial Sloan-Kettering Cancer Institute to learn whether the similarities could be extended from genetics and structure to function. Specifically, could epidermal cells influence the maturation of *T* cells in a laboratory culture?

The experimental design called for cultivating immature human or mouse lymphocytes, but not mature *T* cells, in the presence of epidermal keratinocytes or control cells and then looking for evidence of *T*-cell maturation. In one set of experiments a population of human white blood cells was depleted of mature *T* cells. The blood cells were then cultivated together with either human epidermal keratinocytes, mammary cells, fibroblasts or white blood cells, or in a control medium. In a second set of experiments mouse bone-marrow stem cells were depleted of mature *T* cells and co-cultivated with either epidermal cells or fibroblasts from mice of the same strain, or in a control medium. The co-cultivated blood or bone-marrow cells were then harvested and stained with a fluorescent antibody that would identify an enzyme called terminal deoxynucleotidyl transferase (*TdT*).

*TdT* had hitherto been detected primarily in lymphocytes at a certain stage of their maturation in the thymus. It is lost as the *T* cell matures, and it is not found in normal mature *T* cells in the blood, spleen or lymph nodes. Yet Rubenfeld and her colleagues found that *TdT* was regularly pro-



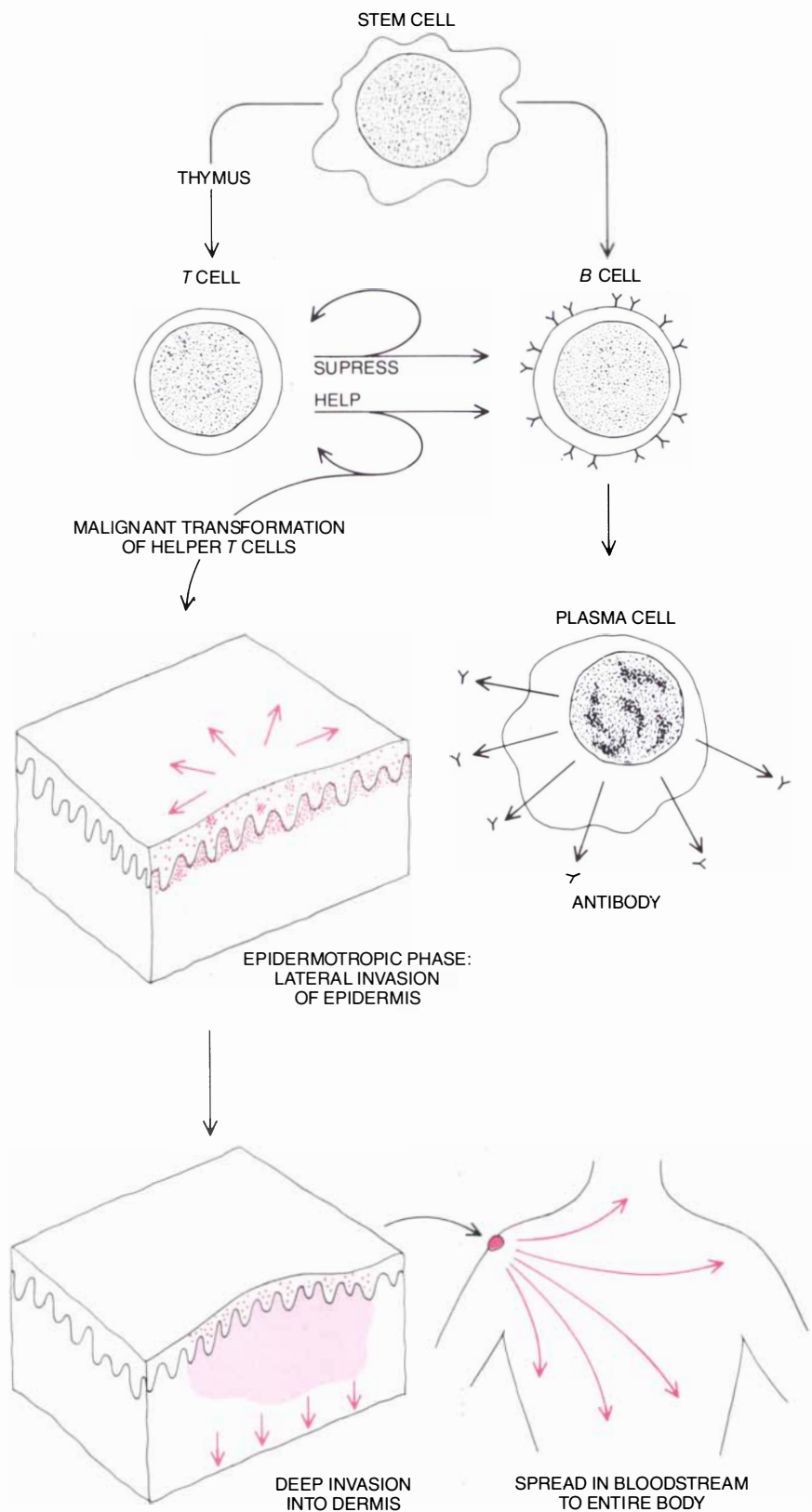
**COMPLEX ANATOMY OF SKIN** is suggested in an idealized drawing. The surface is covered by a horny layer of dead keratinocytes filled with the protein keratin. Living keratinocytes dominate the epidermis and proliferate as dead cells are lost from the surface. Melanocytes, which secrete pigment granules responsible for skin color, are at the base of the epidermis. Langerhans cells, dendritic cells that process antigens applied to the surface of the skin, lie above the basal layer of keratinocytes. The dermis is largely a network of connective tissue and is underlain by fatty subcutaneous tissue. The specialized keratinocytes of the hair follicles form hair. The dermis is richly supplied with nerve fibers, some of which innervate sensory nerve endings, and with blood vessels (red and blue). *T* lymphocytes are scattered throughout the skin, primarily in the epidermis and the upper dermis.

duced by the lymphocytes that had been co-cultivated with human keratinocytes; the enzyme was not present in any of the control cultures. Apparently the epidermal cells had somehow caused incompletely matured *T* cells (which were presumably present in the tested blood or bone marrow, depending on the experiment) to do something they ordinarily do under the influence of the thymus: synthesize *TdT*. The enzyme was found only in the lymphocytes that had synthesized new DNA in preparation for cell division. (It was clearly not the synthesis of new DNA in itself that had signaled the production of *TdT*, since various other stimuli of cell division could be shown not to produce similar results.)

It seemed clear that keratinocytes were the cells responsible for the skin's impact on *T*-cell maturation in these experiments, because the cultures were devoid of other types of epidermal cells. It was astonishing to find that a skin cell previously thought to be important primarily for producing keratin could have a major impact on *T*-cell biology. To be sure, the skin cannot fully duplicate the function of the thymus, because mice in which the thymus gland is removed immediately after birth do not develop normal *T*-cell systems. Rubinfeld's results did suggest, however, that keratinocytes in the skin have some influence on post-thymic steps of *T*-cell maturation.

At about the same time Bijan Safai and his associates at Memorial Sloan-Kettering found that patients with *CTCL* have elevated blood levels of a chemical factor with properties similar to a thymic hormone, and that skin from *CTCL* lesions can produce this factor. Anthony C. Chu and Carole L. Berger in our group, collaborating with Gideon Goldstein of the Ortho Pharmaceutical Corporation, extended Safai's observations. They showed that an antibody to thymopoietin, a thymic hormone that influences *T*-cell maturation, binds to a molecule found in the cytoplasm of keratinocytes in the basal layer of normal human epidermis. It is tempting to speculate that the keratinocyte factor identified by this antibody is an active hormone responsible, at least in part, for Rubinfeld's results. Confirmation of the possibility will require isolation and characterization of the reactive molecule and direct demonstration of its biological activity.

Was the cumulative laboratory evidence of a thymus-skin collaboration relevant to the situation in living human beings? An important clue was forthcoming from the work of Barton F. Haynes, Brian V. Jegasothy and



**CUTANEOUS *T*-CELL LYMPHOMA (*CTCL*)** is a malignancy of helper *T* cells. Both *T* and *B* lymphocytes originate in bone marrow. Some stem cells in the marrow mature into *B* lymphocytes, which synthesize specific antibodies, display them on their surface and (when triggered by a specific antigen) develop into antibody-secreting plasma cells. Other stem cells mature under the influence of the thymus to become *T* lymphocytes. Some of these are "helper" or "suppressor" *T* cells that respectively enhance or inhibit the function of *B* cells and other *T* cells. Helper *T* cells that become malignant often have an affinity for the epidermis, and they accumulate near it and migrate into it. As subclones of more malignant *T* cells arise they lose their affinity for epidermis and metastasize to internal organs.

their colleagues at the Duke University School of Medicine. They worked with a fluorescently labeled monoclonal antibody that binds selectively to a molecule (human *Thy-1*) normally found in the cell membrane of thymocytes but not of mature *T* cells circulating in the blood. They studied malignant *T* cells obtained from the skin lesions of *CTCL* patients and also from their blood. Only the malignant cells from the skin were found to have *Thy-1* in their outer membrane. The implication was that once malignant *T* cells leave the bloodstream and localize in the skin, their outer membrane becomes altered to a form similar to that of *T* cells maturing in the thymus.

More evidence that the thymic epithelium is structurally similar to the epithelial component of the skin was

produced when Tung Tien-Sun and his colleagues at the New York University Medical Center examined epithelial cells from human and rodent thymus in tissue culture. They applied fluorescently labeled monoclonal antibodies that bind selectively to human keratin, the major protein produced by keratinocytes in the epidermis. They were able to identify abundant quantities of this skin-cell protein in the cytoplasm of thymic epithelial cells.

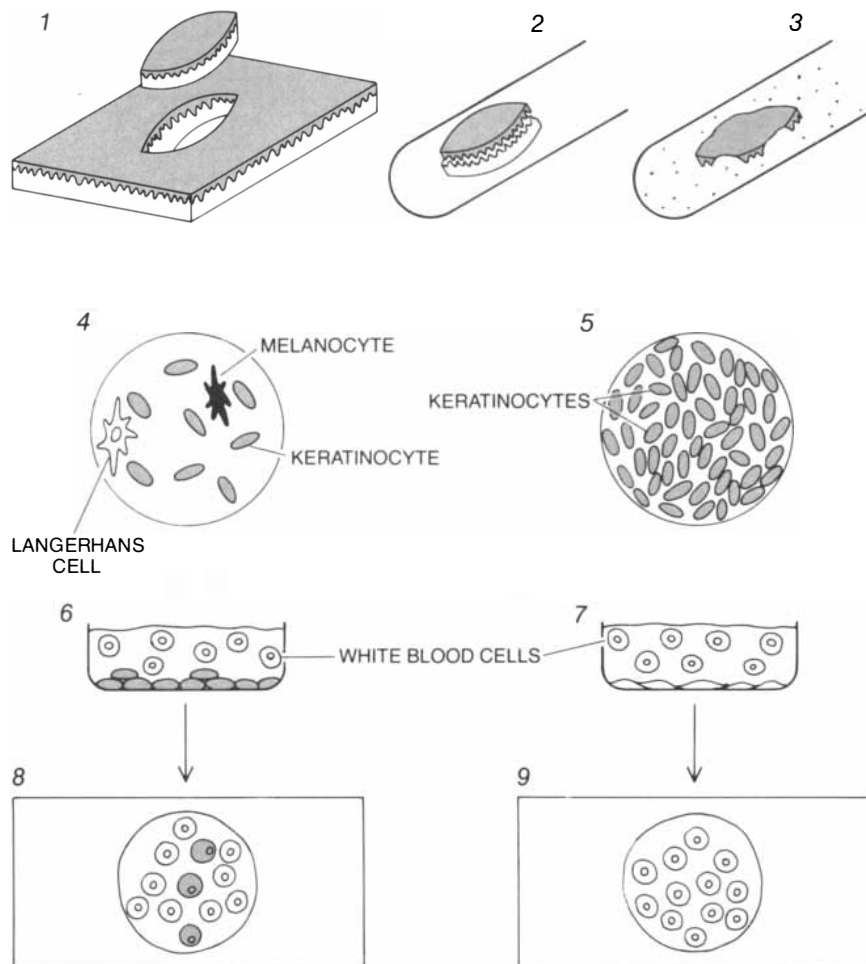
Haynes, Kay H. Singer and their associates then compared the surface of thymic epithelial cells with that of cells in the epidermis by applying a series of fluorescent monoclonal antibodies. They found three distinctive marker molecules, designated *TE-4*, *A2B5* and *p19*, on the outer membrane of the thymic epithelial cells secreting thy-

mic hormones. Amazingly, they found the same set of molecules on the surface of keratinocytes from the basal layer of human epidermis. These were the very epidermal cells that we had shown to produce a substance closely resembling the thymic hormone thymopoietin.

In short, some compelling anatomical, molecular and functional similarities had been found between the epithelial cells of the thymus and those of the skin. It appeared that the skin might indeed be an integral component of the human immune system.

The experiments we have described were done with cells in tissue culture or with intact human tissue. An abundance of relevant information has also accumulated from direct investigation of the immunologic properties of the skin in experimental animals. There are, to be sure, major differences between human and rodent skin. Human skin lacks the thick fur of the experimental animals and is therefore uniquely susceptible to damage from environmental hazards; it has a much thicker and more stratified epidermis and has widespread sweat glands. The animals often fail to develop disorders analogous to human skin disorders. Selective manipulation of the immune system and the skin of inbred strains of animals does, however, make feasible a level of experimentation beyond what is permissible in human beings, and much of what is learned from such studies is surely relevant to human biology.

One elaborate series of studies in mice was done at the University of Texas Health Science Center at Dallas by Paul R. Bergstresser, Robert E. Tigelaar, Craig A. Elmet and J. Wayne Streilein. They examined the possibility that antigen applied to the skin can be "presented" to responding *T* cells by cells residing in the skin. They tested the possibility by applying small amounts of an antigen to the skin of a mouse, excising the affected skin and transplanting it to another mouse of the same inbred strain or of another strain; the recipient mouse was then tested to see whether it had been immunized to the specific antigen. At the very low antigen concentrations of their experiments only the transplantation of skin to mice of the same strain led to immunization. In other words, the antigen applied to the skin needed to be presented to *T* cells by cells residing in the skin and genetically identical with the responding *T* cells; simple diffusion of unprocessed antigen to lymph nodes was not immunogenic. This procedure demonstrated that the initial processing of the anti-



**INFLUENCE OF EPIDERMIS on *T* cells** was suggested by tissue-culture experiments. A human skin sample was excised (1). The epidermis was separated (2) and incubated with the enzyme trypsin (3) to dissociate its cells: keratinocytes, melanocytes and Langerhans cells (4). When the cells were cultured for two weeks, the keratinocytes grew to confluence; no other skin cells were detectable (5). The keratinocytes (6) and control cells (7) were cultivated together with human white blood cells depleted of mature *T* cells. The white cells were removed and tested with an antibody to terminal deoxynucleotidyl transferase (*TdT*), an enzyme found only in maturing *T* cells. The antibody bound to some of the white cells co-cultivated with keratinocytes (8) but not to the white cells co-cultivated with control cells (9). The results showed that human keratinocytes can influence *T*-cell maturation.

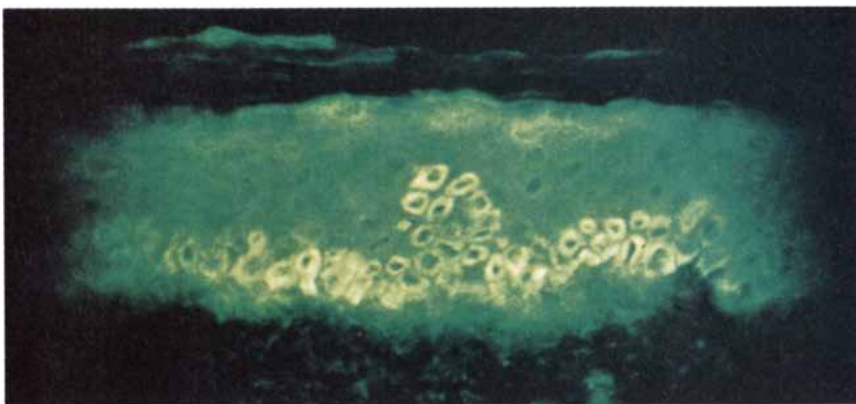
gen did take place directly in the skin rather than in lymph nodes.

In another set of experiments the Dallas investigators found that when an antigen called dinitrofluorobenzene (*DNFB*) was applied to mouse skin previously exposed to ultraviolet irradiation in the sunburn spectrum, the animal was not immunized against the chemical. In contrast, *DNFB* applied to nonirradiated skin was a most effective immunizing agent. Even more striking, application of *DNFB* to ultraviolet-irradiated skin actually induced a state of lasting immunologic nonreactivity specifically to *DNFB*: an irradiated mouse could not subsequently be immunized to *DNFB* even though it could be immunized to unrelated antigens. The nonreactivity to *DNFB* could be transferred to other mice by transfusion of *T* cells from the irradiated mouse, revealing that the nonreactive state was caused by the activity of specific suppressor *T* cells.

All of this suggested that the skin normally incorporates a class of cells that are extremely efficient in presenting antigen to *T* cells and whose function can be nullified by ultraviolet energy. The identity of the cells remained to be discovered. The findings also suggested that when the presenting cells are thus disabled, an antigen somehow bypasses them and directly stimulates the activity of specific suppressor *T* cells, thereby inducing a specific immunologic paralysis.

By exposing mice to larger doses of ultraviolet energy (much more than is needed to cause severe sunburn) Margaret L. Kripke and Warwick L. Morison of the NCI and Raymond A. Daynes of the University of Utah College of Medicine were able to create a general rather than a specific state of immunosuppression. They showed that the immunosuppression was caused by stimulation of circulating suppressor *T* cells and was associated with a diminished capacity of cells in the spleen to respond to antigens. Injury to the skin, in other words, can have a profound impact on distant parts of the immune system; immunologically reactive cells in the skin must be in communication with such cells in the rest of the body.

The identity of those immunological cells in the skin was revealed by the work of Georg Stingl and Klaus Wolff of the University of Vienna, who collaborated with Shevach, Stephen I. Katz and Green of the NIH. Their experiments were inspired by a significant but previously overlooked finding by Rudolph L. Baer and I. Silberberg-Sinakin of the New York University Medical Center concerning Langer-



**KERATINOCYTES** at the base of human epidermis are shown to contain large amounts of a molecule that is indistinguishable from the thymic hormone thymopoietin, which influences *T*-cell maturation. A frozen skin section was treated with an antibody to thymopoietin. The antibody was labeled by the binding of a second antibody linked to fluorescein. The antithymopoietin antibody bound to the cytoplasm of the basal keratinocytes.



**LANGERHANS CELLS** in human epidermis are shown to display a molecule, designated *T6*, that is ordinarily displayed on *T* cells maturing in the thymus. Here an indirectly labeled antibody to *T6* has bound to three Langerhans cells scattered among keratinocytes.

hans cells: a small population of dendritic cells in the epidermis, to which they are now known to have migrated from the bone marrow. The N.Y.U. investigators had noted that in the course of an allergic skin reaction the Langerhans cells can be seen to associate physically with lymphocytes.

Stingl and his co-workers first demonstrated that Langerhans cells have cell-membrane receptors for certain immunologically important molecules. Similar observations were made by Geoffry Rowden of the McGill University Faculty of Medicine and his associates and by Lars Klareskog and his colleagues at Uppsala University. Then, by enriching the concentration of Langerhans cells in suspensions of guinea-pig epidermal cells to about 33 percent (the remaining two-thirds being keratinocytes), Stingl's group showed that only cell suspensions containing Langerhans cells were capable of presenting antigen to responsive *T* cells. It became evident that Langer-

hans cells are responsible for the immunizing capacity of topically applied antigen, and that it is this function of Langerhans cells that can be abolished by sufficient quantities of ultraviolet energy.

Wlodzimirz Ptak of the Copernicus Medical School in Cracow, working with colleagues at the Yale University School of Medicine, found a clue to the apparent paradox implied by the experiments of the Dallas group: How can a particular antigen applied to the skin normally induce a helper (positive) *T*-cell immune response and yet induce a suppressor (negative) response in the absence of functional Langerhans cells? They showed that Langerhans cells characteristically present antigen in a way that preferentially activates a helper *T*-cell circuit; other antigen-processing cells tend, under certain circumstances, to present antigen in a way that preferentially activates a suppressor circuit.

Do the Langerhans cells function



**HUMAN SKIN CELLS IN SUSPENSION** are stained differentially. The single Langerhans cell reacted with the antibody to T6, which in this experiment is indirectly labeled with rhodamine, a red dye. The cytoplasm of the three keratinocytes reacted with the antibody to the thymopoietinlike molecule, which here is indirectly labeled with fluorescein.

alone or is there a contribution from the keratinocytes? After all, the observations described above had demonstrated that human and mouse keratinocytes in culture are able, in the absence of Langerhans cells, to induce the presence of *TdT*, a *T*-cell marker, and that they contain a molecule resembling the thymic hormone thymopoietin. Gerald G. Krueger and Daynes found that the mouse keratinocyte can be induced to express on its membrane a molecule designated *Ia* (for immune-associated). Other workers noted that human keratinocytes in *CTCL* skin lesions and some other skin lesions also express membrane *Ia*. The molecules had originally been thought to be present in the epidermis solely on Langerhans cells, where they were known to be important in the presentation of antigen to *T* cells. The Utah group obtained evidence indicating that once mouse keratinocytes have been induced to express *Ia*, they can enhance the Langerhans cell's presentation of antigen to *T* cells.

Thomas A. Luger, Daniel N. Sauder, Joost J. Oppenheim and Katz of the NIH made a surprising observation that further clarified the picture. They set out to show that Langerhans cells synthesize interleukin-1 (*IL-1*), a factor known to be secreted by other antigen-presenting cells, the macrophages. *IL-1* is critical in the initiation

of *T*-cell-mediated immunity. It binds to receptors on the surface of *T* cells that are programmed to react against a particular antigen. The *T* cells are thereby stimulated to release interleukin-2, which in turn induces *T*-cell proliferation to deal with the specific immunologic challenge.

The NIH investigators found, as they expected, that cultured mouse epidermal cells (which contained Langerhans cells) did produce *IL-1*. In an effort to prove that the *IL-1* was being generated specifically by the Langerhans cells, they selectively depleted mouse epidermal-cell cultures of these cells. Much to their surprise, the depleted cultures produced just as much *IL-1* as the cultures incorporating Langerhans cells. It turned out that the *IL-1* was being secreted by keratinocytes. Luger's group went on to show that normal human keratinocytes also secrete *IL-1*.

These results complemented the earlier findings from our laboratory and from the Duke group. Keratinocytes, the silent-majority cells of the epidermis, are apparently important immunologically. In addition to generating the protective outer layer of keratin and hair, they produce hormonelike molecules capable of profoundly affecting *T* lymphocytes passing through the skin. Their potential impact on these *T* cells is broad: it ranges from influencing the *T* cells' maturation to

enhancing their specific immunologic response to antigens.

In the past year yet another type of immunologically active epidermal cell has been identified. Richard D. Granstein, Adam Lowy and Mark I. Greene of the Harvard Medical School exposed suspensions of mouse epidermal cells to ultraviolet radiation, thereby depleting them of functional Langerhans cells. In the depleted suspensions they could detect another type of dendritic antigen-presenting cell. The Granstein cell, as it has come to be called, is more resistant to ultraviolet radiation than the Langerhans cell, and it has a tendency to interact with suppressor *T* cells rather than with helper *T* cells.

The cellular and molecular components of the response to an antigen in the skin's immune subsystem are coming into focus. A scheme based on what has been learned to date can be summarized as follows. An antigen binds to either of the two kinds of dendritic antigen-presenting cells in the epidermis: Langerhans or Granstein. Langerhans cells present the antigen to specific helper *T* cells, which have a tendency to migrate to the epidermis as they percolate through the tissues. Granstein cells may interact similarly with suppressor *T* cells. The helper and the suppressor response are in rough balance, but normally the net result is a helper, or positive, signal; this is generally the appropriate response to the potentially harmful foreign invaders the skin often encounters. If Langerhans cells have been damaged (as by ultraviolet energy) or circumvented (perhaps by certain antigens that interact directly with the suppressor circuit), a suppressor signal may instead predominate.

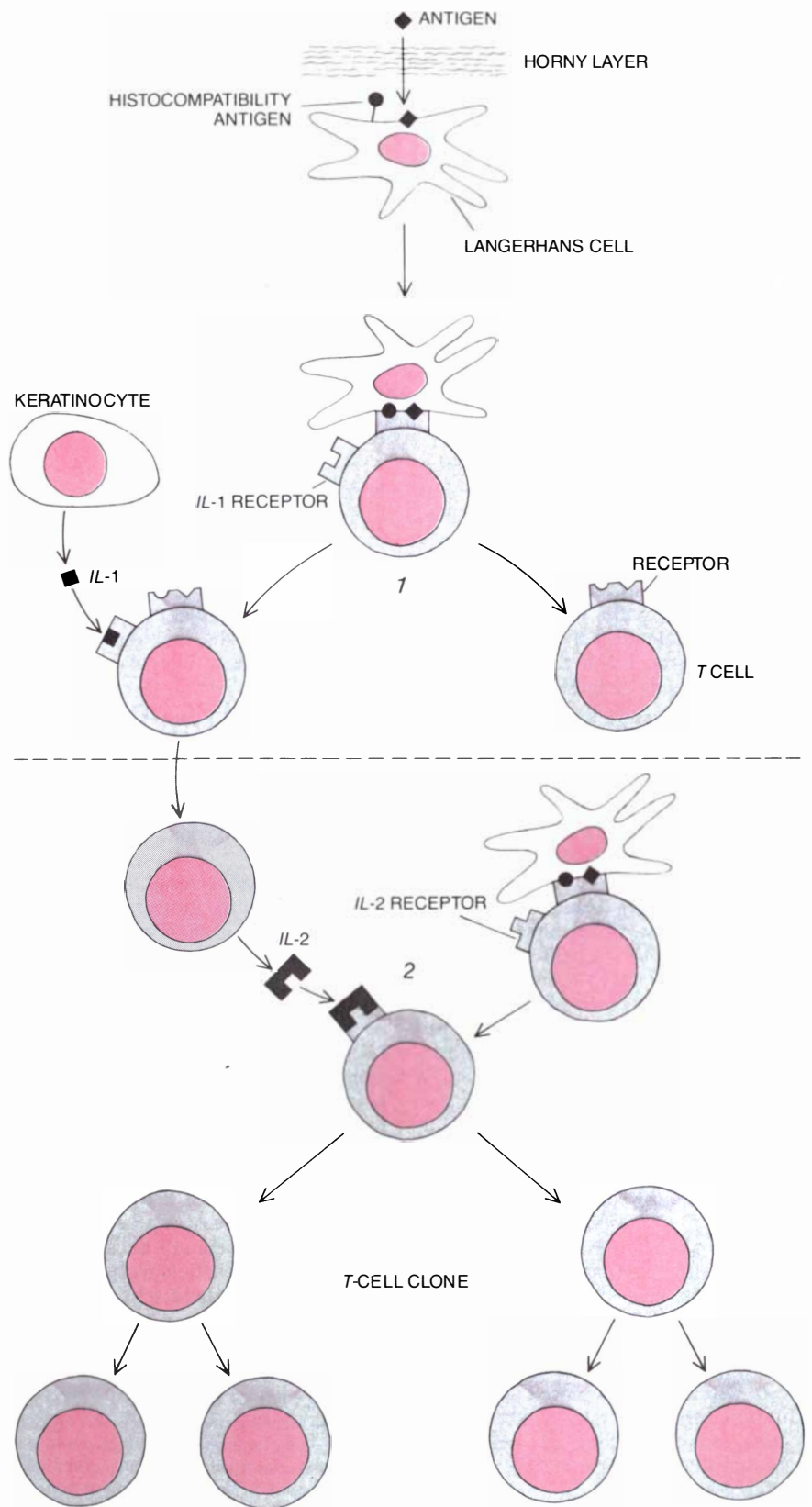
In addition to being presented with an antigen, a *T* cell programmed to respond to that antigen receives a second, complementary kind of signal in the form of *IL-1* secreted by keratinocytes. This prompts the *T* cell to secrete *IL-2*, which binds to additional antigen-responsive *T* cells and causes them to proliferate, dramatically increasing the number of *T* cells capable of responding to the antigenic challenge. These *T* cells enter the lymphatic system and spread throughout the body. Streilein has suggested grouping these various epidermal components of the total immune system under the designation "skin-associated lymphoid tissue," or SALT.

This elaborate system for triggering a cutaneous *T*-cell response to an antigen is probably only one component of a complex relation between *T* cells and skin cells. It is clear that certain types

of *T* cells have a natural tendency to localize in the skin and that the epidermis generates a hormone (or more than one) similar to thymus hormones and able to influence the maturation of cutaneous *T* cells. Investigators throughout the world are exploiting experimental animal systems to decipher further the dynamic relation between the epidermis and *T* cells. A serendipitous finding in our laboratory may provide a way to extend these animal studies to human beings.

We were examining intact human skin for evidence of a molecule called *T6*. The molecule is found on the surface of *T* cells undergoing maturation in the cortex of the thymus; once *T* cells leave the thymus they no longer display *T6*. We hypothesized that malignant *T* cells might reexpress *T6*, which would serve as a marker and thereby facilitate the early diagnosis of *CTCL*. Ellen Fithian applied a monoclonal antibody to *T6*, expecting to show that normal human epidermal cells do not display the molecule. To our surprise she found the antibody reacted with scattered normal epidermal cells lacking the other properties of *T* cells. The dendritic shape and the distribution of these *T6*-positive cells suggested they were Langerhans cells; the identification was confirmed when they were shown to have membrane *Ia*, a characteristic of normal Langerhans cells. Shinichiro Takezaki and Sherie L. Morrison of our group then demonstrated that the *T6* molecule on the Langerhans cells is chemically indistinguishable from the one found on thymocytes.

The presence of *T6* on Langerhans cells (which was also noted by George F. Murphy and his colleagues at Harvard at about the same time) provides a convenient marker for quickly identifying these cells. More important, it also gives rise to some provocative questions. Why do human Langerhans cells express a membrane marker characteristic of immature *T* cells? Langerhans cells migrate to the skin from the bone marrow in the bloodstream, and the blood of normal adults does not contain a significant number of *T6*-positive cells. Are the Langerhans cells induced to synthesize the molecule by cells in the epidermis? Alternatively, is it *T* cells, under the influence of the skin, that synthesize and secrete the molecule, which then adheres to Langerhans cells? Does the *T6* (which is present on the surface of a Langerhans cell in very high concentration) perform some function related to the processing of antigen by the cell? The answers to these questions should be quite readily obtainable with current technology.



**COORDINATED RESPONSE** of immunologically competent cells in the skin begins when a foreign antigen penetrates the horny layer of keratin. It encounters Langerhans cells, which present it, along with a histocompatibility antigen, to *T* cells (1) programmed to respond to this antigen. The *T* cells, now activated, display a receptor for interleukin-1 (*IL-1*). The binding of this factor, which is secreted by keratinocytes (and perhaps by Langerhans cells), induces the activated *T* cell to secrete interleukin-2. *IL-2* binds to a receptor displayed on other *T* cells responsive to the antigen, in the dermis, lymph nodes or spleen (2), and triggers their proliferation to form a population of *T* cells directed against the antigen.

# The Search for Proton Decay

*Physicists have been keeping watch over an 8,000-ton underground detector, waiting for a sign that all matter has a finite lifetime. So far no proton has been observed to decay, but the vigil will continue*

by J. M. LoSecco, Frederick Reines and Daniel Sinclair

Life is fleeting and kingdoms fall, and even stars and galaxies may someday fade, yet one might think the basic stuff of matter—the protons, neutrons and electrons of the atom—would endure forever. In the case of the electron it is probably so: both experiment and the elegance of physical theory suggest that electrons are immune to decay. For the proton and the neutron, however, immortality is far from certain. Indeed, the neutron, when it is not stably bound inside the nucleus of an atom, is known to decay. It breaks down spontaneously to yield three lighter particles: a proton, an electron and a neutrino. What keeps the proton from decaying into lighter particles? There has never been a secure basis for the belief that it does not. A proton might break down, for example, into a positron (a positively charged electron) and a pair of neutrinos, and no general or fundamental law of physics would be violated. It seems there is nothing in nature to prevent this process; on the other hand, it has never been observed.

The possibility of proton decay has been a nagging question in physics since the 1930's, but in the past 10 years it has become a matter of more than passing interest. A new class of theories has been developed in which the decay of the proton is not only allowed but also definitely predicted. With the simplest of the theories it is even possible to calculate the proton's lifetime: the estimate is about  $10^{30}$  years, many orders of magnitude greater than the age of the universe (roughly  $10^{10}$  years), but finite all the same. Observation of the decay and measurement of the lifetime are the chief means of testing the theories.

Experimental physicists have taken up the challenge. The methods needed for the search are quite different from those of most experiments in elementary-particle physics. Instead of bom-

barding a detector with a beam of high-energy particles, every effort is made to shield the apparatus from stray particles that might strike it. Accordingly, enormous detectors have been set up deep underground in tunnels and mine shafts. No one has yet observed an unequivocal instance of proton decay, and as a result the experimental lower limit on the proton's lifetime is now greater than the theoretically predicted lifetime. This does not mean that the proton is stable, however; it means the search for proton decay must go on.

The rules governing the decay of particles are conservation laws, which state that a certain property or quantity must remain forever unchanged. The most important laws for our purposes require the conservation of energy, linear momentum, angular momentum and electric charge. In general any particle will decay unless it is prevented from doing so by one of these laws. The decay of an isolated neutron described above is consistent with all four laws; for example, the total electric charge is zero both before and after the decay. The electron, on the other hand, cannot decay because it is the lightest particle with an electric charge; any imagined scheme of elec-

tron decay would violate the law of charge conservation.

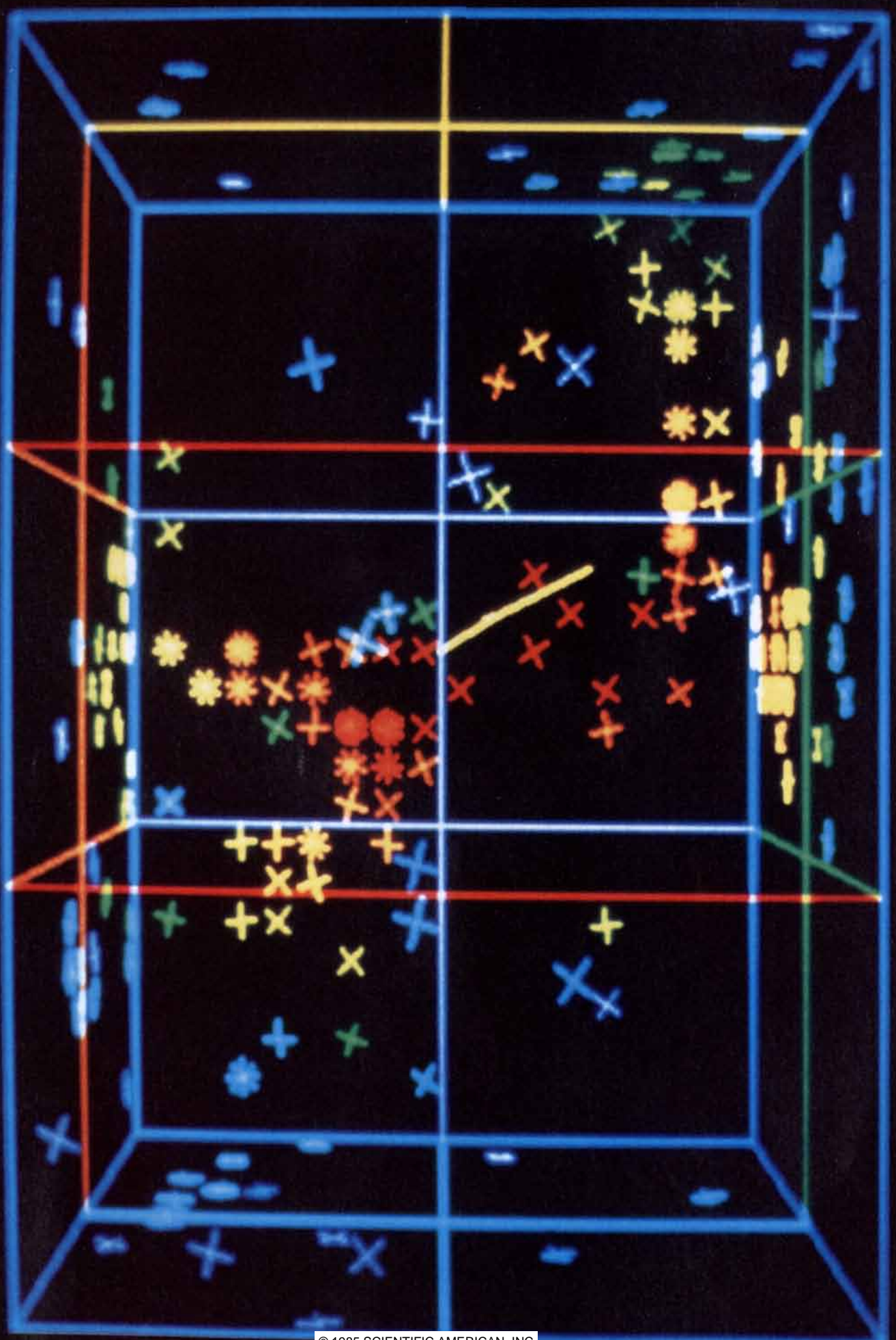
It should be emphasized that the four conservation principles cited here are quite general and are grounded in fundamental concepts. They are useful in all domains of physics, and their validity has been proved. The puzzling thing about the proton's apparent stability is that no such general law accounts for it. The hypothetical decay of a proton into a positron and two neutrinos would conserve energy, linear momentum, angular momentum and electric charge. Many other possible decay modes would also obey the conservation rules.

In the 1930's Hermann Weyl and E. C. G. Stückelberg, and later Eugene P. Wigner, attempted to explain the proton's stability by postulating a new conservation law. They put the proton and the neutron in a class of particles called baryons and assigned a "baryon number" of +1 to them; they then proposed that baryon number is a conserved quantity in nature. The neutron can decay into a proton, an electron and a neutrino because the total baryon number is unchanged by the process (it is +1 both before and after the decay). The proton cannot decay because it is the lightest baryon.

The introduction of baryon number

**SIMULATED DECAY OF A PROTON** creates a splash of colored stars on a graphic display representing the structure of a large detector. The detector is a rectangular tank filled with clear water, and in the display one is looking down into it from overhead. Photomultiplier tubes mounted on the walls and floor of the tank and near the surface of the water detect light emitted by fast-moving particles. In this simulation a proton decays into a positron and a neutral pion, which fly in opposite directions. When light from the moving particles strikes a photomultiplier, the tube "fires," producing an electrical pulse. Here each tube that fires is marked by a star. The number of points in the star indicates the amount of light received and the color indicates timing; tubes marked by a red star are the first to fire, followed by yellow, green and blue. The reconstructed back-to-back tracks of the decay products are indicated by heavy yellow lines near the middle of the display. A simulated proton decay is shown because no actual event of this kind has been seen. Events of other kinds have been observed at the rates expected, and they match the simulations. Hence experimenters are confident that if a proton had decayed in this way, they would have detected it.





does not really explain what inhibits the proton's decay but merely gives it a name. The proposed law of baryon conservation is not a general law of physics: it has no application beyond the field of elementary particles, and it is not founded on any fundamental concept. Furthermore, the law would be suspect even if it were not arbitrary. The universe is composed almost entirely of protons and electrons. For each positive electric charge (proton) there is evidently a negative charge (electron), so that the universe as a whole is electrically neutral. It is not neutral with respect to baryon number, however. There are far more protons (baryon number +1) than antiprotons (baryon number -1). It appears that in the extremely hot first moments of the big bang more protons than antiprotons were created, and so some process operating then must have violated the law of baryon conservation. If the law could be broken then, why not now?

Although the conceptual underpinnings of proton stability have always been shaky, the problem was given little attention until 1974, when new theoretical models called grand unified theories suddenly made it an is-

sue of immediate concern. The goal of the new theories is to unify three of the four fundamental forces of nature. The strong and the weak nuclear forces and the force of electromagnetism are to be made part of a single framework, leaving only gravitation as a separate entity. There are precedents in physics for such a unification. In the 19th century James Clerk Maxwell unified the theories of electricity and magnetism, and in the 1960's a deep connection was found between the weak force and electromagnetism.

The simplest of the grand unified theories was developed by Howard Georgi and Sheldon Lee Glashow of Harvard University; it is called minimal SU(5). The designation SU(5) refers to the mathematical group of symmetries on which the theory is based; it is minimal in that it is the theory with the fewest "adjustable parameters," which must be assigned a value by experiment. According to minimal SU(5), the strong, weak and electromagnetic forces, which seem very different under ordinary circumstances, become indistinguishable when particles interact with an energy of approximately  $10^{15}$  billion electron volts (GeV). Moreover, at this enormous energy the law of baryon conserva-

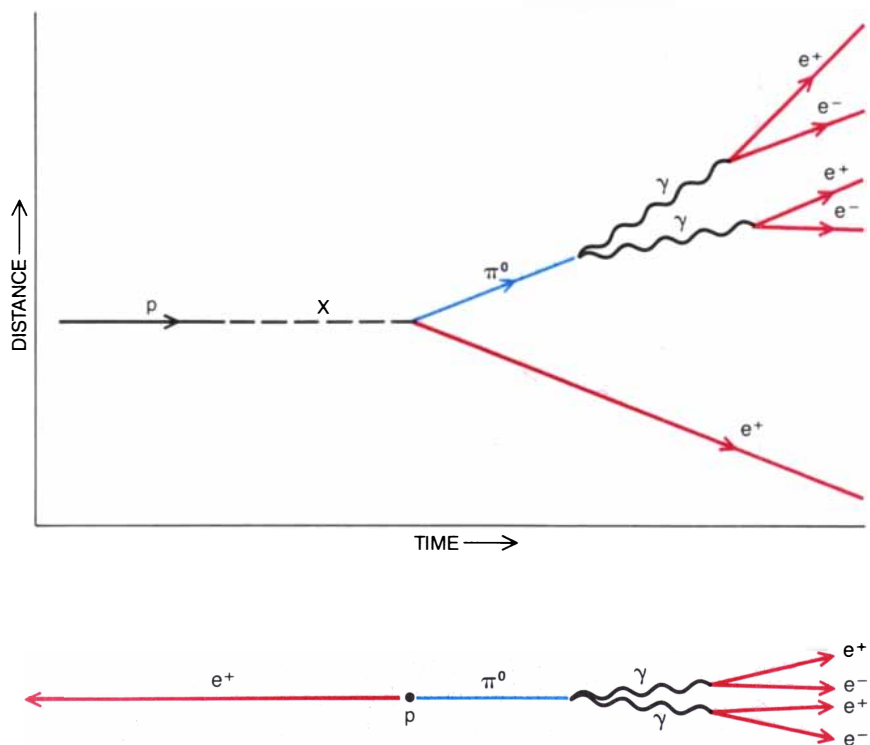
tion is repealed: events that change baryon number can take place as readily as events that conserve it.

The unification energy of  $10^{15}$  GeV is far beyond the reach of laboratory experiments: the largest particle accelerators have yet to surpass 1,000 GeV. The failure of baryon conservation has consequences even for matter at rest, however. In particular, minimal SU(5) predicts that a proton can decay by means of an intermediate state with an energy, or mass, of  $10^{15}$  GeV; the intermediate state decays in turn to yield light particles. It might seem this process would violate the conservation of energy, since a proton with a mass of less than 1 GeV gives rise to an intermediate particle of much greater mass. The intermediate state is so short-lived, however, that it cannot be detected even in principle; from the point of view of observation it does not exist, and energy is conserved.

Even though the great mass of the intermediate state does not forbid proton decay, it does greatly diminish the probability of the event. At any given instant a proton is most unlikely to emit a particle with a mass of  $10^{15}$  GeV. Because the decay process is highly improbable, the lifetime of the proton is extremely long. In minimal SU(5) the estimated lifetime is approximately  $10^{30}$  years. Other grand unified theories also predict the decay of the proton, but the theories are too complicated to support a calculation of the lifetime.

Although we have been speaking exclusively of the proton's fate, the neutron is subject to decay by the same mechanism. An isolated neutron can decay (as outlined above) to yield a proton, an electron and a neutrino, but when the neutron is bound in an atomic nucleus, the process is suppressed. (The reason is that adding the positive charge of a proton to a nucleus that already has a positive charge costs more energy than is released in the decay.) Thus a bound neutron is stable against all modes of decay that conserve baryon number, but it can decay if baryon conservation is violated. In the grand unified theories the lifetime of a bound neutron is similar to the lifetime of a proton, but the decay modes would be different because of the difference in electric charge.

How can one measure a lifetime of  $10^{30}$  years in a universe that is only  $10^{10}$  years old? There is no need to wait for a particular, selected proton to decay. The estimate of  $10^{30}$  years represents the half-life of the proton: the time in which half of the protons in any sample of matter can be expected to decay. With a large enough sample



**PREDICTED MODE OF PROTON DECAY** yields a positron ( $e^+$ ), or positively charged electron, and a neutral pion ( $\pi^0$ ). The pion then decays into two photons ( $\gamma$ ), which in turn produce pairs of electrons and positrons. The upper diagram shows the evolution of the system of particles; distance in space is represented along one axis and time along the other. The event is mediated by an extremely massive, short-lived particle designated  $X$ . The lower diagram shows the geometry of the event as it would be observed in the laboratory. The positron and the decay products of the neutral pion separate on back-to-back trajectories.

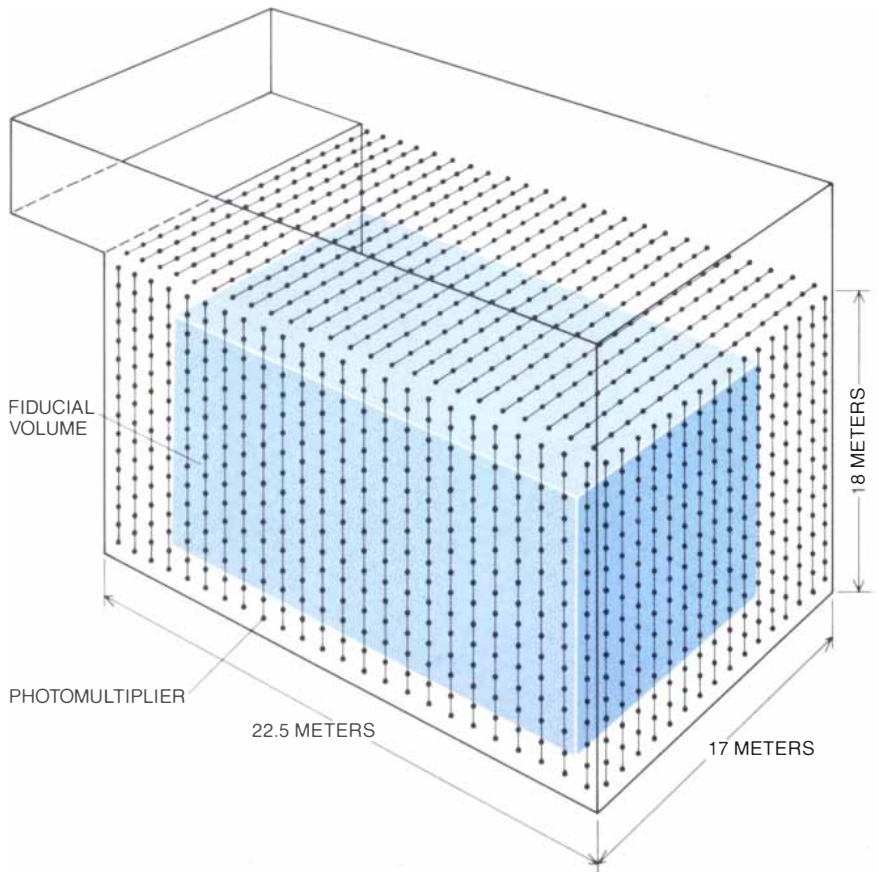
some events will be seen even in a much shorter period; monitoring  $10^{30}$  protons ought to yield an average of one decay per year.

The most straightforward approach to detecting the decay of the proton would be an experiment based on simple counting. All the protons in a large sample of matter would be counted and then the matter would be set aside for a year or so. If a second count revealed that some protons were missing, they could be assumed to have decayed. Such an experiment would be independent of all assumptions about the mode of the proton's decay: the set of particles emitted is immaterial, since what is detected is the absence of the proton itself. Unfortunately the experiment is totally impractical. It would require counting  $10^{30}$  protons without error.

Another approach is not only practical but also easy. Proton decay constitutes a kind of radioactivity, and so it must contribute to the background radiation at the earth's surface. The background can be measured with a Geiger counter or a similar instrument; if the contributions from known sources were then subtracted, one could assume that whatever is left is due to proton decay. This crude procedure sets a lower limit on the proton lifetime of  $10^{17}$  years, or 10 million times the age of the universe.

