SCIENTIFIC AMERICAN

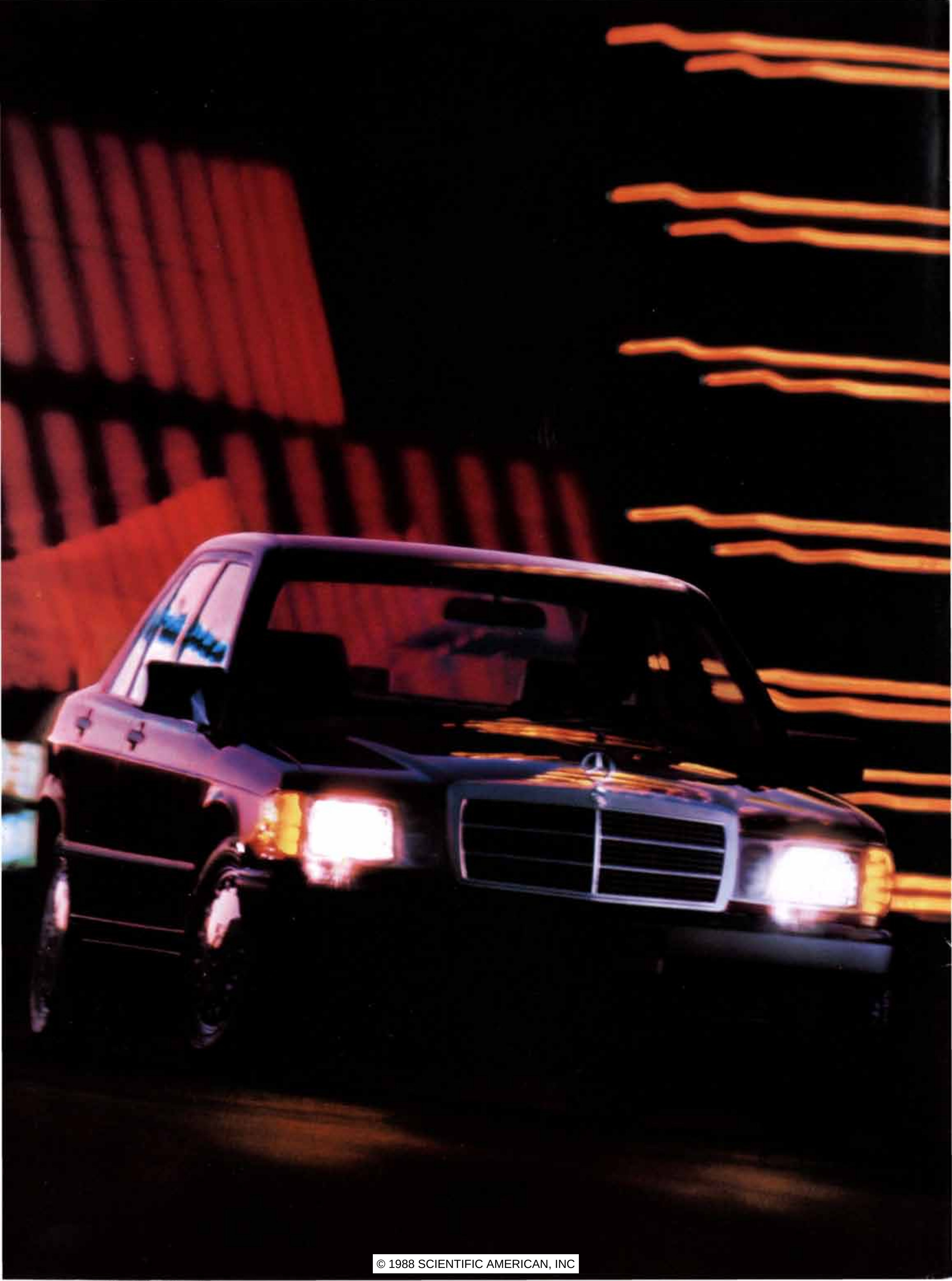
JULY 1988
$2.50

Is the North Atlantic a temporary ocean?

Randomness in simple arithmetic.

How the human eye focuses.

Gravitational lens, *foreseen by Einstein, can bend light from a far-distant object, distorting the object's shape.*



SCIENTIFIC AMERICAN

July 1988 Volume 259 Number 1

DEPARTMENTS