Until the development of the grand unified theories the only incentive for mounting more elaborate experiments was the conviction that a principle, such as baryon conservation, is no better than the experiments that test it. In accordance with this precept a measurement of the proton lifetime was undertaken in 1953 by Clyde L. Cowan, Jr., of the Los Alamos Scientific Laboratory, Maurice Goldhaber of the Brookhaven National Laboratory and one of us (Reines). It was the first experiment to employ a massive detector to search for proton decay.

The detector was a 300-liter tank filled with a liquid scintillator, a substance that emits a flash of light when a charged particle passes through it. Ninety photomultiplier tubes registered the flashes. The detector had originally been designed for another purpose (experiments with an intense beam of neutrinos) and the instruments available could not distinguish the decay of an individual proton from other events that might mimic it. Installing the apparatus 30 meters underground, however, shielded it from most cosmic rays, the chief source of extraneous particles. By assuming all events that could not be accounted for otherwise were due to proton decay, a



**8,000-TON DETECTOR** in the Morton Thiokol salt mine near Cleveland is operated by the authors and their colleagues in a group called the IMB collaboration. The detector was built by excavating a cavern at a depth of 2,000 feet, lining it with plastic, filling it with water and installing 2,048 five-inch photomultiplier tubes. Even at this depth some cosmic rays reach the detector, causing reactions that can be confused with the decay of a proton. Because of the cosmic-ray background, an event is counted among the potential proton-decay candidates only if its apparent point of origin is at least two meters from the detector's walls. The inner, "fiducial" volume thus defined (color) has a mass of 3,300 metric tons.

lower limit of  $10^{22}$  years was set on the proton's lifetime.

Over the next two decades further experiments pushed the lower limit back several more orders of magnitude. Some of the experiments were based on a radiochemical analysis of geologic samples. For example, the disintegration of a proton within a nucleus of potassium 39 converts the nucleus into argon 38, which is unstable and promptly emits a neutron to become argon 37. This last nucleus is radioactive, so that its concentration can be measured by simple radiation counters. The key to the experiment is to find specimens of rock that have remained buried and undisturbed since they formed. From the ratio of argon 37 to potassium 39 one can then deduce the number of protons that have decayed over a geologic period. By this method it was shown that the half-life of the proton is at least  $10^{26}$  years.

Other experiments, more like the ones now under way, attempted to count proton decays as they took place. The measurements were made

in conjunction with studies of cosmic rays done deep underground, where all but the most penetrating radiation was screened out. In 1974 one of us (Reines) reported the results of work with a detector set up two miles below the surface in a South African gold mine. The device included 20 tons of liquid scintillator as well as some 84,000 flash tubes, in which an electrical discharge in a gas is triggered by the passage of a charged particle. Again a limit on the lifetime was calculated by isolating those events whose origin was unknown and assuming they could all be attributed to proton decay. The limit set in this way was greater than  $10^{30}$  years, but it applied only to certain decay modes, namely those in which the decay products included either a muon or a positively charged pion. (The muon is a particle related to the electron but with a mass about 200 times larger; the pion is the lightest member of the family of particles called mesons.)

The entire character of the study of proton decay changed with the ad-

vent of minimal SU(5). Earlier a few theorists had been enthusiasts for experiments testing proton stability, notably Jogesh C. Pati of the University of Maryland and Abdus Salam of the International Centre for Theoretical Physics in Trieste. Now many more took an interest. In part the change in outlook came from the simple fact of having a definite prediction that the proton does decay. Equally important, however, minimal SU(5) offered an es-

timate of the lifetime, and one that did not seem beyond the range of possible measurement. The theory also supplied a preferred mode of decay: in the commonest events a proton would break down into a positron and a neutral pion. These predictions gave a goal and a direction to experiments.

The decay mode favored by minimal SU(5) happens to be one that leaves a distinctive signature. The positron and the neutral pion both have

comparatively high energy—about .5 GeV—and they fly off in opposite directions. Such “back to back” events at high energy are unlikely to arise from any cause other than proton decay. Hence the events yielding a positron and a neutral pion should be easier to identify than those resulting from many other decay modes.

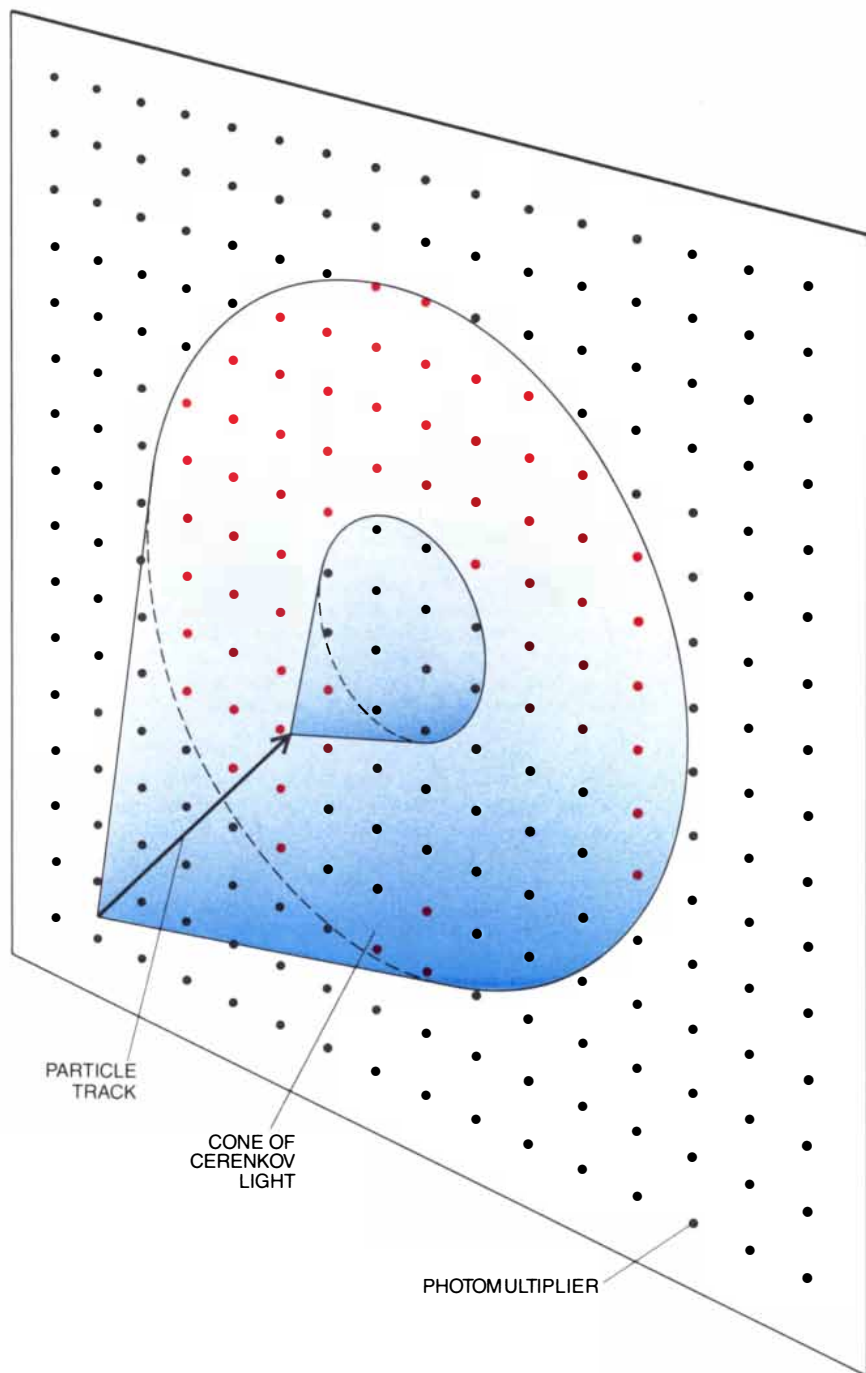
Even when one knows what to look for, however, testing the predictions of minimal SU(5) is a formidable task. A half-life of  $10^{30}$  years corresponds to the decay of about one proton or neutron per day in 1,000 metric tons of matter. To allow for uncertainties in the theory, an experiment ought to be able to detect the decay even if the half-life is as long as about  $10^{33}$  years. Thus in the worst case  $10^{33}$  protons, or roughly 3,000 tons of matter, must be kept under observation for a year merely to detect a single decay. Furthermore, that decay must be recognized against a background of many extraneous events.

Two strategies can be adopted for dealing with background events: the events can be excluded from the detector or identified and discounted in the data recorded. In practice both techniques are important.

One source of background is radiation from natural radioactivity, which cannot be screened out entirely. Any shielding material chosen would itself include some nuclei susceptible to radioactive decay; for that matter, so would the material of the detector. On the other hand, emissions from radioactive nuclei are relatively easy to identify. The energy of a typical radioactive decay is less than 1 percent of the energy released by the decay of a proton, and so a simple energy measurement serves to discriminate between the two kinds of event.

Cosmic rays are more troublesome. They come in all energies, and at the earth's surface they include a wide variety of particle types. At sea level the flux is about one particle per square centimeter per minute, so that a 10,000-ton detector would be pelted by more than  $10^{12}$  particles per year. To find a single proton decay in the course of a year, a trillion cosmic-ray events would have to be identified and rejected.

It is to reduce the flux of cosmic rays that detectors are put underground. Many of the incident particles, such as protons, neutrons and pions, can be absorbed by a few meters of heavy shielding, but excluding muons calls for more extreme measures. Muons lose energy very slowly as they move through matter, so that for the shielding to be effective it must be thousands



**CERENKOV RADIATION** is the basis of the water-filled detector. The radiation is emitted when a charged particle moves through the water faster than the speed of light in water (which is about three-fourths of the speed of light in a vacuum). The Cerenkov light forms a cone centered on the particle's path, and so the path can be reconstructed from a recording of the light's intensity and time of arrival at the photomultipliers in the detector array.

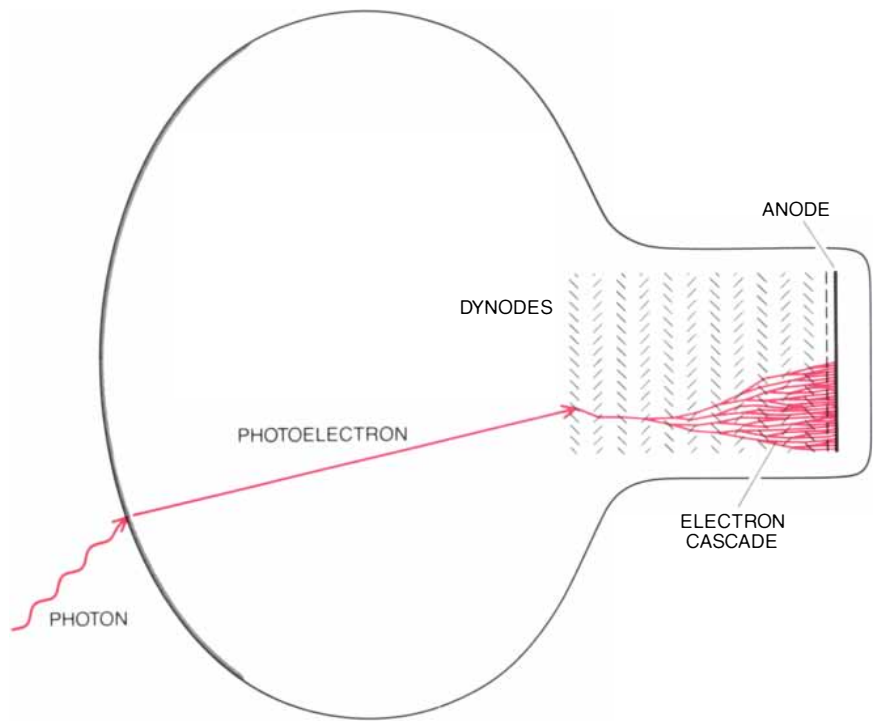
of meters thick. Even then a few high-energy muons leak through.

In the case of the neutrino, shielding is quite impossible. Neutrinos in the energy range characteristic of proton decay interact with matter so seldom that they readily pass through the entire earth. Of course the rarity of neutrino interactions implies that most of the neutrinos also pass through the volume of the detector without causing any disturbance, but now and then a neutrino does collide with a particle of matter. The frequency of the events is proportional to the mass of the detector (just as the frequency of proton decays is); in 1,000 tons of material there is one neutrino interaction every few days. In some cases the debris from the interaction includes muons, electrons and pions with energies comparable to what would be expected from proton decay.

Since neutrino-induced events cannot be prevented, they must be distinguished from proton decays. The property that allows the distinction to be made is the angular distribution of the products of the event. When a neutrino strikes a particle, the neutrino's momentum causes the collision products to be thrown forward. In contrast, when a proton at rest decays, the particles emitted move in opposite directions, with a net momentum of zero. The need to make such distinctions has had an important influence on the design of detectors for proton-decay experiments. It is not enough for the detector to record the total energy of the event, as some earlier instruments did; the position and direction of the particles must also be determined.

The detectors planned and built over the past decade are of two general types: layered tracking detectors and water Cerenkov detectors. In a tracking detector plates of iron or steel are interleaved with electronic "counters" sensitive to the passage of a charged particle. The iron provides the stock of protons and neutrons whose decay is awaited. The counters record the successive positions of the decay products as they pass from plate to plate. The spatial resolution of a tracking detector (and hence the accuracy with which a trajectory can be reconstructed) depends in part on the thickness of the plates. Thin plates give better resolution, but they also increase the cost of the device, since more particle counters are needed.

The water Cerenkov detectors are based on an effect discovered in 1934 by the Russian physicist Pavel A. Cerenkov. The effect is the emission of light by a charged particle moving through a transparent medium, such as



**PHOTOMULTIPLIER TUBES** respond to exceedingly faint flashes of Cerenkov light. When a photon strikes the tube, the photocathode on the inner surface of the glass envelope emits an electron, which is attracted to a series of electrodes called dynodes. The dynodes are in a "venetian blind" arrangement, with each row at a progressively higher voltage. A dynode struck by an electron emits several more electrons; as a result the number of electrons is multiplied by a factor of  $10^9$  and a single photon can give rise to a measurable pulse of current. The photon's time of arrival can be determined to within five nanoseconds.

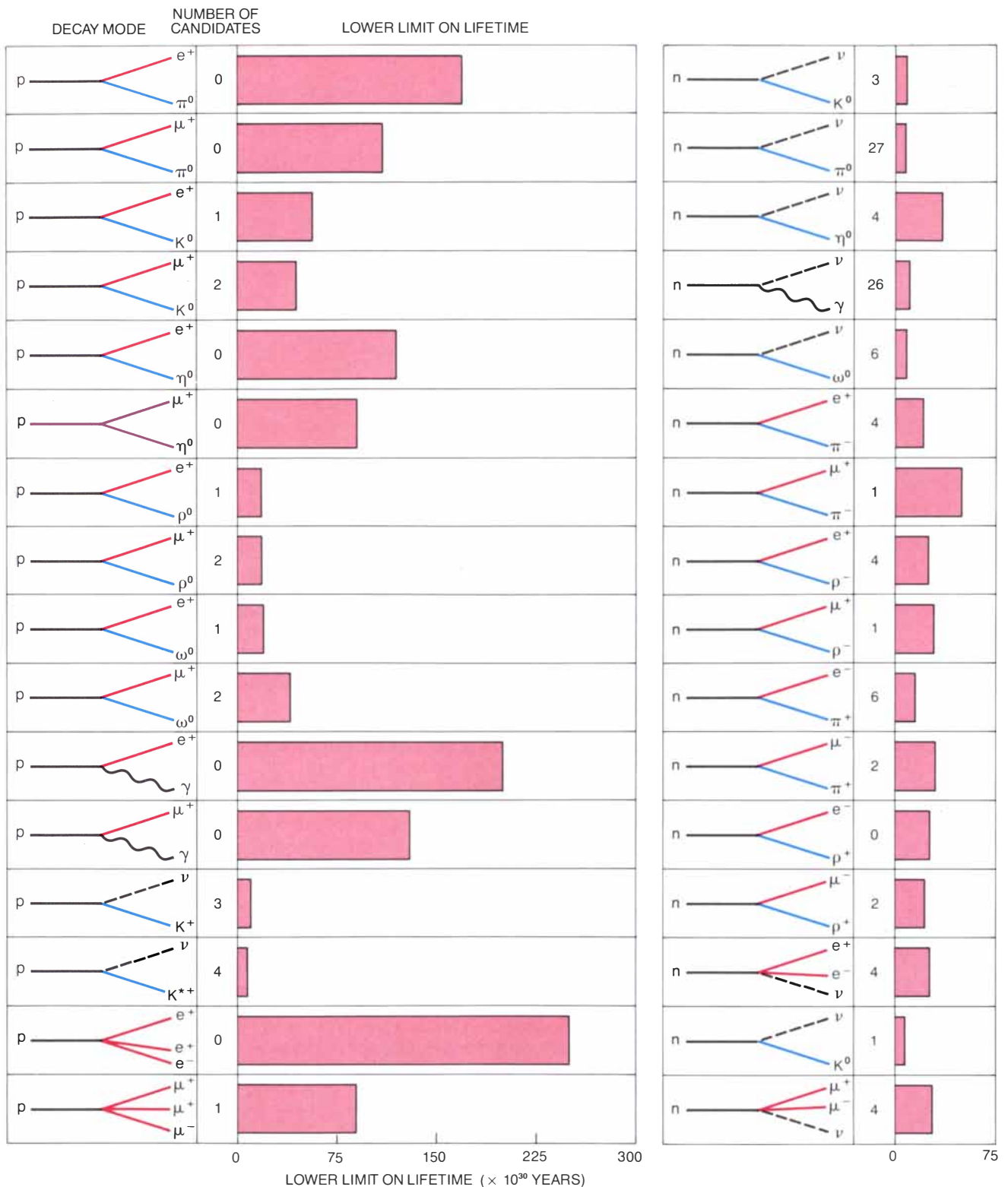
water, faster than the speed of light in that medium. (Nothing can move faster than the speed light has in a vacuum, but in water light itself is slowed to about three-fourths of its vacuum speed.) The Cerenkov light is an electromagnetic shock wave analogous to the sonic boom formed when an aircraft exceeds the speed of sound. Like the sonic boom, the Cerenkov radiation is emitted in a cone. The angle between the particle's direction of motion and the direction of light emission depends on the ratio of the particle's speed to the speed of light in the medium. For a particle moving through water at nearly the vacuum speed of light the angle is 42 degrees.

In a Cerenkov detector a large volume of clear water is surrounded by photomultiplier tubes. Each charged particle created by the decay of a proton or a neutron in the water gives rise to a brief flash of Cerenkov radiation, which causes some of the photomultipliers to "fire," or produce an electrical pulse. The amplitude of the pulses and their times of arrival are recorded on magnetic tape for later analysis. The pattern of photomultipliers exposed to the light and the sequence in which they fire provide the information needed to reconstruct the path of the particle. The photomultiplier tubes must be

extraordinarily sensitive. At a distance of five meters the pulse of Cerenkov light from a single charged particle is about as bright as that from an ordinary flashlight seen at the distance of the moon.

Because Cerenkov radiation is emitted only by charged particles, it might seem that electrically neutral particles would escape unseen. Some of them do—notably neutrinos—but others can be detected through their secondary decay products. A neutral pion, for example, generally decays quickly into a pair of gamma rays, or high-energy photons. The gamma rays are also electrically neutral, but they interact to produce pairs of electrons and positrons, which can be detected. The neutral pion is therefore observed as a cascade of electron-positron pairs.

One advantage of the water Cerenkov detector is that water, being made up partly of hydrogen atoms, includes protons that are not bound up with other particles in a complex atomic nucleus. The signal of proton decay should be clearest in the case of such free protons. Another advantage of the water Cerenkov detector is that its cost does not increase in proportion to its mass. The major cost is not the active medium (water) but the photomultipliers. Since they are mounted at the



**LIMITS ON THE LIFETIME** of the proton and the neutron are given for 32 possible decay modes. The values are based on data collected in 204 days of operating the IMB detector. All events that could not definitely be identified as cosmic-ray interactions were considered possible candidates for each of the decay modes. The minimum lifetime for each mode was then calculated from the number of possible candidate events and from the detector efficiency: the percentage of the events of each type that the detector could be expected to identify. A lower limit on the lifetime for a particular

mode represents the minimum lifetime that would be observed if that decay mode were the only one possible. For the favored decay mode yielding a positron and a neutral pion the theoretically predicted lifetime is no more than  $2.5 \times 10^{31}$  years; no candidates have been observed for this mode, and the experiment sets a lower limit of  $1.7 \times 10^{32}$  years. Although there are candidate events for some of the other modes, those modes are harder to distinguish from cosmic-ray interactions. In the authors' estimation there is no compelling evidence that any instance of proton decay has been observed.

surface of the enclosed volume, their number depends on the surface area, which is proportional to the two-thirds power of the mass. In short, the cost per unit of mass diminishes with increasing detector size. For this reason the water Cerenkov technique has been adopted for the largest proton-decay detectors. Ultimately the size of such instruments is limited by the absorption of light in water, but the limit has not yet been reached.

We have taken part in the design and operation of the largest of the water Cerenkov detectors. It is a rectangular vat of water 23 meters long, 17 meters wide and 18 meters deep and has a total volume of more than seven million liters. On all six sides photomultiplier tubes are mounted one meter apart on a square grid. There are 2,048 tubes in all.

Planning for the experiment was begun in 1979 by physicists (including us) from the University of California at Irvine and the University of Michigan. We were later joined by Maurice Goldhaber of Brookhaven, and so the project came to be known as the Irvine-Michigan-Brookhaven, or IMB, collaboration. The aim of the experiment was to observe proton decay or, failing that, to extend the lower limit on the lifetime to  $10^{33}$  years. The choice of a detector technology was the first major decision. Layered tracking detectors had been in use (for slightly different purposes) for 10 years or more, whereas the concept of a large water Cerenkov detector, capable of recording enough information to trace the path of a charged particle, was as yet untried. Simulations done with a computer, however, showed that the idea was sound.

There remained the difficult question of where to put the detector. At the earth's surface 100,000 cosmic-ray events per second would swamp the electronic system. The flux of cosmic rays would have to be reduced to no more than a few per second, which meant going underground about 2,000 feet. How could such a huge, water-filled apparatus, the size of a six-story building, be assembled at that depth at acceptable cost?

We considered excavating a cavern in a mine or tunnel and building a tank within it, but the cost was too high. We searched for existing caverns, but we found none that were suitable. The solution came from an unexpected quarter. Members of the group had worked on cosmic-ray experiments in a mine operated by the Morton Salt Company (now Morton Thiokol, Inc.) near Cleveland. Through this connection

we learned that the management of the mine was interested in testing a new continuous-mining machine and would share the cost of bringing it underground. More important, they agreed to provide at cost the mining skills and labor needed to excavate the large cavern our experiment required. To reduce the cost further we decided to forgo a freestanding tank and instead line the walls of the cavern with plastic.

The pool was filled by July, 1982, and the installation of the 2,048 photomultiplier tubes was completed soon after. The tubes are arranged along the walls and floor of the cavern and near the surface of the water, facing inward. Each tube has a five-inch hemispherical face and is mounted in a watertight plastic enclosure. The total mass of the water in the cavern is 8,000 metric tons, but events originating in a two-meter shielding zone nearest the walls on all six sides are eliminated from consideration. Excluding the shielding zone leaves a "fiducial" mass, in which candidate proton decays are identified, of 3,300 tons.

An event is recorded whenever 12 or more photomultipliers fire within a period of 50 nanoseconds. This is the time needed for light to travel 10 meters in water, the largest distance likely to be covered by the products of proton decay. The time resolution of the photomultipliers is five nanoseconds, and so the sequence in which the tubes fire can be used to determine the direction from which the light came.

Even at a depth of 2,000 feet the detector is triggered by penetrating muons about 2.7 times per second. In passing through the detector from top to bottom a muon gives off enough Cerenkov radiation to fire about 600 photomultipliers; a proton decay, on the other hand, would fire no more than about 250 tubes. Many of the muon-induced events can therefore be identified and rejected on this basis alone. About a third of the muons, however, cut across a corner of the detector and fire fewer than 300 tubes. These events cannot be rejected until the recorded data are analyzed. The discrimination is done by a computer program that excludes from further consideration all events initiated by a particle entering the detector from outside. The events remaining, known as contained events, are those originating within the fiducial volume. If examples of proton decay are to be found, they must be among the contained events.

By no means are all the contained events plausible candidates for proton decay. Neutrinos (and occasionally other neutral particles) can enter

the detector unseen, then interact with matter to generate an event that appears to originate within the fiducial volume. These events must be rejected by a detailed analysis in which the geometry of the event is the most useful criterion for judgment. As noted above, the debris created when a neutrino collides with a particle of matter usually moves in the general direction of the incident neutrino. In proton decay the commonest expected pattern consists of particles moving in opposite directions.

Unfortunately the distinction between these patterns is not perfectly sharp. On rare occasions one of the products of a neutrino interaction can bounce backward, at a large angle to the direction of the incident neutrino. Moreover, the back-to-back alignment of the products of proton decay can also be disrupted. If the decaying particle happens to be one of the isolated protons that constitutes a hydrogen nucleus, the angle between the products in a two-body decay should be precisely 180 degrees. In the water molecule, however, 16 of the 18 protons and neutrons present are bound in the oxygen nucleus, where scattering of the decay products can weaken their back-to-back angular correlation.

Only a small subset of the contained events satisfy the geometric criterion. These few events are examined further in an analysis drawing on all the information supplied by the detector, including the total energy of the event, the number of particles, the paths they follow and in some cases the kinds of particles emitted. Each event is considered separately as a possible candidate for each hypothetical decay mode of the proton and the neutron. The number of candidates for each mode that cannot be rejected determines the lower limit on the lifetime for that mode. In calculating the lower limits we also take into account the detector efficiency: the percentage of events in each mode that we could expect to recognize in the data, given a large sample.

The members of the IMB collaboration have completed an analysis of the data recorded during 204 days of operating the detector. In that time there were 169 contained events. Are any of them consistent with proton decay? In the case of the distinctive mode favored by minimal SU(5), yielding a positron and a neutral pion, the answer is unequivocal: no events were seen with these particles in the characteristic back-to-back arrangement. The lower limit established by the experiment for decay in this mode is a half-life of  $1.7 \times 10^{32}$  years. The theory

had predicted that the lifetime would be found to lie in the range between  $10^{28}$  and  $2.5 \times 10^{31}$  years, and so there is clearly a disagreement. When only this mode is considered, the proton lives at least seven times longer than the predicted maximum.

For other decay modes the results are not as clear-cut. There is no compelling evidence that we have observed proton decay in any mode, but there are a number of events whose classification is to some degree uncertain. They could have been induced by neutrino interactions, but the possibility cannot be excluded that they are genuine proton or neutron decays. In setting limits on the lifetime we have adopted the conservative policy of considering each event a candidate decay unless it can definitely be explained otherwise. The resulting limits are given in the illustration on page 60.

The question of whether the proton decays has not been settled by this experiment. In one respect the situation is much as it was before the introduction of minimal SU(5): experimen-

talists have no quantitative goal to set the scope of their efforts. The proton lifetime could lie just beyond the range of the present detectors, or it could be many orders of magnitude greater. In another way the situation is quite different. Even though minimal SU(5) appears to be inadequate, it is now generally acknowledged that theoretical considerations imply the instability of the proton. Moreover, observation of proton decay would be considered a unique and persuasive test of the idea of grand unification.

If the decay is to be seen, or if the limits are to be extended, larger and more sensitive detectors will be needed. The existing instruments are being modified and new ones are being built. We are upgrading our own detector by installing larger photomultiplier tubes, and other schemes for increasing the amount of light collected are under consideration. The changes will improve our ability to distinguish the various decay modes from the patterns generated by cosmic-ray neutrinos.

Does the proton decay? Will the question ever be answered? The pres-

ent lower limits on the lifetime can doubtless be extended somewhat, but this process cannot continue indefinitely. Perhaps someday a detector 10 times as large as the IMB device might be built, but a detector 100 or 1,000 times as large is not feasible. Cost is not the only constraint (although it is a formidable one). If the proton lifetime is much greater than  $10^{33}$  years, the irreducible background of neutrino interactions would probably obscure many decay modes no matter how large the detector was. Adding to the mass would simply increase the number of background events in the same proportion as the decay events.

Although the question of proton decay remains unsettled, the history of the search provides an interesting commentary on the relations between theory and experiment. When the question was first raised, the limits on the proton lifetime were not very stringent, and yet most physicists expected the answer to be no. The limits have now been increased by 15 orders of magnitude, but most physicists have come to believe the answer is yes.

	SPONSORING INSTITUTIONS	LOCATION	DEPTH (EQUIVALENT METERS OF WATER)	DETECTOR MASS (METRIC TONS)	DETECTION METHOD
WATER CERENKOV DETECTORS	University of California at Irvine, University of Michigan, Brookhaven National Laboratory, Cleveland State University, University of Hawaii, California Institute of Technology, University College Warsaw	Morton Thiokol salt mine, Painesville, Ohio	600 1,600	8,000 TOTAL 3,300 FIDUCIAL	2,048 FIVE-INCH PHOTOMULTIPLIERS ON ONE-METER SURFACE GRID
	KEK, University of Tokyo, University of Tsukuba	Kamioka metal mine	825 2,400	3,000 TOTAL 1,000 FIDUCIAL	1,000 20-INCH PHOTOMULTIPLIERS ON ONE-METER SURFACE GRID
	Harvard University, Purdue University, University of Wisconsin	Silver King mine, Park City, Utah	525 1,500	700 TOTAL 420 FIDUCIAL	704 FIVE INCH PHOTOMULTIPLIERS ON ONE-METER LATTICE, MIRRORED WALLS
LAYERED TRACKING DETECTORS	Tata Institute, Osaka City University, University of Tokyo	Kolar gold fields, South India	2,500 7,600	140 TOTAL 100 FIDUCIAL	1,600 PROPORTIONAL GAS COUNTER TUBES
	CERN, Frascati Laboratory, University of Milan, University of Turin	Mont Blanc tunnel, French-Italian border	1,850 5,000	150 TOTAL 100 FIDUCIAL	47,000 LIMITED STREAMER TUBES
	Orsay, École Polytechnique, Saclay, Wuppertal University, Tufts University	Fréjus tunnel, French-Italian border	1,550 4,200	160 TOTAL	1,500 PLASTIC FLASH-TUBE PLANES; 200 GEIGER TUBES
	Argonne National Laboratory, University of Minnesota, University of Oxford, Rutherford Laboratory	Soudan iron mine, Soudan, Minn.	675 1,800	30 (PROTOTYPE)	HEXAGONAL DRIFT TUBES

**SEVEN EXPERIMENTS** installed in deep mines and tunnels are continuing the quest for proton decay. Three of the experiments employ water Cerenkov detectors. The other four experiments are

based on detectors in which iron plates are interleaved with particle counters. The depth of each detector is given, as is the depth of water that would offer equivalent shielding from cosmic rays (color).



# A REPAIR SHOP LIKE NO PLACE ON EARTH.

In low orbits around the Earth, satellites gather, analyze and transmit critical information—for scientists, for corporations, for countries.

Because of the high cost of getting them up there in the first place, some of these satellites are designed to be repaired where they are if something goes wrong.

A manned space station could act as the neighborhood garage—the local spare parts and repair facility for these valuable satellites. That's one reason why

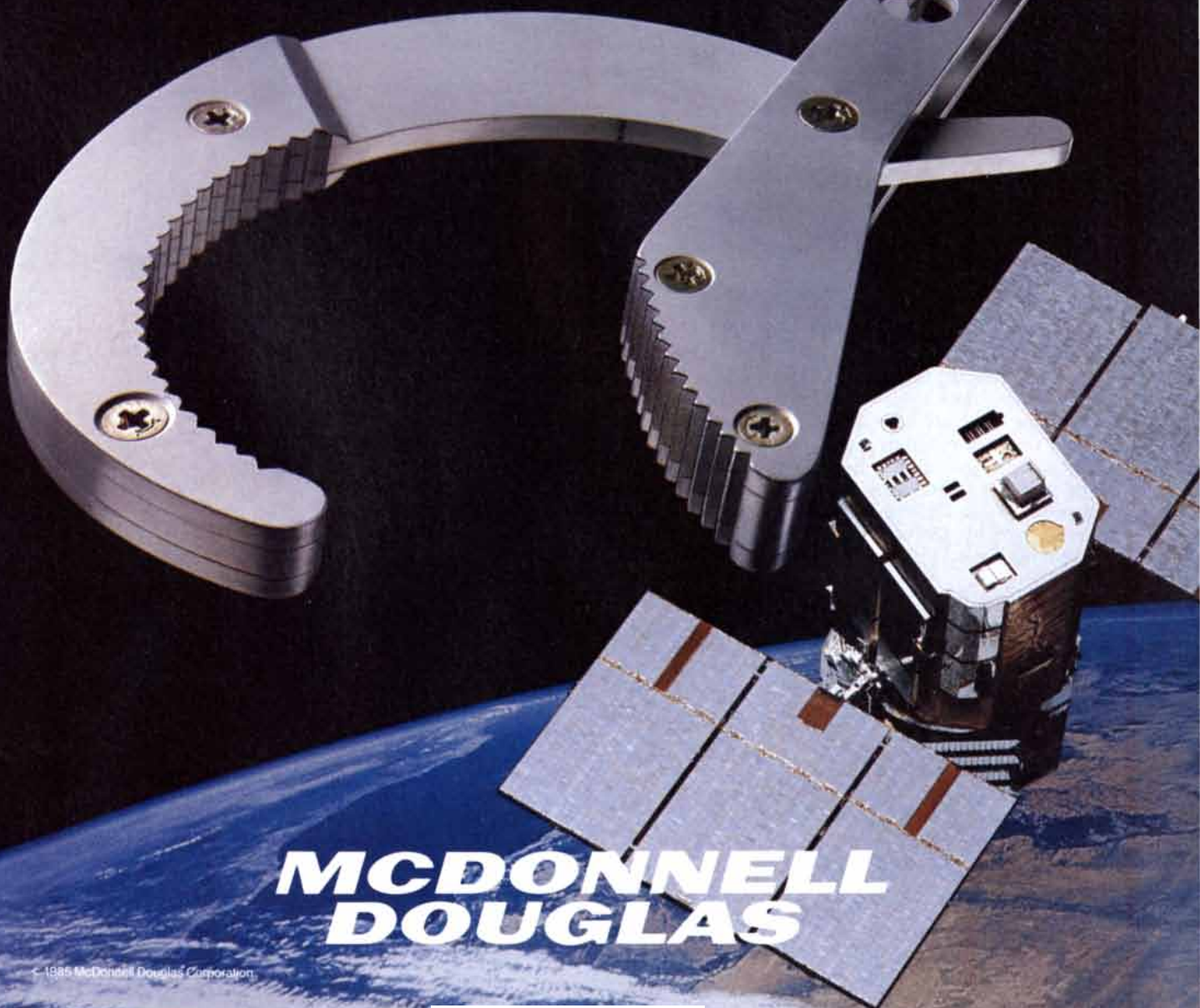
McDonnell Douglas is working to put a space station in orbit by the 1990s.

Since 1960, when we did the first Manned Orbiting Research Laboratory studies, we've been a world leader in space systems. We built the original space station—Skylab. And we've been a pioneer in space systems operations, from launches to payload integration.

Now we're devoting our space experience, systems expertise, and technological ingenuity to taking America the next step into

space—with a manned station.

Such a station could lower the cost of technological upkeep—for science, for industry. We think that's a very good reason to build a repair shop like no place on Earth.



## MCDONNELL DOUGLAS

# SCIENCE AND THE CITIZEN

## Summing Up

In his State of the Union Address, President Reagan declared: "Despite budget restraints, we will seek record funding for research and development." Indeed, the Administration's proposed budget for the fiscal year 1986 includes \$60.3 billion for research and development, a 12 percent increase over fiscal 1985. The large overall figure, however, masks trends that disturb some observers.

The fraction of the nation's science budget that is devoted to defense-related R&D has climbed swiftly since the last year of the Carter Administration; in the proposed budget it reaches 72 percent, the highest proportion since the early 1960's. The total sought for military projects is \$43.6 billion—a 22.6 percent increase over the 1985 figure. According to the Department of Defense, that level of spending is needed to maintain the technological advantage of the U.S. military over Soviet forces, to keep pace with rapidly increasing Soviet investments in defense R&D and to embark on the Strategic Defense Initiative: the plan to construct a shield against enemy ballistic missiles, which alone is slated to receive \$3.7 billion next year.

Science outside the Defense Department and the military programs of the Department of Energy, in contrast, would receive no increase in funding over 1985. When the expected rate of inflation is taken into account, the proposed funding of civilian R&D amounts to a decrease of 4.4 percent. The same contrast is evident when money for basic research is considered separately from funds for development: funding for basic research under the aegis of the Defense Department would grow by 12.8 percent, whereas funding for civilian basic research would decline.

The largest contributor to the decline is a proposed cut in funding for the National Institutes of Health, which finances nearly 40 percent of Federally supported basic research. The 1986 research budget allots \$4.9 billion to the NIH, a drop of nearly 6 percent from 1985.

The NIH budget is in part the legacy of an effort by the Administration's Office of Management and Budget to reduce the number of research grants awarded by the NIH in 1985. In brief, the OMB instructed the NIH to set aside funds to cover the full terms of several hundred of its multiyear grants, thereby tying up a part of the

funding authorized by Congress for 1985 and leaving the NIH with money for only 5,000 new awards rather than the 6,500 that Congress had intended it to fund. The "forward funding" of grants reduced the amount of money that, in the Administration's view, the NIH will need in 1986 to sustain its commitments and fund new grants at an annual level of 5,000.

As the NIH cutbacks indicate, the life sciences do not fare as well in the proposed budget as physics, engineering and mathematics. Another reduction in funding for the life sciences is a drop of 6.4 percent in the research budget of the Department of Agriculture. Of the cutbacks affecting other disciplines, the largest would affect civilian projects within the Department of Energy, including fossil-energy, energy-conservation and magnetic-fusion research. The department's programs in high-energy and nuclear physics would also suffer a cut, of 6.2 percent.

How might congressional actions alter the proposed budget? NIH funding for 1986 will depend in large measure on the outcome of the ongoing wrangle over the 1985 budget. Congress may strike a compromise with the Administration on the number of new grants or may simply overturn the OMB's action and insist that the 6,500 grants originally authorized for 1985 all be awarded. Traditionally Congress has been responsive to the concerns of the biomedical community, and many members view the OMB's effort to limit the number of grants as an attempt to thwart the will of Congress. If the full number of grants is restored for 1985, there will be a strong incentive for Congress to add to the Administration's proposed level of funding for 1986. Whatever adjustments Congress makes, they are not likely to undo the general tightening of the budget for civilian science.

Some of the effects of straitened Federal support for R&D are already evident. Universities are increasingly seeking research funding from industry. Some institutions have tried to bypass traditional review procedures and appeal directly to Congress for research facilities. A third trend is the development of closer relations between universities and the cash-swollen military.

In 1986, for example, the Defense Department plans to spend \$25 million to foster interchange between university and defense laboratories and to strengthen university research pro-

grams in areas of interest to the military. The department will also devote \$30 million to buying equipment for such programs. Meanwhile Strategic Defense Initiative officials have invited university scientists to submit \$70 million in proposals in areas of research affecting missile-defense technology. Yet some workers are approaching the Pentagon's largess warily, fearing that defense grants could entail restrictions on academic freedom.

## Speech Impediments

"It is like being pecked to death by sparrows," says Donald N. Langenberg, chancellor of the University of Illinois at Chicago, referring to a trend that worries investigators and university administrators alike: slowly increasing efforts by the Federal Government to exert greater control over the flow of information and ideas. In general the step-up in Government intervention has resulted not from the promulgation of new regulations but from the more stringent application of existing ones.

One example is the Export Administration Act of 1979, which can be used to exclude unclassified papers from conferences where foreign nationals are present and to limit attendance at certain unclassified meetings to U.S. citizens. According to Leo Young, director of research and laboratory management in the Department of Defense, the department has been applying such restrictions much more often now than in the past. "Lately," he says, "papers have been held back that would have been let through without a second thought three or four years ago."

Aggressive use of export controls has led to a kind of self-censorship: several professional societies have voluntarily excluded foreign citizens from meetings and conferences without having been asked to do so. Conference organizers hope thereby to avoid the disruption that can ensue when papers must be retracted or sessions closed at the last minute.

Another regulation controlling the publication of unclassified material is a 1981 amendment to the Atomic Energy Act that authorized the Department of Energy to limit the dissemination of what is designated "unclassified controlled nuclear information." Critics say the amendment gives the department the power to keep the public from learning about administrative errors or safety hazards. They argue that

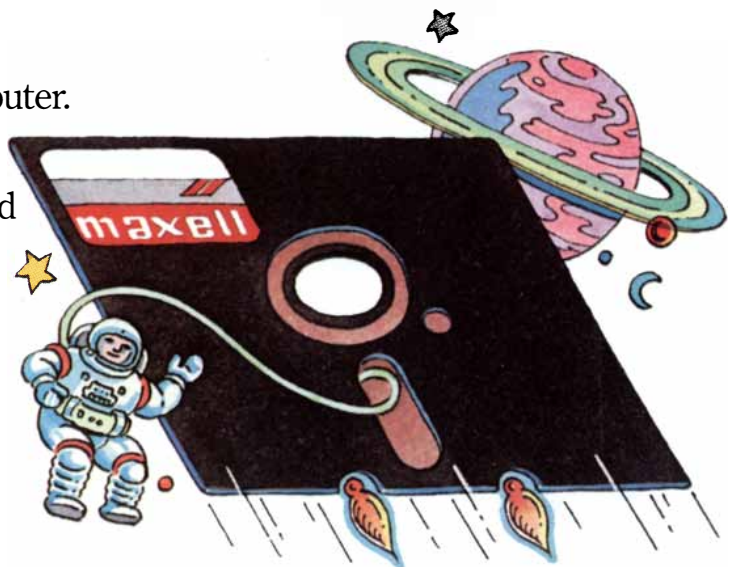


# Maxell Gold.

The floppy disk that  
packs more facts  
into Compaq,<sup>®</sup>  
sets HP<sup>®</sup> free,  
and takes IBM<sup>®</sup>  
Portable where it's  
never gone before.

It's great to have a portable computer. Especially when your data stays put. For error-free performance at home or abroad, trust Maxell. The Gold Standard in floppy disks. There's a Maxell disk for virtually every computer made. Each is backed by a lifetime warranty. Maxell. Accepted everywhere, without reservation.

**maxell**<sup>®</sup>  
IT'S WORTH IT.



any information sensitive enough to be controlled should be classified.

A major concern is that restrictions on unclassified information may have a harmful effect on scientific progress within the U.S. by hampering the free exchange of information. One proposal from within the Administration would directly threaten such exchange. In January, Secretary of Commerce Malcolm Baldrige suggested that more severe controls be placed on the National Technical Information Service, which makes unclassified technical studies originating in Government agencies available to the public and to commercial database vendors. Baldrige expressed concern that Soviet science might profit from such information. Critics say restricting the service could actually harm national security by slowing U.S. research and development.

Government control over scientific information is likely to increase as a larger proportion of the U.S. research and development effort comes under the purview of the Department of Defense, but the restrictive actions go beyond the area of national defense. Last December, John Shattuck, vice-president for government, community and public affairs at Harvard University, produced a report summarizing Federal restrictions on the free flow of academic information. Shattuck found that certain forms of restriction originally intended to protect national security have been imposed by agencies that "have no relationship to national security matters." Such restrictions often take the form of "prepublica-

tion review" clauses in grant contracts, which can require a scholar to get approval from a funding agency before publishing the results of Government-supported research.

Other types of restriction include "technical direction" clauses, which require direct participation in the project by a Federal official, and "changes" clauses, which allow the funding agency to change "without notice the content and/or scope of the research contract without the researcher's agreement." Restrictive clauses have been found in contracts issued by the National Institutes of Health, the National Institute of Education, the Department of Housing and Urban Development, the Health Resources and Services Administration and the Food and Drug Administration. Harvard and some other major universities have been able to negotiate changes in such contracts or have refused the grants, but administrators at other institutions confirm that they have had to accept contracts with restrictive clauses.

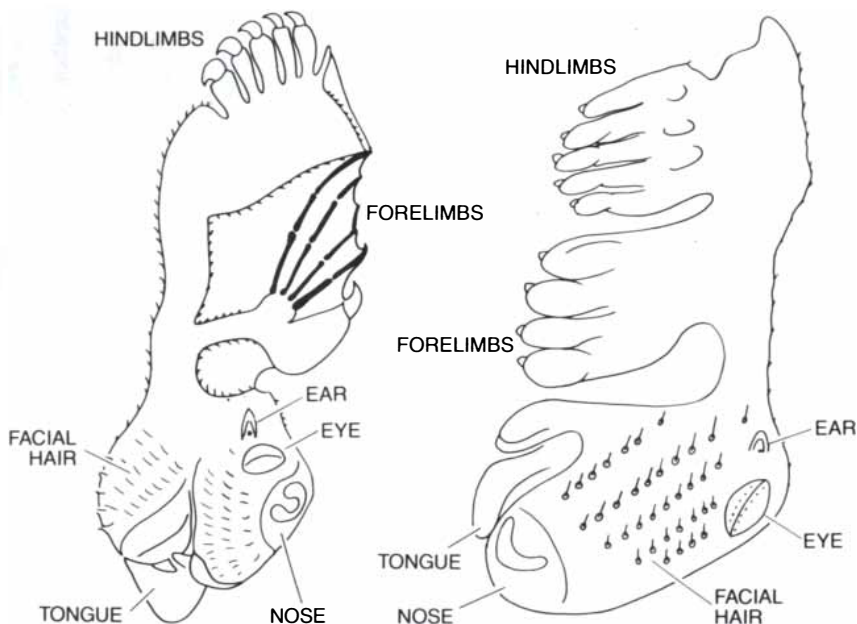
The Shattuck report also mentions National Security Directive 84, a presidential directive signed in March, 1983, but not currently in effect because a Senate resolution returned it to the White House for further consideration. Under NSDD 84 more than 120,000 Federal employees would have been required to sign lifetime agreements to submit to prepublication review any work related to their Government service. Such agreements would effectively silence many who take Federal jobs and then return to academic life. It would also discour-

age academic workers from taking Government posts, since it would limit their future publications. "It would have been enough to turn me off" Government service, says Langenberg, who served as deputy director of the National Science Foundation, because "it is unlikely that any publication I would ever write would have nothing to do with my Government service."

## Bat Map

Every mammalian brain studied during the past half century contains three maps of the mammal's body surface. Each map consists of areas that correspond to specific surface regions. The maps are in the somatic sensory cortex, the part of the brain that receives information from muscles, joints and the skin. The maps are distorted because the area of the cortex devoted to a part of the body is proportional not to the actual size of the part but to the amount of sensory information it provides. For example, a large amount of the cortex is dedicated to the hairs of the face, which are important sensory organs in most mammals. In a wide range of species the maps faithfully reflect the overall configuration of body parts.

The fact that the configuration of each body-surface map corresponds to the configuration of the body may be a coincidence; it is more likely that this arrangement has an undiscovered functional significance. The latter interpretation is supported by research on the fruit-eating bat *Pteropus poliocephalus*. Jon H. Kaas of Vanderbilt Uni-



Fruit-eating bat (left); forelimb placement in its body-surface map (center) is reversed from that of mammals such as the rat (right)



Georgian Bay

## Blue Horizons

Miles and miles of glistening granite isles, whispering pine and the siren call of fair winds and wide open waters – these are the lures of Ontario's northern horizons. A sparkling vista of 400,000 lakes and rivers, wilderness parks and resorts are yours to discover! Ask about the 7% sales tax rebate, 5% room tax rebate and the advantageous exchange on American currency, up to 35%.\* For resort reservations, call 1-800-461-0249. For information call TOLL FREE 1-800-268-3735 or write: Ontario Travel, Queen's Park, Toronto M7A 2R9, Ontario/Canada.

*\*Subject to change*



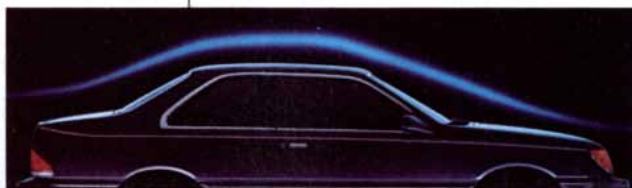
# Forward

## At last, a family car with the

When we developed Ford Tempo, we didn't forget your family's need for room, comfort and trunk space. But since we don't think that a family car has to be a boring car, we added some special refinements. One of which is Tempo's advanced aerodynamic shape.

### Round vs. Square.

A round-object, of course, is much more aerodynamic than something square-shaped. And that's why Ford Tempo's lines are rounded rather than squared-off. This kind of forward thinking results in a distinctive design. And just as

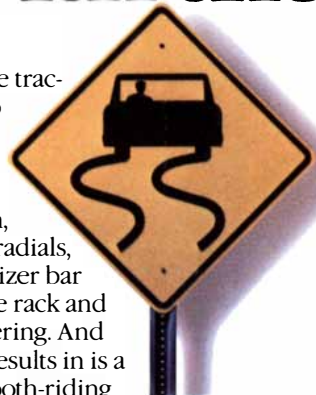


importantly, it results in a functional shape that actually reduces lift for improved directional control and stability. In short, Tempo's shape improves the way it drives. Which brings us to the next paragraph which deals with handling.

### Excellent reflexes.

As you'd logically expect from a forward thinking car, Tempo offers front-

wheel drive traction. It also offers four-wheel independent suspension, all-season radials, front stabilizer bar and precise rack and pinion steering. And what that results in is a stable, smooth-riding car that helps the driver handle the idiosyncracies of a winding road. Good news for the driver. And the passenger.



### Forward thinking under the hood.

Tempo is powered by a specially



# Ford Tempo.

# Thinking true instincts of a driver's car.

developed 2300 HSC (High Swirl Combustion) engine. And to keep Tempo's thinking current, we've added Electronic Fuel Injection this year. A forward thinking 2.0 liter diesel engine is available. And the optimum operating efficiency of your Tempo will be maintained by the EEC-IV

Computer, a state-of-the-art microprocessor engine control system.

**State-of-the-art thinking for five.**

The end result is a five-

passenger, state-of-the-art family car that thinks and acts like a driver's car. Any car that offers you less, is backwards by comparison.

**Best-Built American Cars.**

"Quality is Job 1." A 1984 survey established that Ford makes the best-built American cars. This is based on an average of problems reported by owners in the prior six months on 1981-1983 models designed and built in the U.S.



**Ford Dealer Lifetime Service Guarantee.**

As part of Ford Motor Company's commitment to your total satisfaction, participating Ford Dealers stand behind their work, in writing, with a Lifetime Service Guarantee. No other car companies' dealers, foreign or domestic, offer this kind of security. Nobody. See your participating Ford Dealer for details.

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



Get it together—Buckle up.

# The forward thinking car.

versity and J. D. Pettigrew of the University of Queensland and their colleagues write in *Nature* that they delineated the body-surface maps of the bat by stimulating small areas of the animal's skin with probes. Microelectrodes planted in localized regions of the somatic sensory cortex detected the resulting electrical activity.

They found that the body-surface maps of the bat are laid out much like those of other mammals. There is a significant variance, however: the relative placement of the forelimb digits is reversed in the maps of the bat. In the maps of most mammals the digits point in a forward direction, whereas in those of the bat the digits point to the rear. The investigators suggest the difference is related to the fact that most mammals hold their forelimbs under or in front of their heads. In contrast the bat holds its forelimbs behind its head while flying and above its head while resting in its familiar upside-down position.

### User-friendly Antibody

The first laboratory cells producing a pure human antibody of therapeutic promise have been developed by a group of investigators at the Stanford University School of Medicine and the University of California at San Diego. The antibody represents a major advance because all active antibodies previously made in the laboratory (without the use of genetic engineering) have been animal ones. Such antibodies carry a relatively high risk of being rejected by the human immune system. The human antibody was developed for the treatment of gram-negative sepsis. This bacterial illness is a leading cause of death from infection in the hospital, killing about 75,000 Americans annually.

The standard procedure of making antibodies begins by injecting mice with an antigen to stimulate the multiplication of antibody-producing *B* lymphocytes. The *B* cells are then isolated and fused with malignant myeloma cells to form immortal antibody-producing cells called hybridomas. The desired hybridoma is selected and cloned to supply large amounts of a specific, "monoclonal" antibody.

Writing in *Proceedings of the National Academy of Sciences*, Nelson Teng and Henry Kaplan of Stanford and Abraham Braude of San Diego explain their procedure for making a pure human monoclonal antibody. They vaccinated two patients who were suffering from Hodgkin's disease with a mutant strain of the bacterium *Escherichia coli*. *B* lymphocytes were extracted from the patients' spleens (which had to be

removed because of the disease). The workers made hybridomas by fusing *B* cells making an antibody against an endotoxin on the bacterial surface with heteromyeloma cells (which are themselves the result of a fusion of human and mouse myeloma cells). Because the antibody-producing genes in the resulting hybridoma cells are human genes, the antibodies they encode are human too.

The new antibody offers hope for the treatment of gram-negative sepsis, an illness caused by the release of endotoxin. *E. coli* normally colonize the human digestive tract, but in healthy people the bacteria are harmless because they are kept under control by the body's immune system. Disease or surgery weakens the immune system and sometimes allows the bacteria to break through a patient's defenses, releasing endotoxin.

Teng and his colleagues have found their antibody is highly effective in protecting rabbits and mice from bacterial endotoxin-induced sickness and death. Teng speculates that the antibody neutralizes endotoxin by binding to certain segments of its core, although no one fully understands the mechanism. He thinks a test of the antibody in humans, which he plans for sometime within the coming year, should reveal few complications.

### Spacetimequake

A design for a pair of instruments that could detect tremors in the structure of spacetime is expected to receive \$1 million in additional funding from the National Science Foundation for the coming fiscal year. The tremors, which are called gravitational waves, are predicted by Einstein's general theory of relativity; they would alternately stretch and compress any region of spacetime through which they pass. The two proposed detectors, being developed jointly by the California Institute of Technology and the Massachusetts Institute of Technology, could register changes in distance caused by gravitational waves to less than a few parts in  $10^{21}$ .

Each detector is to be a large laser interferometer, housed in two pipes set at right angles to each other and evacuated to a pressure of about  $10^{-11}$  atmosphere. Each pipe would be 48 inches in diameter and about five kilometers long. A beam of laser light would be split in two near its source at the intersection of the pipes, and each part of the beam would be directed along one of the pipes to a mirror mounted on a freely suspended mass of metal at the other end. The two reflected beams would then be recom-

bined at another freely suspended mass near their source, and their interference pattern would be observed. A passing gravitational wave would slightly alter the distance between one or both pairs of masses, thereby changing the observed interference pattern of the recombined beams.

Such gravitational ripples are so small that there is only indirect evidence for their existence: Joseph H. Taylor of Princeton University has observed a loss of energy from one binary-star system that agrees accurately with the loss predicted for the system through gravitational radiation.

The current lower limit on the amplitude of gravitational waves is set by so-called resonant-bar detectors. The resonant bar is a cylinder, usually made of aluminum, that vibrates like a gong in response to the tidal forces caused by a gravitational wave. The gravitational strain in such bars has been shown to be less than one part in  $10^{17}$ .

The laser interferometer has two main advantages over the resonant-bar detector. One is the sensitivity of an interferometer to a relatively broad range of frequencies. The second is the comparative ease with which its sensitivity can be improved: longer arms on an interferometer lead to better sensitivity, but a resonant bar cannot be made longer without changing the frequency of its response. To rule out spurious signals two detectors of comparable sensitivity are needed. One laser interferometer would probably be built in California and the other would probably be built in Maine.

Gravitational-wave astronomy is likely to open a window on the universe even more intriguing than the window opened by radio astronomy in the 1950's. Gravitational waves could carry information about the coalescence of binary-star systems made up of neutron stars and black holes, the evolution of supernovas, the earliest stages of the big bang and perhaps other phenomena not yet envisioned.

### Bee Picture

To human beings a rose is demonstrably a rose because it matches an image stored in the brain. What is a rose to a honeybee? Most workers would say a bee recognizes a flower by just a few distinguishing features, such as the relative proportion of edges and surface area on its blossom. Unlike vertebrates, bees and other invertebrates are thought to be incapable of remembering images.

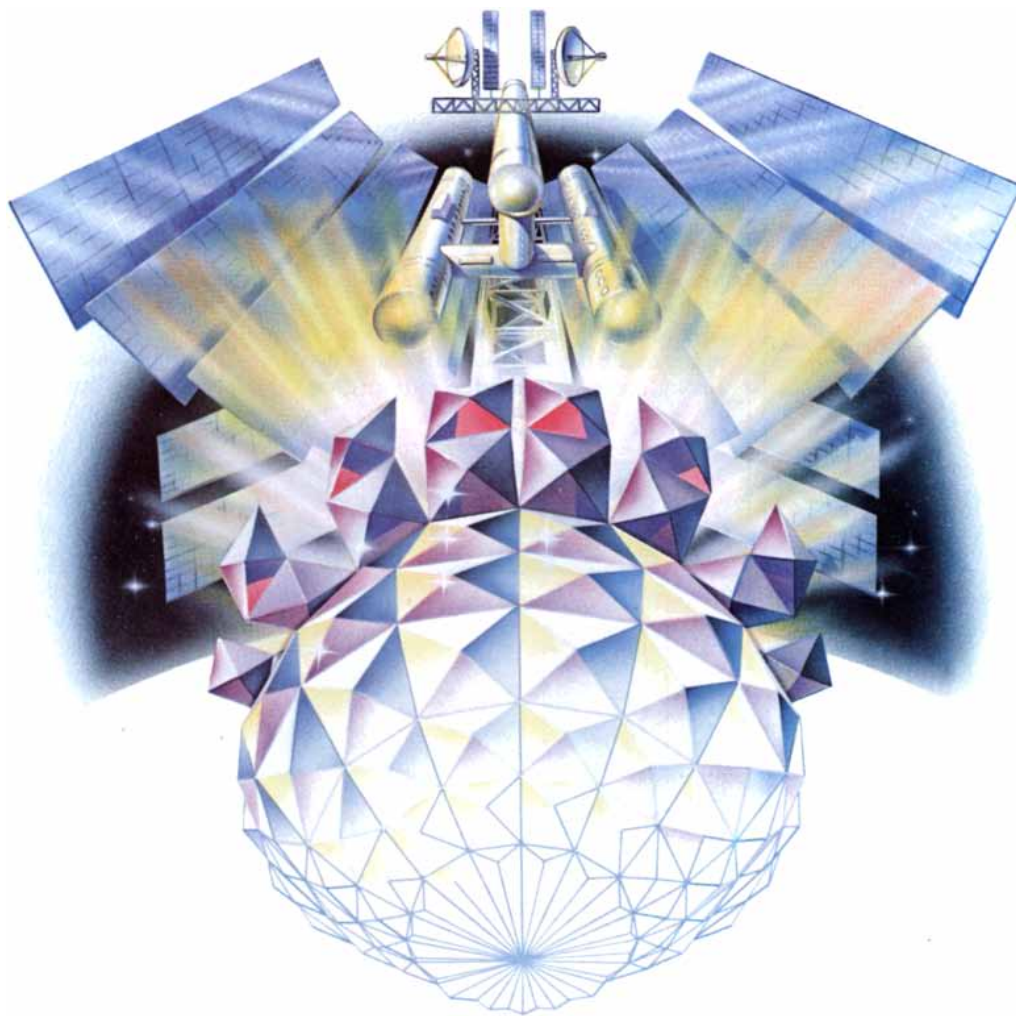
Experiments described by James L. Gould of Princeton University in a recent issue of *Science* suggest that this



# 'bal·əns

## Balance

- A well-ordered integration of elements; e.g., combining the sales from diverse businesses into nine years of uninterrupted earnings growth.
- Stability produced by distribution of weight; e.g., planned diversification between government and commercial business.
- A harmony, one encouraging divisional entrepreneurship with strong financial support—like \$2.2 billion in capital expenditures over the last five years.



Ours is a strategic balance—between down-to-earth management and options as limitless as space. This balance has been the key to our consistent financial growth.

Now we have added dramatically to that balance. Our new subsidiary, the Allen-Bradley Company, brings us nearly \$1 billion in sales in industrial automation and electronics equipment and systems. That moves us up to a \$3.5 billion share in electronics markets which are rapidly expanding. And positions Rockwell for \$11 billion in overall sales. To learn more about us, write: Rockwell International, Department 815S-56, 600 Grant Street, Pittsburgh, PA 15219.



**Rockwell International**

...where science gets down to business

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



**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 141 F Jonathan Blvd. N Chaska, MN 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.

view may be an unfounded prejudice. Gould presented foraging honeybees (*Apis mellifera ligustica*) with a choice between two vertically oriented artificial flowers differing only in the spatial relation of their "petals." For instance, the second floral pattern was often simply a rotation of the first one; to a bee, which keeps its head vertical in flight, the two patterns would always look different.

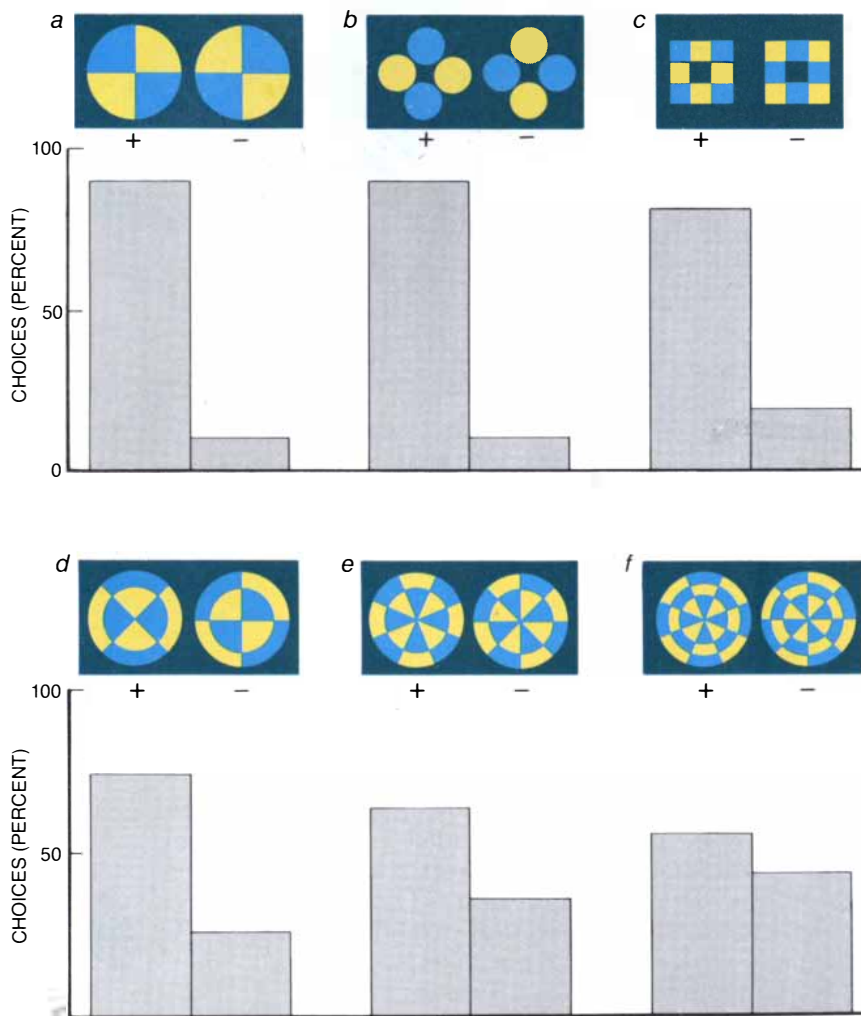
Initially one of the plastic flowers contained an unscented sucrose solution, whereas the other contained no food. After training the bees in this manner Gould again presented them with the same two patterns, but this time neither flower held sucrose. If bees remember only isolated features, he reasoned, they should show no preference for one of the flowers; only an eidetic, or photographlike, image can preserve in memory the spatial relations among the components of an object.

In repeated trials with many pairs of patterns the bees consistently showed a strong preference for the pattern that

had initially offered the sucrose reward. When the plastic flowers had only four petals, the bees chose the "correct" pattern more than 90 percent of the time; even when the number of petals was increased to 16, the preference was a statistically significant 65 percent.

Gould concludes that the bees must have learned to recognize the sucrose-associated pattern by storing it as an eidetic image. The stored image, however, appears to be fairly crude. When the number of petals was raised to 24, the insects no longer seemed to remember the difference between the flowers. Although their visual system is almost certainly able to distinguish between patterns of such complexity, the resolution of their eidetic memory is probably much lower.

The fact that bees can store images at all indicates to Gould that the "presumptive vertebrate-invertebrate dichotomy is false." The enhanced status of bees raises an interesting possibility: the investigation of their relatively simple memory mechanism



Bees remember the sucrose-bearing pattern (+) unless the choice is too complex (f)

may produce insights applicable to the study of learning in higher animals, including human beings.

### *Hypnotic Misrecall*

Where memory fails in the course of a criminal or civil investigation hypnosis is being used with increasing frequency to help witnesses remember forgotten events. Testimony thus elicited has even been admitted as evidence. A panel of the Council on Scientific Affairs of the American Medical Association questions the usefulness of hypnotic recall for this purpose. In its report the panel notes that although subjects generally reveal relatively more information while they are hypnotized, the recollections include many inaccurate details.

The panel based its conclusion on a two-year review of studies of the effect of hypnosis on the ability to remember. In one of these studies, the panel reports in *The Journal of the American Medical Association*, hypnotized subjects seemed to remember passages of poetry they had learned many years earlier but were unable to quote when not hypnotized. On comparing the quoted passages with the text of the original poem, however, the researchers found the subjects had improvised freely or filled in forgotten sections by fabricating passages in the style of the author. Numerous other studies support the conclusion that although a subject seems to remember more while hypnotized, the increase in reported detail is generally accompanied by loss of accuracy. Preconceptions held by the subject or the examiner can further distort recall. For instance, a hypnotized subject is more easily swayed by leading questions than a subject who is not hypnotized.

The panel recommends that the use of hypnosis in the judicial process be limited to the investigative stage. At this juncture it would not matter much if some of the recollections of a hypnotized subject were incorrect; they could be checked for consistency with other evidence. A single correct recall, on the other hand, might open an effective line of inquiry.

### *Strategic Metallurgy*

Cobalt and manganese are strategic metals the U.S. obtains almost entirely (90 percent) through imports, mostly from Africa; the remainder of the supply comes from recycling. Two workers at the Argonne National Laboratory have now discovered a simple, two-step process that extracts the metals from low- and medium-grade ores mined primarily for other met-

## A FATHER'S DAY TOAST TO THE FATHER OF THE MARTINI.



Jerry Thomas was a colorful 19th-century bartender in San Francisco. He concocted the first martini to cheer a weary traveler on his way to an outlying village called Martinez—which is how the martini got its name.

Thomas' creation was not confined to California for long. Today it is the most highly esteemed cocktail in America. Especially when called for by its first name: Beefeater.®

So: a salute to Jerry Thomas, from the gin that's done the most to immortalize his genius.

To send Dad a gift of Beefeater anywhere, dial 1-800-238-4373 (Void where prohibited.)



Imported  
**BEEFEATER® GIN**  
The Crown Jewel of England.™

**“The Boys Club helped me run my life.”**  
Pres., O.J. Simpson Enterprises  
O.J. Simpson

“When I was growing up, I was the quickest kid on the block. But the streets were catching up with me. I’m sure glad there was a Boys Club around to help keep me a step ahead.

“You know, a Boys Club shows kids there are lots of ways to reach goals, besides scoring touchdowns. It gives them every

chance to be leaders. And encourages something every bit as important as good leadership—good citizenship. They sure pointed me in the right direction.

“Hey, I’m not saying a Boys Club can turn a kid into a star. But it sure can teach ’em how to reach for one.”



**The Club that beats the streets.**

# Discover the most powerful

---

## The IBM Personal Computer AT.

---

Hold on to your hat.

The IBM Personal Computer AT (for Advanced Technology) is based on the advanced 80286 16-bit microprocessor. This remarkable computer will run many of the programs written for the IBM PC, up to three times faster. You'll be able to recalculate large spreadsheets in seconds and retrieve files in a flash. And it's ideal for IBM TopView, the new kind of software program that lets you run and "window" several other programs at once.

The IBM Personal Computer AT has got the power (and price) to surprise you. In many ways.

---

## Compatibility, expandability, networking too.

---

With the IBM Disk Operating System, the IBM Personal Computer AT can use many programs from the fastest-growing library in the personal computer software industry.

The IBM Personal Computer AT is also available with up to 3 million bytes of user memory to run multiuser, multitasking operating systems such as XENIX™. Volume upon volume of information is available at your fingertips. You can customize your system to store up to 20,000 pages of information at one time. And its keyboard helps you use all of this computing power more easily.

This member of the IBM PC Family is a powerful stand-alone computer that can also be both the

primary file server and a station on your network. With the IBM PC Network (which is so easy to

### IBM Personal Computer AT Specifications

<b>User Memory</b> 256KB-3MB*	<b>Diagnostics</b> Power-on self-testing* Parity checking* CMOS configuration table with battery backup*
<b>Microprocessor</b> 16/24-bit 80286* Real and protected modes*	<b>Languages</b> BASIC, Pascal, FORTRAN, APL, Macro Assembler, COBOL
<b>Auxiliary Memory</b> 1.2MB and 360KB diskette drives* 20MB fixed disk drive* 41.2MB maximum auxiliary memory*	<b>Printers</b> Supports attachment of serial and parallel devices
<b>Keyboard</b> Enlarged enter and shift keys 84 keys 10-foot cord* Caps lock, num lock and scroll lock indicators	<b>Permanent Memory</b> (ROM) 64KB Clock/calendar with battery*
<b>Display Screen</b> IBM Monochrome and Color Displays	<b>Color/Graphics</b> Text Mode Graphics Mode
<b>Operating Systems</b> DOS 3.0, XENIX* PC/IX 1.1	<b>Communications</b> RS-232-C interface
	<b>Networking</b> High-performance, high-capacity station on the IBM PC Network*

\*Advanced Features for Personal Computers

connect you can do it yourself), the IBM Personal Computer AT can share information with IBM PCs, PC/XTs and IBM *Portable* PCs.

---

## Get a hands-on, hats-off demonstration.

---

The IBM Personal Computer AT has the power, compatibility and expandability many PC users need, at a very appealing price.

For more information contact your authorized IBM PC dealer, IBM Product Center or IBM marketing representative. For a store near you call 1-800-447-4700. In Alaska or Hawaii call 1-800-447-0890.





personal computer IBM has ever made.



XENIX™ is a registered trademark of Microsoft Corporation.  
UNIX is a trademark of AT&T Bell Laboratories. PC/IX is based on UNIX System III, which is  
licensed to IBM by AT&T Technologies, Inc. Developed for IBM by INTERACTIVE Systems Corp

© 1985 SCIENTIFIC AMERICAN, INC

als. They think the process will be much cheaper on an industrial scale than conventional methods of obtaining cobalt and manganese.

The key to the process is a molten salt that dissolves more than 90 percent of the cobalt or manganese found in common, low-grade ores mined chiefly for such other metals as nickel and copper. The salt is a mixture of the chlorides of sodium, potassium and magnesium. It melts at a temperature of about 400 degrees Celsius (750 degrees Fahrenheit). Less than four pounds of the salts (mostly recyclable) will process a pound of ore.

After the metals have dissolved, a low electrical voltage (1.5 volts or less) is applied across the mixture. The voltage deposits the dissolved metal on a carbon electrode, from which it can be removed by chipping; it can simply be lifted off if the electrode shrinks enough in cooling.

The laboratory's work on the process was started by Victor A. Maroni. Later Samuel von Winbush of the State University of New York at Old Westbury came to the laboratory on a sabbatical leave. The two foresee no major problems in scaling the process up to an industrial level. They also foresee the possibility of taking ore from the sea and extracting the cobalt and manganese on specially built ships or floating platforms.

## Smoking Gun

Someone who smokes is far more likely to get cancer than someone who does not. What is the underlying physical mechanism that links tobacco smoke and carcinogenesis? Investigators at the Japanese National Cancer Center Research Institute in Tokyo may have found a clue: they have discovered that cigarette smoke can damage the DNA of a human lung cell.

Specifically, tobacco smoke can cut one of the two strands of the DNA double helix. Such a break would not by itself cause any cancer-inducing rearrangement of the DNA, because the cut strand and the intact one would still be bound together. A single-strand break could, however, make the genetic material more susceptible to influences that do generate such changes. Other groups have already found tobacco smoke can act as a promoter of carcinogenesis: it can enhance the carcinogenic effect of other substances.

The investigators, Tsutomu Nakayama, Motohisa Kaneko, Masahiko Kodama and Chikayoshi Nagata, report their work in *Nature*. They treated a culture of human lung cells with cigarette smoke dissolved in a liquid. Then they exposed the cells' DNA to a

solution that unwound the strands and pulled them apart from one another. Once the strands were no longer bound together, any single strand that had been cut was free to break up into fragments. The group found that DNA treated with cigarette smoke broke into a greater number of fragments than either untreated DNA or DNA subjected to gamma radiation.

## Colliding Interests

A \$20-million annual program is under way to plan and design the proposed particle accelerator known as the Superconducting Super Collider (ssc). There is general agreement in the community of high-energy physics about the energy and rate of interactions required if the machine is to meet one of its basic goals: the experimental resolution of some of the deepest questions that remain for a unified understanding of the electromagnetic and the weak nuclear forces. Several basic issues, however, such as the location of the ssc, are still to be decided. Moreover, some hard choices must be made about the design of the instruments that will detect new particles once they are produced. The choices are hard because the instruments are expensive, political support for the machine appears to be limited and each choice will almost certainly have a substantial effect on developments in elementary-particle physics for the next two decades.

The basic design of the ssc calls for a machine that accelerates two beams of protons in opposite directions around a large, closed loop. The two beams will then be made to collide, and the energy released by the collisions will create new particles. The main goal of every proposed design is to achieve a beam luminosity high enough to make the experimental findings statistically unambiguous. Luminosity, which is roughly a measure of the cross-sectional density of the particles in an oncoming beam, determines the number of collisions per second between particles in the two beams and hence the rate of physically interesting interactions. The luminosity of each beam of the ssc is to be  $10^{33}$  per second per square centimeter.

Another central aim is to achieve energies in each beam that are high enough to reach the threshold of unexplored physical processes with a high degree of confidence. Physicists argue on general grounds that new processes will be observed if the protons in each beam are accelerated to an energy of 20 trillion electron volts (20 TeV).

The outstanding issue for the accelerator itself is the design of the mag-

nets that will steer and focus the proton beams. One design, a compromise between groups at the Brookhaven National Laboratory and at the Fermi National Accelerator Laboratory (Fermilab), calls for superconducting magnets with a field strength of about six teslas. Such magnets could bend each beam of protons into a relatively tight circle 100 kilometers in circumference. A different design, endorsed by the Texas Accelerator Center (an organization of universities in Texas), calls for magnets with a field strength of three teslas. The magnetic field generated by the weaker magnets could be less sensitive than the stronger field to accidental shifts in the coils of wire. The circumference of the accelerator, however, would have to be 160 kilometers. Such a design might well argue in favor of a Texas site.

One major issue in the design of detectors is calorimetry, or the measurement of the energy carried away from the collision site. There is agreement that the calorimeters must be able to measure the energy of particles emitted in practically every direction, but the material that will stop some of the particles has not yet been determined. Uranium is more effective than iron or lead for stopping neutrons, and the measurement of energy by a uranium calorimeter can be twice as accurate as the measurement by a calorimeter made of iron. Uranium is much costlier, however, than iron or lead.

A related question is whether or not to measure the charge and momentum of emitted muons, which are expected to be the final decay products of several exotic particles. Such measurements depend on the extent to which a known magnetic field deflects the flight path of the particle. The muons emitted by the collisions at the ssc will be so energetic that a magnetized iron detector the weight of a battleship, mounted outside the calorimeter, would be needed to cause a measurable deflection in their flight paths.

Many other questions must be resolved: Will the calorimeter itself incorporate a magnetic field in order to measure the momentum of charged by-products? At what angle to the main loop will the colliding beams be made to cross? How will the experimental stations, where the beams will collide, be arranged around the loop? Will there be any provision for generating polarized beams of particles? Will it be possible to direct secondary beams of collision by-products, such as neutrinos or *B* mesons, into fixed targets in order to study secondary collisions? The implications for physics of such decisions will be thoroughly debated in the next two years.

Room. That was the idea behind the Camry. Legroom. Headroom. Family of five room. But along the way, something happened. Advanced technology made the Camry more than a roomy, comfortable sedan. It made it exciting, innovative; a whole new car was born. And the critics loved it. They loved its gripping front-wheel drive. They applauded its Electronic Fuel Injection and its sophisticated, electronically controlled 4-speed automatic overdrive transmission. They declared the Camry, every gorgeous inch of it, a star. The Camry's space and weight saving design helps make it fuel efficient, too. With all that responsiveness, you can still expect 29 City MPG, 34 Highway MPG!\* A performance like that deserves reviews like this...

### CAR AND DRIVER

"Toyota has fielded the most elaborate electronically controlled transmission to date." (And it's roomy. With almost 39" of headroom, even a 6'4" basketball player feels at home.)

### CONSUMER'S DIGEST

"Toyota Camry has been elected by our readers to the Consumer's Digest Hall of Fame, 'for the consistent high quality of its entire model line, its high owner satisfaction and commitment to safer design.'" (And it's roomy. The Camry seats five with room to spare.)

### THE CHRISTIAN SCIENCE MONITOR

"The Toyota Camry is a landmark vehicle... the Camry is fun to drive and think of all the gasoline

## OH WHAT A FEELING! TOYOTA

you save to the volume of space inside the car. If Toyota were to get a letter grade, it would have to be an A." (And it's roomy. Camry's got 93 cubic feet of passenger space... that's real stretch-out room for a family of five.)

### MOTOR TREND

"...even with an electric sunroof installed, the five-passenger Camry will accommodate six-footers in all locations." (See, we told you it was roomy!)

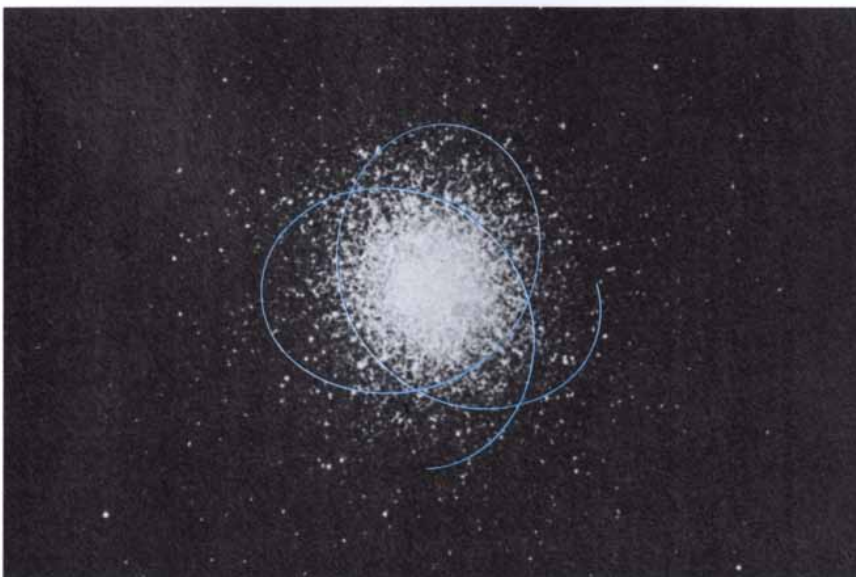
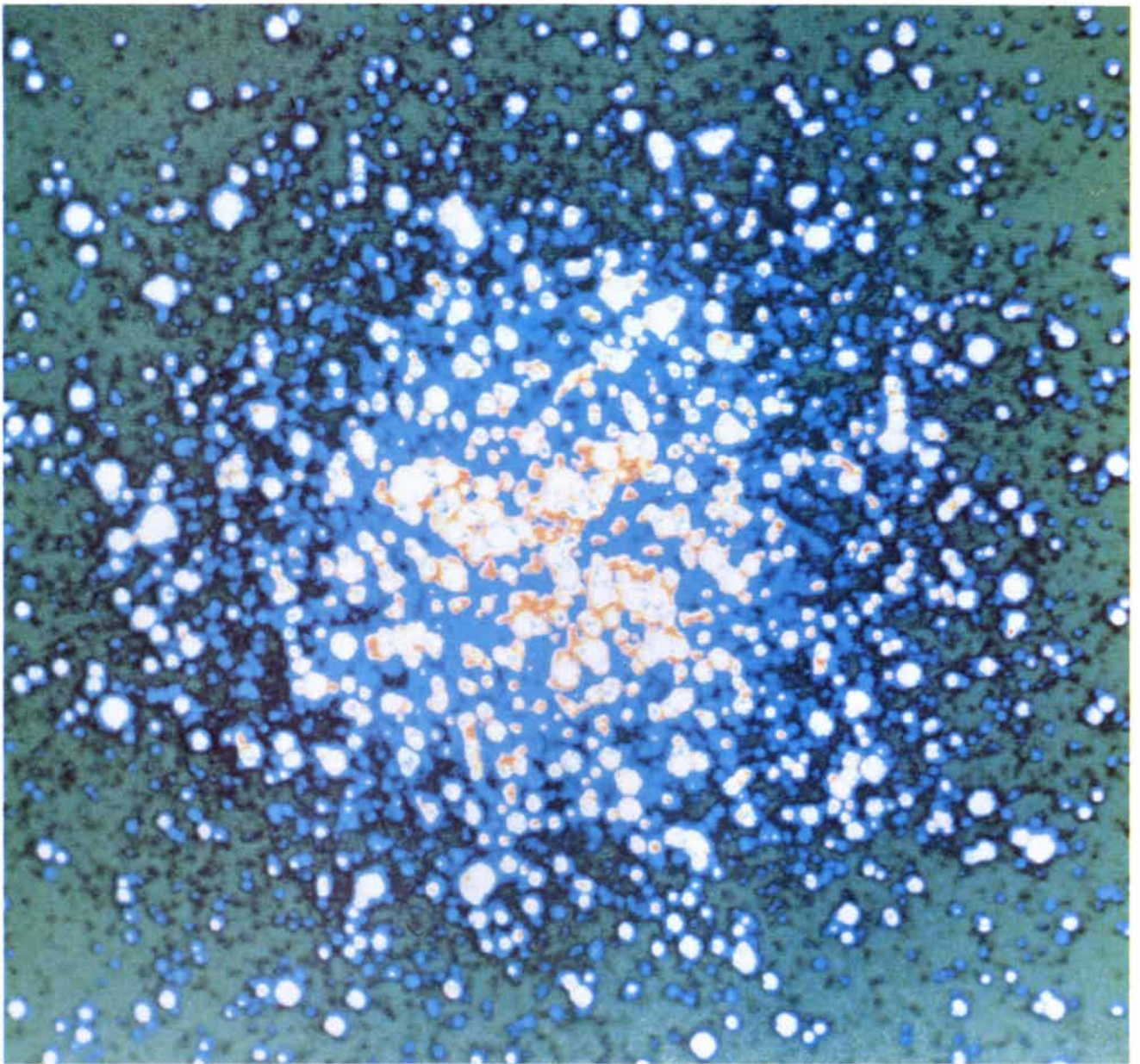
\*Remember: Compare these estimates to the "EPA Estimated MPG" of other cars. You may get different mileage, depending on how fast you drive, weather conditions and trip length.

BUCKLE UP—IT'S A GOOD FEELING!

# THE 1985 FAMILY CAMRY. IT'S DEVELOPED QUITE A ROOMY REPUTATION.



# STAR!



**MESSIER 13** in the constellation Hercules is the brightest globular cluster in the northern sky; on summer evenings it is overhead and just visible to the unaided eye. It contains some 500,000 stars, and the star density at its center is about 20,000 times that of the solar neighborhood. On the false-color image of the central region (*top*) the brightest colors represent the regions of most intense emissions. The image was made with a charge-coupled device (an electronic, silicon-based light detector) attached to the 200-inch telescope on Palomar Mountain. Because a CCD can register a much larger range of brightnesses than a photographic plate, some of the detail in the dense core of M13 can be resolved. The area covered by the photograph (*bottom*) is 15 minutes of arc wide, about three times as wide as that in the CCD image. Stars orbit the cluster center with a period that is on the order of a million years. The orbit of a typical star (*blue line*) lies nearly in a plane, but the star does not return to its original position.



# Globular Clusters

*They are dense crowds of ancient stars bound together by their own gravitation. For decades the study of clusters has yielded insights into the evolution of stars, of galaxies and of the universe as a whole*

by Ivan R. King

To a person looking through a large telescope a globular cluster is one of the most beautiful objects in the sky. Stars fill the field of view by the thousands; as many as a million, most of them too faint to be visible, may be packed into a spherical space whose diameter is typically less than 150 light-years. For decades astronomers have pondered how such crowds of stars may have been formed and how the interaction of their gravitational fields holds them together in a stable cluster.

Throughout the 20th century, moreover, the study of globular clusters has led to fundamental advances in many branches of astronomy. In part because the clusters are so luminous their spatial distribution has helped investigators to stake out the frontiers of the Milky Way and of other galaxies; it has even been suggested that the clouds of gas that engendered globular clusters were the building blocks from which galaxies were made. Furthermore, all stars in a given cluster can be assumed to be the same age, and so the types of stars in clusters offer general insight into how stars evolve and why they differ in color and brightness. Finally, globular clusters bear on the evolution of the universe itself. They are the oldest objects known, dating perhaps from just after the big bang; as a result their ages impose severe observational constraints on cosmological models, and their chemical composition is evidence of the composition of galaxies at the earliest stage of development.

At a time when each of these subjects—galactic structure, stellar evolution and classification, and cosmology—has become the concern of a quasi-independent discipline, the study of globular clusters still conveys, in addition to its intrinsic interest, a sense of the underlying unity of astronomy. Indeed, if astronomers could answer all the questions about globular clusters that continue to perplex them,

they would know a great deal more than they do now about the nature of the universe.

## Galactic Structure

At the turn of the century our stellar system was thought to consist of a disk only a couple of thousand parsecs in diameter centered on the sun. (One parsec is 3.26 light-years.) This heliocentric conception came into serious question in 1918, when Harlow Shapley used the telescopes of the Mount Wilson Observatory to measure the distance to several dozen globular clusters. He found that they make up an extended system centered behind the brightest star clouds of the Milky Way, in the constellation Sagittarius, where the clusters are commonest. Shapley made what he later described as a “bold and premature assumption”: that the globular clusters constitute a kind of “bony frame” whose centroid lies at the center of the entire stellar system. The sun, he argued, actually lies far from the center, toward one edge of the disk.

Shapley had in effect discovered the Milky Way galaxy. The galactic system defined by the spatial arrangement of the globular clusters was much larger than the apparent “local system,” which was later shown to be an illusion caused by the murkiness of space. Interstellar dust absorbs starlight, making the stars appear more distant than they really are. When interstellar absorption is neglected, astronomical distances are overestimated; the magnitude of the error increases with the actual distance of the object. As a result fainter, faraway stars appear to be distributed much more sparsely in space than bright, nearby stars, producing the illusion of a systematic fall-off in star density in all directions from the earth. It was this illusion that buttressed the heliocentric conception.

Ironically, Shapley himself did not

take interstellar absorption into account; but he had the good fortune to be observing clusters well outside the absorbing layer of dust, which is largely confined to the thin, flat disk of the galaxy. Nevertheless, his neglect of absorption led him to overestimate greatly the cluster distances and hence the size of the galactic disk, to which he assigned a radius of 50,000 parsecs. The error was corrected only in 1930, when Robert J. Trumpler of the Lick Observatory showed that interstellar absorption is a general phenomenon. Today the radius of the galactic disk is put at about 15,000 parsecs.

Shapley’s “premature assumption” that the centroid of the globular-cluster distribution defines the center of the galaxy is now accepted as fact. A strong source of radio emissions in Sagittarius clearly marks the direction to the center, but an accurate value for its distance from the sun has been elusive. Shapley’s basic approach of plotting globular-cluster positions has been applied repeatedly. A comprehensive survey in 1976 by William E. Harris of McMaster University in Ontario yielded a value of 8,500 parsecs. More recently, however, Carlos Frenk of the University of Sussex and Simon D. White of the University of Arizona have argued that globular-cluster distances continue to be overestimated and that the center of the galaxy is actually only about 6,800 parsecs away. The correct figure probably lies somewhere between these two values. Although other methods of finding the center have been used, with similar results, globular clusters still seem to offer the best hope of settling the issue.

The “bony frame” made of globular clusters is also a good tracer of the outline of the Milky Way. The reason has to do with the distinctive stellar content of the clusters, which Shapley discovered when he measured the color and magnitude of individual cluster stars. He noted that the distribution of

the stars on a Hertzsprung-Russell diagram (on which the vertical axis represents magnitude, or luminosity, and the horizontal axis represents color) is quite different from that of ordinary stars of the solar neighborhood. For example, the brightest nearby stars are blue, whereas those in globular clusters are red.

### Stellar Populations

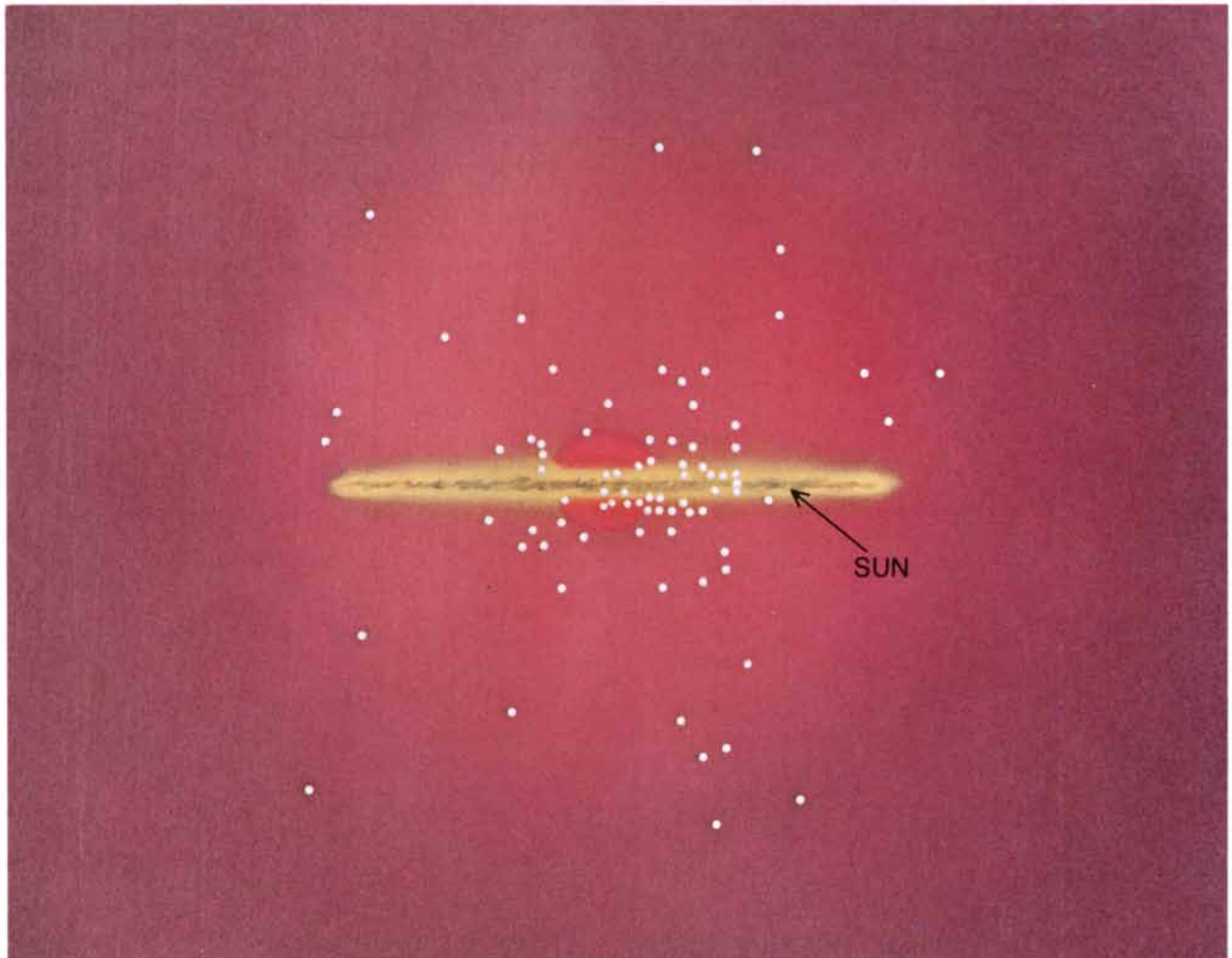
No one made much of this curiosity until 1944. Then Walter Baade of Mount Wilson observed that the brightest stars in the central region of the Andromeda galaxy are also red giants. He thereupon proposed, in a great imaginative leap, that all stars are divided into two fundamentally different "populations." Population I consists of the solar-neighborhood stars and in general of the stars found

in the disk of the Milky Way and of other galaxies. Population II stars are scattered throughout an almost spherical "halo" surrounding the disk, although like the disk stars their concentration is greatest toward the center of the galaxy. Globular clusters are simply luminous, readily observable aggregations of halo stars. Thus their distribution not only points the way to the galactic center but also traces the extent of the halo; it is now thought the Milky Way halo may reach as much as 100,000 parsecs from the center.

When Baade conceived the notion of stellar populations, the physical basis of the observed differences in their color-magnitude distributions was not immediately obvious. It began to become clearer after World War II, as the technique of photoelectric photometry came into wide scientific use. Photoelectric studies of star clusters

produced much more accurate color-magnitude diagrams. A Hertzsprung-Russell diagram of a cluster, whose stars were presumably all formed at roughly the same time and place, in effect shows the track of stellar evolution: the stars are spread out on the upper part of the diagram because the brighter and more massive ones evolve faster. Most of a star's life is spent on the "main sequence" of the diagram, during which time it radiates energy by fusing hydrogen into helium in its core. When the supply of hydrogen in the core is exhausted, the star "turns off" the main sequence, continues to burn hydrogen in a thin shell around the core and evolves into a red giant.

The position of the turnoff point is an index of the cluster's age: the brightest stars turn off first, and as the cluster ages, the turnoff point moves down the main sequence into regions of lower



**GLOBULAR CLUSTERS OF THE MILKY WAY** have helped to reveal the structure of the galaxy. Some 125 clusters (*white dots*) are known; those whose positions are known are shown here. Dust may obscure our view of many more on the far side of the galactic plane from the sun. The cluster distribution defines the center of the galaxy. Globular clusters are aggregations of Population II stars:

stars belonging to the galactic halo (*red*), which is thought to have formed as the protogalactic gas cloud collapsed not long after the big bang. The young stars of the spiral-armed disk (*yellow*), where gas is still present and star formation continues, belong to Population I. The radius of the disk is roughly 15,000 parsecs; the halo may extend as far as 100,000 parsecs from the center of the galaxy.

luminosity. In the early 1950's a number of workers estimated the ages of both globular clusters and "open" clusters, which are much less dense aggregations of Population I stars in the galactic disk. These studies revealed the first physical difference between the two stellar populations. Individual open clusters turned out to span a range of ages, but the globular clusters were all older—indeed, they seemed to be the oldest objects in the universe.

Further research showed that age is not the only peculiarity of Population II stars: they are also different in chemical composition. In the mid-1950's Joseph W. Chamberlain, then at the University of Chicago, and Lawrence Aller, then at the University of Michigan, observed that certain dark absorption lines in the spectra of halo stars are weaker than the comparable lines of Population I stars. The weaker lines indicated lower abundances in the halo population of the chemical elements that absorb radiation at those particular frequencies.

The elements in question turned out to be the "heavy" elements: all elements except hydrogen and helium. In calculating the first stellar-evolution tracks for globular clusters, Fred Hoyle, then at the University of Cambridge, and Martin Schwarzschild of Princeton University found that the observed red-giant tracks could be explained only by assuming a deficiency of heavy elements. In astronomical shorthand such elements are called metals, even though the most abundant ones are carbon, oxygen and nitrogen. Hence globular clusters, in addition to being old, were identified as "metal-poor."

Most astronomers now believe globular clusters, and halo stars in general, are metal-poor precisely because they are old. It is generally accepted that the big bang with which the universe began created only hydrogen and helium. The heavy nuclei are thought to have been synthesized at a later time inside stars, where the prevailing temperature and pressure are high enough, and then ejected into space by supernova explosions; the heaviest elements may have been formed during the explosions themselves. According to this scenario, the globular clusters and the halo stars formed early on, while the abundance of heavy elements was still low throughout the universe. The protogalactic gas clouds then collapsed to form the disks of the Milky Way and of other similar galaxies. By that time nucleosynthesis in dying stars of the halo population had raised the heavy-element abundance of the gas to its present level. The Milky Way halo is probably a good indicator of the orig-

inal size of our galaxy at the time stars began to form; today star formation continues only in the gas clouds of the thin disk.

### Population Complications

The foregoing explanation of stellar population differences is appealingly simple, but unfortunately the truth is a bit more confusing. Globular clusters are not all alike. To begin with, although they are all metal-poor, the specific abundance of heavy elements varies from cluster to cluster. This was first noted nearly 30 years ago by William W. Morgan of the Yerkes Observatory, who observed differences among clusters in the strength of their spectral lines. Since then investigators have made quantitative estimates of heavy-element abundances by analyzing in detail the spectra of individual stars. Such analyses have shown that metal concentrations in globular-cluster stars range from about one two-hundredth of the levels observed in the sun (a typical Population I star) to only slightly less than solar values. The precise upper limit is still uncertain, primarily because the spectra of individual cluster stars are quite faint.

Certain stars ordinarily assigned to the halo population even seem to have metal abundances equivalent to those of the sun. The RR Lyrae stars are variable stars (ones whose brightness changes periodically) that are common in globular clusters and throughout the Milky Way halo, and some of them have heavy-element spectral lines as strong as the corresponding solar ones. If there were indeed a continuum of metal abundances in halo stars extending right up to the levels characteristic of disk stars, then the sharp distinction between the two populations would be undermined. Two circumstances, however, suggest that the strong-lined RR Lyraes are not true halo stars. First, no strong-lined RR Lyrae has ever been found in a globular cluster. Second, their orbits around the galactic center are closer to those of disk stars than to those of halo stars. In spite of their striking visual resemblance to their brothers in the globular clusters, the strong-lined RR Lyrae stars may belong to a separate stellar class, perhaps even to a population intermediate between the disk (I) and halo (II) populations.

In any case globular clusters do exhibit a range of metal abundances; might this be evidence of a range of ages? Although it seems very likely that the globular clusters of the Milky Way halo all predate the more numerous stars of the disk, it is not clear just how old they are, and whether the pe-

riod of their formation spanned an appreciable fraction of the early history of the galaxy. The method of estimating a cluster's age has remained essentially unchanged since it was developed in the 1950's: one looks for the age that, in conjunction with a theoretical model of stellar evolution, best reproduces the observed distribution of cluster stars on a color-magnitude diagram, particularly at the turnoff point. Don A. Vandenberg of the University of Victoria in British Columbia has recently computed an impressively precise set of evolutionary tracks for a number of globular clusters. He has concluded that the clusters are all approximately 16 billion years old, but even these calculations contain an uncertainty of about three billion years.

Some astronomers contend that, whatever the age of the Milky Way globular clusters, they must all be equally old, because dissipation of energy in the star-forming gas cloud would have prevented it from maintaining the spherical shape of the halo for a long period. According to this argument, the spinning cloud would have collapsed quickly into the thin disk, leaving behind the globular clusters and the other halo stars. The range of metal abundances could be attributed to the fact that different globular clusters formed in different local regions of the gas, whose chemical composition was not uniform.

Yet there is evidence that globular clusters differ from one another in respects other than their heavy-element content: clusters with the same metal abundances often have noticeably different Hertzsprung-Russell diagrams. For example, the "horizontal branch," which follows the red-giant stage on the evolutionary sequence, may contain blue stars or red stars or both. There must be a second variable parameter that accounts for these differences, and some workers maintain it is age. The differences might also be explained by a variation among clusters in helium content, in the detailed proportions of the individual heavy elements or in the rate at which the stars spin on their axes. The nature of the second parameter is still a mystery.

### Cosmology

Even more of a mystery is how globular clusters acquired any heavy elements at all, given that the big bang is thought to have produced only hydrogen and helium. The observed metal abundances, while quite low compared with those of Population I stars, are not insignificant. Consequently there must have been an earlier generation of stars inside of which the heavy

elements now found in cluster stars were synthesized. So far no traces of a primordial stellar population, in the form of stars older than globular clusters, have ever been detected.

Because the globular clusters of the Milky Way halo are the oldest objects known, their age sets a lower limit on the age of the universe itself. This constraint on cosmological theory is particularly valuable now, at a time when observational cosmology finds itself in a sorry state. Allan R. Sandage of the Mount Wilson and Las Campanas Observatories once described cosmology as "the search for two numbers": the present expansion rate of the universe and the rate at which the expansion is decelerating. (That the universe is in fact expanding is proved by the red shift in the spectra of distant galaxies, which shows that they are receding

from our own galaxy at a velocity proportional to their distance.) Neither number is at all well known from observation. The present expansion rate, usually referred to as the Hubble constant, is in principle the easier of the two to calculate, and yet a dispute over its value has divided cosmologists into rival camps. One camp, led by Marc Aaronson of the University of Arizona and Jeremy R. Mould of the California Institute of Technology, puts the Hubble constant between 80 and 100 kilometers per second per million parsecs; Sandage and his colleague Gustav A. Tammann say the correct value is just over half that large, roughly 55 kilometers per second per megaparsec.

The value of the Hubble constant is directly related to the age of the universe, because by extrapolating the expansion rate back into the past one ar-

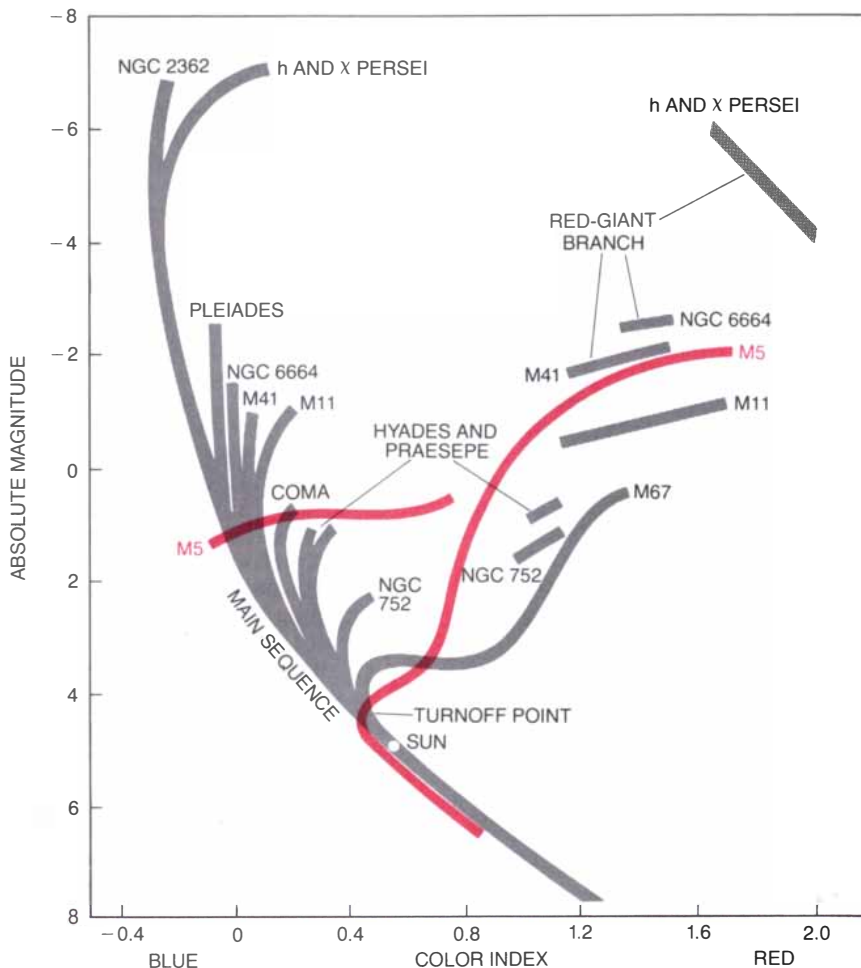
rives at the time the expansion began. It is virtually certain that the mutual gravitational attraction of galaxies has slowed the expansion, which must have been even faster in the past, and so the age extrapolated from the Hubble constant is an upper limit. The higher of the two proposed values implies that the universe can be no more than 10 to 11 billion years old; it thus conflicts sharply with the best globular-cluster data, which indicate cluster ages of at least 13 billion years.

If Sandage and Tammann's lower value for the Hubble constant is correct, the big bang could have taken place as much as 20 billion years ago. Many cosmologists now believe, however, as a result of recent developments in particle physics, that the rate at which the expansion is decelerating—the second cosmological number—is large [see "The Inflationary Universe," by Alan H. Guth and Paul J. Steinhardt; *SCIENTIFIC AMERICAN*, May, 1984]. In that case even the lower value for the Hubble constant implies that the universe is only 12 to 13 billion years old. This is still uncomfortably low considering the globular-cluster ages. It is not clear how the conflict will be resolved.

### Galaxy Formation

Globular clusters do not bear only on the date of the big bang; they may also offer clues to how the galaxies formed. Soon after the primordial fireball the uniform, diffuse mass of hydrogen and helium began to fragment into vast clouds. The size of the clouds must have been determined by a balance between gravity, which tended to pull the gas together, and heat, which tended to disperse it. P. J. E. Peebles and Robert H. Dicke of Princeton have suggested that the pregalactic clouds, which formed in great numbers, are most likely to have been the size of globular clusters. The clouds drifted together under their mutual gravitational attraction. Although most of them coalesced into the larger agglomerations that formed galaxies, some of them escaped collision while remaining gravitationally bound to the larger galactic structures. Such clouds, according to Peebles and Dicke, went on to form the globular clusters of the galactic halo.

If this scenario is correct, it is likely that it will have to be modified to explain the formation of galaxies other than the Milky Way. In addition to the 125 or so globular clusters known in our own galaxy, thousands have been identified in each of a number of other galaxies. Most of these distant clusters are quite faint, and the study



**HERTZSPRUNG-RUSSELL DIAGRAMS** show the distribution of stars in a cluster according to color and magnitude; they provided the first evidence for the existence of two different stellar populations. This schematic composite illustration contrasts the H-R diagrams of open clusters (gray) in the galactic disk with the diagram of the globular cluster M5 (color), which belongs to the halo. In an individual cluster all the stars are the same age, and so the H-R diagram reflects the path of stellar evolution. Massive, bright stars (those with negative magnitudes) evolve the fastest; they are the first to exhaust the hydrogen in their core, turn off the main sequence and become red giants. As the cluster ages, the turnoff point moves down the magnitude scale. Open clusters (Population I) vary in age. One of the youngest, NGC 2362, was formed roughly a million years ago. Globular clusters (Population II), including M5, are all thought to have been formed at least 13 billion years ago.

In the 25 years since the birth of the world's first laser at Hughes Research Laboratories, the "light fantastic" has grown from a laboratory curiosity into an indispensable tool in medicine, industry, electronics, data processing, communications, and scientific research. That first laser, built by Dr. Theodore H. Maiman, was operated on May 15, 1960. It used a flash lamp coiled around a solid ruby crystal to produce an intense pulse of red light with a wavelength of precisely 6943 angstroms. Lasers today employ various gases or crystals and operate throughout the electromagnetic spectrum. They are used as tools for cutting, welding, drilling, and marking metals; as alignment and measuring devices; as the sources of signals in fiber-optic communications systems; and as rangefinders and target illuminators in military systems. Promising new medical uses include advanced eye surgery techniques, internal cauterization, and treatment of cancer. Already used in some computer printers, lasers one day will be widely used in high-speed optical computers to process and store data.

An advanced computer system for air traffic control is being designed to serve the U.S. into the 21st century. The new Advanced Automated System (AAS) will consolidate existing en route facilities and approximately 130 terminal facilities into 23 area control facilities throughout the country. It will automate many routine air traffic control activities now done manually. Computers will monitor and evaluate air traffic situations and offer solutions to potential conflicts between airplanes in flight. AAS will include controller consoles to display radar data, weather information, and flight plan data; powerful modern computers; and new software to run the new system. Hughes is designing AAS for the Federal Aviation Administration under a competitive contract. Hughes has built air defense systems for more than 20 nations, including the U.S., Canada, and NATO countries.

The first full-scale development AMRAAM missile was fired successfully at the White Sands Missile Range in New Mexico. The missile was launched from a U.S. Air Force F-16 at 40,000 feet at a speed of Mach 1.2. It flew a preprogrammed course designed to evaluate the missile's control system and separation from the launch aircraft. It did not have a seeker but instead was programmed through its autopilot to fly a prescribed route. The Advanced Medium-Range Air-to-Air Missile is in full-scale development at Hughes for the U.S. Air Force and Navy.

An advanced factory management system model, developed by Computer Aided Manufacturing-International and Hughes, will help optimize use of manufacturing resources. The model will address interactions of all work areas within every level of the organization. It will precisely identify department production capacities, queue bottlenecks, and resource flow. Managers now must make decisions without knowing all interactions among workstations, cells, and departments.

Hughes Research Laboratories needs scientists for a spectrum of long-term sophisticated programs, including: applications of focused ion beams; electron beam circuit testing; liquid-crystal materials and displays; nonlinear optics and phase conjugation; submicron microelectronics; plasma applications; computer architectures for image and signal processors; gallium-arsenide device and integrated circuit technology; optoelectronic devices; and growth, characterization, and process technology development for new electronics materials for high-speed, infrared detection and optoelectronic applications. Send your resume to Professional Staffing, Hughes Research Laboratories, Dept. S2, 3011 Malibu Canyon Road, Malibu, CA 90265. Equal opportunity employer. U.S. citizenship required.

For more information write to: P.O. Box 11205, Dept. 69-9, Marina del Rey, CA 90295



of them has begun in earnest only recently. Nevertheless, some differences between galaxies have already emerged. For example, in the Milky Way all globular clusters are old, and the young open clusters of the galactic disk are much poorer in number of stars; the Clouds of Magellan, in contrast, contain young, star-rich aggregations that much resemble globular clusters. It is not known why the Magellanic clouds, our nearest galactic neighbors, should still be making rich clusters when the Milky Way is not.

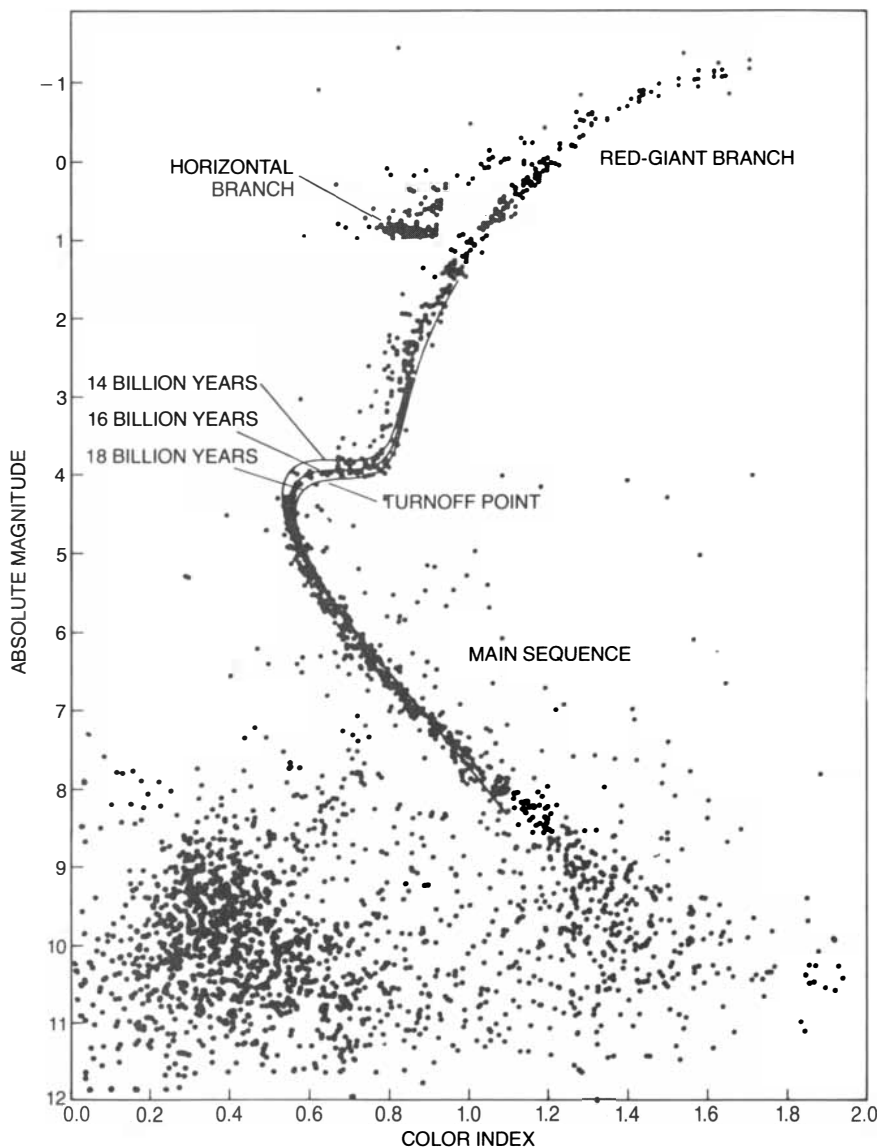
Nor is it understood why elliptical galaxies seem to have many more globular clusters per unit of mass than spiral galaxies. The observation is of particular significance because it argues against a popular theory of how the ellipticals formed. Alar Toomre of the Massachusetts Institute of Technology and other investigators have proposed that elliptical galaxies are formed when spiral galaxies collide and merge. The strongest evidence against this hypothesis is the higher proportion of clusters in the ellipticals.

Every bit as fascinating as the implications of globular clusters for the structure and formation of galaxies is the structure of the clusters themselves. How does the interaction of thousands of stars produce an overall form of such simplicity and regularity? Each star is held in the cluster by the joint gravitational attraction of all the others; it loops inward and outward in a regular rose-shaped orbit whose period is on the order of a million years. On the average at a given moment half of the stars are moving outward and half are moving inward. Their velocities are just large enough to balance the gravitation that would otherwise draw them into the center. More precisely, there is an exact correspondence between the distribution of stellar velocities and the radial distribution of stars, which determines the cluster's density profile and thereby its gravitational field.

In principle many different pairings of these two quantities are possible, but the structural similarity of most globular clusters suggests that certain velocity and density distributions are favored. The favored distributions arise from the nature of stellar interactions in a cluster. Although the motion of each star is governed almost completely by the rather smooth gravitational field of all its cohorts, on rare occasions two stars pass close enough to each other to affect each other's motion individually. The exchange of energy arising out of such random stellar encounters tends to produce what is called a Maxwellian distribution of velocities, after the Scottish physicist James Clerk Maxwell, who derived a statistical formula to describe the motions of molecules in a gas.

A globular cluster cannot achieve a full Maxwellian distribution, which would include objects of all velocities, because the cluster has a finite escape velocity; stars accelerated to a higher speed by stellar encounters acquire enough energy to escape the cluster's gravitational field. Below this cutoff, however, the distribution of stellar velocities in a cluster closely approximates Maxwell's formula. The velocity distribution in turn determines the radial density profile.

Of course globular clusters are not all structurally identical. Two decades ago I studied many of them and found that their structural differences could be adequately described by three parameters: the radius of the central core, the outer radius and the number of stars in the cluster. The most important difference among clusters is in the core radius, which is defined as the radius at which the density of stars on the cluster's image has fallen to half of its



**AGE OF A GLOBULAR CLUSTER** can be estimated by comparing a theoretical model of the evolution of stars in the cluster with their observed distribution on an H-R diagram. The distribution calculated from an evolutionary model, particularly at the turnoff point, depends sensitively on the cluster age assumed in the calculations. The H-R diagram shown here is for the giant globular cluster 47 Tucanae. Most of the data were obtained by James E. Hesser of the Dominion Astrophysical Observatory and William E. Harris of McMaster University with the four-meter telescope at the Cerro Tololo Inter-American Observatory in Chile. The theoretical models (black lines) were computed by Don A. Vandenberg of the University of Victoria. The "best fit" at the turnoff point is offered by the model corresponding to a cluster age of 16 billion years. Most of the stars in the lower left corner of the diagram actually belong to the Small Magellanic Cloud, a galaxy that lies behind 47 Tucanae.

# TANDY... Clearly Superior™

The Ultra-High Performance Tandy 2000 runs the best programs better and faster.

Ordinary personal computers—like IBM's PC—are slow and dull compared to the Tandy 2000.

## Unmatched Power, Graphics and Software

In fact, in actual benchmark comparisons, the Tandy 2000 performed almost three times faster than the IBM PC\*.

The key to the Tandy 2000's speed and power is its more advanced 80186 16-bit microprocessor, more detailed

graphics and industry-standard MS-DOS operating system. Together in the Tandy 2000, they let you use the most popular software—and finish in a hurry. Lotus 1-2-3, Symphony, Framework, dBase II and III, SuperCalc<sup>3</sup> . . . whatever your business demands, there's probably a Tandy 2000 program to suit your needs.

## Get the Tandy 2000 for Less than the IBM PC

Best of all, the Tandy 2000 costs less than the IBM PC.

A Tandy 2000 with two 720K floppy disk drives is just \$2499 (26-5103). For performance comparable to the IBM PC AT, but at a better value, get our Tandy 2000 with a 10-megabyte hard disk for just \$3950 (26-5104).



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

**Radio Shack**  
COMPUTER CENTERS

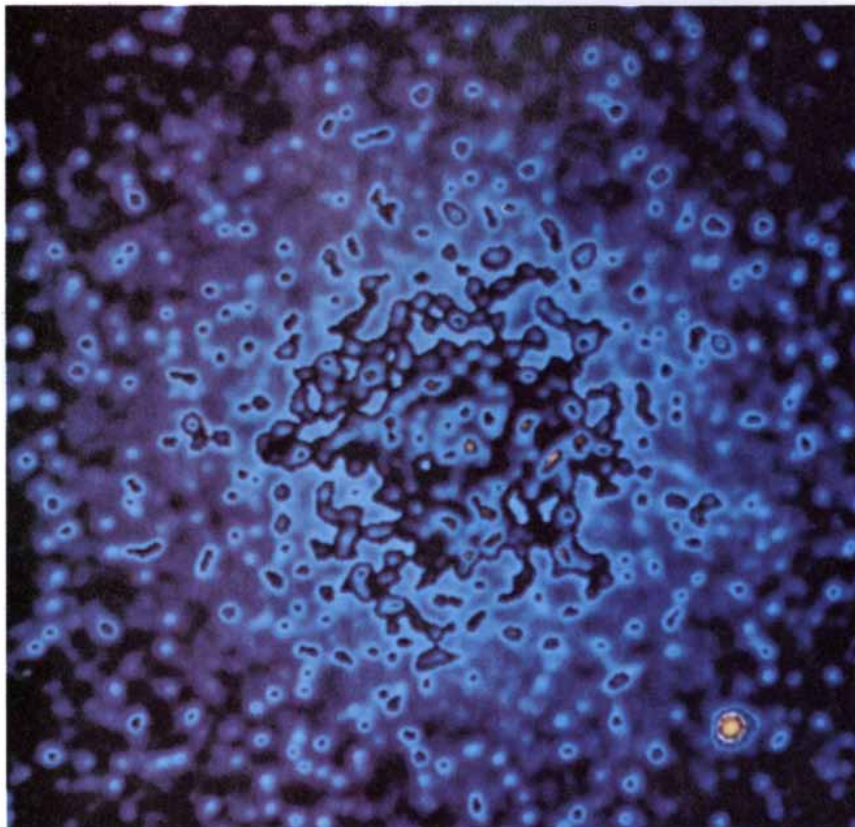
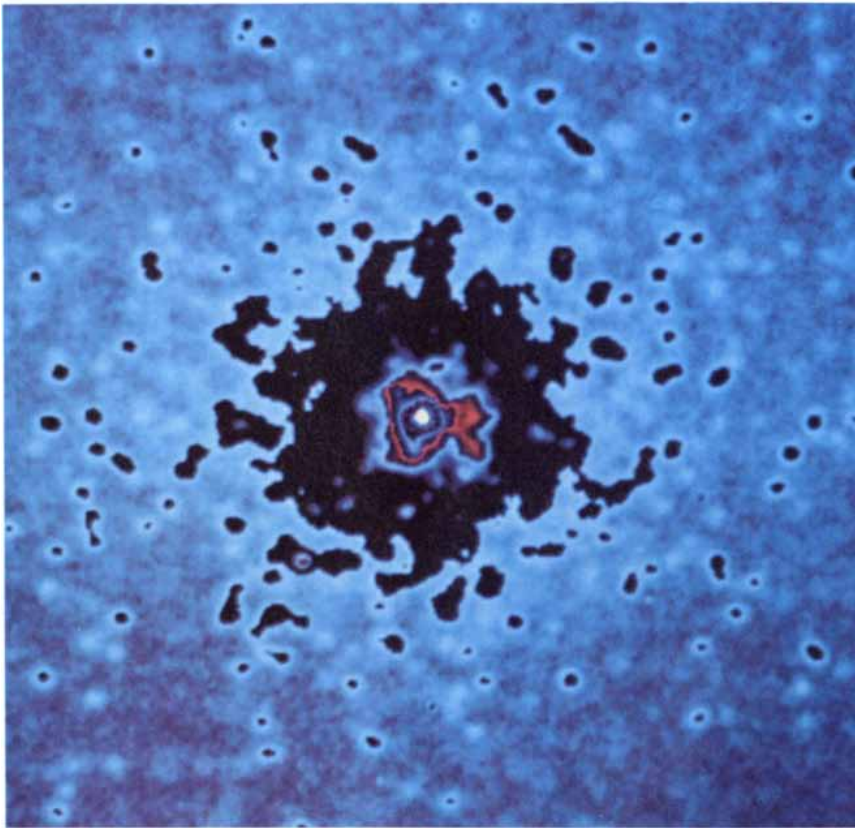
A DIVISION OF TANDY CORPORATION

### Send Me a 1985 Computer Catalog.

Mail To: Radio Shack, Dept. 85-A-567  
300 One Tandy Center, Fort Worth, Texas 76102

Name \_\_\_\_\_  
Company \_\_\_\_\_  
Address \_\_\_\_\_  
City \_\_\_\_\_  
State \_\_\_\_\_ Zip \_\_\_\_\_  
Telephone \_\_\_\_\_

\*80 Micro Magazine, May, 1984. Prices apply at Radio Shack Computer Centers and at participating Radio Shack stores and dealers. IBM/TM International Business Machines Corp. Lotus 1-2-3 and Symphony/TM Lotus Development Corp. dBase II, dBase III and Framework/TM Ashton-Tate. SuperCalc/TM Sorcim Corp.



**CORE COLLAPSE** produces a sharp peak in star density at the center of a globular cluster and a corresponding peak in brightness. Such a peak is visible on the computer-generated false-color map of the brightness distribution in the cluster NGC 6624 (*top*). No evidence of core collapse is discernible on the map of 47 Tucanae (*bottom*), although the brightness of that cluster also increases toward its center. The author and his colleague Stanislav Djorgovski have found at least half a dozen globular clusters that have undergone core collapse.

value at the center. Some clusters have smaller, denser cores than others; these clusters are more tightly bound, and their escape velocities are correspondingly higher. The escape velocity is just the velocity a star must have to reach the outer radius of the cluster. Unlike the boundary of the core, the outer radius is not determined solely by the cluster's gravitational binding energy. Rather, it is primarily a tidal limit defined by the gravitational field of the galaxy, which tends to pull stars out of the cluster.

### Dynamical Evolution

Although globular clusters are obviously long-lived, they are not immutable. Slowly but steadily stars "evaporate" from a cluster as they reach the escape velocity. A theory that adequately predicts the resulting evolution of the cluster is basically simple, but many of the details have remained elusive.

The binding energy of a cluster is really an energy deficit: the energy it would take to accelerate all the stars to their escape velocity and tear the cluster apart. To reach escape velocity a star must acquire enough positive kinetic energy to overcome the cluster's negative gravitational energy. Thus the stars that escape are those with the most kinetic energy, whereas the amount of gravitational energy they contribute to the cluster is no more than average. As a result the evaporation of stars increases the amount of binding energy per star remaining in the cluster, and the cluster contracts.

According to current theory, it does not reach a steady state. Instead, energy from the contraction is converted into the kinetic energy of stellar motion, thereby "heating" the core. More stars evaporate, and the core continues to contract and heat without bound until it is infinitely dense. Donald Lynden-Bell of Cambridge, a proponent of the theory, has dubbed this positive feedback phenomenon the "gravothermal catastrophe."

When the theory was first put forward in 1960 by Michel Hénon of the Nice Observatory, there was little observational evidence to support it. Only one globular cluster, M15, showed any sign of the sharp density peak one would expect in a collapsed core. Recently, however, my colleague Stanislav Djorgovski and I have been making more careful observations of a large number of globular clusters. We have observed at least half a dozen displaying central density peaks we believe to be evidence of core collapse. Still, half a dozen is not very many; theories of cluster evolution predict



# Control yourself.

Introducing the first Volkswagen you can buy for *how your heart feels*—as well as *what your head says*: the new Jetta GLI. \$9,995\*.

It's designed with clean aerodynamic lines. High-performance low-profile tires. Alloy wheels. Sport seats. And a multi-function computer display.

German engineered and built, it has a high-performance 1.8-liter fuel-injected engine. A close-ratio 5-speed transmission. Four-wheel independent sport suspension. Rack-and-pinion steering. And 4-wheel disc brakes.

On the test track, it accelerates 0-50 mph in 7.1 seconds. Top speed: 115 mph. Lateral acceleration: 0.85g.

The new GLI: No dream. No illusion. It's the real thing.

Introducing the new  
Jetta GLI. \$9,995\*.



It's not a car.  
It's a Volkswagen.

For details, call 1-800-85-VOLKS. \*Mfr.'s sugg. retail pricing, excluding tax, title, dealer prep and transportation. The Jetta GLI is covered by the new Volkswagen 2-year Unlimited-mileage Protection Plan: 2-year unlimited-mileage, limited warranty on entire car, except tires; 3-year unlimited-mileage, limited warranty on corrosion perforation. See U.S. dealer for details. [Seatbelts save lives.](#)



that a much larger fraction of the ancient globular clusters in the Milky Way halo should have collapsed by now. (The various analytic and numerical models also concur in predicting that once core collapse begins it proceeds so rapidly one would be unlikely to observe it in progress.) Why have more central density peaks not been detected?

One possibility is that the predicted time scales for core collapse are too short. A more likely explanation is that some mechanism halts the collapse and even causes the core to reexpand to a normal size.

### Binary Stars

Binary stars—pairs of stars gravitationally bound to each other in a close orbit—could serve as such a mechanism. As long ago as 1959 numerical simulations by Sebastian von Hoerner of the National Radio Astronomy Observatory demonstrated that binaries tend to form in star clusters as a result of chance encounters involving three stars. In later simulations by Sverre Aarseth of Cambridge the contraction of the core was almost invariably stopped by the formation of a massive central binary, whose encounters with other stars gave them “kicks,” boosting them into higher orbits.

Aarseth’s models were of open clusters containing no more than about 500 stars; Lyman Spitzer, Jr., and Michael Hart of Princeton subsequently showed that three-body encounters are

much less likely to produce binaries in globular clusters, which generally contain 100,000 stars or more. In part the reason is that stellar velocities in globular clusters are much greater. As the core of a globular cluster collapses, however, it effectively detaches itself from the surrounding envelope. Thus eventually it might contain sufficiently few stars to allow binaries to form. In addition Andrew Fabian, James Pringle and Martin J. Rees of Cambridge have found that two-star encounters can also result in the formation of a close binary. Either or both of these mechanisms could operate in a small, dense cluster core and produce the binaries needed to halt its collapse.

By the time the core is small enough it would probably be almost indistinguishable, at current levels of resolution, from a completely collapsed core; it would probably exhibit the kind of central density peak Djorgovski and I have observed. To explain why such peaks have not been detected in more clusters, one must further postulate that the energy contributed by binaries to the stars in a contracted core is enough to reexpand the core. The idea is attractive, but its plausibility has not been fully established.

Unfortunately the centers of globular clusters are too dense to allow any hope of observing binary stars directly with ground-based telescopes. The high-energy X rays emanating from several clusters, however, may be indirect evidence of double stars. At one time it was popular to suggest the

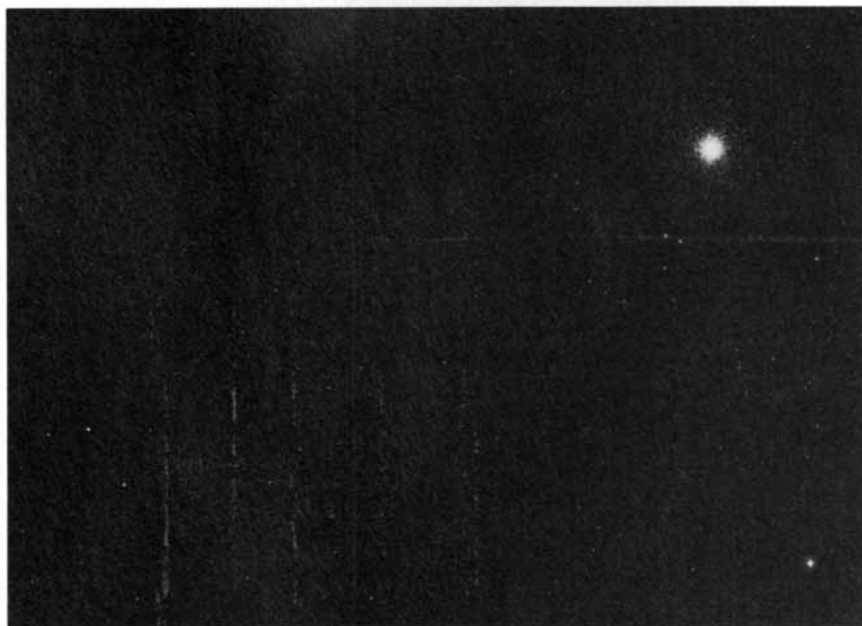
X rays come from material crashing through the tremendous gravitational field of a black hole. Yet one would expect to find such a massive object precisely at the center of a cluster, and recent studies by Jonathan E. Grindlay of Harvard University and his co-workers have shown that the X-ray sources in globular clusters are slightly off-center. It now seems most likely that the sources are close binary systems in which material is sucked from a distended star by the strong field of a neutron star or of a white dwarf.

Nearly all the X-ray sources in globular clusters are situated in dense core regions, and so it is natural to suggest they are the binary stars responsible for halting core collapse. Actually there is no reason to assume that the binary stars stabilizing the core would have to be X-ray binaries; indeed, several of the collapsed-core clusters we have found do not have X-ray sources. Conversely, several of the known X-ray clusters lack collapsed cores. A dense core may simply favor the formation of binaries in general because it promotes stellar encounters.

The crowded centers of globular clusters may yield many of their secrets to the Space Telescope, which is scheduled for launching in 1986. Observations from the ground are limited in their resolution primarily by the unsteadiness of the earth’s atmosphere. The 2.4-meter-diameter orbiting telescope will have a resolving power about 20 times as great as that of the best ground-based instruments.

Today, in the pre-Space-Telescope era, the most rapid observational advances are occurring in the study of the motions of individual cluster stars. With new digital spectrographs it is now possible to measure a star’s motion along the line of sight by means of the Doppler shift in its spectrum. This technique is more accurate than the cruder method of measuring transverse velocities from the tiny displacements of stars on photographs taken decades apart. Doppler measurements are adding a new dimension to knowledge of cluster structure.

Like all endeavors in astronomy, the study of globular clusters has benefited tremendously from such technological improvements. Yet globular-cluster investigations in particular have profited from their position at the intersection of different avenues of research, just as they have been essential to many fundamental developments in astronomy. It is this unique position, more than any particular technology, that allows the student of globular clusters to be sanguine about the future. New insights could come from just about anywhere.



**RANGE OF PROPERTIES** among globular clusters is illustrated by a photograph of two that happen to lie in the same general direction. M53 (*upper right*), in the constellation Coma Berenices, is typical of dense, star-rich clusters that are tightly bound. In contrast, NGC 5053 (*lower left*) is a relatively loosely bound cluster containing far fewer stars.

# Unravel the mysteries of the brain



Produced in conjunction with WNET/New York's eight-part documentary series *The Brain*

## Brain, Mind, and Behavior

In the last several decades, science has mounted a major assault on the mysteries of the brain. Armed with new techniques from nuclear medicine, genetics, immunology, and other fields, researchers are finding fresh answers to some of the central questions of our being.

Share in the excitement of this scientific adventure


*Brain, Mind, and Behavior*, by Floyd E. Bloom, Arlyne Lazerson, and Laura Hofstadter, presents an in-depth picture of the very latest advances in our understanding of the brain—its cellular structure, its chemical signals, and its operations. Exploring such topics as brain rhythms, brain malfunctioning, learning and memory, sensing and movement, emotions, and thinking and consciousness, the book emphasizes the underlying biological basis for complex behavioral phenomena and attempts to demystify some of the lingering mysteries of the brain.

Watch for the rebroadcast of *The Brain*

This fall, the Public Broadcasting Service will rebroadcast the highly acclaimed series *The Brain*. Eight one-hour programs combine advanced graphics, intimate case histories, and personal accounts by

the world's foremost brain scientists to convey the essence of modern brain research.

Be sure to have your copy of *Brain, Mind, and Behavior* for this fall. Order now by completing this coupon and mailing to:

 W. H. FREEMAN AND COMPANY  
4419 West 1980 South  
Salt Lake City, Utah 84104

Please send me \_\_\_ copy(ies) of *Brain, Mind, and Behavior*, (ISBN 1637, 1985, 322 pages, 315 illustrations, clothbound, full color) @ \$23.95 each.

\_\_\_ My check made payable to W. H. Freeman and Company is enclosed.

\_\_\_ VISA \_\_\_ MasterCard Expiration Date \_\_\_ / \_\_\_

Account No. \_\_\_\_\_

Signature \_\_\_\_\_

Name \_\_\_\_\_

Address \_\_\_\_\_

City, State, Zip \_\_\_\_\_

Please include \$1.50 for postage and handling. NY, UT, and CA residents add sales tax.

 0204

# The First Organisms

*The very first systems able to evolve through natural selection are likely to have been made differently from today's organisms, and of different materials. They may have been crystals of clay*

by A. G. Cairns-Smith

A curious similarity underlies the seemingly varied forms of life we see on the earth today: the most central molecular machinery of modern organisms has always been found to be essentially the same. This unity of biochemistry has surely been one of the great discoveries of the past 100 years. And surely too it illuminates evolutionary history. But in an inquiry into the origin of evolution itself I believe the unity of biochemistry is of no direct help.

In this I am at odds with the most generally favored view of how life started. It is generally supposed that before there were organisms of any kind, that is to say before there were systems able to evolve indefinitely under natural selection, there was another kind of evolution, a "chemical evolution," whose effect was to build up a stock of the types of molecules that form the universal "construction kit" from which extant organisms are made: amino acids, sugars and so on.

This view took root in the 1920's when the Russian biochemist A. I. Oparin and the British biologist J. B. S. Haldane introduced the idea of a "primordial soup" of organic molecules existing in the oceans of the pre-vent earth. The soup was imagined as having been formed through geochemical processes and through the action of various energy sources on an atmosphere that was somewhat like Jupiter's, in which unoxidized gases such as methane, ammonia and hydrogen predominated. This was Harold Urey's view of the primordial atmosphere. It was supported by an experiment that Stanley L. Miller, then a research student of Urey's, decided to carry out in the early 1950's. Miller passed sparks ("lightning") through a mixture of the kinds of gases thought to have been present in the primitive atmosphere. Water-soluble organic molecules were found. No less than 15 percent of the original carbon added as methane

turned up in a fairly limited set of small molecules that included four of the 20 amino acids making up proteins. Impressive, one must admit. So was Juan Oro's experiment of the early 1960's showing that cyanide (HCN) molecules could join to form adenine in one step. Miller had shown that cyanide was produced in his sparking experiments. Another small molecule, formaldehyde (CH<sub>2</sub>O), was also produced, and it had been known for 100 years that formaldehyde molecules also tend to join and form into sugars such as ribose, a constituent of RNA.

It seemed only a matter of time before an adequate "beginner's construction kit" would be built up. If skeptics were to say the early atmosphere of the earth was probably not Jupiter-like (a rather general view now), that seemed to be no great worry since experiments on other simulated atmospheres with other energy sources often produced similar mixtures of amino acids.

Yet the initial promise has not been maintained. Miller's experiment has hardly been improved on. Even the simpler molecules are produced only in small amounts in realistic experiments simulating possible primitive earth conditions. What is worse, these molecules are generally minor constituents of tars: it remains problematical how they could have been separated and purified through geochemical processes whose normal effects are to make organic mixtures more and more of a jumble. With somewhat more complex molecules these difficulties rapidly increase. In particular a purely geochemical origin of nucleotides (the subunits of DNA and RNA) presents great difficulties. In any case, nucleotides have not yet been produced in realistic experiments of the kind Miller did.

In spite of all of this does it not remain common sense that the construction kit must have come first? And are

there not still these two incontrovertible facts in support?

1. The most central molecules of life are the same in all organisms on the earth today.

2. At least some of these molecules can be made under conditions that might well have existed on the primitive earth.

I think these statements are Red Herrings, and are all the worse for being a pair.

Red Herring 1 reveals itself when you analyze the unity of biochemistry—when you remember that what is common to all organisms now is much more than a construction kit of small molecules. There is also a system, an entire design approach, that is always the same. The machinery in this system is exceedingly complex. Even one protein molecule is a complex object—a specific arrangement of thousands of atoms—and hundreds of well-made protein molecules are needed. One of the places where proteins are most needed (and where they have to be most competent) is in the machinery for making proteins. This is a typical convolution, just one example of another aspect of the unity of biochemistry: the critical interdependence of all the components of the central machinery. Finally, there is an arbitrariness about some features of the central machinery. For example, the code for converting RNA messages into protein sequences is nearly the same everywhere, and the set of amino acids is always exactly the same. It is hard to believe there could be only one code or one amino acid set that would work, or that would be best for every kind of organism under all circumstances.

Surely the proper conclusions to be drawn from such a detailed look at the unity of biochemistry are that (1) all life now on the earth is descended from a common ancestor, (2) this ancestor was quite high up the evolutionary tree and (3) the central biochemical system

was already fixed by that time. That it should have remained fixed for so long is surely because of its curious interdependent kind of complexity. This is the complexity of "high tech" engineering where many well-chosen components depend so much on each other that they cannot, any of them, be changed. That kind of cleverness could only have been a product of evolution. It is at least on the cards that the choice of the components that became fixed was also a product of evolution. To conclude, the unity of biochemistry does not refer to the start of evolution but to a much later stage.

Red Herring 2 can be paraphrased: "Some of our central biochemicals are easily made, full stop." This is what the post-Miller work has amounted to: there is a subset of biochemicals, particularly the simpler amino acids, that are easy to make under all conditions, not just those that might have been peculiar to the primitive earth. The full stop spoils any simple historical interpretation, because you would expect that an extended evolutionary process would also have settled on components some of which were not too diffi-

cult to put together and were reasonably stable. As with Red Herring 1, there is no particular reference to the base of the evolutionary tree.

Of course it might still be that the first organisms were made from molecules similar to those in organisms today, but this should be seen as an assumption with no special warrant.

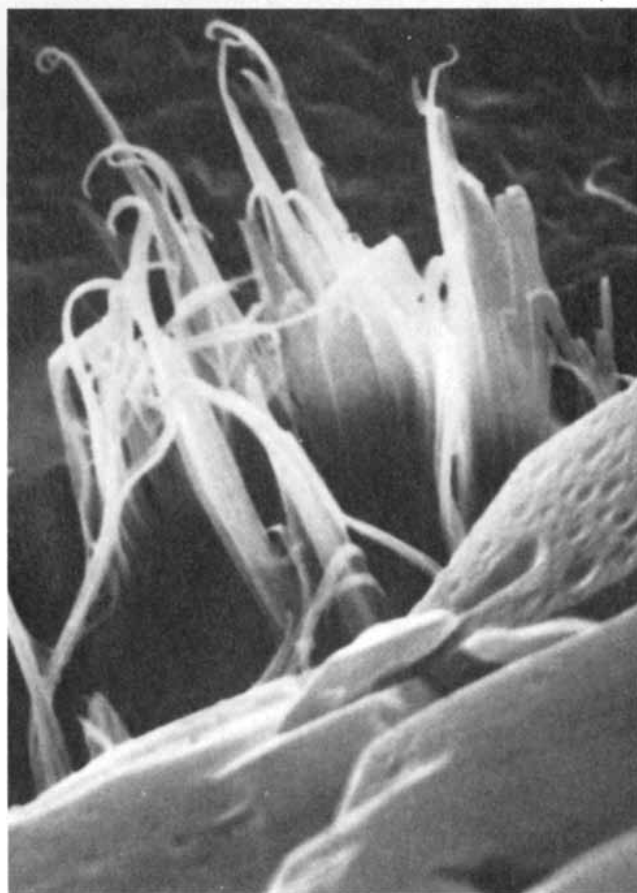
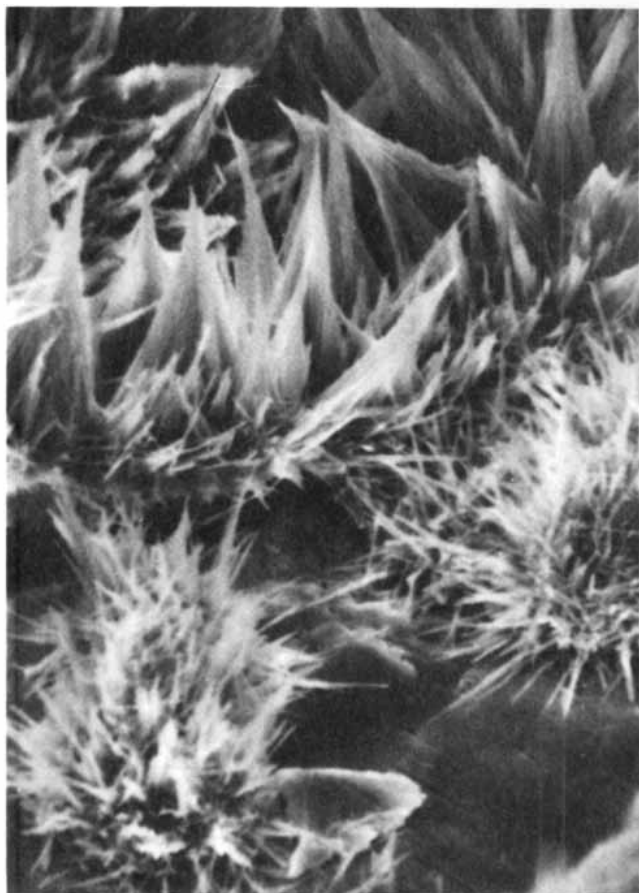
There are indeed good reasons for doubt. They arise from that high-tech interdependent complexity of central biochemistry. The first organisms could not have been like this. They must have been "low tech" machines of the kind that are fairly easily put together and for which there are simple versions that work, more or less (spears, not machine guns). There is a difference in approach here that might lead you to suspect the first organisms would have been made differently from today's, and with different materials. Certainly it is true of human artifacts that high tech and low tech almost invariably call for different kinds of components and materials. Search in vain for one wood abacus bead in a pocket calculator. (Search in vain for wood.)

We should doubt, then, whether amino acids, which are so good for making catalysts (given the technology), would have been good for catalysts to begin with. We should doubt whether amino acids or any other of the now critical biochemicals would have been at all useful right at the start.

Indeed, contemporary organisms are full of high tech at all levels. The eye is the classic example of the precision-built multicomponent machine that has to be just so to be of any use at all. "How could such a thing evolve in small steps?" ask the anti-Darwinians, sure that they are onto something. The anti-Darwinians can stand down: there need be no paradox. High tech can appear through gradual evolutionary processes.

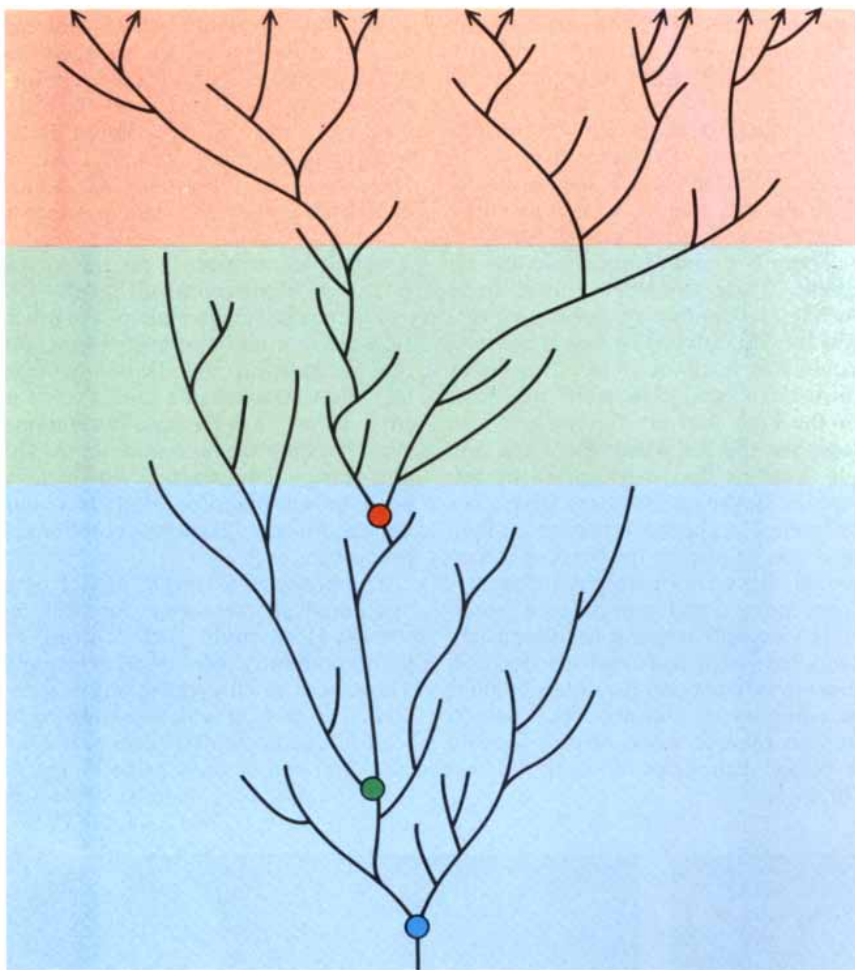
Think about a simple model of a "paradoxical" structure: an arch of stones. How could such a thing be made gradually, one stone at a time? The answer is with scaffolding of some kind. To start off with there has to be scaffolding that is itself "nonparadoxical," that can be built piece by piece.

I think this must have been the way

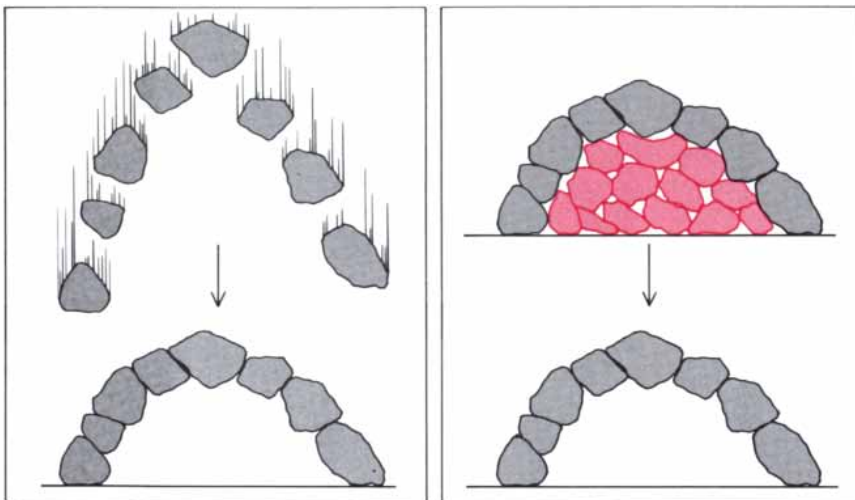


**CLAY CRYSTALLIZES** out of dilute solutions formed as water percolates through weathering rocks. Halloysite clay crystals that grew in water seeping from cracks in granite are enlarged 3,750 diameters in a scanning electron micrograph made by W. D. Keller

of the University of Missouri at Columbia (*left*). Crystals of illite (*right*) that grew in the pores of a sandstone are enlarged 16,000 diameters in a micrograph by David W. Houseknecht of Missouri. Such processes may have had a critical role in the origin of life.



**IMAGINARY EVOLUTIONARY TREE** was generated by a procedure that introduced branchings and extinctions at random. The pattern is typical of trees thus produced in that extant species (*arrows at top*) are all related to an ancestral branch point some distance from the base. (As the tree grows, these universal branch points may change, but only to higher levels.) Real evolution proceeds through branchings and extinctions, and so it is not surprising that all organisms now on the earth seem to have a rather highly evolved common ancestor. All organisms now have similar, sophisticated molecular machinery, but that fact does not imply that the same machinery was characteristic of the first organisms.



**ORIGIN OF A SYSTEM** whose parts cooperate (such as an arch of stones) might well be assumed to have depended on fortuitous, extraordinary circumstances (*left*). It may not be apparent that such a system is much more likely to have been built successfully on a scaffold, which is not seen because it was later removed (*right*). Inorganic clays may have provided the scaffolding within which today's molecular machinery subsequently evolved.

our amazingly "arched" biochemistry was built in the first place. The parts that now lean together surely used to lean on something else—something low tech. Perhaps there are still some pieces of that earlier scaffolding in our present biochemistry, but the scaffolding as such is gone.

What is the way forward? What can there be to say about something that is missing?

Ask another question: How would you try to decide what kinds of weapons a primitive people might have used, supposing there were no material traces known of their activities? You would not try to think of the nearest thing to a machine gun that might be made out of sticks and stones. You would simply try to think of the easiest way in which a primitive people might have been able to make weapons of some kind. You would be guided by what you know of the requirements, of the level of technological development and of the materials available.

Thinking along these lines, we can say of the first organisms:

1. They could evolve.
2. They were low tech.
3. They were made of geochemicals.

These I take to be Green Herrings, statements that not only are likely to be true but also are worth following.

Certainly Green Herring 1 is pretty safe since I am defining organisms as systems that are able to evolve. But one needs to be very fussy about what the word "evolve" should mean, and this leads to some quite tight specifications of the kind of systems the first organisms must have been. An organism cannot evolve; only organisms in the plural—successions of organisms—can evolve. Even that is not quite fussy enough. What can evolve is what connects organisms in lines of succession, what is passed on from generation to generation. That is not actually a material but genetic information; not substance but form.

Admittedly genetic information has to be held in some material substance: in genes of some kind. And the genetic information must have some kind of effect (called its phenotype) that helps it to survive and propagate. This is likely to involve other materials. But the only long-term survivor is the information itself. By the time a few hundred generations have passed every atom in some set of founder organisms will have been mislaid; the original substance will be gone. Only forms survive, modified or unmodified. To play this game, to have forms persisting in this curious way through copies of copies of copies, is an essential requirement for evolution.

For evolution actually to happen

# Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine*.

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

**Comprehensive.** Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

**Complimentary CME.** At no extra cost, we offer a CME program in which you can earn 32 top credits\* per year.

**Trial Offer.** We invite you to try SCIENTIFIC AMERICAN *Medicine* for two

months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us \$245 for the first 12 months (current renewal is \$185); otherwise, return the books; there is no charge.

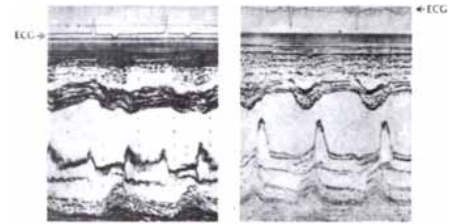
Please mail the coupon today so that we can put medical progress to work for you.

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

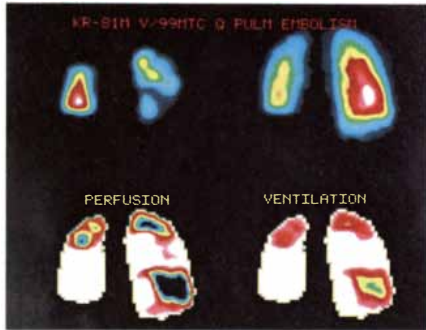
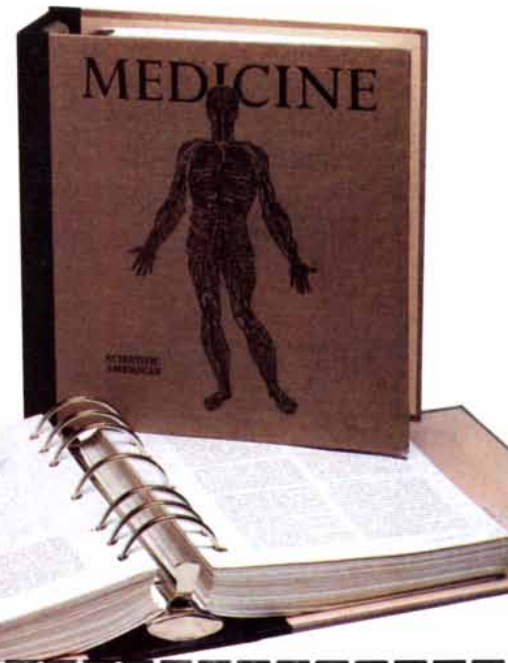
\*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

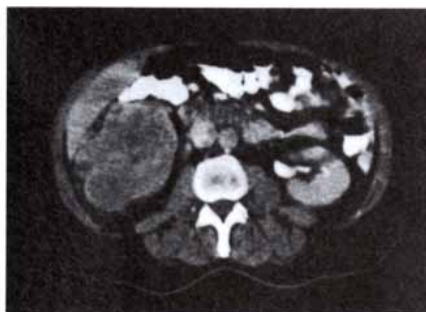
This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.



Echocardiograms from patients with aortic regurgitation.



Computerized scintigraphy reveals pulmonary thromboembolism.



Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.

## SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of \$245 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a complete refund.

- Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
- I shall also enroll in the CME program
- I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for \$245\*
- Bill me  VISA  MasterCard

Expiration Date \_\_\_\_\_ Account Number \_\_\_\_\_

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

Name \_\_\_\_\_ MD Specialty \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Signature \_\_\_\_\_

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an a © 1985 SCIENTIFIC AMERICAN, INC every is prepaid.

there are additional requirements. There must be occasional random changes in genetic information—mutations—and these changes too must be inheritable by offspring and able to produce altered phenotypes so that there can be a selection of the altered genetic information. Hence, over many generations, lines of succession may transform, the genetic information being adjusted to produce phenotypes that are particularly effective within specific environments.

This is not all there is to say about evolution by any means, but we have arrived at a *sine qua non*. Whatever else they contained, those first low-tech organisms we are trying to imagine had to include genes of some kind.

What else apart from genes would have been needed in principle in those first organisms? H. J. Muller gave an answer in 1926. His answer was “Nothing else”: the minimum specifications needed to account for the then known properties of genes in organisms were sufficient in principle for genes to be able to evolve on their own. Muller went further. Not only is it possible to imagine the first organisms as simply being genes but it is in fact probable that the first organisms were something close to this.

Muller argued first that a gene or genes would be absolutely required. Suppose, then, there had to be something else actually within the first organisms. For the organisms to reproduce, such companion structures would have to be remade or reacquired. This implies that additional information for such remaking or reacquisition would have to preexist in the genes. Much better to be able to do without “help” of this sort, or with as little of it as possible.

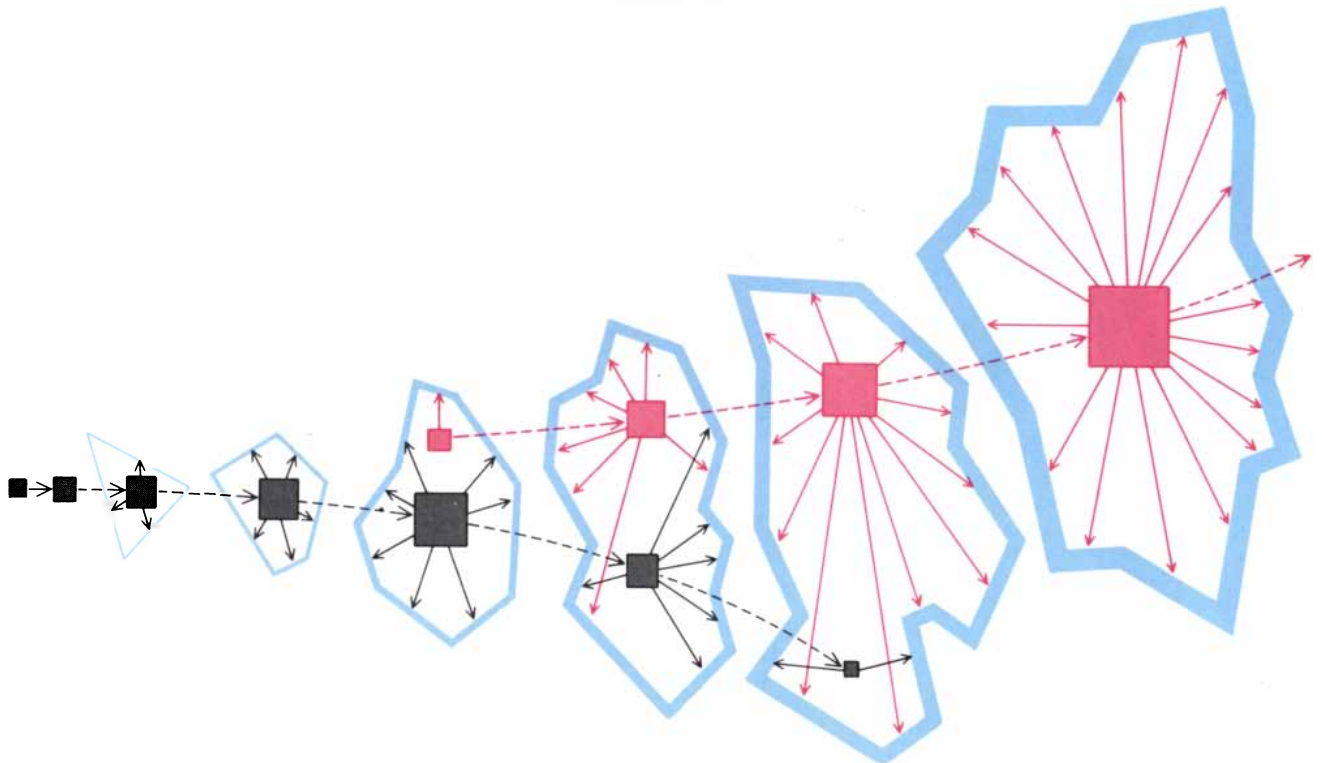
Thinking along these lines, RNA has been suggested as the original genetic material, although it seems to me that RNA is too high tech. Yet RNA molecules have indeed been shown to evolve in the test tube. A key feature of RNA here is that although, like DNA, its sequence information is replicable, this information does not have to be translated to be effective. As with a protein chain, the way a single RNA chain folds up may depend on the information it contains. These are most interesting experiments—but do they relate to the very start of evolution? The enzyme catalyzing the RNA experiments is far too complex to be a conceivable product of geochemical processes on the primitive earth. Even if much simpler catalysts prove to be adequate, there is the further difficulty

that the replication of RNA requires clean supplies of special high-energy nucleotide units.

In any case, Green Herring 2 contained a strong hint: in looking for plausible designs for the first organisms one should not look for some cutdown version of present-day life. The first organisms in being low tech would have been different, and probably made from different materials altogether. One might suspect in particular that designs for gene materials required to work without companion structures would be different from designs that are possible once evolved companion structures are allowed.

It is not difficult to imagine an evolutionary process through which a first, geochemical genetic material could be gradually replaced by an altogether different organic chemical one. I call such a process genetic takeover [*see illustration below*].

If indeed one or more genetic takeovers were involved in the early evolution of the central biochemical control machinery, one should not expect to find the components of the first genetic material anywhere in the modern molecular construction kit. This might seem to make genetic takeover a rather negative, spoilsport idea. But there is a positive aspect too in this point of



**GENETIC TAKEOVER** is seen by the author as a key phase in early evolution. Originally there were naked genes of some unknown first genetic material (*gray boxes at left*), which evolved to control their immediate environment by specifying the production of increasingly elaborate surrounding phenotypes (*blue shapes*). A

new kind of gene appeared (*red boxes*) that could function only within a rather sophisticated phenotype but that proved to be more effective there than the original genes. The new genes gradually took over control of the phenotype, which came to be converted to their exclusive service. Eventually the original genes were dropped.



# **BREAKTHROUGH: A COMPUTER THAT UNDERSTANDS YOU LIKE YOUR MOTHER.**

Having to learn letter-perfect software languages can be frustrating to the average person trying to tap the power of a computer.

But practical thinkers at our McDonnell Douglas Computer Systems Company have created the first computer that accepts you as you are—human.

They emulated the two halves of the human brain with two-level software: One level with a dictionary of facts and a second level to interpret them. The resulting Natural Language processor understands everyday conversational English. So it knows what you mean, no matter how you express yourself. It also learns your idiosyncrasies, forgives your errors, and tells you how to find what you're looking for.

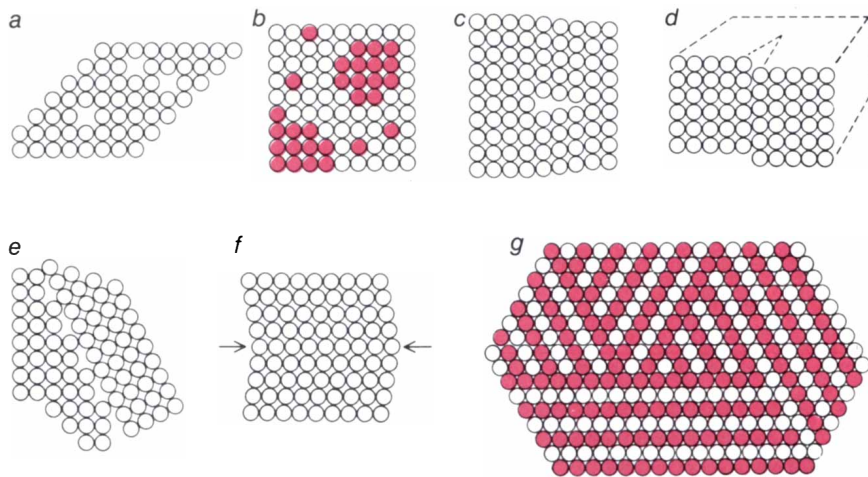
Now, virtually anyone who can read and write can use a computer.

We're creating breakthroughs not only in artificial intelligence but also in health care, space manufacturing and aircraft.

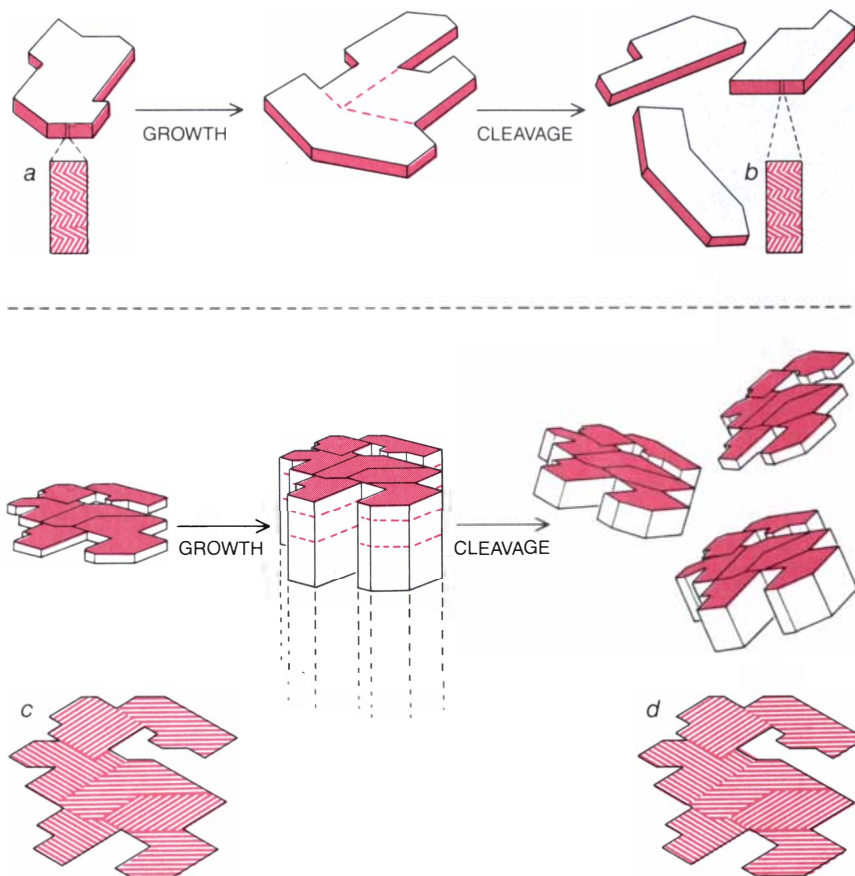
We're McDonnell Douglas.

# **MCDONNELL DOUGLAS**





**DEFECTS IN CRYSTALS** could supply multiple, stable alternative configurations, the sine qua non for information storage. The drawings show some common crystal defects: vacancies in a lattice (a), substitutions of individual units or of domains (b), an edge dislocation (c), a screw dislocation (d) and a grain boundary between lattices (e). In a twinned crystal (f) differently oriented parts share a common plane of units (arrows). In some crystals (g) large domains have the same overall composition but differ in the alignment of units.



**CRYSTAL GENES** would need to display the right combination of structural, growth and cleavage characteristics. Information might be stored in one or in two dimensions in crystal genes. In a one-dimensional gene (top) information would be held in the detailed structure of a sequence of stacking layers (color), which remains constant (a, b) as the gene replicates. Growth takes place only on the colored faces and cleavage takes place only parallel to those faces. The information-carrying layers could vary physically (there might be differently aligned crystal structures, for example) or in their chemical composition. In two-dimensional crystal genes (bottom) information would be held as a pattern (again either physical or chemical) on one face of the crystal (color); that pattern remains constant (c, d) as the gene replicates by growing on the colored face and by being cleaved in a plane parallel to it.

view: it provides a brand-new field of chemical possibilities to explore. We have Green Herring 3 to focus our attention on the mineral world, and we still have considerations of a general sort to suggest what any genetic material must be like.

Here is what Muller said about the nature of a genetic material a quarter of a century before the role of DNA was known: "'Gene' material is any substance which, in given surroundings—protoplasmic or otherwise—is capable of causing the reproduction of its own specific composition, but which can nevertheless change repeatedly—'mutate'—and yet retain the property of reproducing itself in its various forms."

That there should be some kind of templating in the process of gene replication is strongly suggested by this description. One can hardly help seeing "specific composition" (genetic information) as some kind of specific pattern that is copied through the assembly and linking of new units in contact with that pattern. (This is the kind of thing that goes on in DNA or RNA replication.) If templating is not the only conceivable way complex mutable patterns might replicate, it seems to be the easiest and most direct way.

But we have to think of a genetic material whose units are much simpler than those of DNA. We have to think of units the earth could have manufactured cleanly and consistently over quite long periods. And there could be no elaborate enzymes to help; the components of the first genetic material had more or less to self-assemble.

Reach for another Green Herring:  
4. Genes must incorporate a large number of atoms.

A gene could never be a very small assemblage of atoms if it is to contain more than a trivial amount of information and hence have any potential for evolution. Moreover, the genes must surely be well-structured objects.

Were the first genetic materials crystals? These are the commonest self-assembled objects. The analogy between crystallization and fundamental living processes has often been drawn, and then usually rejected as much too limited. (The physicist J. D. Bernal went further: "Crystallization is death.")

To my nose there is another pair of Red Herrings behind this kind of objection:

3. Crystal structures are boring.
4. Carbon is best for life.

Red Herring 3 owes its irrelevance to a concern with "perfect" crystals, which do not exist. It is true that a crystal has a basic crystal structure, which is highly repetitive. But any real crystal also has a defect structure superim-

posed. Simply to be finite—to have a shape and size—is already a “defect,” but many other features are almost invariably present. There may be units that are missing or are replaced by others; large or small sections of the “wallpaper” may be misaligned in various ways. Such features can be very small in scale. They provide real crystals with a large potential capacity for information.

Can one imagine defect structures of any kind that might replicate as a crystal grows? The answer is yes. Several classes of crystals might have a suitable combination of structural characteristics, growth patterns and cleavage properties.

Red Herring 4 still has to be dealt with. Again the issue is not truth but relevance. We can agree that organic molecules are the best materials for life. But the best is what you might expect evolution to arrive at; what you expect it to begin with is instead the easiest. And the easiest form of self-assembly is a spontaneous crystallization from simple, available units. Which leads us to clay.

All around us, all the time, clay minerals are crystallizing from dilute solutions of silicic acid and hydrated metal ions formed by the weathering of hard rocks. Indeed, the earth's surface has been described as a huge factory that manufactures clay minerals.

Two great wheels drive this factory. First there is a geologic cycle that derives its energy from radioactive heating inside the earth; it is a set of processes that buries sediments, cooks them at high temperatures and pressures deep within the earth and then pushes the transformed materials back to the surface. Here they are no longer quite stable. They are ready to dissolve in water, to break into those small silicic acid and metal-ion units and to crystallize into quite new materials: clay minerals of various kinds. Sooner or later these materials, more or less transformed, return to the sediments to be buried again. The second cycle maintains the supply of water. It is a cycle driven by the sun: evaporation from the sea, the formation of clouds, rain, groundwaters, streams, rivers, and back to the sea.

Of course the earth when life originated would have been different from the earth now; what evidence we have suggests, however, that the probiotic earth may not have been very different. Metamorphosed sediments are among the oldest rocks known, suggesting a weathering cycle was already in action 3.8 billion years ago. It is quite possible that life is even older than these rocks. Perhaps it originated under conditions

where clays could not have formed, but there is no particular reason to think it did.

It is also possible that primitive genes were microcrystalline minerals other than layered silicates—“clays” more loosely defined. I shall nonetheless stick with standard clay mineral types for the sake of discussion.

One of the implications of this line of thinking is that the primitive genetic material, or something similar, should be forming here on the earth now. How could we start to look for a clay genetic material? From the abstract specifications for different kinds of genetic crystal, and from what we know about clay minerals, we should be able to imagine what the stuff ought to look like and then see if there are real clays that fit.

Consider first a one-dimensional clay gene. It is common indeed for layers in clays to be stacked in a variety of sequences. The layers may be of the same kind differently superposed, or there may be a sequence of different kinds of layers. Remembering that a clay crystallite—even one with several layers—can be quite flexible, and that a one-dimensional crystal gene would grow exclusively sideways, we begin to picture a mass of folded (perhaps branched) membranes or laths of constant thickness. There are many clays that both have an irregular stacking and look something like this.

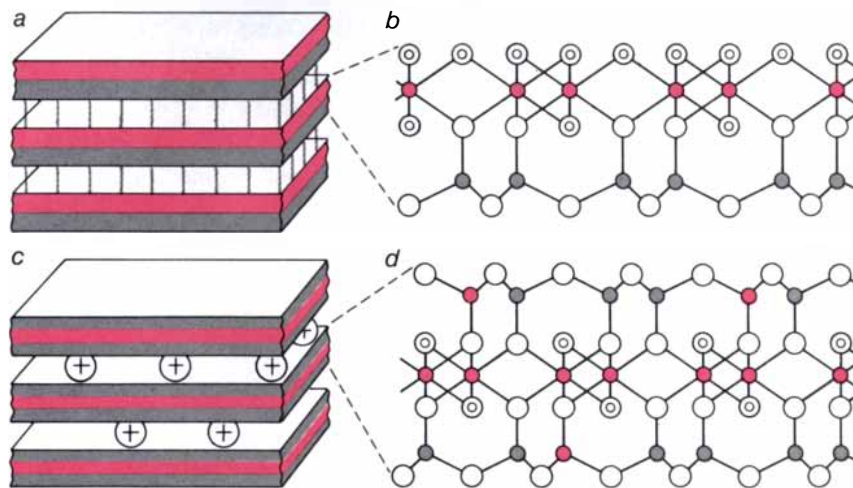
Admittedly the above is only mildly suggestive. (But what an interesting kind of gene it would be that could

spread its message indefinitely without having to divide!) Let us move now to another distinctive-looking clay that might be an example of a two-dimensional clay gene.

Sturges W. Bailey and C. F. Mansfield of the University of Wisconsin at Madison made an X-ray study of large crystals of vermiform kaolinite. They found an interesting defect structure. The individual kaolinite layers consist of a mosaic of small domains that resembles crazy paving. In any given domain all the aluminum atoms are arranged in one of three possible orientations. A structure such as this could hold quite a lot of information, and such information would be replicated provided the aluminum orientations in a newly forming layer were always determined by the orientations within the layer on which it was growing. Ideal kaolinite crystals have their aluminum orientations maintained between layers, although in real crystals there are often “mistakes.”

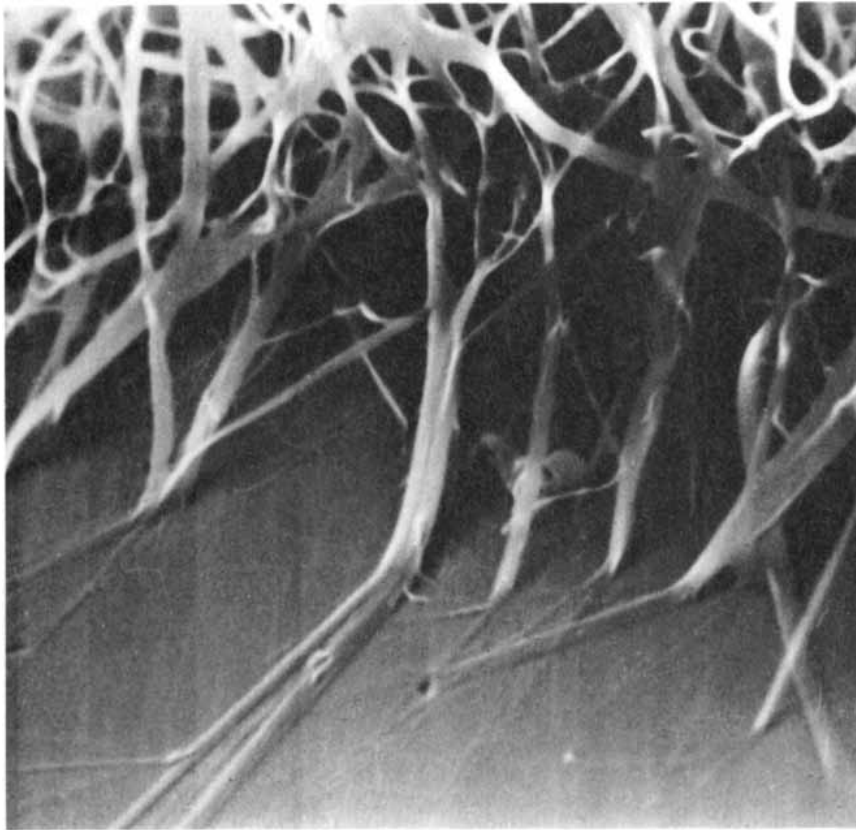
There are signs in any case of a similar patterning of some kind in the individual layers of typical vermiform crystals, some of which have complex but constant cross sections. This, and particularly deep grooving, is indicative of a domain structure [see top illustration on opposite page].

In a preliminary account Armin Weiss of the University of Munich has reported laboratory studies of the growth of smectite crystals. Weiss says that new layers, growing between pre-existing layers of the crystals, pick up information from the original layers—



- OXYGEN
- ⊙ HYDROXYL GROUP
- SILICON
- ALUMINUM
- ⊕ POSITIVE ION

**MOST CLAYS** are made up of stacks of layers. In kaolinite (a) asymmetric layers are linked by hydrogen bonds. Each layer consists of a net of aluminum atoms and hydroxyl (OH) groups fused to a net of silicon and oxygen atoms (b). Other clays have symmetric layers, in which a silicon-oxygen net is fused on both sides to a metal-hydroxyl net; these layers are negatively charged and linked by positive ions (c). In illites (d) much of the negative charge arises from the substitution of aluminum atoms for silicons.



**ILLITE LATHS**, firmly attached to a grain of sandstone, are enlarged 10,000 diameters in a scanning electron micrograph made by W. J. McHardy of the Macaulay Institute for Soil Research in Aberdeen, Scotland. Such a clay is a conceivable one-dimensional genetic crystal but (being only a few silicate layers thick) would have limited information capacity.



**VERMIFORM KAOLINITE**, enlarged 1,350 diameters in a micrograph made by Keller, was formed by weathering. Such clays are conceivable two-dimensional genetic crystals.

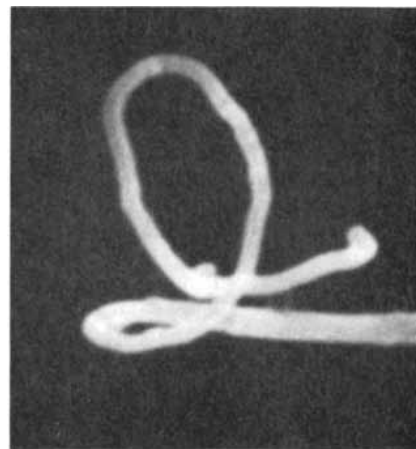
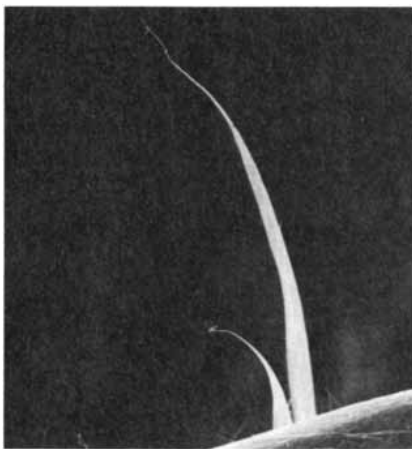
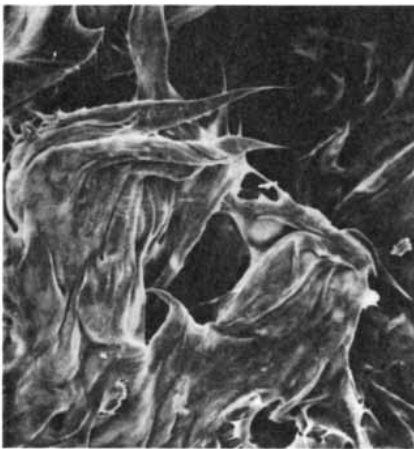
information in the form of negative charge density arising from aluminum substitutions in the silicon-oxygen net.

Clearly there are further observational and experimental clarifications to be made of the big question: Do mineral crystal genes exist? At this point I can only answer "Quite possibly" and go on to the next question: Could mineral crystal genes evolve? The answer to this, it seems to me, is "Yes, they could hardly help it."

**T**hink of a common microenvironment for clay formation: the pore space of a sandstone. Weathering solutions are percolating through the sandstone, and two-dimensional replicating clays may be crystallizing in the pores. There are a number of zones, each filled with millions of copies of crystals, each of which has some characteristic defect structure. One zone might have crystals of such shape and size that they fit together tightly to form an impervious plug. The movement of solutions into this region is diverted; the crystals in the region stop growing. In another zone small, loosely cohering particles allow solutions to flow, but these particles are easily washed away when the rains come. This is not a great success either. In a third zone the crystals are so shaped that they are snagged in pits in the walls of the pores; these crystals both stay in place and allow the flow of nutrient solutions. In yet another zone rather long crystals are replicating that have little holes and projections that cause the crystals to lock together, forming a rather open framework: another way of staying in the right place without stopping the flow of nutrient solutions.

Even such a simple environment as this could generate some quite subtle selection pressures, factors that determine why some defect structures—in this case mainly shapes and sizes—are more successful than others. Mutations would create variants, and so different limbs of a sprawling zone would often be populated by crystals having somewhat different defect structures. As a result some limbs might grow faster than others or be more likely to survive periodic bad conditions.

Now let us think about a somewhat more complex environment. Groundwater solutions are emerging from a sandstone into a fast-flowing stream that, although too acid and too dilute for clay synthesis, contains one of the components for a genetic clay that is in rather short supply in the sandstone. Chemically the ideal place for the genetic clay to grow would be at the interface between the two milieus. Physically this would seem far from ideal: the interface would be narrow and



**VARIETY OF FORMS** assumed by clay minerals is one factor suggesting they might have been suitable materials for early organisms. A leathery halloysite is seen at successive magnifications to be a mass of fibers, at least some of which are hollow tubes, in a se-

ries of scanning electron micrographs made by Keller. A very small spike that is barely visible at an enlargement of 130 diameters (left) can be seen at 1,000 diameters to incorporate a loop (middle), which is perceived as a hollow tube at 26,000 diameters (right).

variable. Any crystals forming would be in danger of being swept away or redissolved, or both. Yet the interface could be stabilized and extended through a suitably coherent mass of (replicating) crystallites attached to the sandstone; it would be a sticky kind of paste. Mutations that change the shapes of the interweaving crystals forming this paste would change its porosity, and hence acidity gradients within it, as well as concentration gradients of other ions. The suitability for clay synthesis within the paste could be adjusted through natural selection. The successful gene paste would be a compromise between being a good hanger-on and providing a good environment for clay synthesis.

The giraffe is said to have its neck length adjusted through natural selection so that members of the species can eat leaves from treetops without actually falling down in a faint. We should expect that varieties of replicating clay particles should adjust their shapes and sizes appropriately in particular circumstances too. Forget about complicated physiology. The logic is much the same. Replicating, mutating structures become optimized in ways that depend on circumstances. They can hardly help it.

So far we have been thinking of the first kinds of organisms as having been made of nothing but genes. This has been possible because even pure gene stuff can have a phenotype, that is to say, it can have information-dependent physical properties that can affect its success. More realistically, clay gene stuff would become contaminated with other clays growing under the conditions the clay genes had contrived. Such cocrystallizing material might sometimes help to improve

some property such as the porosity or the ability of a paste to stay in place. Mutations in the clay genes would then be selected that encouraged suitable secondary clays. There is a particularly direct way one could imagine a defect structure in a genetic crystal exerting control on the growth of another material: through epitaxy, that is, through secondary clays crystallizing on the surfaces of genetic clays and being affected by specific defect characteristics.

We are now moving beyond consideration of the first organisms. I shall try very briefly to sketch in a connection between those organisms and us. The connection would start with the introduction of organic molecules. There is a "Why?" and a "How?" here.

Why were organic molecules introduced? There are many reasons. Some smallish organic molecules (for example amino acids, di- and tricarboxylic acids) can make metal ions, such as aluminum, more soluble. In this way they can act as catalysts for clay synthesis. Other classes (for example heterocyclic bases and polyphosphates) are particularly adept at sticking to clays, often altering physical properties of a clay paste. Organic molecules can also exert powerful effects on the shape and size of inorganic crystals by inhibiting the growth of certain faces. This might have been particularly important for the controlled replication of crystal genes. Then again organic polymers might have had structural effects, such as holding clay particles together.

My guess would be that precursors of RNA appearing in well-evolved clay organisms would have served as structural materials in the first place. (Indeed, RNA is still used somewhat in this way.) An RNA-like polymer

with its negatively charged backbone would tend to stick to the edges of clay particles (which are often positively charged). On the other hand, heterocyclic bases (molecules such as adenine) have a tendency to stick between the layers of clays. One might imagine some RNA-like polymer as having evolved specifically to interact with clays (perhaps even to "read" the information exposed at the edges of one-dimensional clay genes).

The genetic takeover that was to bring today's biochemical control machinery into existence would have started, according to this story, when RNA became a replicating molecule—a new kind of collaborating genetic material with (at first) a minor role. It would have been a long haul before the clay "scaffolding" could be disposed of—perhaps only after the necessarily elaborate machinery for competent protein synthesis had appeared. The evolution of such machinery becomes thinkable, however, in the context of an already evolved organism. The machinery would evolve as a subsystem; at first an "optional extra," it would gradually become more useful and sophisticated. Then, with the scaffolding gone, it would emerge as a necessity.

Why might evolution move from what is inorganic and crystalline to what is organic and molecular? Because organic structures can be built much more finely (once you have the technology), thereby achieving more intricate control.

How were the organic molecules introduced? My prejudice here is for photosynthesis from the start, using carbon dioxide from the atmosphere to make, at first, molecules such as formic acid.

The really interesting question, though, is how somewhat more com-

plicated molecules could have been put together long before there were enzymes. How could nucleotides have been made? There would have had to be manufacturing procedures: many chemical reactions and other processes such as purifications would have had to be properly sequenced. That kind of thing does not just happen. It has to be organized. For evolved clay organisms the organizer would have been natural selection. Through what physical means would it have operated before there were protein molecules to work on? Before there were enzymes I think there must have been a more old-fashioned kind of apparatus in organisms: apparatus more like what one would find in the organic chemist's laboratory or in an industrial chemical plant. Along with containers, tubes, pumps, filters, ion exchangers and adsorption columns there are rather unspecific catalysts.

That brings me to the final part of the case I am making for clay minerals as the main materials out of which the first organisms were formed. If you want (rather unspecific) catalysts, you will often find them among clay minerals. More particularly, if you want plain apparatus of the kind just described, there are plenty of ordinary clays that seem to have the propensity to form such things. How it all got put together and under what selection pressures is murky history. But is this not the right kind of stuff one ought to be thinking about?

Three skeptical questions:

*Why are crystal genes not everywhere obviously around us if indeed they are of*

*common stuff and if their evolution could hardly be prevented?*

I can think of four answers to this question: (1) There are, perhaps, no such things as crystal genes. (2) Mineral genetic materials are in fact rare. (3) Suitable conditions for the replication of mineral genes are rare. (4) Evolved mineral genes are common but unrecognized. Take your pick.

In any case I would not imagine that modern clay-based organisms could again reach the point of exploiting organic molecules. The competition from DNA microorganisms would be too fierce. The same would apply to ancestral forms, which are unlikely in any case to be with us.

*How could little clay crystals possibly be described as alive?*

The first organisms would have been pretty unimpressive and not, I think, alive. You need organisms as a prerequisite for evolution, but "life" is something else. It is a rather vague idea, a kind of oddness, a seemingly purposeful complexity that would gradually emerge as a product of evolution. But later primary organisms would have been alive, I think, in anyone's book.

*What experiments are there to do?*

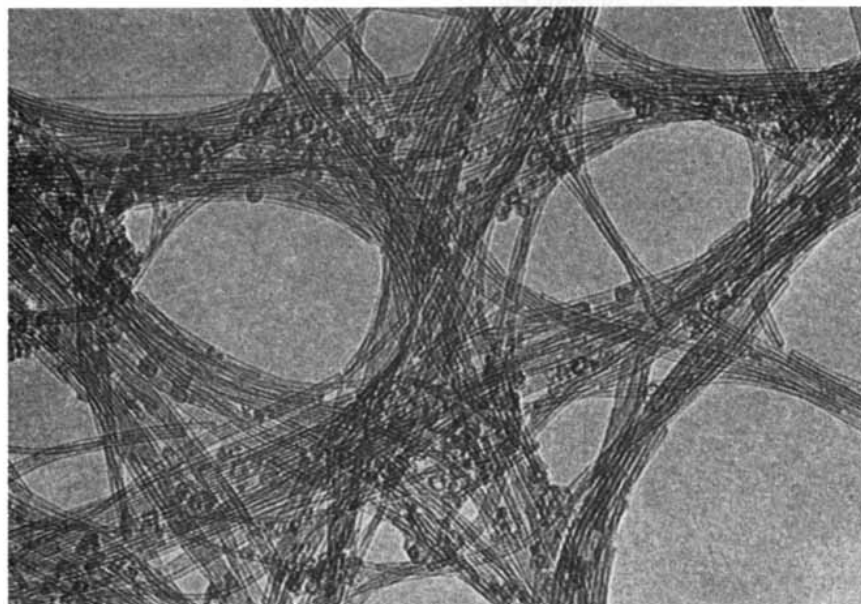
The interface between clay and organic molecules is the focus of vigorous exploration, much of it being done at the National Aeronautics and Space Administration Ames Research Center near Mountain View, Calif. There James G. Lawless and his co-workers have shown how metal ions such as zinc and copper can mediate the binding of nucleotides to clays. They also find that ions in clays exert selective

catalytic effects on amino acids. While working at Ames, Max M. Mortland of Michigan State University found that the coenzyme pyridoxal phosphate can perform enzymelike functions if it is combined with copper-containing montmorillonite clays. Noam Lahav of Hebrew University, together with David White of Santa Clara University and Sherwood Chang of Ames, showed how clays subjected to cycles of wetting and drying can link molecules of the amino acid glycine. The cycling transfers energy from the environment to the organic molecules.

Energy-handling machinery would have been a necessary part of all but the very simplest organisms. Lelia M. Coyne of the University of California at San Jose has found that kaolinite clays might well have supplied such machinery. They can gather energy from the environment (from radioactive processes), store it and then release it when the clay is suitably disturbed, for example wetted or dried.

Trying to take a more direct look at the clay-organic molecule interface, Lawless, Chang and their colleagues have studied carbon-containing meteorites about as old as the solar system, which provide clues to what organic chemistry may have been like early in the earth's history. It is intriguing that clays and organic molecules coexist in such meteorites. The surface of Mars may also have something to reveal about early conditions on the earth. Amos Banin of Hebrew University developed the view that the Martian surface is dominated by iron-rich montmorillonite clays. This would account for the outcome of an experiment performed by the Viking lander, in which carbon dioxide and carbon monoxide were converted into organic molecules (of which formic acid was probably a major component) in the presence of ultraviolet light. Jerry Hubbard of the Georgia Institute of Technology had shown experimentally that iron minerals, including clays, could produce similar effects.

The most critical experimental challenge now is surely to discover crystal genes—not just one kind but many kinds, and not just minerals either. Imagine doing experiments with crystals that could evolve, setting them problems—applying selection pressures—and seeing how they cope. This would be an interesting thing to do anyway, whatever the crystals are made of. We would soon find out whether mineral versions of replicating systems are plausible, although we might lose interest in our ultimate ancestors once we had in our hands the first organisms of another kind: the first organisms of our own contriving.




**IMOGOLITE AND ALLOPHANE** are enlarged 500,000 diameters in an electron micrograph made by Naganori Yoshinaga of Ehime University in Japan. The long, thin structures are seamless tubes of imogolite, which are studded with tiny hollow pods of allophane.



**ExperTelligence**<sup>™</sup> has the “tools” to transform your Macintosh<sup>™</sup> into a powerful Artificial Intelligence workstation. **ExperLisp**<sup>™</sup> is the first complete implementation of LISP on a microcomputer. Developed on a Symbolics 3600,<sup>™</sup> the compiler generates efficient MC68000 code providing speed and function ideal for the development and delivery of sophisticated AI applications. **ExperOPS5**<sup>™</sup>, by Science Applications International Corporation, is a complete implementation of the well-known OPS5 expert systems

building tool. It provides a fast and efficient method for constructing complex Expert Systems. **ExperLogo**<sup>™</sup> features 3-D and spherical graphics, English-like commands and shares the speed and function of ExperLisp. In the classroom or in the lab, ExperLogo provides an environment for discovery and exploration for children and developers alike.

**Call today** for more information about these and other innovative AI products.

 **ExperTelligence, Inc.**  
559 San Ysidro Road  
Santa Barbara, CA 93108  
Tel: 805/969-7871

Symbolics 3600 is a registered trademark of Symbolics, Inc.  
Macintosh is a trademark licensed to Apple Computer, Inc.

# The Social Ecology of Chimpanzees

*Wild chimpanzees have rarely been studied without the lure of food, which can distort their social relations. A study of unprovisioned apes shows their social structure is shared only with human beings*

by Michael P. Ghiglieri

Among all the species of mammals that have been studied in depth chimpanzees have a unique social structure. Each chimpanzee community, which may consist of 50 or more members, occupies a territory from which other male chimpanzees are excluded. Within their territory the members of the community are constantly on the move, searching for fruit-bearing trees and other sources of food. When fruit is sparse, the community members have the option of leaving their party and striking out on their own to forage. When fruit is plentiful, however, the apes tend to congregate in large parties to feed, mate, groom each other and rest. This "fusion-fission" form of organization, in which the community continually fragments and reassembles, is rare among social animals.

Another feature of the chimpanzee community is even rarer: female exogamy, or mating outside the home group. When females reach sexual maturity, they emigrate to the territory of a new community to mate. In contrast, males spend their entire lives in the territory where they were born. Ultimately they become part of the male collective that patrols the territorial boundaries and sires the community's next generation. Female exogamy results in a genetic division between the males and the females of the community; the males are closely related genetically, whereas the females may or may not be related to one another.

The distinction within a community between a group of closely related males and a group of unrelated females has significant implications for understanding the evolutionary history of both chimpanzees and human beings. Among animals in the wild only chimpanzees display the combination of a fusion-fission society, territoriality and female exogamy. Anthropological research, however, suggests that this form of organization is typical of hu-

man societies in the hunting-and-gathering phase. Chimpanzees are the closest living relatives of human beings: the DNA of the two species differs by only 1.2 percent. Thus human beings and chimpanzees not only are close genetic relatives but also share a unique social structure. Understanding the evolutionary forces that shaped the chimpanzee community may shed light on how human hunting-and-gathering societies evolved.

Much time and effort has been spent studying the behavior of chimpanzees. It is only recently, however, that field studies done by me and by several other workers have provided detailed observations about the social structure of the chimpanzee community in the wild. One reason information has been accumulated slowly is that wild chimpanzees are shy and elusive. When they detect a human presence, they vanish. To overcome the apes' shyness, most previous studies have relied on offering them food, generally sugarcane or bananas. Two notable long-term studies on the eastern shore of Lake Tanganyika have been based on the provisioning approach. Jane Goodall and her colleagues carried out their study in Gombe National Park; Toshisada Nishida of the University of Tokyo and his colleagues carried out theirs in the Mahale Mountains. Both studies have run continuously for 20 years. During that time they have provided many crucial observations of chimpanzee behavior along with tantalizing suggestions regarding social structure and evolution.

Unfortunately such work cannot offer a complete understanding of chimpanzee social organization. Although providing chimpanzees with abundant food makes them easier to study, it is also known to distort their normal social and ecological patterns. As we shall see, the distortions themselves are not without interest. Knowledge of

the chimpanzee community in its undisturbed state, however, depends on studies that are not based on provisioning. The desire to study the undisturbed chimpanzee band brought me to the Kibale Forest Reserve in western Uganda for a two-year study in which I accustomed the apes to my presence by appearing repeatedly at naturally occurring sources of food.

The central block of the Kibale Forest Preserve has been little affected by human society, and so it provides an excellent site for studying *Pan troglodytes schweinfurthii*, the eastern long-haired subspecies of chimpanzee. My study area was Ngogo, the center of the nature reserve within the forest. In 1977 Thomas T. Struhsaker and William J. Freeland of the New York Zoological Society began research on red colobus monkeys, red-tail monkeys and gray-cheeked mangabeys at Ngogo. They constructed a rectilinear grid of trails for observing the monkeys and recording the distance they traveled within an area of about six square kilometers. The grid system proved to be an invaluable tool for carrying out a quantitative study of the chimpanzees at Ngogo. I used the grid and expanded it where necessary to observe the daily activities of the Ngogo community, which included about 55 chimpanzees.

To understand the social ecology of an animal group it is necessary to identify the group's resource base and observe how the group arranges itself to exploit those resources. For chimpanzees the primary resource is fruit: at Ngogo the apes spent 78 percent of their feeding time consuming fruit or seeds. That proportion tallies with what observers have noted in other habitats, although the percentage varies and may be considerably lower in marginal habitats. Wild chimpanzees also eat insects and hunt mammals, including monkeys. In addition they consume a variety of plant foods: bark, pith, blossoms and young leaves. Their





**FORAGING CHIMPANZEE** is shown in a *Ficus* tree. *Ficus* trees bear a fruit that is one of the chimpanzees' preferred foods in the equatorial rain forest. Chimpanzees spend more than half of their waking hours foraging. The apes often forage alone, as this one is doing, but groups of 20 or more can gather in a single tree. The un-

usual structure of the chimpanzee community, which makes changes in party size possible, is an evolutionary adaptation that maximizes the efficiency of the search for fruit. The photograph was made by Constance S. Ghiglieri in the Kibale Forest in southwestern Uganda during the author's two-year study of wild chimpanzees.

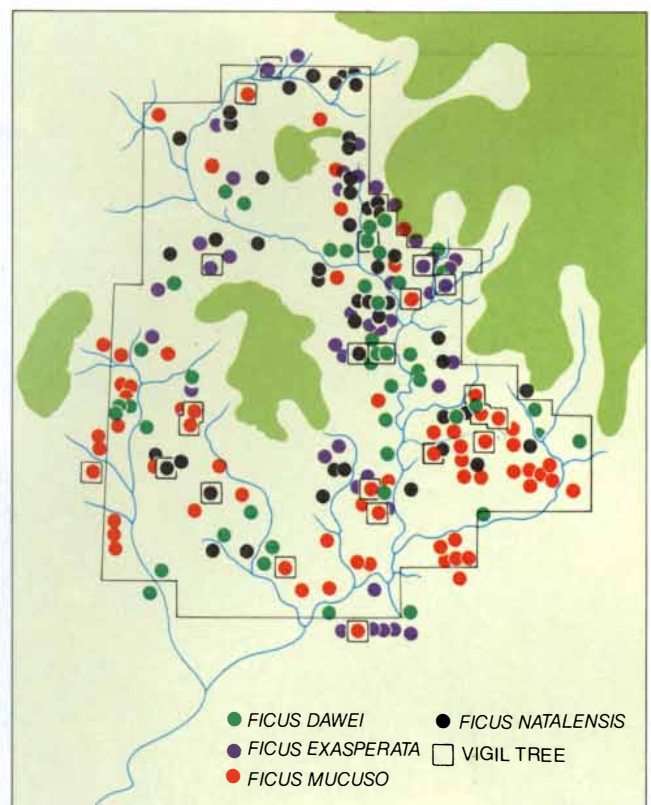
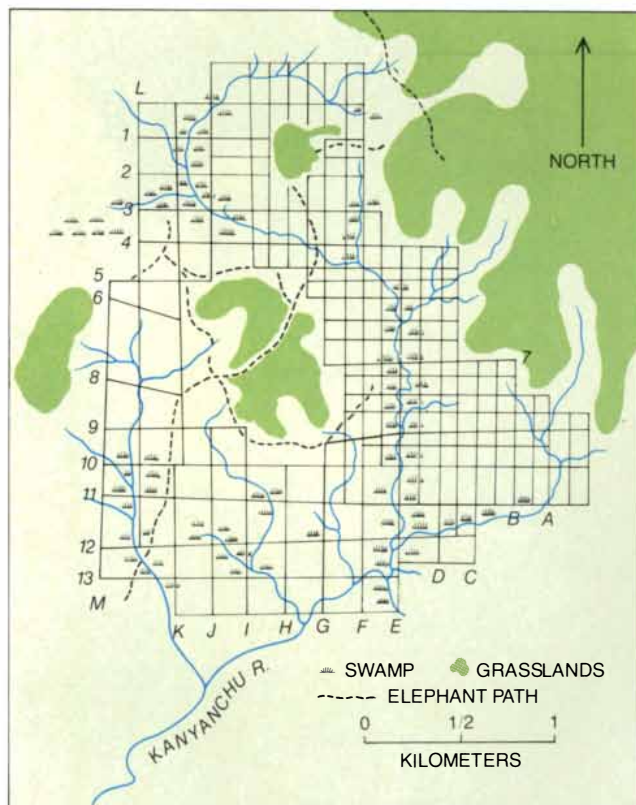
most sought-after food, however, is fruit. At Ngogo the chimpanzees feed on tens of species of fruit trees, favoring in particular several species of the genus *Ficus*, which yields the protein-rich fig.

The exploitation of fruit as a primary food places constraints on social structure; these constraints flow from the uneven distribution of fruit in time and space. More than 100 species of trees may grow within the 10 to 30 square kilometers that make up the territory of a chimpanzee community. Of the 100 species, however, perhaps only one-fourth ever provide edible

fruit. Analysis of the distribution of the 12 most important food species at Ngogo showed the apes depend on rare species that tend to grow in clumps rather than being uniformly distributed. Moreover, finding these small clumps does not guarantee a meal. Tropical trees rarely bear fruit on a regular annual schedule. Some trees fruit at unpredictable intervals. Others, including several *Ficus* species, bear two crops a year on an irregular schedule. Within a community's territory ripe fruit on the trees of a particular species appears and disappears in a matter of days. The total volume of edible fruit can vary by a factor of

eight between the wet season and the dry season.

Chimpanzees are not the only animals searching the forest canopy for this ephemeral commodity. During the daylight hours at Ngogo I often observed the chimpanzees confront fruit-eating birds or monkeys. The competition with the seven species of monkeys is particularly robust. According to my censuses, the monkeys together outnumber the chimpanzees some 200-fold. None of the seven species has a diet that includes as high a proportion of fruit as the diet of the apes, but all the monkeys compete with the apes to some extent. During con-



**KIBALE FOREST** lies in the equatorial rain-forest belt of eastern Africa (left). Several long-term studies of chimpanzees, including a well-known study by Jane Goodall, have been carried out on the eastern shore of Lake Tanganyika. The author did his work in a part of the Kibale Forest called Ngogo, where a community of about 55 chimpanzees lived. He observed the chimpanzees by means of a rectilinear grid of trails (upper left). Thirteen species of *Ficus* grow at Ngogo. Of these, the four major species tend to grow in clumps (upper right). The trees enclosed by squares are "vigil trees," from which the foraging behavior of the apes was observed over extended periods. The uneven distribution of the fruit trees of the rain forest and the irregular timing of their fruiting imply that the amount of available fruit fluctuates greatly, which has a profound effect on chimpanzee social organization.

# HOW TO REACH THE PEOPLE WHO MAKE THE FUTURE HAPPEN IN JAPAN



SAIENSU is the Japanese edition of SCIENTIFIC AMERICAN which attracts a young audience of affluent professionals who have come to the top because of their technical expertise. Almost half of them hold top management job titles. The fortunes of major corporations in Japan increasingly depend on technically sophisticated people. SAIENSU keeps these people in touch with the advances in science that drive the growth of Japan's industry.

SCIENTIFIC AMERICAN speaks the languages of more than half the world's population. We are the one publication in the world today providing efficient coverage of technology-based management. We reach men and women in industry, whose qualification to make technical decisions places them in key positions in their country's government and industry.

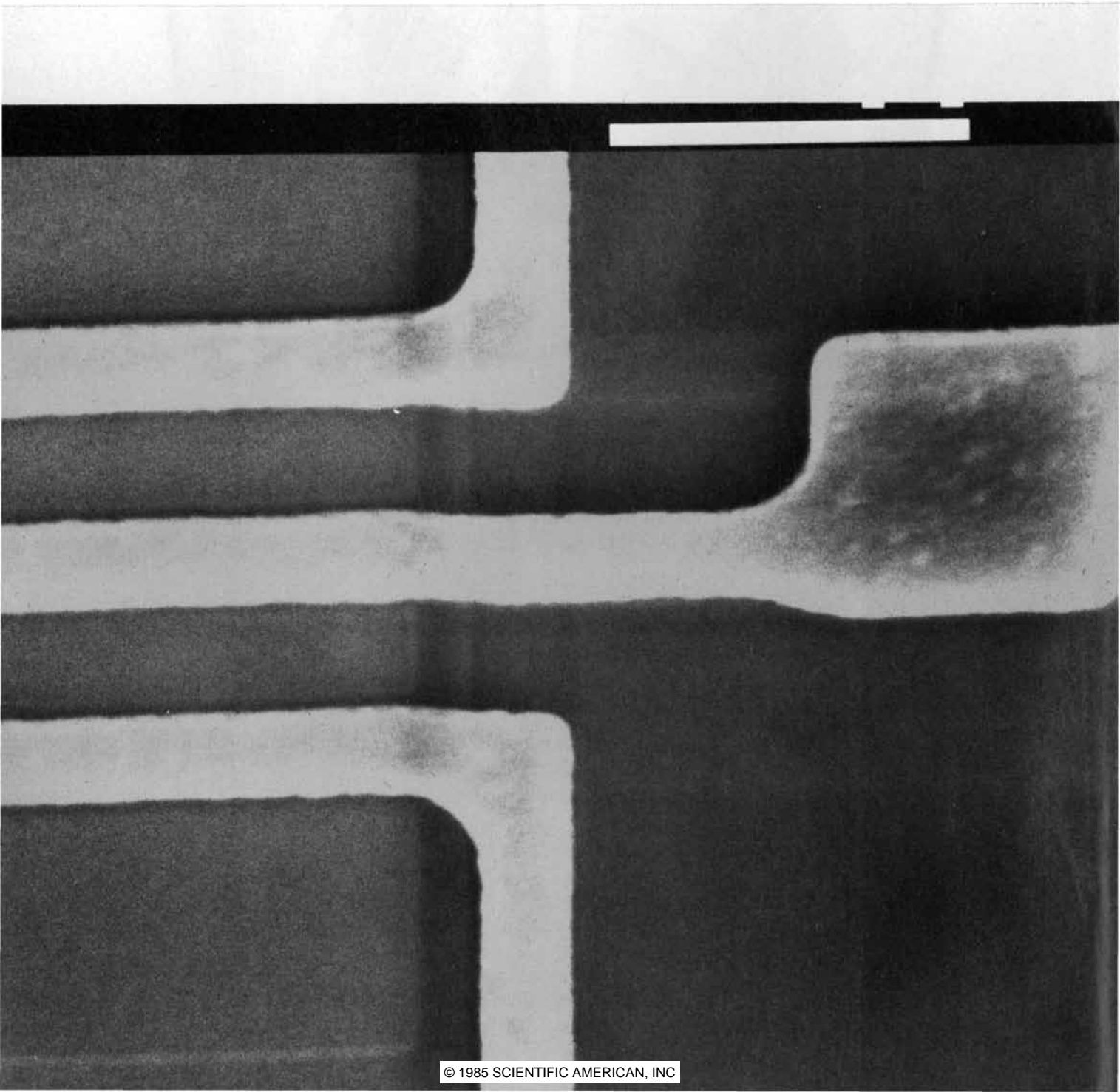
SCIENTIFIC AMERICAN in eight languages: English, Spanish, Italian, French, German, Japanese, Chinese and now Russian has gathered in its audience three million people who make the future happen around the world.

For more information on our Japanese-language edition contact:

Michiaki Akasu  
Nikkei International  
The Nihon Keizai Shimbun  
No. 9-5, 1-Chome  
Otemachi, Chiyoda-ku  
Tokyo 100, JAPAN  
Telephone 011-813-270-0251

*or in New York*  
John Kirby  
V.P./Advertising Director  
SCIENTIFIC AMERICAN  
415 Madison Avenue  
New York, New York 10017  
Telephone 212-754-0262

# **INTEL TAKES THE LEAD IN CMOS. BY HALF A MICRON.**



Low power dissipation of CMOS devices has typically meant low performance, but with Intel's 1.5  $\mu\text{m}$  design rules this limitation is virtually eliminated.

These half-micron-narrower design rules are critical to a process we call CHMOS—a process that yields CMOS products with performance on par with HMOS technology.

What's equally significant is that we provide a complete family of CHMOS 1.5  $\mu\text{m}$  products—not just an isolated product here and there. Including the likes of the highest density DRAM available in CMOS (256K), a high-performance 35ns static RAM, EPROMs and world standard microprocessors and controllers in CMOS.

Intel's CHMOS process is the amalgam of low-power CMOS with high-performance HMOS. Switching speeds of 200 picoseconds at 3 nanowatts power per gate can now be achieved.

Despite the continuous improvement of HMOS by scaling, the delay-power product for CHMOS is more than an order of magnitude lower than its HMOS counterpart in the typical integrated circuit. For example, in a VLSI part with 50,000 gates operating at greater than 10 MHz, the CMOS version would dissipate less than 2 Watts compared to over 25 Watts for an HMOS part.

A special use of double-metal layers with CHMOS can produce a large savings in real estate, making package densities of over 500K transistors on a chip possible.

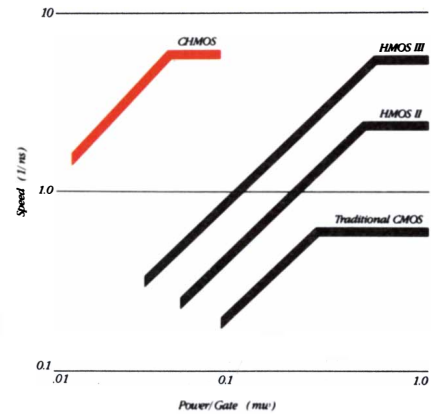
As critical dimensions of the technology are reduced, reliability becomes more difficult to achieve. With Intel's CHMOS technology, reliability is assured by the following:

First, by using epitaxial substrates, the latch-up problem is eliminated over a wide range of operating con-

ditions. For example, no latch-up is observed at supply voltages of 9 volts at 125°C for a part that nominally operates at 5V. And input pins can withstand currents up to 200 mA without initiating latch-up.

Secondly, the lower die temperatures that Intel's CHMOS provides substantially diminish the electromigration problem typical of VLSI devices.

Finally, by locating the storage node in the CMOS well, a natural barrier is provided against soft errors for DRAMs and SRAMs.



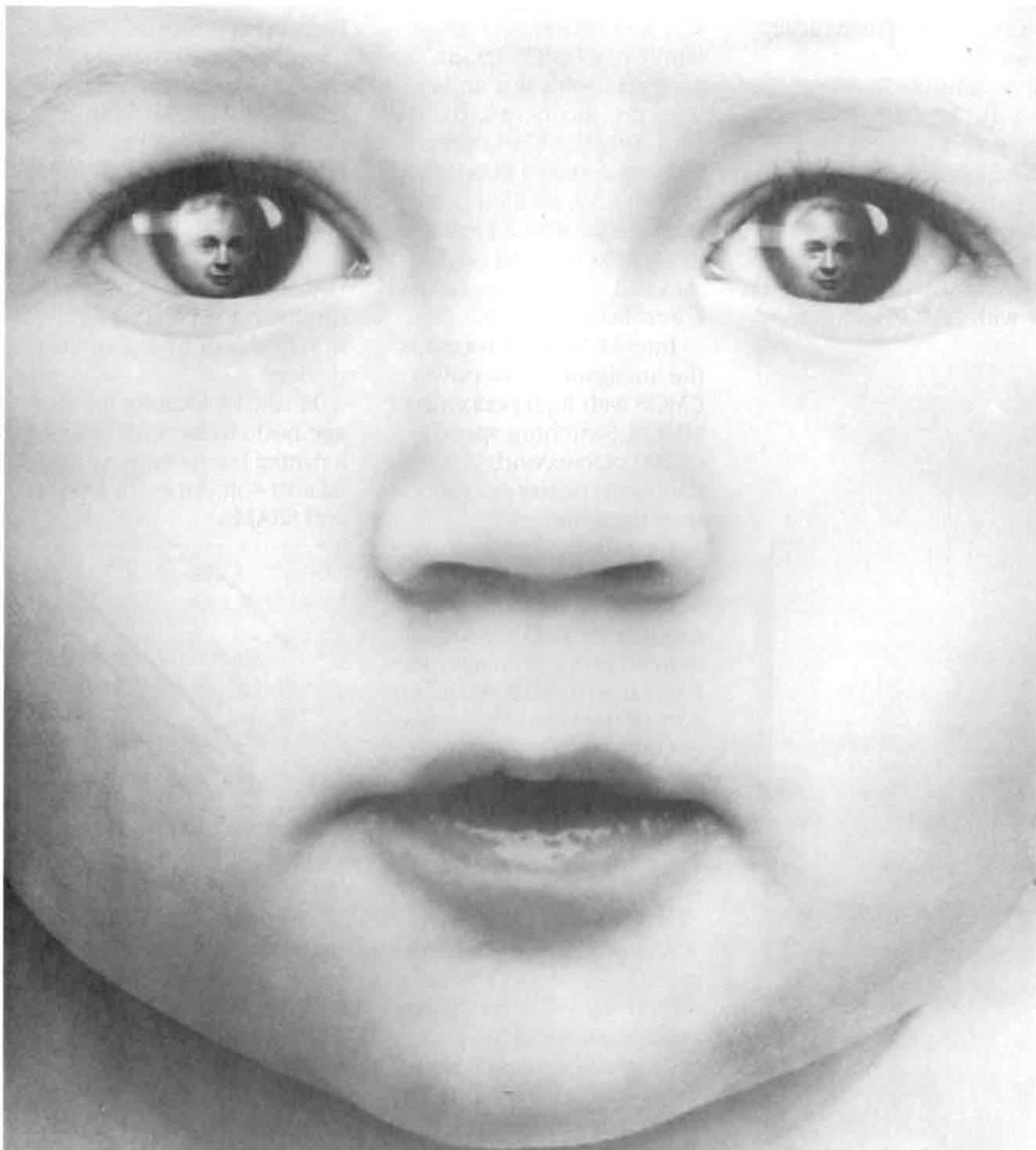
*Intel's CHMOS gives you the high-performance of HMOS but uses only one-tenth the power.*

The result is a low-power, high-performance CHMOS technology with high reliability. And an array of 1.5  $\mu\text{m}$  products unmatched in technical properties.

And that's only part of the story. For more information, call (800) 538-1876. In California, (800) 672-1833. Or write Intel Corp., Lit. Dept. W-203, 3065 Bowers Ave., Santa Clara, CA 95051.

And see just how big half a micron can be.

**intel**<sup>®</sup>



## Be Immortal.

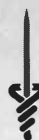
If you could look into the eyes of generations yet to come, you would be there.

Because immortality lies not in the things you leave behind, but in the people that your life has touched, for good or bad.

By including the American Cancer Society in your will, you can have a powerful effect on those who come after you.

You see, cancer *is* beatable. The survival rate for all cancers is already approaching 50% in the United States.

You'll be leaving behind a legacy of life for others. And that is a beautiful way of living forever yourself.



**AMERICAN CANCER SOCIETY®**

For more information, call your local ACS unit or write to the American Cancer Society, 4 West 35th Street, New York, NY 10001.

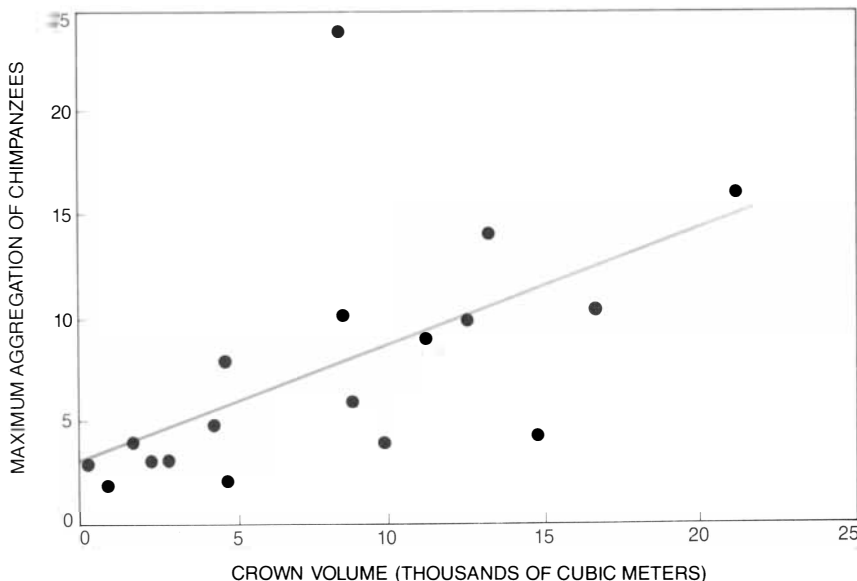
© 1985 SCIENTIFIC AMERICAN, INC

frontations the apes often displaced all monkeys except the red colobus. A concerted attack by the males of a large red-colobus troop frequently chased a party of chimpanzees from a favored fruit tree. One reason for the monkeys' success is that they are much more agile climbers than the apes. Male chimpanzees have been seen to sustain serious injuries or die as the result of falls from the rain-forest canopy. Therefore in the treetops most apes retreat before the aggressive colobus. On the ground or in low foliage, however, male chimpanzees hunt and kill the red colobus along with other species of monkey.

Thus the chimpanzees' way of life is based on finding a rare and quickly vanishing food before their more numerous competitors do. From such a perspective it is somewhat surprising that chimpanzees survive at all. Evolutionary adaptation, however, has equipped chimpanzees well for survival both individually and collectively. One of the main advantages of the individual chimpanzee is intelligence. Laboratory tests of captive apes have shown that if food is the reward, chimpanzees are capable of considerable mental prowess: task learning based on insight, on the principles of arithmetic and on discrimination, symbolic communication, the efficient planning of a food-gathering itinerary and the ability to communicate the location of hidden foods to other chimpanzees.

Among the components of intelligence, the chimpanzees' excellent sense of spatial relations and acute memory for those relations are of particular significance in foraging. Richard W. Wrangham of the University of Wisconsin, who studied chimpanzees at Gombe, concluded that the apes there were "good botanists" who could pick out a plant species in fruit from among all the surrounding plants and then search each plant of that species for ripe fruit. Chimpanzees outscore monkeys on all intelligence tests, and they are undoubtedly better than monkeys at finding fruit in the wild. Furthermore, chimpanzees travel on the ground, which is more efficient than clambering through the canopy as monkeys do. Their greater mobility enables the chimpanzees to range over a much larger area in search of food than monkeys can cover.

The chimpanzee's physical and mental abilities make it possible for the individual ape to hold its own against monkeys and other fruit eaters. The flexibility of the fusion-fission society is also crucial to success in this competition. No single fruit tree could supply the foraging needs of the entire Ngogo community of some 55 apes; it is rare



**SIZE OF FORAGING PARTY** is strongly influenced by the amount of fruit in a particular tree. The crown volume of the tree (*horizontal axis*), or the volume defined by the foliage, is assumed to be proportional to the amount of fruit. As the crown volume increases, so does the maximum number of apes observed foraging in the tree (*vertical axis*). The chimpanzees' capacity to travel in groups varying in size from one to 20 or more enables the apes to maintain social ties while specializing on fruit that can be quite sparsely distributed.

for even a clump of trees to satisfy the entire community. Hence the chimpanzees must forage in smaller subgroups. As the size of the foraging party increases, the number of trees that must be visited to satisfy the hunger of all its members increases proportionally. A group of trees with enough fruit for three adults to forage for an hour can accommodate 30 adults for only six minutes. Therefore the party of 30 must travel 10 times as far as the party of three to meet its nutritional needs. Such an increase in traveling would not only escalate each ape's metabolic requirements but also raise the time spent on the move from the normal 10 to 12 percent of daylight hours to 100 percent or more, thereby occupying all the available time.

The result is that apes traveling in search of food move in relatively small parties. After many hours of vigils at Ngogo I found that the average feeding party included 3.6 members. The parties were by no means uniform: I often saw solitary apes as well as larger feeding aggregations that included as many as 24 members. In an effort to identify the factors that influence party size I measured the volume of the crown (the rounded region defined by leaves and fruit) of several fruiting trees. I could then monitor visits made by the chimpanzees to trees with a known crown volume. It turns out that the size of the feeding group is proportional to the crown volume of the tree. Large trees, which have more

fruit, attract more chimpanzees and the chimpanzees stay longer than they do in small trees. Moreover, the chimpanzees are attracted to the large trees for more repeat visits and the large trees are foraged more intensively.

These results may seem obvious, even pedestrian: more fruit attracts more apes, which feed together. Yet this pattern is in striking contrast to the behavior of the orangutan, the chimpanzee's fruit-eating cousin. The two species are closely related: their DNA differs by only 2.2 percent. In addition their feeding ecology and habitat are similar and females of both species are about the same size. Yet chimpanzees and orangutans differ greatly in social structure. One reason for the difference is the degree of the orangutan's anatomical specialization for tree climbing. Orangutans literally have four hands and the bones of all the digits are curved to aid in gripping tree limbs. On the ground they curl their handlike feet into clumsy clubs unsuitable for traveling long distances. A group of orangutans would have difficulty traveling to enough trees to feed them. One result is that the orangutan leads a solitary life. The massive adult male occupies a home range that overlaps the ranges of two or more females. The male repels rivals with long calls or fights them to retain the exclusive right to mate with the females in his home range.

Orangutan social structure is a basic model shared by many other species of mammals. In contrast, that of the

chimpanzees is unique. Clearly the difference between chimpanzees and orangutans in the anatomy of the hands and feet is related to the disparity in social organization. The differences in anatomy, however, are reinforced by differences in psychology. When social contact is available to chimpanzees, they choose it. It was noted above that the provisioning carried out by Goodall and her colleagues at Gombe distorted the chimpanzees' social behavior. Specifically, lifting the normal nutritional constraints by supplying the apes with abundant food had the effect of increasing the average party size. Furthermore, at Ngogo I observed that friendly interactions between chimpanzees were more frequent by a factor of 10 than antagonistic ones.

Both observations show that chimpanzees favor the company of other chimpanzees. Under normal conditions parties must often be small to feed efficiently. The fusion-fission society is an individual adaptation that enables the chimpanzee to maintain social ties without sacrificing foraging efficiency. In contrast to almost all other primates that live in groups, adult chimpanzees always have the option of splitting off from their party to forage alone or with another foraging party. The fission reduces competition in times of scarcity. Over the course of evolution chimpanzees that could split off from large parties and forage on scarce resources without losing their ties to the community had a nutrition-

al advantage over those remaining in large parties when food was inadequate. An abundance of food reverses the fission, producing fusion: in times of plenty at Ngogo the chimpanzees gathered in relatively large groups to feed, travel and socialize.

If chimpanzees do well on their own or in small parties, it might seem there was little reason for them to rejoin other members of their community at all. Currently accepted theories of evolution hold that natural selection operates through the individual rather than through the group or the species. By exploiting all reproductive opportunities an individual maximizes the number of copies of its genes that appear in subsequent generations. Exploiting the greatest number of reproductive opportunities depends on having superior nutrition and access to adults of the opposite sex. The reproductive strategy of the males of most mammalian species is based on defending food and excluding other males from females that are in estrus. Yet male chimpanzees in the wild do just the opposite: males at Ngogo displayed cooperative behavior in both feeding and mating.

Some of the most striking examples of cooperative behavior among male chimpanzees are based on the vocal signals known as pant-hoots. Pant-hoots include stereotyped shrieks, hoots, wails and roars that can carry as far as two kilometers through the rain forest. They can be made by a lone ape or by a party of apes in chorus. The chimpanzees of Ngogo most often pant-hooted when traveling, ap-

proaching a source of food, watching other chimpanzees approach or responding to the calls of another party. More than half of all the calls at Ngogo occurred as part of an exchange with other apes. Peter Marler and Linda Hobbett of Rockefeller University analyzed sonograms of pant-hoots and found enough cues in each call for an individual pant-hooter to be recognized. Thus when a party calls across the rain forest, their pant-hoots may communicate the identity of the party's members along with their number and location.

The most intriguing function of the pant-hoot is to alert other members of the community to the presence of fruit. Roughly one-fourth of the times a group arrived at a large fruit tree one or more males pant-hooted. Their calling sometimes produced an impressive din that lasted for at least 10 minutes. After many of the calls other chimpanzees appeared at the tree and began to forage for fruit. To see whether the effect was a significant one I pooled all my data on the arrivals of subsequent parties at food trees. The data showed that subsequent arrivals came significantly more often if the males of the first party pant-hooted on reaching the tree. Food calls attracted both male and female apes, and the new arrivals shared the fruit with its finders.

Noncompetitive behavior among male chimpanzees even extends to reproduction. Unlike orangutans, the male chimpanzees of a community spend little time attempting to exclude





one another from mating opportunities. It is common for the males of a party to ignore one of their number copulating only a few meters away. Moreover, a female frequently copulates with several members of the same party in quick succession. More recent observations at Gombe, however, suggest that female chimpanzees are not always promiscuous and that males are not always tolerant. A dominant male can sometimes exclude his rivals by taking a female "on safari" in order to breed exclusively with her. In the Kibale Forest a male occasionally tried to assert dominance and exclude other males from a particular estrous female. Yet when an estrous female presented herself to a subordinate male for mating, dominant males rarely interfered.

The apparent sexual tolerance of the male chimpanzee is even more striking when the rarity of reproductive opportunities is considered. A normal adult female is sexually receptive for only a few weeks every five years. Most communities include about 15 adult females, along with an equal number of males. Therefore three females will be receptive per year on the average and only three males will succeed in siring offspring that year. The pressure of natural selection favors behavior that maximizes opportunities to reproduce. Yet the male chimpanzee does nothing while other males compromise his reproductive success by mating with the female he has just mated with. How can this be?

The answer lies in female exogamy.

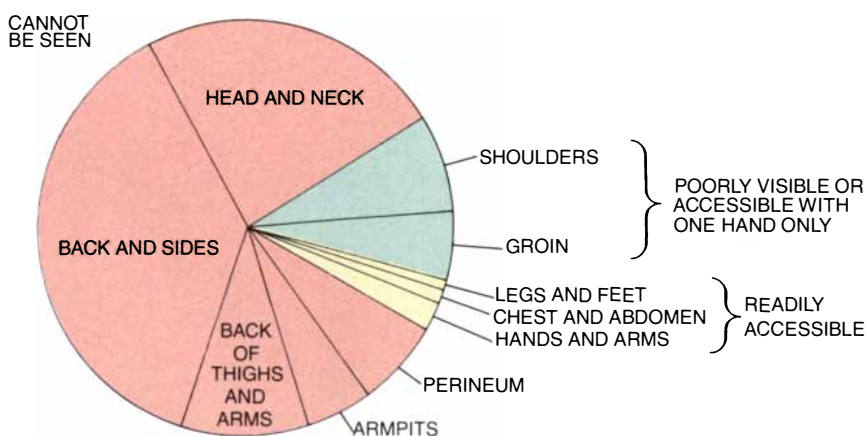
The females enter the chimpanzee community as strangers from another territory. Although the new females may be genetically related to one another (if they have migrated from the same community), it is more probable that they will not be related. The males of the community, on the other hand, are all closely related genetically because they are descended from the same line of "patriarchs." The genetic relatedness of the males appears to underlie their apparent altruism. Descent from the same small group of patriarchs implies that any two males in the community share some, if not many, of the same genes. Hence if one male reproduces, some of the genes of the other male are also replicated in the next generation. The degree of vicarious success depends on the extent of the genetic overlap: the more genes the unsuccessful male shares with the successful one, the greater the vicarious genetic success.

The degree of reproductive success that an organism shares when a genetic relative reproduces is referred to by W. D. Hamilton of the Imperial College of Science and Technology in London as inclusive fitness. The concept of inclusive fitness helps greatly to explain why male chimpanzees cooperate in feeding and mating. The food call that males give on arriving at a fruit tree attracts others of their band. If the later arrivals are other males or immature females, they are likely to be genetically related to the caller. Hence any improvement in their nutrition and their capacity to reproduce is

shared by the caller as an increase in his inclusive fitness. (In addition, if the later arrival is an estrous female, the caller may mate with her.) Of course, if the quantity of food in the tree is so small that the caller cannot be adequately nourished after others arrive, his inclusive fitness will decrease. I observed, however, that male chimpanzees almost never pant-hooted when they arrived at a small tree. Cooperation among members of the chimpanzee community is not based on altruism but on a complex form of self-interest.

Such self-interest also has a conspicuous role in mating. When a male sires offspring, the other males of his community share in his success by way of increases in their inclusive fitness. The size of the increase varies with the number of shared genes, and it is significant that the males of the chimpanzee community are quite closely related. In an extreme instance it is even possible for a male to increase his reproductive success without breeding at all. Goodall and David C. Riss of the Stanford University School of Medicine observed a partially paralyzed male that was a poor contender for mating opportunities. This male supplied his younger brother with assistance that was critical for the younger sibling in attaining the rank of alpha male. The alpha male is the dominant male of the community and the one most successful in taking estrous females "on safari." Thus by an apparently altruistic strategy the paralyzed sibling got an indirect genetic reward: his inclusive fitness was greatly increased. On a larger scale such rewards are essential in holding the community of males together and preventing it from splitting into atomistic, orangutanlike fragments. Genetic relatedness and inclusive fitness are key factors in the evolution of a community maintained by males that cooperate to feed, mate and defend their territory.

The division between the closely related males and the females of their community extends to every aspect of life. Males tend to choose other males as traveling companions and also as partners in the extended grooming sessions that often follow the morning's foraging; females choose females as traveling companions and grooming partners. Males and females even have different patterns of overall daily activity. Three basic activities fill the day of a chimpanzee in the wild: traveling, foraging and resting, which includes grooming and all other forms of socializing and self-maintenance. Males at Ngogo spend more of their time traveling than females (12 v. 10 per-



**GROOMING** is a social activity that appears to serve a crucial hygienic function. Chimpanzees groom in pairs, alternating active and passive roles. Grooming requires close visual inspection to remove lice from the skin. As indicated at the left, the chimpanzee's body includes areas that are visible and readily accessible to the ape itself (yellow), those that are marginally visible or are accessible with one hand only (green) and those that are not visible to the ape itself (red). The author's observations show that in mutual grooming sessions the greatest amount of time is spent on the nonvisible areas (above). Some observers have argued that grooming is a means of social facilitation, specifically the placating of a more dominant partner. Such facilitation could be carried out by grooming anywhere on the body. The concentration on nonvisible areas suggests that the primary basis of grooming's social function is the removal of parasites, which has a significant role in maintaining health.

cent). They also spend more time foraging than their female counterparts (62 v. 52 percent). The males spend less time resting than females (26 v. 38 percent).

It is not surprising that the sexes should have different daily activity budgets, because they have quite different roles in producing the next generation. The concept of parental investment, as defined by Robert L. Trivers of Harvard University, yields an illuminating way to think about parental roles in an animal community. In an attempt to quantify parental investment, Trivers defined it as the fraction of a parent's life span that is spent raising each offspring beyond the age of dependence. For a male chimpanzee direct parental investment is negligible, since the male's role may end after copulation. In contrast, the female invests the nine months of gestation, followed by at least four or five years of extrauterine dependence for each offspring. During that time the mother is constantly with the young chimpanzee, protecting it, demonstrating survival skills to it and often carrying it when the pair travel. Statistics from Gombe indicate that a female typically lives 35 to 40 years, with breeding beginning at about age 15. Thus each offspring surviving to maturity represents an investment of about one-fourth of the adult life span. If the female has three or four offspring that survive, most of the mother's adulthood is taken up.

As a result of the different reproductive strategies of each sex, their activity budgets may be fundamentally incompatible. A female does not need to travel to find mates. She needs to travel only enough to provide adequate nutrition for herself and her in-

fant. Any travel beyond what is strictly necessary for that purpose exacts a high metabolic toll because she is carrying a young ape. Males, on the other hand, need to travel, often widely, to find mates and defend their territory. In addition adult males do not generally need to consider the nutrition of their offspring, except when they are accompanied by juvenile males. Even then the adult male does not carry the juvenile. Therefore the cost of additional travel for males is not nearly as high as it is for females. In view of the lower cost of male travel and the much greater incentive for travel (in the form of finding opportunities to mate), it is to be expected that females rest more and travel less than males.

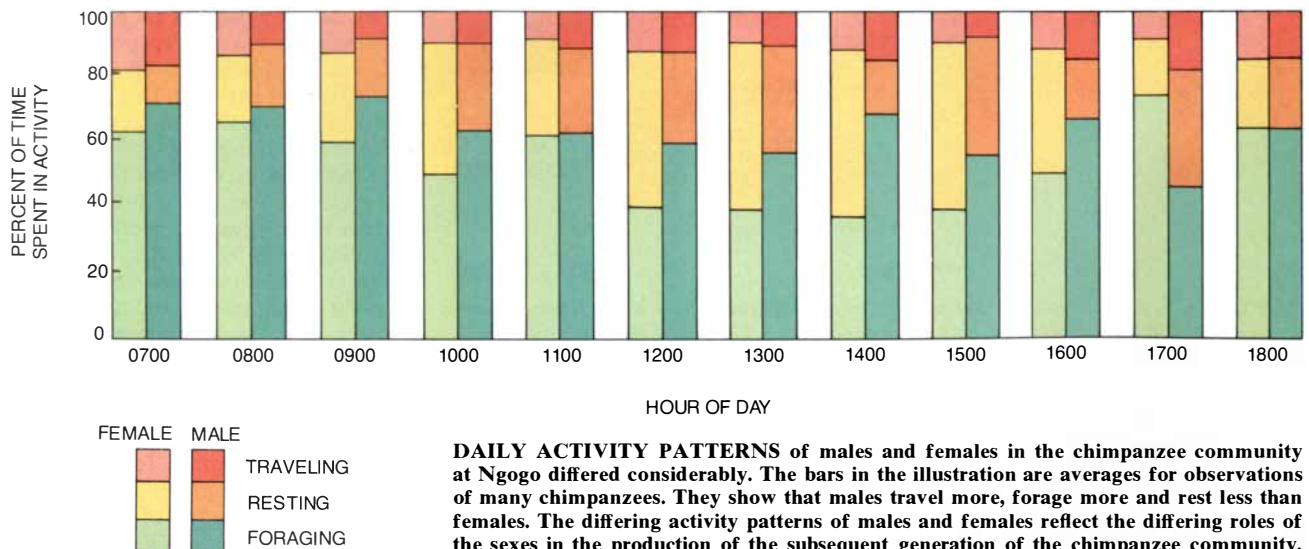
As a result of the recent work by Goodall, Nishida, Wrangham, me and others the structure of the chimpanzee community is coming into sharper focus. There is an emerging consensus that cooperation among genetically related males in defense of the home range is one of the foundations of the community. At times the defense can be murderous. After the heyday of banana provisioning at Gombe had passed, bands of males made stealthy patrols along the boundaries of their home range. The patrols, which did not include foraging, were apparently intended for monitoring the boundaries of the community's territory. On two occasions patrols were seen to attack strange females that entered the area with their infants. In both instances the infants were brutally killed. (Similar instances of infanticide have been observed in other study areas.)

If these infants had lived, they would have competed with the offspring of the male band for community resources. The females, on the other hand, offered reproductive opportunities to

the territorial defenders. Once their infants were dead the females quickly came into estrus. If they had not fled, they might have mated with the males of the community and thereby increased the males' reproductive success. Thus the male patrols treat other chimpanzees quite differently depending on what competition the other apes offer. In a series of savage forays the same males at Gombe killed the males of a small chimpanzee community to the south. That community ceased to exist and the victors absorbed their territory.

Although it is agreed that cooperation among males is fundamental to the structure of the chimpanzee community, complete agreement on that structure has not been achieved. For example, Wrangham interprets the chimpanzee band as a strictly male phenomenon. Because the pattern of female travel does not correspond with that of males, Wrangham does not identify the females as part of the territorial community. In the Wrangham model the home ranges of the females are dotted across the male territories like raisins in a pudding. The fact that the boundaries of the females' home range lie within one male community rather than within another is more or less accidental. Wrangham argues that the female chimpanzees are in essence like orangutan females, the main difference being that the chimpanzee females are bred by a coalition of males rather than by a solitary overlord.

The model of a community based on the travel pattern and genetic interests of males alone is compelling because it is simple. One drawback of such a hypothesis, however, is that it omits several factors of great significance in the lives of the female chimpanzees. A fe-



male cannot simply take up a position in the landscape at random, as the "males only" model implies. When she is sexually mature, she must leave her natal community to avoid inbreeding with her genetic relatives. Once she has an infant she must stay well within the territory of her mate; otherwise alien males patrolling the border may kill her infant. Thus reproductively successful females with infants will confine their range within the limits of a single community of males.

Furthermore, the behavior of the female chimpanzees suggests that they have community interests. Within her adopted community a female tends to socialize only with a subset of the other females. Some of these companions may have come from her original community and therefore may be her genetic relatives. Unlike orangutans, female chimpanzees prefer to travel with one another. Most intriguing of all, in some instances females collectively repulse strange females that are attempting to enter the females' home range. All this behavior implies a female community identity. If such observations are given substantial weight, they could imply a separate female community superimposed on that of the male chimpanzees.

Both Goodall and I hold that the community structure of the chimpanzee is based evenly on the behavior of males and females and serves the sexes equally. Because the reproductive strategies of males and females diverge considerably, their ranges and daily activities cannot be expected to coincide; yet the two groups form a single entity. The fusion-fission social structure is quite flexible in overcoming these contradictions. Such a structure also makes it possible for the apes to specialize on fruit and yet repeatedly re-form the bonds needed to maintain a large, unified community.

During most of the time that human beings have existed they have lived in hunting-and-gathering bands. Because the patriarchal fusion-fission structure of the chimpanzee group is similar to the organization of most hunting-and-gathering groups, it is possible that the study of chimpanzees will yield insights into our own warlike tendencies. In order for the study of chimpanzee society to yield any benefits, however, chimpanzees must survive. Human technology is destroying the mature tropical rain forests of the world at a rate that would eliminate them altogether by 2035. As a result chimpanzees are a seriously threatened species. I hope human society will leave the apes enough of their habitat for them to continue sharing the earth with us.

## QUESTAR® 12 on the QUESTAR® MOUNT

The Questar 12, latest addition to the Questar family of fine optical instruments, now has its own Questar-designed mount. A German equatorial type, it is notable for its 360° continuous tracking in R.A. with precision tracking to better than 4 arc seconds. The Questar Mount is designed with over-sized components so that it can accommodate any Questar up to 18 inches. The standard mount shown is straightforward in design but can be modified so as to be compatible with more sophisticated tracking devices or other special equipment.

The Questar 12 is a superb instrument for the serious astronomer, for the university astronomy department or the engineer seeking sophisticated tracking and surveillance equipment for which Questar Corporation has a noted reputation.

Questar Corporation Box 59, Dept. 20, New Hope, Pa. 18938 (215) 862-5277



*Let us send you our literature  
describing Questar telescopes,  
the world's finest optical  
systems. Please send \$2  
for mailing in N.A.; by  
air to S.A. \$3.50;  
Eur.; N. Africa \$4;  
elsewhere \$4.50*

© Questar Corporation 1981

# Siphons in Roman Aqueducts

*To carry an aqueduct across a valley the Romans built either a bridge or a siphon. Their siphons relied on the principle that water in a pipe will always return to its original height*

by A. Trevor Hodge

A remarkable engineering accomplishment by the Romans was the system of aqueducts with which they delivered millions of gallons of water daily to major cities of the empire. A typical aqueduct ran for many miles over varied topography. In order to carry an aqueduct across a valley the Romans relied on two solutions: a bridge, which merely maintained the gently declining slope of the aqueduct, and a siphon, which carried the water in a steep plunge down one side of the valley and a steep climb up the other side, relying on the principle that water in a pipe will always rise to its original height. A bridge was the solution if the valley was fairly shallow, a siphon if the valley was so deep that a dangerously high bridge would be required.

Most people today are familiar with what is the true siphon, a pipe or tube that carries liquid from one level to another over an intermediate elevation along a path resembling the letter *n*; in other words, the liquid first moves upward, so that the motion must be started by a pump or by some other outside force. Subsequently atmospheric pressure on the surface of the originating pool keeps the liquid moving. In a typical application one end of a tube is pushed down into the gasoline tank of an automobile; someone sucks on the other end to set the liquid in motion and then pushes that end into a container. As long as the container is no higher than the surface of the gasoline in the tank, the liquid will flow.

The Roman structure is properly called an inverted siphon: the path followed by the liquid is a U, and the siphon starts as soon as liquid is introduced into one arm of the U. In a simple inverted siphon the liquid entering one end of the U will rise to the other end; in the Roman siphons, because of friction in the pipes, the receiving end had to be somewhat lower than the originating end.

Although more than a score of these siphons have been identified, the role of the siphon in Roman hydraulics is generally unrecognized. The siphons offer little in the way of imposing remains; unlike the spectacular aqueduct bridges, they are at ground level and are therefore more easily destroyed by looters. Moreover, they played only a minor part in the aqueduct system of metropolitan Rome, which is the system modern scholars have studied the most intensively. (Siphons seem to have been built mainly in France, particularly around Lyon, which had a total of nine siphons on the four aqueducts serving it.) Both causes have led to their neglect.

A measure of the neglect is the long line of propagators of the doctrine—still standard fare in handbooks on Roman hydraulics, along with other errors—that the Romans built aqueduct bridges in preference to siphons because they could not make pipes strong enough to contain the pressure in an inverted siphon. The fact is that their siphon pipes successfully carried water at considerable pressure. Indeed, in 1875 the French engineer Eugène Belgrand made replicas of Roman pipes and tested them to destruction. They failed only when the pressure reached 18 atmospheres. This means, roughly speaking, that the pipes would have served in a siphon dipping 180 meters below the natural water level. Such a siphon would be deep enough to replace more than three Gard bridges built one on top of the other. (The Pont du Gard, a spectacular Roman aqueduct bridge near Nîmes in southern France, is 50 meters high.)

A typical siphon began at the point where an aqueduct, running as usual in an open masonry channel, reached the edge of the valley to be crossed [see illustration on page 116]. There the water entered a header tank, built of brick and cement, that was set

crosswise to the channel. This structure was in effect a distribution tank, because the siphon was composed not of one pipe (as in modern engineering) but of a battery of as many as nine small pipes laid side by side. Their intakes were arranged in a row on the downstream side of the tank.

The pipes, made out of lead, were formed by bending a flat sheet around a wood core to form a tube, soldering the two edges together and withdrawing the core. This process yielded a pipe of oval or pear-shaped cross section, with a continuous seam along the top. (Strangely, the seam was evidently not a weak spot; in Belgrand's test it was the side wall rather than the seam that failed.) On the other hand, the process made it difficult to manufacture large pipes, which is the reason a Roman siphon consisted of a battery of small pipes. They were normally 25 to 27 centimeters in outside diameter and the lead was three to five centimeters thick. Evidence from the remains suggests that the pipes came in lengths of about three meters (10 feet).

From the header tank the pipes ran down a short ramp to the ground. There they were buried about a meter deep and carried down the side of the valley. Presumably the purpose of laying them underground was to protect them from human meddling, but an additional effect was that they were not strained by expansion in sunlight or on a hot day.

Arriving at the bottom, the siphon could of course follow the profile of the ground, but often a low bridge (*venter* in Latin) was built to flatten the bottom of the U, reducing the drop of the water. The *venter* introduced a sharp bend (*geniculus*) at each end of the bridge and a consequent strain on the pipe joints as the water hit the corner. It did, however, cut down the static pressure by reducing the head (the distance from the top of the U to the bottom).

Even where the *venter* is well preserved, as it is at Beaunant near Lyon, its upper surface no longer shows any trace of the bank of pipes that once crossed it. The Beaunant *venter* is remarkably wide (7.35 meters), far wider than was necessary to accommodate nine 25-centimeter pipes. Probably the extra width was for catwalks that provided access for workmen.

From the *venter* the pipes climbed through a second *geniculus* to the top of the valley. There the water ran into a receiving tank similar to the header tank and thence back into a conventional aqueduct channel. The receiving tank was significantly lower than the header tank; the difference is known as the hydraulic gradient. Water will rise to its own level, but its passage through

the siphon was considerably slowed by the increased friction as it was forced through the nine small pipes. The area of contact in the pipes greatly exceeded the area in the conventional rectangular channel, particularly because the channel was not normally full and the upper water surface flowed without friction. Hence if both ends of the siphon were at the same height, water



**HEADER TANK AND RAMP** marked the beginning of the siphon at Soucieu on the Gier aqueduct, one of four aqueducts serving Roman Lugdunum (Lyon). The tank distributed the water, which arrived in an open channel at the rear, into nine rather nar-

row lead pipes. They were carried down the ramp in the foreground and then (buried about a meter below the surface) down the side of the valley, across a low bridge (the *venter*) and up the other side to a receiving tank. From there the water returned to a single channel.

would come through so slowly that it would block back and overflow at the upstream end. In order to ensure the delivery of water at an adequate volume and velocity a siphon had to lose more height in crossing a valley than a conventional bridge did. The hydraulic gradient of a siphon was something like 10 times the normal slope of the aqueduct.

The variety of topographies in which the Romans resorted to siphons is apparent in the four aqueducts serving Lyon: Mont d'Or, Gier, Craponne and Brevenne. Even a fairly short aqueduct that had a moderate overall drop might require several siphons; the key to the number of siphons was the number of deep valleys the aqueduct had to cross. In a drop of 90 meters the Mont d'Or aqueduct had two siphons. The Gier aqueduct had an even and gentle slope but in its total drop of 110 meters needed four siphons. Craponne had a precipitous drop of 420 meters and two siphons, one of them apparently of heroic scale. The Brevenne aqueduct ran in a steplike alternation of cascades and plateaus. Although it dropped a total of 350 meters, it needed only one siphon.

It is even more instructive to compare some of the siphons with one another. The two large ones of the Gier

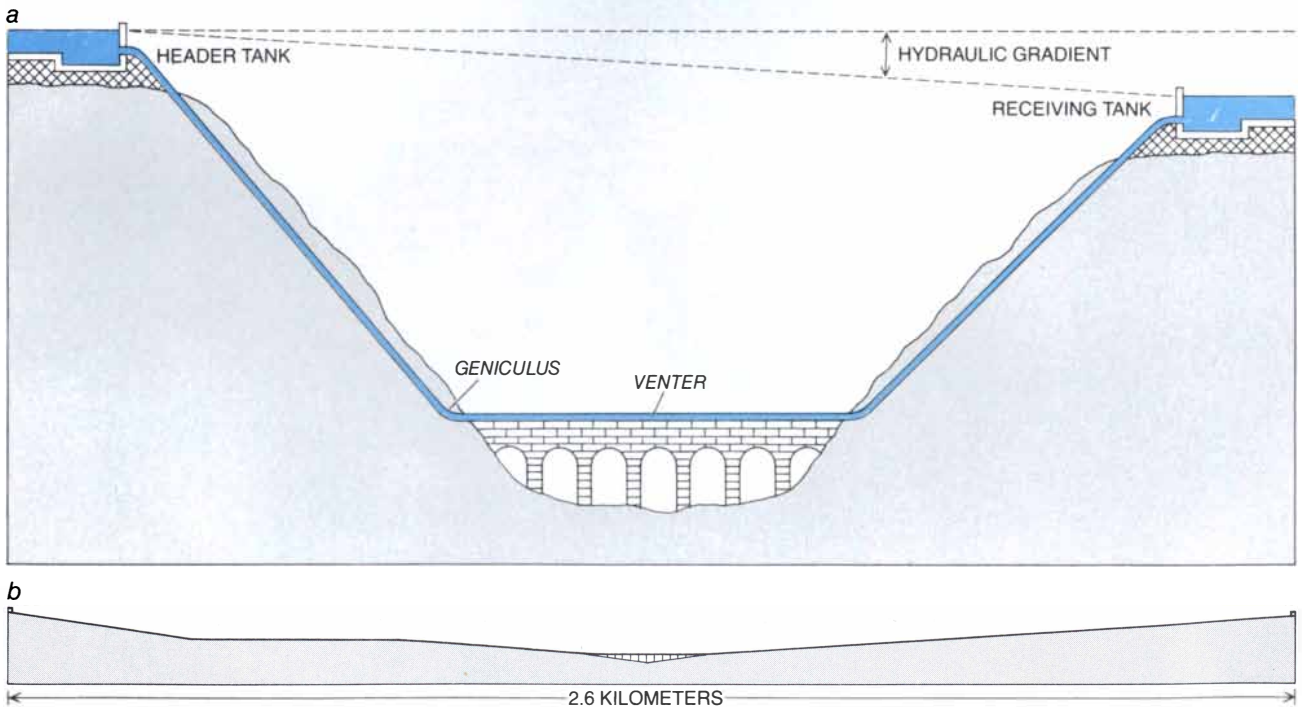
system are at Soucieu and Beaunant. Soucieu is 1.2 kilometers long and 93 meters deep, Beaunant 2.6 kilometers long and 123 meters deep. Coincidentally each had a loss of height of about nine meters, which meant that at Soucieu with its shorter length the hydraulic gradient was steeper. The Craponne aqueduct gives evidence of a truly enormous siphon some six kilometers long, which dipped some 100 meters below the hydraulic gradient. There are few remains on the site and the evidence is largely topographical: the aqueduct is known to have crossed the valley, and since the valley is too wide and deep for a bridge, there must have been a siphon.

By ancient standards the amount of water delivered by the four Lyon aqueducts was not particularly spectacular. It has been estimated at 80,000 cubic meters per day, compared with between 700,000 and one million cubic meters per day from the system serving Rome. (All figures for ancient water consumption sound high to modern ears. The ancients seldom used taps, and everything ran continuously. The resultant flow kept the sewers constantly flushed.)

Still, as engineering works the siphons command respect by their sheer magnitude. The total length of the nine siphons in the Lyon system came to

about 16.6 kilometers. At nine pipes each that represents 150 kilometers of piping—enough to stretch almost from Rome to Naples, easily from New York to Philadelphia. The pipes required some 12,000 to 15,000 tons of lead—a monumental undertaking in mining and transportation. Every foot of this piping was under pressure, on occasion as high as 12 atmospheres (12 kilograms per square centimeter). No doubt the system leaked a bit, but it also worked, and it spanned valleys far beyond the scope of the largest Roman viaducts and bridges.

In addition to the fact that water rises to its own level, Roman engineers had to reckon with three other forces governing siphons. First, friction in the pipes retards the water so that the siphon has to lose height to maintain the flow. Second, the static pressure within a pipe depends on the depth of the pipe below the natural water level, that is, on the vertical column of water being supported. Static pressure is created by the simple presence of the water and is exerted in all directions equally. The pressure remains the same whether the water is moving or standing. Third, water exerts an inertial thrust only when it is moving and only at bends in the pipeline. There the thrust is in only one direction: toward the outside of



**ROMAN SIPHON** is depicted schematically and with the vertical scale somewhat exaggerated (a). It is called an inverted siphon because the water follows the path of a U rather than the initially upward course (resembling an n) of a true siphon. The force of the water was particularly strong at the *geniculus*, or bend, at each end of the *venter*, so that the Romans typically reinforced the pipes there

by embedding them in a mass of masonry. The purpose of the *venter* was to reduce the head, or the drop of the water from the distributing tank to the valley. The receiving tank had to be somewhat lower than the header tank because friction in the pipes slowed the water; the difference was the hydraulic gradient. A profile (b) of the Beaunant siphon in the Gier aqueduct shows the actual gradients.

the curve. The second force operates any time the siphon is full; all three operate only when the siphon is full and running.

From time to time a siphon must be drained for cleaning or repair. Inertial thrust becomes a crucial force as a pipe is refilled. This is potentially a dangerous operation. The water has to be run in quite gradually until the pipe is full. If the sluices are simply thrown open, the water—hitting the first bend after an unchecked downhill run—can wreck the pipe. Conversely, shutting the water off in preparation for draining must also be done gradually, because a suddenly closed valve or sluice can generate water hammer: a shock wave transmitted back along the abruptly arrested moving column of water. This force too can severely damage the pipe.

How fully the Romans understood these principles is uncertain. Plainly they could apply them empirically, because the siphons were built and did work. If one seeks an explicit statement of the theory, however, ancient writings are disappointing. Sextus Julius Frontinus, who was appointed water commissioner of Rome in A.D. 97 and whose treatise on the system survives, does not mention siphons, probably because they were not prominent at Rome. Vitruvius (Marcus Vitruvius Pollio) in Book 8 of *De Architectura* does give a description that provides the only written account.

On some points Vitruvius is clear. He realizes the need for care in draining and filling the siphon, and he recommends that to resist the inertial thrust at the bends the pipes should be embedded in large masses of masonry for reinforcement. On the basic principles, however, he is confused; quite possibly he did not fully understand how a siphon worked.

Nowhere in his exposition of siphon hydraulics is the confusion more evident than in the sentence "Etiam in ventre colliviaria sunt facienda par quae vis spiritus relaxetur." Translation is simple: "In the bottom of the siphon we must put colliviaria to release the air pressure." Interpretation is almost impossible. The Latin is clear, but what it says is nonsense.

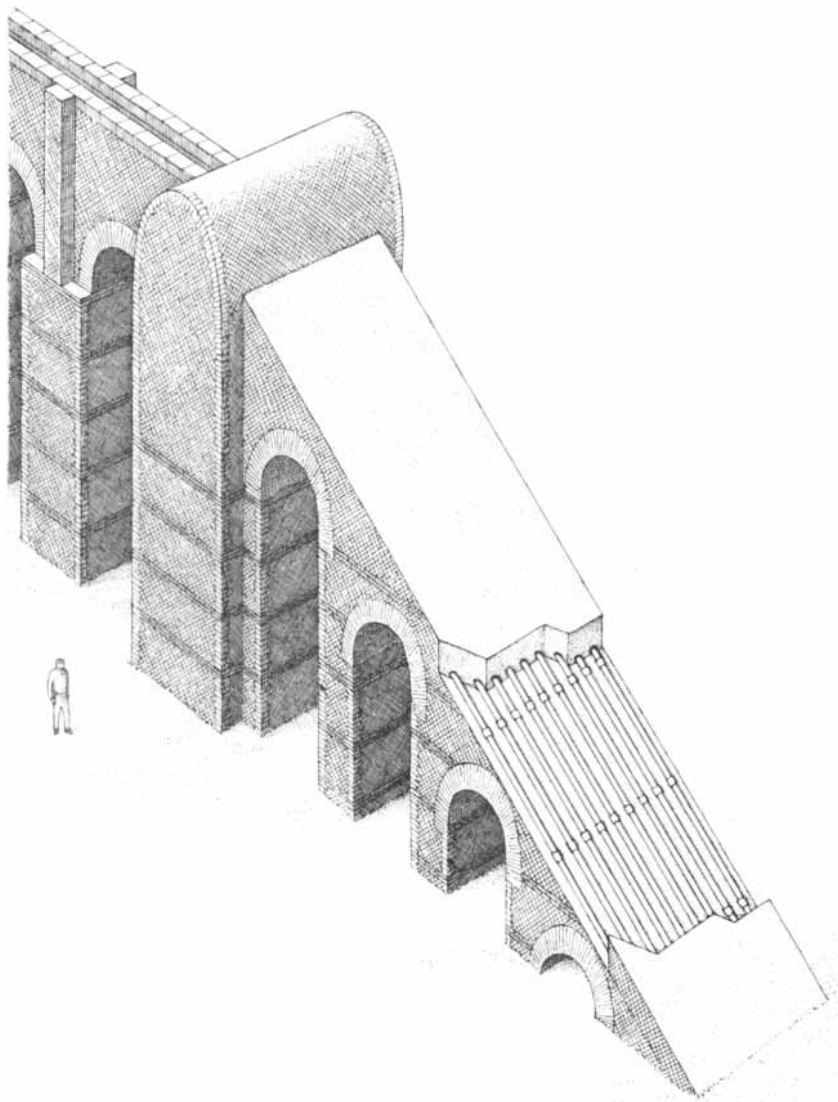
There are two problems. One is the *colliviaria*. The word does not appear elsewhere in Latin, so that there is no telling from the name what the devices were. One must guess from the context.

The second problem is the reference to air pressure. Modern pipelines are often fitted with release valves to prevent the formation of air pockets, and so it is sometimes suggested this is

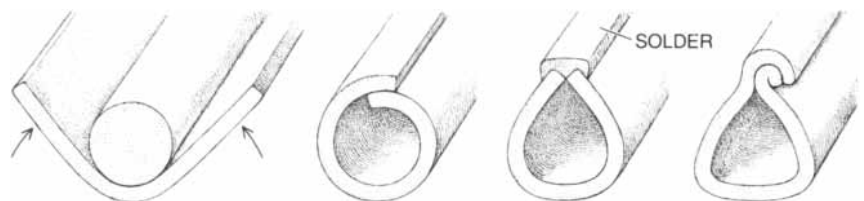
what the *colliviaria* were. The suggestion overlooks several points.

First, air pockets form at the high points of a pipeline, and that is where the modern valves are installed, not "in the bottom." Second, a Roman siphon had no air. The pipe was full of water. Even air coming out of solution will do so only under low pressure, and

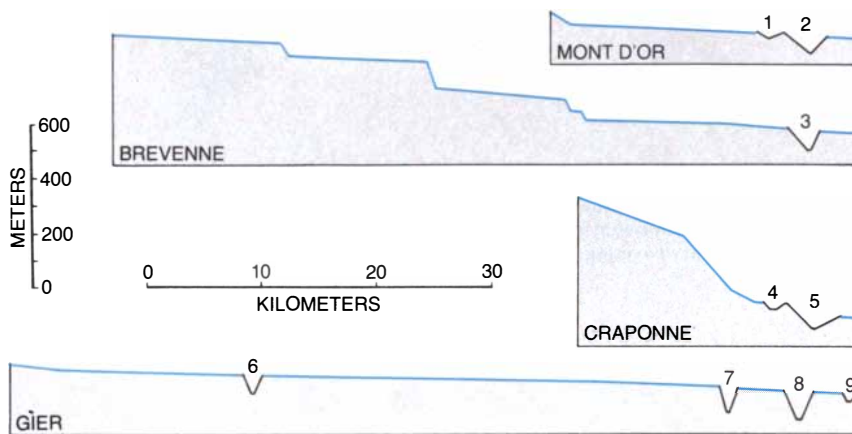
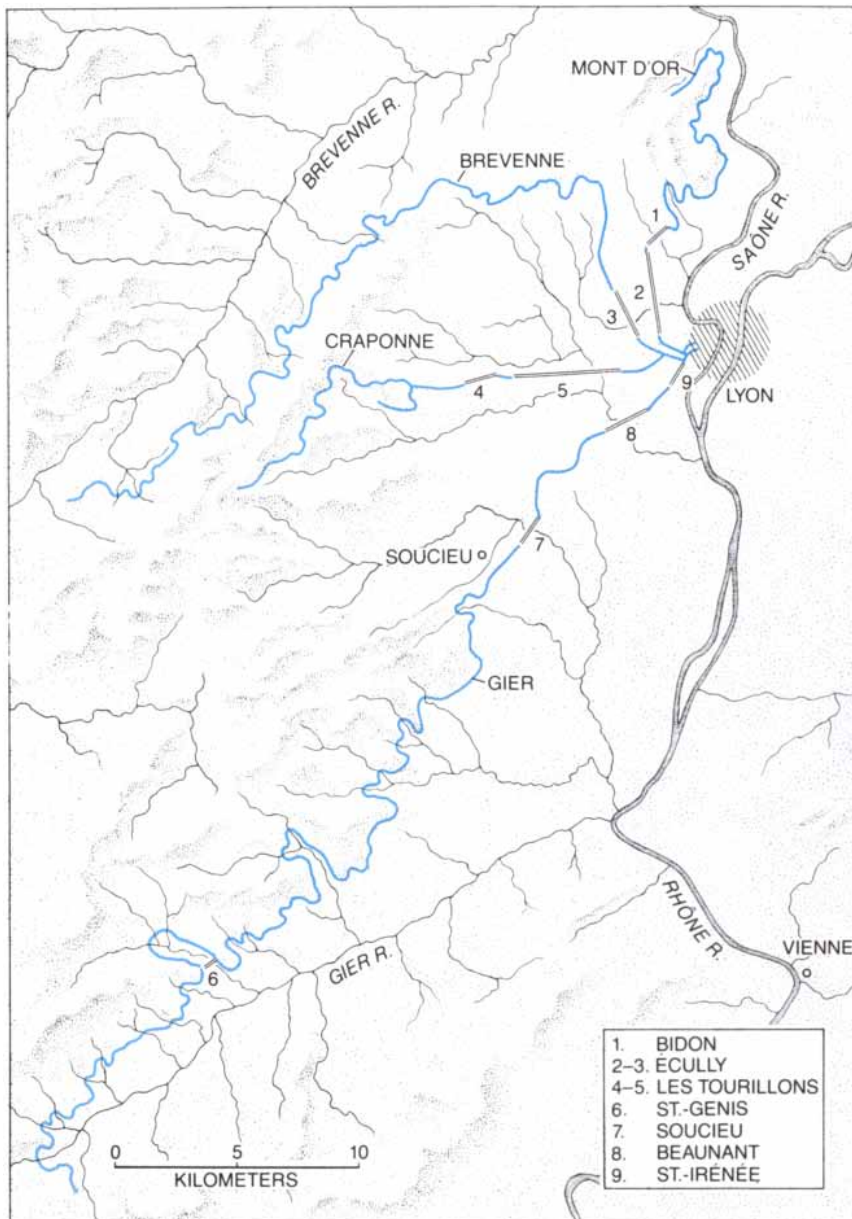
a Roman siphon was under high pressure throughout its length. Entrained air normally forms pockets by expanding under low pressure and lodging in the peaks of the siphon, not in its depths. Again the Roman siphon had the wrong shape: air entrained in the form of bubbles would simply be carried down around the bend by the cur-



**RECONSTRUCTION** of the beginning of a siphon was made by Waldemar Habery of the Rheinisches Landesmuseum in Bonn. The header tank and the pipes on the ramp were protected by a masonry covering. The pipes, three meters long, were 25 to 27 centimeters wide.



**SIPHON PIPES** were made by bending a sheet of lead around a wood core. The core was withdrawn and the joint at the top was hammered or soldered to be watertight, as is shown by the three drawings at the right. The pipes were oval or pear-shaped in cross section.



**LYON SYSTEM** consisted of four aqueducts: Mont d'Or, Brevenne, Craponne and Gier. They had a total of nine siphons; the total was determined by the number of deep valleys each aqueduct had to cross. (If a valley could be crossed by a bridge no more than 50 meters high, the Romans did it that way.) At the bottom the aqueducts of the Lyon system are shown in profile. The numbers by the names of the siphons also appear on the map to show the location of each siphon. The system delivered 80,000 cubic meters of water per day.

rent and would rise to escape at the far end. Third, since air pockets are formed by low pressure and are therefore partial vacuums, the function of the release valves is to equalize pressure by admitting air from the outside.

The absence of air in an inverted siphon has sometimes even led commentators to suppose Vitruvius must have been alluding to a valve that released not air but water pressure. This proposal is even worse. Although it would not have been beyond the ability of the ancients to devise a spring-loaded or counterweighted valve that would open under a set pressure, there would have been no way to reduce the static pressure (except by changing the course of the siphon to reduce its depth), and so the valve would have remained permanently open. It would have functioned not as a valve but as a hole in the pipe.

In sum, there was no air in a siphon and no way of reducing the water pressure. The most likely explanation of the *collivaria* is that they were drain cocks or access holes for cleaning the pipe, probably by some pull-through device. Water in Roman cities was usually hard, and so the incrustation of aqueducts was a fact of life. The pipes required constant chipping and cleaning if the channel was not to become choked. The narrow siphon pipes in particular would have collected these deposits and must have been regularly cleaned out or replaced.

The Greeks too understood and used siphons. Indeed, the best-known siphon of antiquity is certainly the notably large one at Pergamum in Asia Minor. Dating from the reign of the Hellenistic monarch Eumenes II (197–159 B.C.) and therefore clearly pre-Roman, it consisted of a single pipe three kilometers long that reached the remarkable depth of 190 meters. Water in the siphon generated a static pressure of some 19 atmospheres.

For years in the modern era this siphon clouded the issue. Because the more numerous Roman siphons were neglected and unfamiliar, the Pergamum siphon created the misleading impression that the Greeks were more advanced than the Romans in hydraulic theory and that they were even better engineers, able to make strong pipes to withstand the pressure whereas the Romans could not do so.

It is now evident that the impression is wrong. When the depths of the largest Roman siphons are compared with the height of the highest Roman bridges, it is observable that the siphons were all very large, generating a high static pressure in their pipes. This in itself disposes of the belief the Romans





**BEAUNANT VENTER** is shown from the point (rear) of the first *geniculus*, where the pipes emerged from the ground and started across the floor of the valley. The *venter* was much wider (seven me-

ters) than was necessary to carry nine small pipes; the extra width probably provided space for catwalks that gave access to workmen. Because the water was hard, the pipes had to be cleaned often.

did everything to avoid pressure. One could well argue that on the contrary, they installed pipes (as opposed to open channels or aqueduct bridges) only where pressures would be high.

A second conclusion may also be drawn. On comparing the relative heights of siphons and bridges it can be seen that there is no overlap. The siphons take over where the bridges stop, at a height of about 50 meters. Below that height the Romans crossed a valley by bridging it, above that height they built a siphon. One therefore infers they favored bridges and fell back on siphons as a second-best expedient when their engineers could not build a bridge or viaduct high enough. Evidently they thought 50 meters was about as high as a bridge could safely go.

Because the Romans built only siphons that were difficult to construct and avoided easy ones, it seems plain that whatever kept them from resorting oftener to siphons, it was not the difficulty of the engineering. The most likely reason has been proposed by Norman A. F. Smith of the Imperial College of Science and Technology in

London, who says the issue was one of economics. The plain fact is that siphons cost the Romans more than bridges. Stonework was cheap, particularly if the stones were quarried locally, and so were brick and cement. Lead was cheap too; the ancient world had a glut of it as a by-product of refining silver.

Moreover, lead poisoning apparently was not a problem, even though some modern scholars have seen in the Roman use of lead piping the cause of infertility and other aspects of the decline and fall. The Romans did know of this danger. What is more, with a continuous flow and the absence of taps the water was in the pipes for only a short time. In addition the thick crust of calcium carbonate in the pipes acted as insulation, so that the water never touched the lead after the pipe had been in use for a while.

The problem with lead was in transporting it. The cost and hard work of hauling 15,000 tons of lead to the Lyon siphons were perhaps the best arguments for not repeating the experience oftener than necessary.

In modern times the position is re-

versed, thanks to cast iron. This material did not exist in antiquity because the ancients could never attain a furnace temperature high enough to melt iron. All ancient iron was therefore wrought iron, which could not be made into pipes. Today cast-iron pipes make running a siphon cheaper than building a bridge. It is noteworthy that the French in North Africa have often brought water to coastal cities over the routes of the old Roman aqueducts, sometimes trying to renovate them, and often where the Romans spanned a valley with an aqueduct bridge the French put in a siphon.

The message is plain. The Romans built the towering arches of the Pont du Gard not simply as a grandiose display (although they no doubt relished its magnificence), nor because their engineers lacked the hydraulic insight of the Greeks. Even less was it because they could not make strong pipes. Cost was the governing factor, and it explains why time and again along the Roman aqueducts one sees brick arches where one might have admired the technical audacity of a bold, plunging siphon.

# The Topology of Mirages

*A mathematical operation called transfer mapping relates properties of mirage images to topological ideas. Certain features of mirages are thus understood without detailed knowledge of atmospheric conditions*

by Walter Tape

**T**he ship's stern resembled a giant pair of scissors. In a few minutes the scissors disappeared as the upper part of the vessel apparently split into two copies of itself, one of which floated upside down in the sky. As I watched the passing Great Lakes ore carrier from shore the images evolved from one to another, sometimes slowly, sometimes in seconds. When I raised or lowered my vantage on the shore, the mirages often changed dramatically.

Although such mirage images can be complex, many of their features can be explained rather simply by applying topological principles. If you imagine that the photographs of a ship and its mirages on the opposite page are printed on a rubber sheet, then in theory each photograph could be distorted smoothly in such a way that it looks identical with the vessel's undistorted image. Some mirage photographs would have to be distorted only slightly, but for the ones shown here it would be necessary to fold parts of the sheet into several layers.

Bearing in mind this theoretical rubber-sheet model, we can begin to make sense of the photographs. For example, the white curve attached to the ship's bow in the top photograph consists of two merged images of a straight pole; the lower image is upright and the upper one is inverted. Directly above the merged images there is a faint third image, which is straight. To make a rubber sheet bearing this photograph identical with an undistorted image of the ship, the sheet would have to be folded along the horizontal line where the two merged images meet. It would then have to be folded once more, along a horizontal line between the inverted image and the uppermost image.

The rubber-sheet model, in a more complete form, will provide a framework for using abstract topological principles to understand observed mi-

rage images. Topological methods are impressive when they are applied to mirages: they provide insight that does not require quantitative knowledge of the complex atmospheric conditions responsible for the mirage. Conversely, mirages can furnish graphic illustrations of topological ideas.

Mirages occur when atmospheric conditions cause light rays to bend as they pass from the observed object to the eye [see "Mirages," by Alistair B. Fraser and William H. Mach; *SCIENTIFIC AMERICAN*, January, 1976]. The observer mistakenly interprets the light path as a straight one and thus sees a displaced image [see *top illustration on page 122*]. If several rays pass from the same point on the object to the eye, the observer sees multiple images of that point.

**I**n the more complete version of our rubber-sheet model the deformable rubber sheet can be thought of as a piece of a small sphere, called the image sphere, which is centered on the observer's eye. The surface of the im-

age sphere bears an exact likeness of the mirage seen by an observer. Every ray passing between the eye and the object will cross the image sphere at some point. That point on the image sphere will bear an image of the point on the object that lies on the same ray. When this process is repeated for all the light rays, every point in the mirage image seen by an observer will appear on the image sphere.

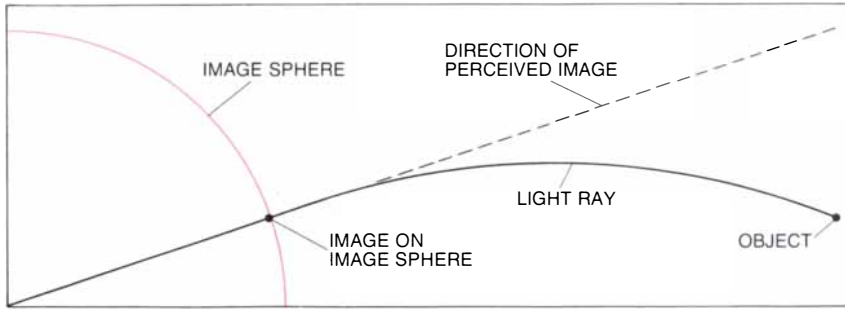
The image sphere is complemented by an "object sphere," also centered on the observer's eye, whose boundary coincides with the object being observed. The surface of the object sphere can be regarded as bearing a life-size picture of the object as it really is (that is, an undistorted picture). The object sphere need not be precisely spherical; it can be bent, so that every ray from the eye reaches it. If the object is a ship, for example, the object sphere would be flattened on the bottom so that part of its surface coincided with the water between the ship and the observer.

The paths followed by light rays be-

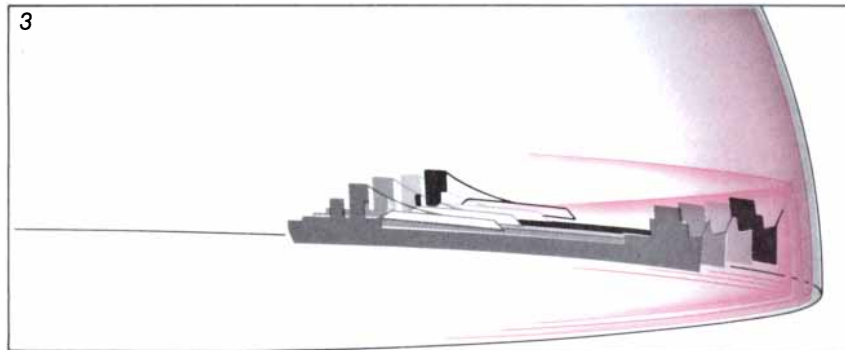
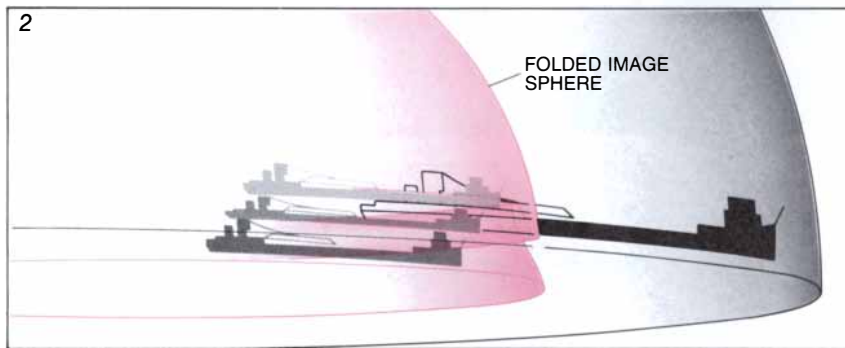
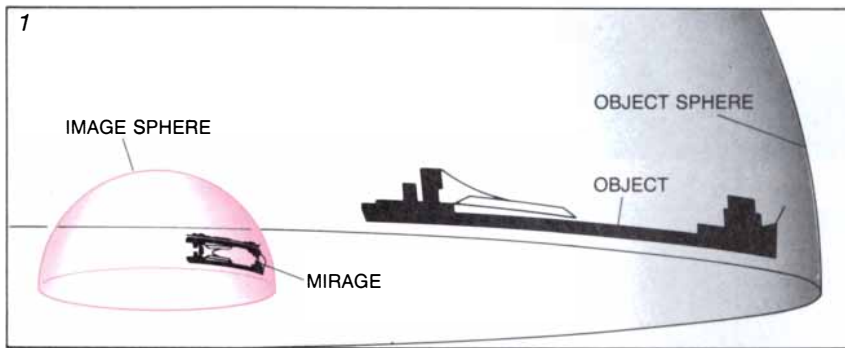


**MIRAGES** of a Great Lakes ore-carrying ship have common features that can be explained by principles derived from abstract topology. One such principle is the so-called odd-number theorem, which states that every part of the object (the ship in this case) must have an odd number of mirage images. The red letter C on the smokestack, for example, has three images in the top photograph; two images are merged and the third lies above them. Several photographs seem to have only two images of the hull. In most cases close examination reveals that the upper image actually consists of two mirror images, one greatly demagnified. The fourth photograph, however, genuinely seems to have only two images, an indication that the odd-number theorem requires qualification. Actually under certain atmospheric conditions it is invalid and under others some images are demagnified nearly to invisibility.





**CURVED LIGHT RAYS** are responsible for mirages. When a ray (solid line) curves, it enters the eye from a direction other than the direction in which the object actually lies. The observer mistakenly interprets the light path as a straight one (broken line) and thus sees a displaced image. The direction of the observer's perceived image can be represented as a point lying on a small sphere, called the image sphere, which is centered on the observer's eye.

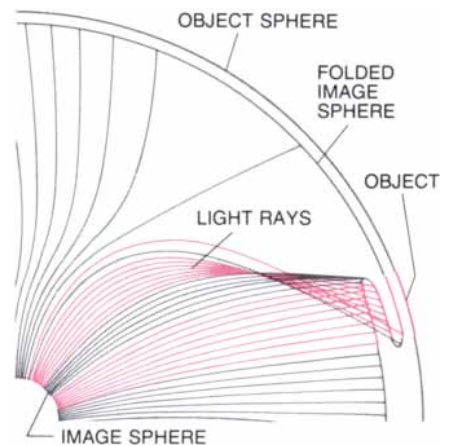


tween the eye and the object provide a way to associate each point on the image sphere with a point on the object sphere. Imagine that the image sphere is rubber and that it physically expands to meet the object sphere. Each point on the image sphere moves outward along the path of a light ray until it meets the object sphere. If no mirage conditions exist, the rays are straight and the image sphere merely expands uniformly. If the light rays are bent so that mirages form, the image sphere may develop folds or pleats as it expands, and hence there will be regions where more than one layer of the distorted image sphere clings to the same part of the object sphere.

I call this process, in which each point on the image sphere is "mapped" along a light ray to a point on the object sphere, the transfer mapping of the mirage. The transfer mapping starts with the mirage scene on the image sphere and distorts it to coincide with the undistorted scene on the object sphere.

If we imagine that the mirage on the image sphere was erased, the transfer mapping would provide a way to recreate it. Once the image sphere was distorted and pressed up against the inside of the object sphere, each section of the image sphere would be imprinted so that it looked identical with the section of object sphere it was pressed up against. If several layers of the distorted image sphere touched the same region of the object sphere, each would be impressed with the same imprint.

Then the action of the transfer mapping would be reversed: the distorted image sphere would shrink again, moving back along the light rays until it regained its original spherical shape.



**TRANSFER MAPPING** provides the basis for applying topological principles to mirages. The mirage image appears on an image sphere, which is centered on the observer's eye (1). An object sphere, also centered on the observer's eye, represents the object as it would appear without distortion. The transfer mapping expands and dis-

torts the image sphere to meet the object sphere (2); each point on the image sphere follows the path taken by a light ray (paths shown at right). As the image sphere expands, it may develop folds or pleats. In the end (3) the mirage image is distorted in such a way that it coincides exactly with the imprint of the undistorted object.

In doing so, the imprint or imprints on it would be distorted. The resulting distorted image would be the mirage.

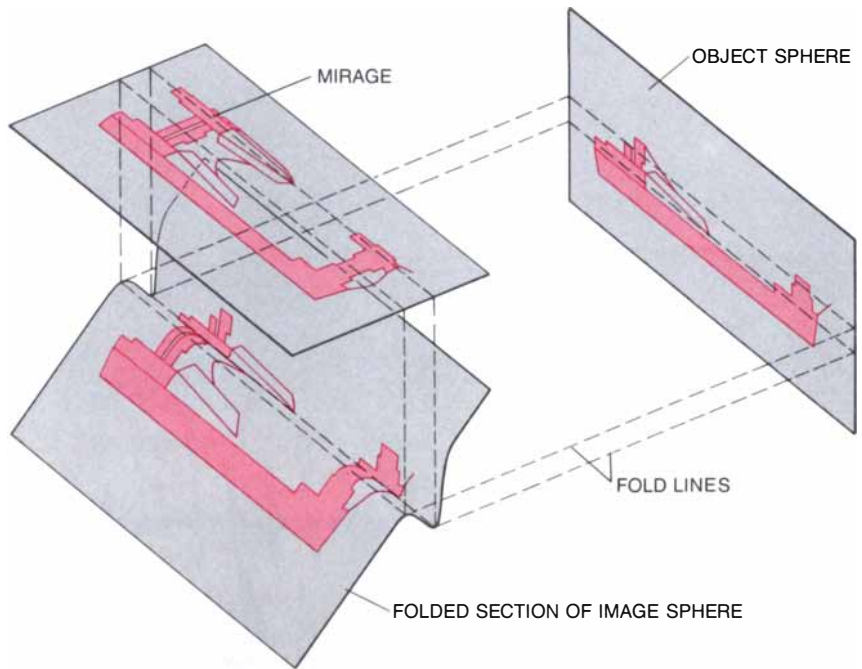
The illustration at the right shows the geometric effect of a transfer mapping that produces two folds. The positions where the fold lines fall on the object (a ship in this case) are critically related to the vessel's miraged appearance: every point between these lines will have three mirage images. The mirage shown in the illustration resembles the second photograph from the top on page 121. To construct the mirage in the top photograph the lower fold line would have to be placed slightly higher on the object. In the fourth photograph from the top the lower fold would be near the waterline. Additional folds are necessary to create the mirage in the third photograph from the top.

Using the more complete version of our rubber-sheet model, we are able to introduce topological concepts. One such concept is the "degree" of a mapping. Suppose there is a smooth mapping from one sphere to another. Each point on the second sphere (the object sphere in the case of our rubber-sheet model) may have antecedents: points on the first sphere that are mapped to it. The region surrounding an antecedent may be upside-down or backward in relation to the region surrounding the original point (as viewed from inside the spheres); in such a case the region surrounding the antecedent is said to have the opposite orientation to that of the original region. The degree of the mapping at any point on the second sphere is equal to the number of antecedents with the same orientation as the original point minus the number of antecedents with the opposite orientation.

A mapping whose degree is relatively simple to compute is the transfer mapping that occurs when no mirage conditions exist. The image sphere is not distorted at all as it expands to meet the object sphere. In this case each point on the object sphere has exactly one antecedent on the image sphere. Furthermore, the antecedent has the same orientation as the original point on the object sphere. This mapping therefore has degree 1.

A more complex example concerns a mapping that could never be the transfer mapping of an actual mirage but that nonetheless illustrates a topological theorem: The degree of any particular smooth mapping is the same in every region of the second sphere.

Suppose a small sphere is deflated and pressed against the concave surface of part of the larger sphere (a useful image is that of a deflated balloon



**FOLD LINES** produced by a transfer mapping have a critical effect on the form of the mirage image. Here the mirage of a ship (top) is folded and distorted to coincide with the image of the undistorted ship (right). Each point of the undistorted ship between the fold lines has three images in the mirage, whereas each point outside the lines has one image.

stuck to the inside of a globe). In this mapping some of the larger sphere has no contact with the small one, whereas the rest has two layers pressed against it. Suppose an arrow pointing to the left is drawn on the region of the larger sphere that is touched by the small sphere and is then transferred directly onto each layer of the small sphere. Then the copied arrows will have opposite orientations as viewed from the inside of the small sphere (once it has been reinflated): one will point to the left and the other will point to the right. Thus the mapping will have degree 0 (one "positive" antecedent and one "negative" antecedent) at points within the portion of the larger sphere that came into contact with the small one. It will also have degree 0 (no antecedents) at those points that had no contact. The degree is the same in both regions.

This property of smooth mappings is known as invariance of degree. The invariance of degree makes intuitive sense if we consider the illustration above. When either of the fold lines is moved across a point of the object, the point gains or loses two antecedents with opposite orientations.

What is the degree of a smooth transfer mapping of a mirage? In our rubber-sheet model the number of antecedents of any object is equal to the number of images of that object that are seen by an observer. Hence if

there is any region of the object sphere that appears exactly once and with its normal orientation in the mirage, then the transfer mapping must have degree 1. In practice this is always the case: most of what the observer sees is not distorted.

Because a smooth transfer mapping of a mirage must have degree 1, there are constraints on the types of image that can appear in the mirage. For example, such a mirage theoretically could not include exactly three images of a ship, all erect and headed in the same direction. In such a case the transfer mapping would have degree 3 or  $-3$  instead of degree 1.

Another such constraint is expressed as the odd-number theorem: if the transfer mapping is smooth, every object will have an odd number of images. The proof is relatively simple. The degree of any smooth transfer mapping associated with a mirage is 1. The number of positively oriented images must therefore be one more than the number of negatively oriented images. Hence the number of positive images plus negative images, the total number of images, must be odd.

The odd-number theorem is nicely illustrated in some of the photographs on page 121. In the top photograph, for example, there are two merged images of the red letter C on the ship's smokestack; these images lie directly below a third, smaller image. The third photograph from the top probably has five

Meeting Japan's Challenge

Twentieth in a Series

**THE AMERICAN  
LEAD WIDENS AS  
MOTOROLA INTRODUCES  
THE MOST POWERFUL  
MICROPROCESSOR  
ON THE WORLD  
MARKET.**

A computer may awe you with its memory, delight you with its graphics and dazzle you with its speed, but deep down inside, the brain behind it all is its microprocessor.

And our new MC68020 can outperform every other merchant market microprocessor in the world at these functions, because it is so much more powerful than any of them.

Essentially, this power will enable computers to vastly improve their capacity for work. Now, they will be able to access even greater pools of data, process them far quicker, and light up the screen with graphs, solutions, possibilities and certainties in even less time.

The MC68020 is a complete 32-bit microprocessor. That is, it can access over four billion bytes of data while providing unsurpassed data processing power.

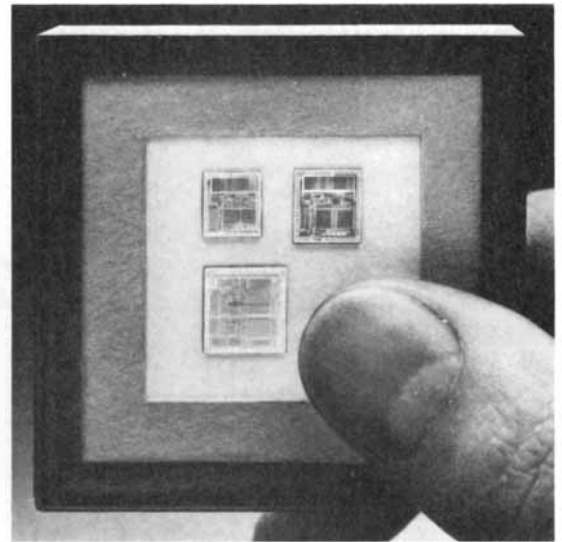
The new unit also happens to be a fully upwardly compatible version of our earlier 16-bit machine. It completes a progression of 8-, 16-, and now 32-bit members of a single Motorola family of microprocessors. This enables product designers to expand the capabilities of their hardware without the interruption of radical re-design.

The MC68020 is already destined for great accomplishments. It seems likely it will be a major factor in the market for use in the next generation of robotics. Its potential for high speed graphics and advanced mathematics make it ideal for high volume data processing, complicated computer-aided design and manufacturing (CAD/CAM) processes, and next-generation general purpose computers.

The U.S. semiconductor industry has always demonstrated a special competence in microprocessors. The advent of the MC68020 widens our country's lead with Motorola in the first rank, for units of power, speed and compatibility.

Today, it is quite safe to say that the Motorola M68000 family of microprocessors holds a prominent world position in design-ins by leading electronics firms in the U.S., Japan and Europe.

It is also quite safe to say that we are not only determined to keep that lead, but to widen the gap again.



*Our family portrait.*



**MOTOROLA** A World Leader in Electronics.

Quality and productivity through employee participation in management.





images of the *C*, although it is hard to be sure from the photograph.

Mirages and the proof of the odd-number theorem constitute a graphic application of the invariance of degree, a topological property. It seems remarkable that the proof should succeed, based as it is on the mildest assumptions regarding light rays and atmospheric conditions.

What the proof does depend on is that the transfer mapping must be a smooth (or at least continuous) mapping of the entire image sphere to the object sphere. For this condition to be met the atmospheric refractive index (the speed of light in the atmosphere) must vary smoothly and every ray that meets the eye must pass through the object sphere transversely; that is, it must not meet the object sphere at a tangent.

One class of mirages, known as the inferior mirages, illustrate the importance of this "condition of transversality." Roughly speaking, an inferior mirage is one in which light rays bend upward as they pass from the object to the eye. The photograph and illustration at the right show an inferior mirage in which a ray from a truck glances tangentially off the ground, thereby violating the transversality condition. (Here the ground forms part of the object sphere.) The small region of the image sphere just above the tangential ray is mapped to the truck, whereas the region just below the tangential ray is mapped to the ground far away from the truck.

This mapping is thus not continuous. A continuous transfer mapping is one that maps neighboring regions of the image sphere to neighboring regions of the object sphere. In order to deform the photograph so that it matched an undistorted image of the truck, we would first have to cut it along the line where the "puddle" seems to start. (There is no puddle, of course; it is merely an image of the distant mountains.) This mirage violates the conclusion of the odd-number theorem: it contains exactly two images of the lower part of the truck.

In a superior mirage (a mirage in which light rays bend downward in relation to the earth's surface), it is normally possible to choose the object sphere so that there are no tangential rays. If atmospheric conditions vary smoothly, the transfer mapping will be smooth and the conclusion of the odd-number theorem will be valid. The ship photographs in this article are of superior mirages.

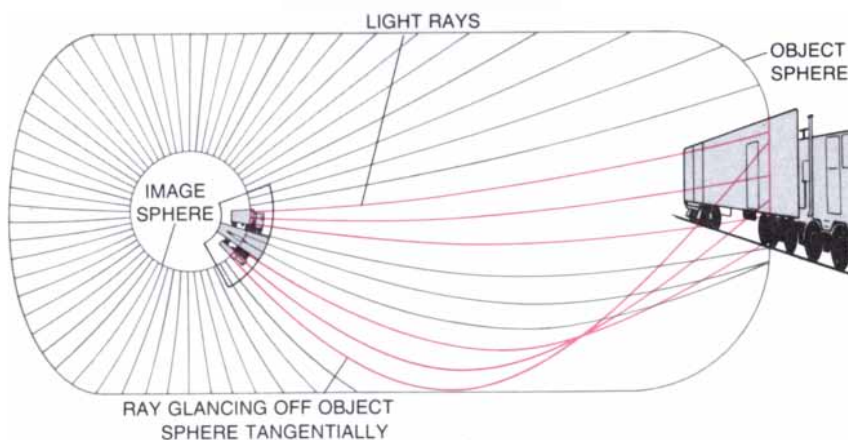
Some superior mirages, such as the fourth photograph from the top on page 121, seem to show an even num-

ber of images. In some cases the reason is that the atmospheric conditions are so chaotic that the transfer mapping is discontinuous. For well-formed superior mirages such as those shown, however, the transfer mapping should not be regarded as discontinuous on the basis of missing images. In these cases missing images are due to another phenomenon: extreme demagnification. Demagnification can be a side effect of the extreme magnification that may occur near broad folds. Such folds are the subject of a theorem proved by Hassler Whitney of the Institute for Advanced Study in Princeton.

Whitney showed that a smooth mapping from one surface to another is rarely more complicated locally (that is, in any small region) than a fold or a pleat. (This is my rough paraphrase of Whitney's more technical conclusion.) One consequence of Whitney's theorem is that close twin images in mirages must nearly always be regarded as results of a fold.

A smooth fold is more subtle than a crease. In a smooth transfer mapping any region of the image sphere that lies sufficiently close to a fold line is greatly compressed. When the compressed region is unfolded, it will expand; the resulting image will be greatly magnified at the fold line. (It sometimes happens, however, that the expanded section occupies such a small region that the magnification is not noticeable.) The top photographs on the next page show a typical example of such magnification. Fold lines can occur in inferior as well as superior mirages, because Whitney's theorem requires only that the transfer mapping be smooth in the region of interest.

Folds probably play a key role in producing the so-called Fata Morgana mirage. The Fairy Morgan (Fata Morgana in Italian) was said to be able to produce castles in the air. I believe many instances of the Fata Morgana mirage are just superior mirages in which a properly positioned broad



**INFERIOR MIRAGES**, in which light rays bend upward as they pass from object to observer, often violate the conclusion of the odd-number theorem. If a ray glances tangentially off the surface of the object sphere (*bottom*), sections of the image sphere that are close together will be associated with sections of the object sphere that are far apart. Here sections just above the tangential ray are associated with the truck, whereas sections just below it are associated with the ground far away from the truck. The odd-number theorem is inapplicable when such discontinuities occur: there are two images of the lower part of the truck.



**GREAT MAGNIFICATION** occurs in mirages near fold lines that are associated with broad folds. Here a mountain range (*top*) is distorted by a fold to look like a palisade of cliffs (*bottom*). Most of the close twin images that appear in mirages are due to such folds.

fold has made a slightly uneven horizon look like "landscapes with towns and towers and parapets." Slight undulations in the fold lines could produce similar results even if the horizon's profile displayed no irregularities.

Intuition suggests that some images

in a superior mirage must inevitably appear compressed to make room if there are any expanded images or multiple images. This is our explanation for the tiny or missing images in some of the photographs.

To make this intuitive idea more

precise, it would be helpful if we could directly compare the size of an undistorted image with that of a mirage. In order to do so, imagine the transfer mapping distorts the image sphere but does not expand it. Visualize the transfer mapping as starting with a rubber



**GREAT DEMAGNIFICATION** results when a mirage contains a large number of multiple images. Demagnified images are horizontal spikes in this distorted view of a mountain range (not the range that is pictured above). Each spike probably consists of two squashed images of part of the range. This mirage, which resembles towers in the air, is a so-called Fata Morgana mirage, named after the Fairy Morgan of Arthurian legend, who created airborne castles. Such mirages appear when fold lines fall on an uneven horizon.

sphere that uniformly covers a rigid globe. Both the globe and the rubber sphere bear images of the mirage. The transfer mapping distorts and rearranges the material of the sphere; the resulting distorted sphere still clings to the globe, perhaps with multiple layers in some places. Because the degree of the transfer mapping is 1, the distorted sphere must still cover the entire globe. The uniform rigid globe still bears an image of the mirage; the surface of the flexible sphere has been distorted to match an undistorted view of the object. Because the distorted sphere clings to the surface of the globe, both have the same radius.

To calculate the magnification at a given point of the mirage, select a small region surrounding that point on the surface of the uniform sphere; then apply the transfer mapping to that region to get the transformed region (that is, the corresponding region of the distorted rubber sphere). The original region bears a mirage image of some part of the object, and the transformed region represents an undistorted view of the object. The magnification at that point of the mirage is the ratio of the area taken up by the region of the uniform sphere (which indicates how large the object appears in the mirage) to the area taken up by the corresponding region on the distorted sphere (which indicates how large the object would appear undistorted).

Demagnification at a point is the reciprocal of magnification. That is, it is the area of a part of the distorted rubber sphere divided by the area of a part of the uniform sphere. Hence the average demagnification of a transfer mapping (the demagnification averaged over all points on the image sphere) will be the total area of the distorted rubber sphere, including the area of any multiple layers, divided by the area of the uniform globe.

If no multiple images occur in the mirage, there will be no multiple layers. Both spheres will have the same area and the average demagnification will be 1. This confirms our intuition that if one area of the mirage appears expanded, some other area must appear contracted.

If there are multiple images, the average demagnification will increase, owing to the areas of the corresponding multiple layers. This confirms the intuitive notion that parts of a mirage must contract to make room for multiple images. Thus the theory predicts that there will be great demagnification in places. A missing image is therefore not necessarily an indication of a discontinuous transfer mapping. Nevertheless, the pragmatic impact of

the odd-number theorem (as well as the impact of the conclusion that mirage transfer mappings have degree one) is diminished by such instances of extreme demagnification.

Magnification and demagnification are particularly noticeable in a mirage whose transfer mapping is almost entirely dependent on height and does not depend on the horizontal position of the object. In such cases most of the distortion may be confined to several narrow horizontal bands of the image sphere. Any magnification or demagnification in the mirage is vertical; if there are multiple images, they must be compensated by vertical contraction within the horizontal band. Hence the mirage may well appear cramped or tightly packed (as in the bottom photograph on the opposite page), particularly if there are many images of the same object. I believe the multiple horizontal spikes sometimes seen on the rising or setting sun may be multiple images of parts of the sun that have been highly demagnified.

Although I have been dealing entirely with terrestrial mirages, most of the ideas I have described can also be applied to cosmic mirages, which occur when massive objects act as gravitational lenses, bending the light rays that come from distant stars or other objects. Perhaps it is not surprising that the odd-number theorem was proved by an astronomer, William L. Burke of the Lick Observatory. Because of the odd-number theorem, gravitational lensing is a less likely explanation of similar quasar images in cases where an even number of similar images is detected.

Most of the conclusions I have reached here rely on two assumptions: that the atmospheric refractive index is a smooth function of position and that rays from the eye meet the object sphere transversely. Naturally, if one of the assumptions fails to be satisfied, as transversality does in the case of many inferior mirages, the conclusions may also fail. Moreover, we have seen that gross demagnification may in practice invalidate predictions made on the basis of the odd-number theorem. Nevertheless, a topological approach provides a useful way to interpret a mirage as a whole. It also gives us new ways to understand some commonly occurring mirage features, such as the extreme demagnification often seen in mirages having multiple images of the same object. Such rules as the odd-number theorem can be helpful as indicators that there are small additional images or irregularities in the atmosphere, even if they cannot guarantee what will actually be seen.

# The Fitness Master LT-35

## TOTAL SYSTEM FOR CARDIOVASCULAR FITNESS.

Simulates cross-country skiing, regarded by fitness authorities as the top cardiovascular exercise. Rated higher than jogging, swimming, biking or rowing.



- Fluid motion — no jarring impact on bones and joints. Avoids running related injuries.
- Excellent for weight control and body tone.
- Stable, unit rests flat on floor. Lightweight for portability. Weighs only 35 lbs.
- Easily stored. Folds to 5 inch height. Slips under a bed or stands upright in closet.
- Can be used by men and women regardless of size or weight. Height and resistance adjustable.
- 30 day home trial. 2 year warranty.

For a Free Brochure call:  
**TOLL FREE 1-800-328-8995**  
 In MN 1-612-474-0992 Mon-Fri 8am-5pm  
**Fitness Master, Incorporated**  
 1387 Park Road Dept. E  
 Chanhassen, Minnesota 55317

## Have you Heard?

Where ancient tradition and modern expression share one roof.

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

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

Where life becomes culture becomes art becomes life.

Have you Heard?

The Heard Museum • Phoenix



# THE AMATEUR SCIENTIST

## *How the sun's reflection from water offers a means of calculating the slopes of waves*

by Jearl Walker

When the sun sets over open water, its reflection can be oval or columnar. What determines the shape? Lorne Whitehead of TIR Systems, Ltd., in Vancouver has looked into the matter and reports (in a manuscript he sent to me) that the shape is related to the maximum slope of the waves on the water. By photographing the reflection one can compute the maximum slope without having to go out on the water.

To follow Whitehead's work I examine first what determines the height and width of the reflection region you observe. For simplicity assume that the sun is a point source of light. (It actually occupies about .5 degree of arc in the field of view.) Assume also that the water surface is smooth.

Suppose the water surface is flat. Since the sun is so distant, the rays arrive at the water parallel to one another. They reflect from the water and some of them reach your eye. You perceive a reflection of the sun as if the surface of the water were a mirror laid flat in front of you.

The reflection of a light ray from a surface is usually described in terms of angles measured with respect to what

is called the normal: a line perpendicular to the surface at the point of reflection. The ray is incident at a certain angle measured with respect to the normal. The reflected ray forms the same angle. This rule applies even if the surface is tilted or curved. Whenever a ray reflects from the surface, mentally construct a normal perpendicular to the surface at that point.

When the water surface is flat, a single spot on the surface reflects rays toward you. One way to determine where that spot is in your field of view is to imagine that you are viewing the scene without the benefit of depth perception, as if you were looking at a photograph.

This flat representation is called a projection plane [see bottom illustration on opposite page]. The horizon lies across the plane. The spot reflecting light rays to you lies below the horizon on the projection plane. That is where you see the sun's image. When the water surface is flat, the image on the projection plane is as far below the horizon as the sun is above it. As the sun sets the image rises toward the horizon. As the sun dips below the horizon the image disappears at the horizon.

When the water surface is curved by smooth waves, the sides of the waves reflect the rays according to the simple rule for the angles in a reflection: Now many spots on the water surface reflect rays to you, however, and their locations constantly vary as the waves reshape the water surface while moving in every possible direction. At any instant you see many images of the sun on the projection plane. The reflection region associated with a setting sun is the composite of those many images.

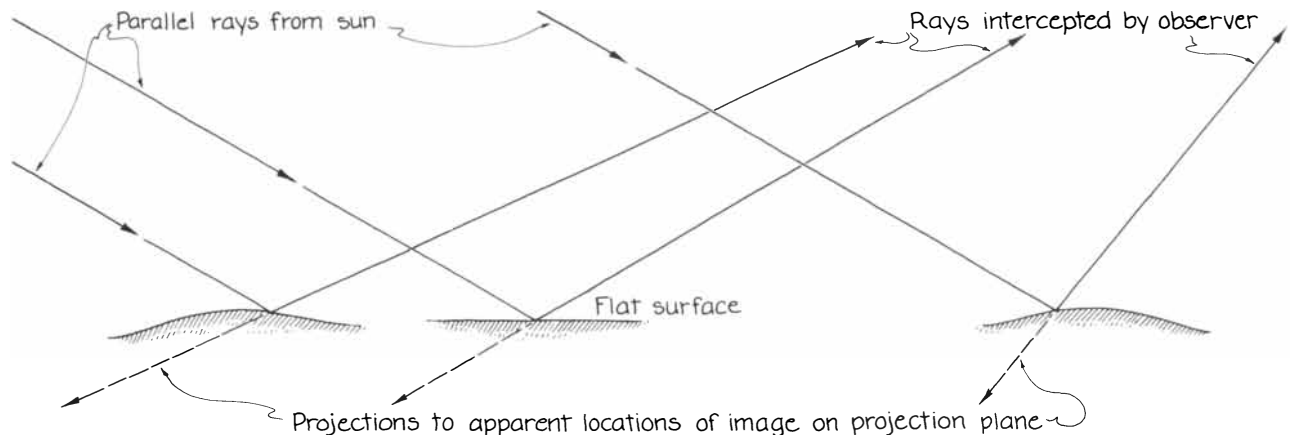
In order to determine how the waves spread the images of the sun I first investigated how an image moves on a projection plane when a reflecting surface is uniformly tilted in one direction. Having taped a pencil upright on a table, I laid a flat, rectangular mirror in front of the pencil and positioned my eyes near the tabletop. The far edge of the mirror, which was at the base of the pencil, functioned as the horizon.

Ignoring any sense of depth, I imagined the pencil and its image to be on a projection plane. To measure distances on the plane I used a transparent ruler, keeping it and my head stationary while I noted where on the ruler various parts of the projection plane were aligned. For example, in my field of view the pencil point was three centimeters above the far edge of the mirror and the image of the point was three centimeters below it, according to the rule for flat surfaces.

I then raised the near edge of the mirror, keeping the far edge on the table. The spot reflecting the pencil point moved toward me along the surface of the mirror. On the projection plane the image of the point shifted downward.

Next I tilted the mirror toward me. This time the spot reflecting the pencil point moved along the surface away from me. On the projection plane the image of the point moved upward.

My demonstration is not entirely



*Reflections of the sun's rays from water.*

faithful to the reflection of rays from the sun by a water surface because the rays arriving at the mirror are not parallel. Still, the general shift of the image on the projection plane is similar. The image of the sun moves downward on the plane as the water surface tilts away from the observer and upward when the surface tilts toward the observer. The slope of the surface determines the extent of the shift.

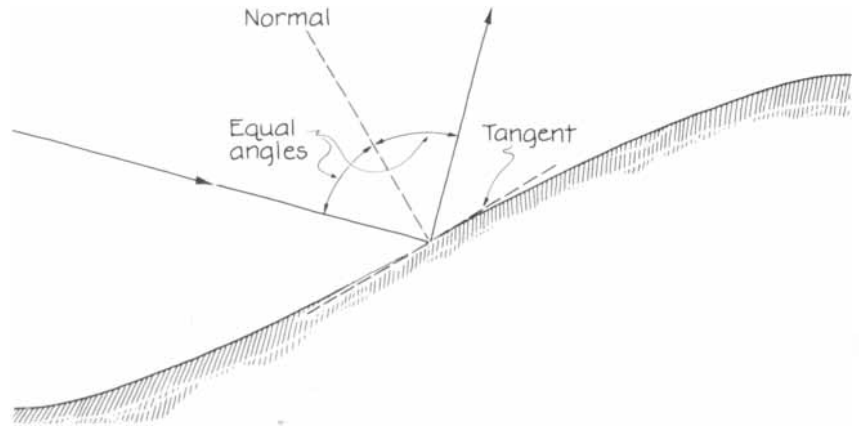
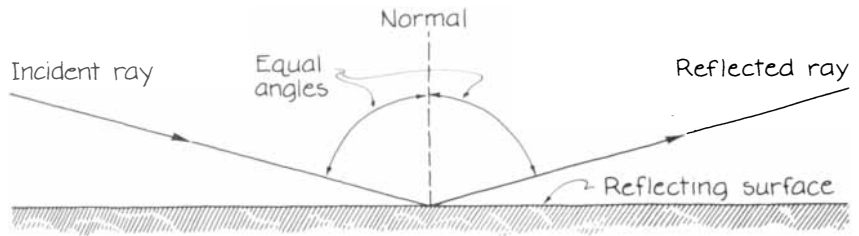
The slope can be defined as the tangent of the angle through which the surface is tilted. For example, if the tilt of the surface is 16.7 degrees with respect to the horizontal, the slope is .3 (a dimensionless number). The slope at a point on a curved surface is that of a line drawn tangent to the surface at that point.

A vertical axis on the projection plane can also be marked off in terms of slope. For example, suppose the sun is 38.7 degrees above the horizon in your field of view. On the projection plane the horizon is at zero height and the sun is in a position that is the tangent of 38.7 degrees, or .8, above it. If the water surface is flat, the image of the sun is .8 below the horizon on the projection plane.

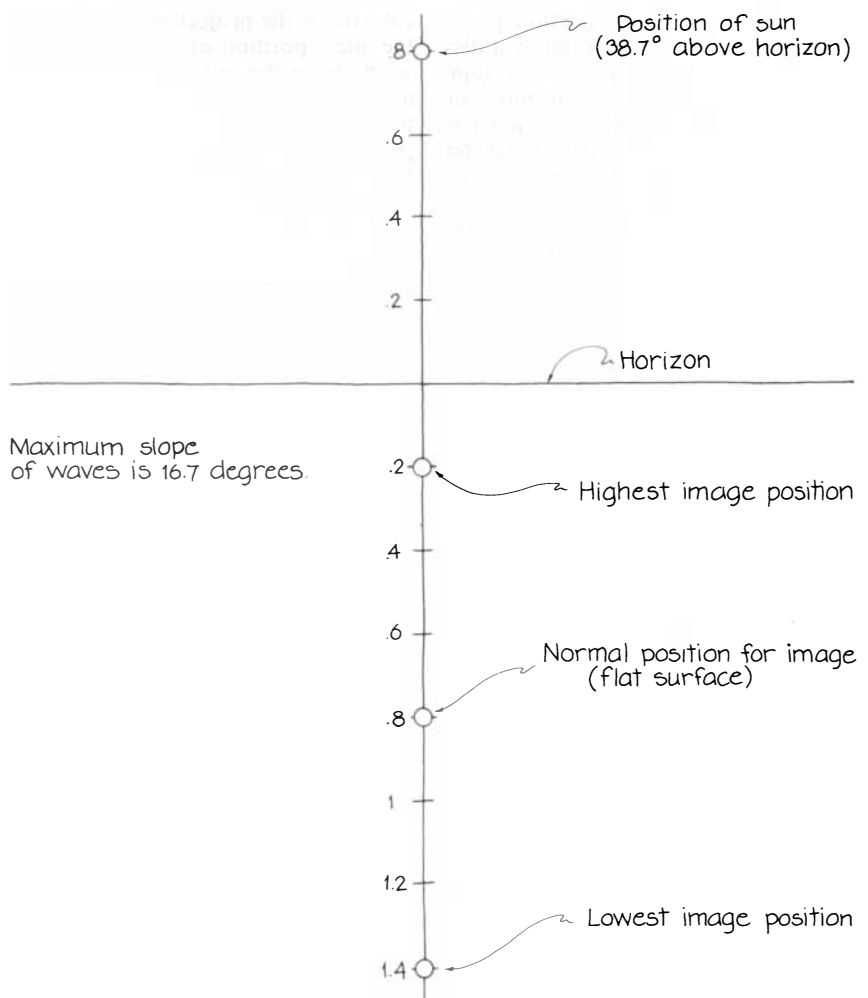
Imagine that the water surface is initially flat and then uniformly tilts away from you until it has a final slope of .3. The image of the sun moves downward on the projection plane by twice the final slope of the surface, that is, by .6. If instead the water surface uniformly tilts toward you with a slope of .3, the image of the sun moves upward on the projection plane by twice the slope of the surface, or .6.

A natural water surface is seldom tilted uniformly. Usually it has a range of slopes because of waves. Whitehead's point is that the maximum slope of the waves determines the dimensions of the sun's reflection region on the water. To follow his argument first assume that the water surface is flat and the sun and its image are .8 above and below the horizon on the projection plane. Now examine how the waves alter the reflections from the spot that creates the normal image of the sun. Sometimes the water surface there is flat and you see the image. At other times the waves tilt the surface and eliminate the image. Thus the image appears and disappears as waves sweep through the corresponding spot on the water.

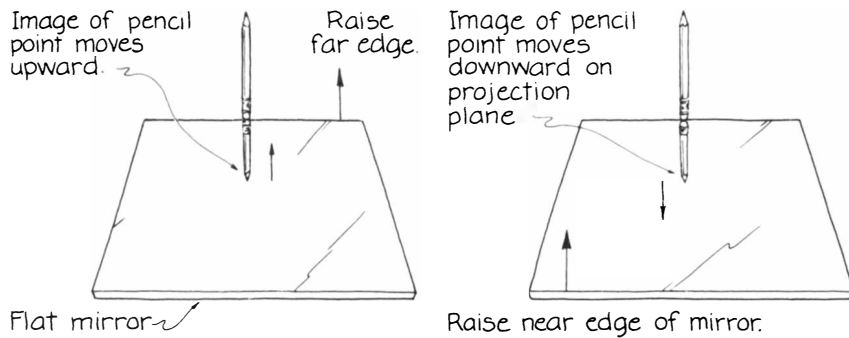
Next imagine a spot on the surface that is slightly higher on the projection plane. If the surface were constantly flat, you would never see the sun's image, but since waves move through it, the spot sometimes tilts toward you enough so that you can see a reflection of the sun. Again the image ap-



The angles encountered in a reflection



The vertical extent of the luminous region



Experiments with reflections from a tilted mirror

pears and disappears as waves move through the spot.

Now think of a spot on the water whose image on the projection plane is lower than the normal image. Sometimes the water surface tilts away from you enough for you to see the sun's image there.

How far above and below the normal position can you see images of the sun? The limits are set by the maximum slope of the largest waves. Suppose the waves have a maximum slope of .3 (16.7 degrees). The highest image position on the projection plane is above the normal position by a distance that is twice the maximum slope of the waves, that is, by a distance of .6. Since in my example the normal position is .8 below the horizon on the projection plane, the image position is .2 below the horizon.

Similarly, the lowest image position is below the normal position by a distance that is twice the maximum slope of the waves, placing the image at a position of 1.4 on the vertical axis of the projection plane. In this example you can see images of the sun throughout the range from .2 to 1.4 on the vertical axis. The images are not constant, because the slopes of the water surface constantly change. Since the visual system averages brightness over time, you perceive a fairly constant illumination. A photograph made with a high shutter speed, however, would reveal areas within the region where there is no image of the sun.

As the sun sinks, the reflection region on the water moves toward the horizon. In terms of the projection plane the region ascends. In my example the far end of the reflection region reaches the horizon when the sun has a slope of .6 in your field of view, which is about 31 degrees. Eventually the region disappears into the horizon.

The vertical extent of the reflection region on the projection plane depends on the maximum slope of the waves. If the maximum slope increases, the vertical extent increases by twice as much on the projection plane. This rela-

tion enables you to measure the maximum slope of the waves. Photograph the scene before the reflection region reaches the horizon. Also measure the angle between the horizon and the sun. Place over the photograph a sheet of clear plastic on which the projection plane is to be marked. Trace the horizon and mark the sun as a point.

Compute the tangent of the sun's angle with respect to the horizon. Suppose the angle is 38.7 degrees (a tangent of .8) when you make the photograph. Use this result to scale the vertical axis on the projection plane. Mark the sun's position on the photograph as .8 above the horizon. With a ruler measure the number of centimeters between the horizon and the sun in the photograph. Suppose the distance is four centimeters. Each centimeter on the ruler corresponds, then, to a distance of .2 on the vertical axis on the projection plane. Scale the axis with this relation.

Next determine the extent of the reflection region along the axis. Suppose it stretches from .5 to 1.1. Since the sun is .8 above the horizon, the normal position of the sun's image must be .8 below the horizon. The lowest point of the reflection region is therefore .3 below the normal position. Since the distance along the axis from the normal position to the lowest point of the region is twice the maximum slope of the waves, the maximum slope must be .15, which corresponds to 8.5 degrees.

Whitehead has also explained the width of the reflection region on the projection plane. To follow his explanation I returned to my pencil and mirror. This time I placed the right-hand edge of the mirror along a line between me and the base of the pencil, with the far-right corner of the mirror touching the base. When I tilted the mirror to the right, the image of the pencil on the projection plane rotated about the base of the pencil by an angle that was twice the tilt of the mirror. When the tilt of the mirror was 45 degrees, the image rotated through 90 degrees and was horizontal on the projection plane.

The image of the pencil point can be found on the projection plane by superposing a line that extends across the plane from the image to the pencil point. The line tilts from the vertical axis by as much as the mirror is tilted. Also superposed on the projection plane is the far edge of the mirror. On the plane the image is as far from the edge of the mirror as the actual pencil point is.

Whitehead employs an analogous construction on a projection plane to determine the width of the reflection region on the water surface [see middle illustration on opposite page]. Suppose the waves have a maximum slope of .3 (16.7 degrees). On the projection plane draw through the position of the sun a line tilted to the right of the vertical by 16.7 degrees. Extend the line toward the bottom left of the projection plane.

Now draw a second line that is perpendicular to the first one and passes through the point on the horizon directly below the sun. This line corresponds to the far edge of the mirror. Along the first line measure the distance between the sun and the intersection of the first and second lines. Starting at the intersection, measure off an equal distance along the first line toward the bottom left of the projection plane. The point you reach is the position of the sun's image in a mirror tilted to the right by 16.7 degrees.

Repeat the entire procedure on the right-hand side of the sun. You now have two lines that diverge toward the bottom of the projection plane. The reflection region lies between those lines and coincides with them at the points corresponding to rightward- and leftward-tilted mirrors. The angle between the lines is twice the maximum slope of the waves.

If you add to the projection plane the highest and lowest points of the reflection region according to my previous explanation, you can sketch the entire reflection region by drawing a smooth curve connecting the extreme points [see illustration at right on opposite page]. If the maximum wave slope increases, the angle between the diverging lines becomes larger and the reflection region grows both wider and higher on the projection plane. When the sun is high, the region is oval, smaller in width than in height. As the sun descends, the region slips toward the horizon, becoming narrower because it is constrained by the diverging lines. The oval may then seem to be a long column that stretches toward the horizon. Because Whitehead and I have treated the sun as a point source of light, the upper end of the column should be a point. Since the sun actually occupies about .5 degree in your

field of view, however, the upper end of the column cannot be narrower than .5 degree.

The fact that the column has unequal width and height is surprising when one recalls that the reflections are from waves moving in every direction. The column is also surprising in seeming to be long. Actually, however, it occupies only a small angle on the projection plane. Some of its apparent length derives from a perception of depth in the scene: the column looks as though it stretches over a long distance to the horizon. Part of the illusion is also due to a misinterpretation of angles near the horizon: the low sun and the upper end of the column look large because the horizon seems to be distant. (The same phenomenon accounts for the apparent enlargement of the moon when it is near the horizon.)

You might enjoy studying reflection regions produced by the sun or the moon under other circumstances. In some cases you might be able to detect surface currents on the water if the maximum slope of the waves in a current differs from the slope of the waves in the surrounding water. Does the reflection region then become asymmetric or otherwise distorted? What happens to it when the waves move uniformly in one direction? How does the statistical distribution of wave slopes alter Whitehead's calculations of the shape of the reflection region? What

happens to the shape of the region and the distribution of the light when the waves become so large that they are no longer sinusoidal?

Last August I described a number of arrangements for stacking dominoes. One of them called for a series of dominoes to be laid one on another so as to extend beyond the edge of a table. Each domino should have a broad face down, with its long dimension perpendicular to the edge of the table. The aim is to find the maximum extension of the overhanging stack for a given number of dominoes.

Several readers (Eugene Wall of the Aspen Systems Corporation in Rockville, Md.; R. F. Tindall of Cambridge, England; David Callway of Fort Collins, Colo., and Wayne Fullerton of Houston) pointed out that the mathematical series I gave to describe the overhang can be written in terms of Euler's constant (.57722). The horizontal distance from the table's edge to the outer edge of the  $n$ th domino is half a domino length multiplied by the sum of Euler's constant and the digamma function for  $n + 1$ . (The digamma function is the logarithmic derivative of the common gamma function.) When  $n$  is large, the digamma function can be approximated as the subtraction of  $1/(2n + 2)$  from the natural logarithm of  $n + 1$ .

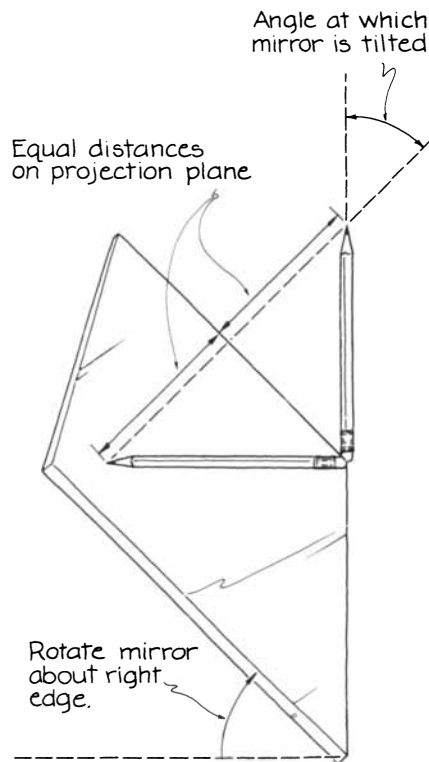
When formulas derived from that

one are employed to compute the number of dominoes needed for an overhang of 50 domino lengths, one finds that  $1.5 \times 10^{43}$  dominoes are needed, not  $1.5 \times 10^{44}$  as I stated.

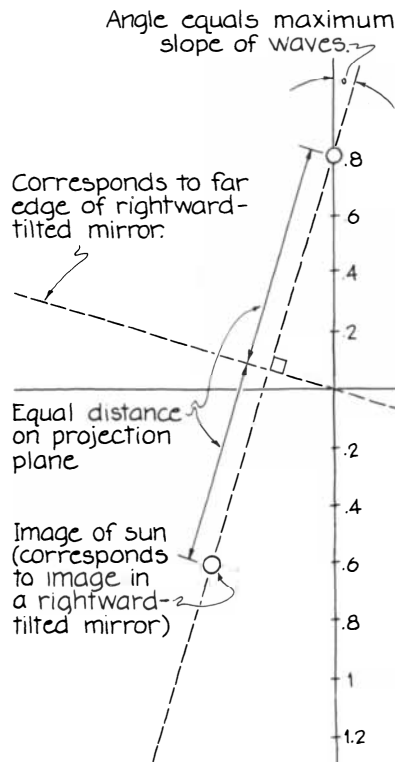
Is there a way to stack dominoes so as to build a large overhang more economically? Tindall and Hans-Hellmuth Cuno of Waldetzenberg in West Germany suggested that a given overhang can be achieved with fewer dominoes by counterbalancing. In their scheme each new domino is positioned so that its midpoint is above the outer edge of the domino just below it. Such an arrangement would of course make the lower domino unstable. Stability is restored by putting a counterweight domino on the inner edge of the lower one [see top illustration on next page].

For example, to achieve an overhang of three domino lengths, align six dominoes in steps from the edge of the table, each with its midpoint over the outer edge of the domino (or table) just below it. To provide counterbalance 31 dominoes are stacked above the inner edge of the domino on the table, 15 above the second domino, seven above the third, three above the fourth and one above the fifth. The sixth and outermost domino requires no counterbalancing. All told 63 dominoes are required for this arrangement, easily beating the 227 dominoes required in my stacking scheme.

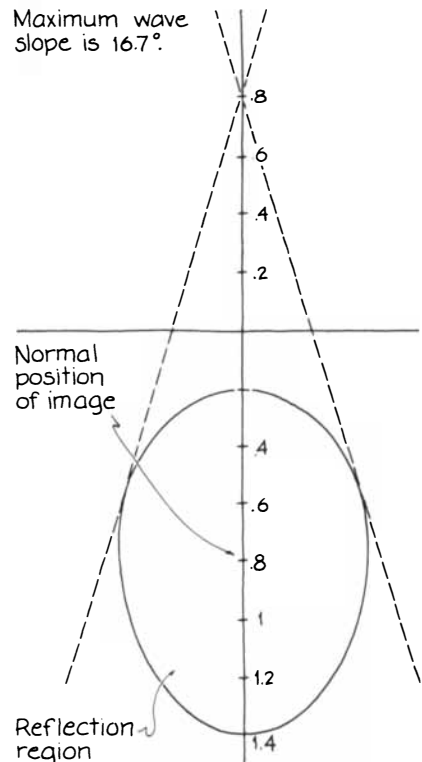
E. James Morton of the John Han-



Mirror rotation



Tilting a surface to the right



Reflections of a high sun

# 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. residents add sales tax.)

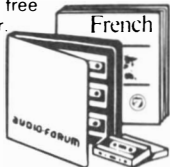
**TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: 1-800-243-1234.**

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.

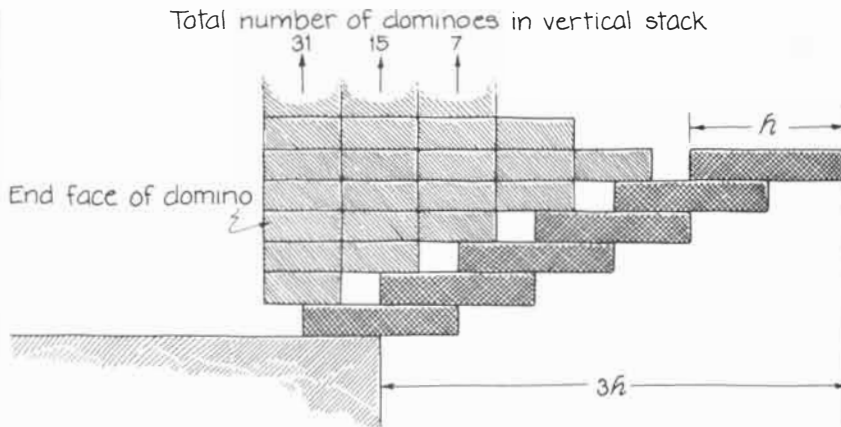
**The Foreign Service Institute's French course is unconditionally guaranteed.** Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn French, return it and we'll refund every penny you paid. Order today!

120 courses in 41 other languages also available. Write us for free catalog. Our 12th year.

**Audio-Forum**  
Room G25  
On-the-Green,  
Guilford, CT 06437  
(203) 453-9794



**AUDIO-FORUM**



*A method of counterbalancing dominoes*

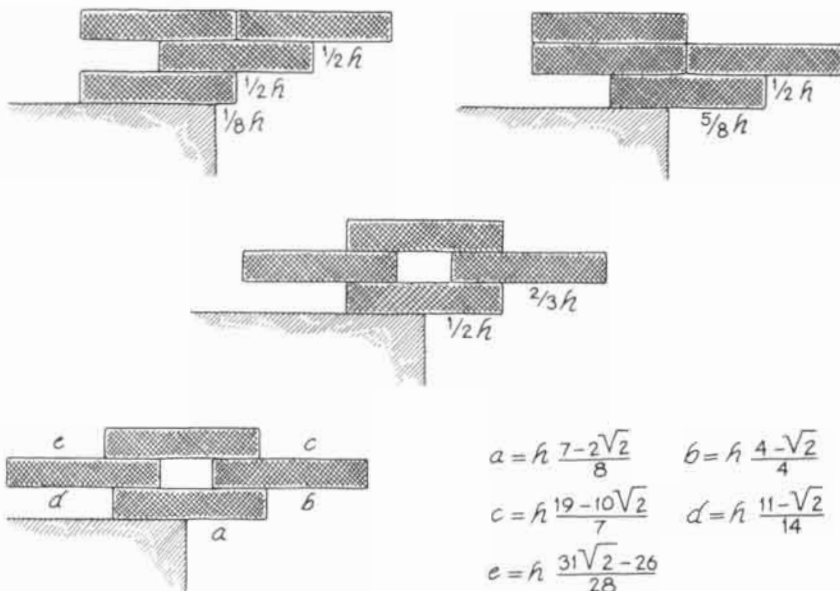
cock Mutual Life Insurance Company in Boston pointed out that some clever stacking schemes for four rectangular objects such as dominoes have been published by Stephen Ainley (*The Mathematical Gazette*, Vol. 63, No. 426, page 272; December, 1979). Each domino's long dimension must be perpendicular to the edge of the table. If the dominoes are stacked in the way I described last August, their maximum overhang is  $1\frac{1}{24}$  times their long dimension. The first two of Ainley's schemes, which are depicted in the illustration below, generate an overhang  $1\frac{1}{8}$  times the long dimension. (The fractions shown are in terms of the domino's long dimension.) One or two dominoes serve as a counterbalance.

In a third arrangement counterbalancing is achieved by a domino on the inner edge of the lowest domino and another domino on top of the stack.

The top domino is positioned directly above the one at the bottom of the stack. This scheme yields an overhang  $1\frac{1}{6}$  times the long dimension.

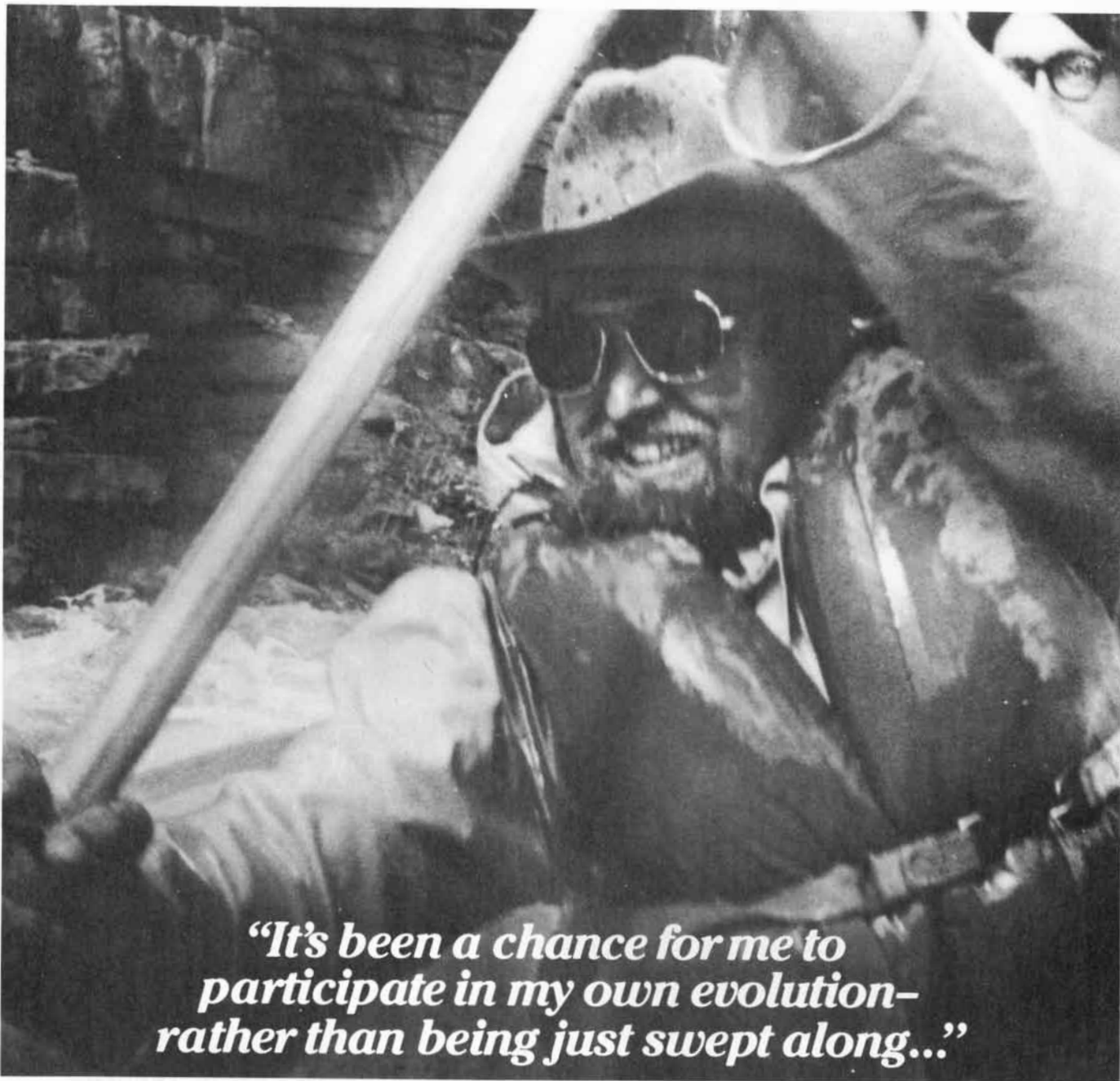
The winner of the balancing act is the last arrangement shown in the illustration. Note that the topmost domino is not exactly above the bottom one. This arrangement gives an overhang approximately 1.1679 times a domino's long dimension, barely beating out the previous arrangement.

Readers interested in such puzzles may enjoy a note published by R. E. Scraton (*The Mathematical Gazette*, Vol. 64, No. 429, pages 202-203; October, 1980). He shows how a complex stacking of 28 dominoes, each two inches by one inch by half an inch, can be built with an overhang of almost eight inches. If the stack is rotated so that the diagonals of the dominoes are at right angles to the table's edge, the overhang is almost 8.9 inches.



*Stephen Ainley's arrangements for stacking four dominoes*





*“It’s been a chance for me to participate in my own evolution—rather than being just swept along...”*



*“Back-home decisions don’t affect me as much— or as quickly —as they do out here.”*

**A**t Outward Bound® it’s not just what you do, but how you feel about it afterwards that counts.

Our courses are tough—they’re meant to be—but not beyond the reach of anyone who tries. They’re fun. And safe as human ingenuity and experience can make them.

At Outward Bound we teach you outdoor skills. From white-water rafting to rock-climbing. But we’re also something of a course in self-reliance (a course in self-reliance where you also have to learn to trust *the group*).

Outward Bound will *not* teach

you to be a man. Nearly half the students, for one, are women. Few are really athletes. Lots are over thirty. What you need is a bit of pluck, and the yen to spend some time in some of this country’s most spectacular settings.

Everyone brings something different to Outward Bound and takes something different away. But whatever your experience—we guarantee it won’t be trivial.

*We’re*  
**Outward Bound!**



For information and brochure:

**800-243-8520**

# BIBLIOGRAPHY

Cairns-Smith. Cambridge University Press, 1985.

## THE SOCIAL ECOLOGY OF CHIMPANZEES

IN THE SHADOW OF MAN. Jane van Lawick-Goodall. Houghton Mifflin Company, 1971.

THE GREAT APES. Edited by David A. Hamburg and Elizabeth R. McCown. The Benjamin/Cummings Publishing Company, 1979.

POPULATION DYNAMICS DURING A 15 YEAR PERIOD IN ONE COMMUNITY OF FREE-LIVING CHIMPANZEES IN THE GOMBE NATIONAL PARK, TANZANIA. Jane Goodall in *Zeitschrift für Tierpsychologie/Journal of Comparative Ethology*, Vol. 61, No. 1, pages 1-60; January, 1983.

THE CHIMPANZEES OF KIBALE FOREST. Michael Patrick Ghiglieri. Columbia University Press, 1984.

## SIPHONS IN ROMAN AQUEDUCTS

VITRUVIUS: DE L'ARCHITECTURE LIVRE VIII. Louis Callebaut. Édition Budé, 1973.

AQUEDUCS ROMAINS. In *Dossiers de l'archéologie*, No. 38; October-November, 1979.

SIPHONS IN ROMAN AQUEDUCTS. A. Trevor Hodge in *Papers of the British School at Rome*, Vol. 51, pages 174-221; 1983.

## THE TOPOLOGY OF MIRAGES

MATHEMATICAL THEORY OF OPTICS. R. K. Luneburg. University of California Press, 1964.

DIFFERENTIABLE GERMS AND CATASTROPHES. Theodor Bröcker. Cambridge University Press, 1975.

DIFFERENTIAL TOPOLOGY. Morris W. Hirsch. Springer-Verlag, 1976.

MIRAGES. Alistair B. Fraser and William H. Mach in *Scientific American*, Vol. 234, No. 1, pages 102-111; January, 1976.

MULTIPLE GRAVITATIONAL IMAGING BY DISTRIBUTED MASSES. William L. Burke in *The Astrophysical Journal*, Vol. 244, No. 1, Part 2, page L1; February 15, 1981.

## THE AMATEUR SCIENTIST

THE POLARIZATION OF LIGHT AT SEA. E. O. Hulburt in *Journal of the Optical Society of America*, Vol. 24, No. 2, pages 35-42; February, 1934.

REFLECTION OF LIGHT. M. Minnaert in *The Nature of Light and Colour in the Open Air*. Dover Publications, Inc., 1954.

*Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.*

## COMPUTER RECREATIONS

SMART SOAP BUBBLES CAN DO CALCULUS. Dale T. Hoffman in *The Mathematics Teacher*, Vol. 72, No. 5, pages 377-385, 389; May, 1979.

AHA! GOTCHA: PARADOXES TO PUZZLE AND DELIGHT. Martin Gardner. W. H. Freeman and Company, 1982.

THE COMPLEXITY OF ANALOG COMPUTATION. Anastasios Vergis, Kenneth Steiglitz and Bradley Dickinson. Technical Report No. 337, Department of Electrical Engineering and Computer Science, Princeton University; February, 1985.

## THE CHOICE OF TECHNOLOGY

INPUT-OUTPUT ECONOMICS. Wassily Leontief. Oxford University Press, 1966.

STRUCTURAL CHANGE IN THE AMERICAN ECONOMY. Anne P. Carter. Harvard University Press, 1970.

THE FUTURE OF THE WORLD ECONOMY: A UNITED NATIONS STUDY. Wassily Leontief, Anne P. Carter and Peter A. Petri. Oxford University Press, 1977.

INPUT-OUTPUT ANALYSIS: FOUNDATIONS AND EXTENSIONS. Ronald E. Miller and Peter D. Blair. Prentice-Hall, Inc., 1985.

THE IMPACT OF AUTOMATION ON WORKERS. Wassily Leontief and Faye Duchin. Oxford University Press, in press.

## THE IMMUNOLOGIC FUNCTION OF SKIN

REACTIVITY OF LANGERHANS CELLS WITH HYBRIDOMA ANTIBODY. Ellen Fithian, Patrick Kung, Gideon Goldstein, Marian Rubinfeld, Cecilia Fenoglio and Richard Edelson in *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 78, No. 4, pages 2541-2544; April, 1981.

CUTANEOUS T CELL LYMPHOMA. Jennifer A. K. Patterson and Richard L. Edelson in *The Medical Clinics of North America*, Vol. 66, No. 4, pages 895-913; July, 1982.

INTERACTION OF T CELLS WITH THE EPIDERMIS. Jennifer A. K. Patterson and Richard L. Edelson in *British Journal of Dermatology*, Vol. 107, No. 1, pages

117-122; July, 1982.

SKIN-ASSOCIATED LYMPHOID TISSUE (SALT): ORIGINS AND FUNCTIONS. J. Wayne Streilein in *Journal of Investigative Dermatology*, Vol. 80, Supplement, pages 12s-16s; June, 1983.

## THE SEARCH FOR PROTON DECAY

EXPERIMENTAL LIMITS ON THE NUCLEON LIFETIME FOR TWO- AND THREE-BODY DECAY MODES. H. S. Park, G. Blewitt, B. G. Cortez, G. W. Foster, W. Gajewski, T. J. Haines, D. Kielczewska, J. M. LoSecco, R. M. Bionta, C. B. Bratton, D. Casper, P. Chrysiopoulou, R. Claus, S. Errede, K. S. Ganzer, M. Goldhaber, T. W. Jones, W. R. Kropp, J. G. Learned, E. Lehmann, F. Reines, J. Schultz, S. Seidel, E. Shumard, D. Sinclair, H. W. Sobel, J. L. Stone, L. R. Sulak, R. Svoboda, J. C. van der Velde and C. Wuest in *Physical Review Letters*, Vol. 54, No. 1, pages 22-25; January 7, 1985.

## GLOBULAR CLUSTERS

GLOBULAR CLUSTERS. Edited by D. Hanes and B. Madore. Cambridge University Press, 1980.

STAR CLUSTERS. Edited by James E. Hesser. D. Reidel Publishing Company, 1980.

THE DYNAMICS OF GLOBULAR CLUSTERS. Ivan R. King in *The Quarterly Journal of the Royal Astronomical Society*, Vol. 22, pages 227-243; 1981.

## THE FIRST ORGANISMS

THE GENE AS THE BASIS OF LIFE. H. J. Muller in *Proceedings of the Fourth International Congress of Plant Sciences*, August 16-23, 1926, edited by B. M. Dugger. George Banta Publishing Company, Menasha, Wis., 1929.

SOME ASSUMPTIONS UNDERLYING DISCUSSION ON THE ORIGINS OF LIFE. N. W. Pirie in *Annals of the New York Academy of Sciences*, Vol. 69, Art. 2, pages 369-376; August 30, 1957.

SPECULATIONS ON THE ORIGIN AND EVOLUTION OF METABOLISM. Hyman Hartman in *Journal of Molecular Evolution*, Vol. 4, No. 4, pages 359-370; 1975.

GENETIC TAKEOVER AND THE MINERAL ORIGINS OF LIFE. A. G. Cairns-Smith. Cambridge University Press, 1984.

SEVEN CLUES TO THE ORIGIN OF LIFE: A SCIENTIFIC DETECTIVE STORY. A. G.

THE · NEW · CHRYSLER · TECHNOLOGY



## The 1985 Turbo New Yorker. Once you drive it, you'll never go back to a V-8 again.

Chrysler introduces the new technology of driving: Turbopower\* in its most advanced luxury sedan.

Here is the confidence of front-wheel drive, the security of advanced electronics and the quiet, smooth ride you expect in a fine luxury car.

And here are the luxuries you demand. Automatic transmission, power windows, power steering, power brakes, power remote mirrors and individual pillow-style reclining seats are all standard.

And finally, here is the new technology of turbopower. More power to move you. To accelerate. To pass. To cruise in serene comfort...yet with remarkable fuel

efficiency. 23 hwy. est. mpg [20] city est. mpg\*\*

Turbo New Yorker merits careful consideration by every luxury car owner. It is backed by a 5-year/50,000-mile Protection Plan covering drivetrain, turbo and outer body rust-through†

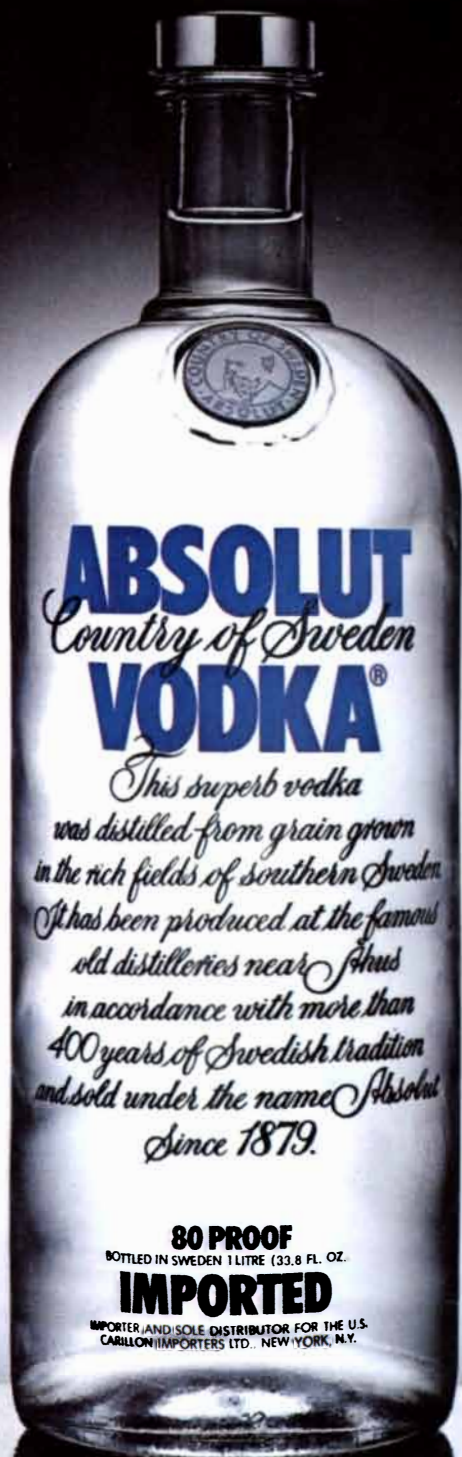
The new technology of driving must be experienced. So Chrysler invites you: test drive Turbo New Yorker. Once you drive it, you'll never go back to a V-8 again.

Purchase or lease your 1985 Turbo New Yorker at your Chrysler-Plymouth dealer. And buckle up for safety.



Division of  
Chrysler Corporation

\*Turbo is optional. \*\*Use these EPA ests. to compare. Actual mpg will vary with options, driving conditions and habits and vehicle condition. CA ests. lower. †Whichever comes first. Limited warranty. Deductible applies. Excludes leases. Dealer has details.



**ABSOLUT**  
*Country of Sweden*  
**VODKA**<sup>®</sup>

*This superb vodka  
was distilled from grain grown  
in the rich fields of southern Sweden.  
It has been produced at the famous  
old distilleries near Åhus  
in accordance with more than  
400 years of Swedish tradition  
and sold under the name Absolut  
Since 1879.*

**80 PROOF**  
BOTTLED IN SWEDEN 1 LITRE (33.8 FL. OZ.)

**IMPORTED**

IMPORTER AND SOLE DISTRIBUTOR FOR THE U.S.  
CARILLON IMPORTERS LTD. NEW YORK, N.Y.

**ABSOLUT PERFECTION.**

80 AND 100 PROOF (100% GRAIN NEUTRAL SPIRITS) ABSOLUT COUNTRY OF SWEDEN. © 1984 CARILLON IMPORTERS LTD. NY.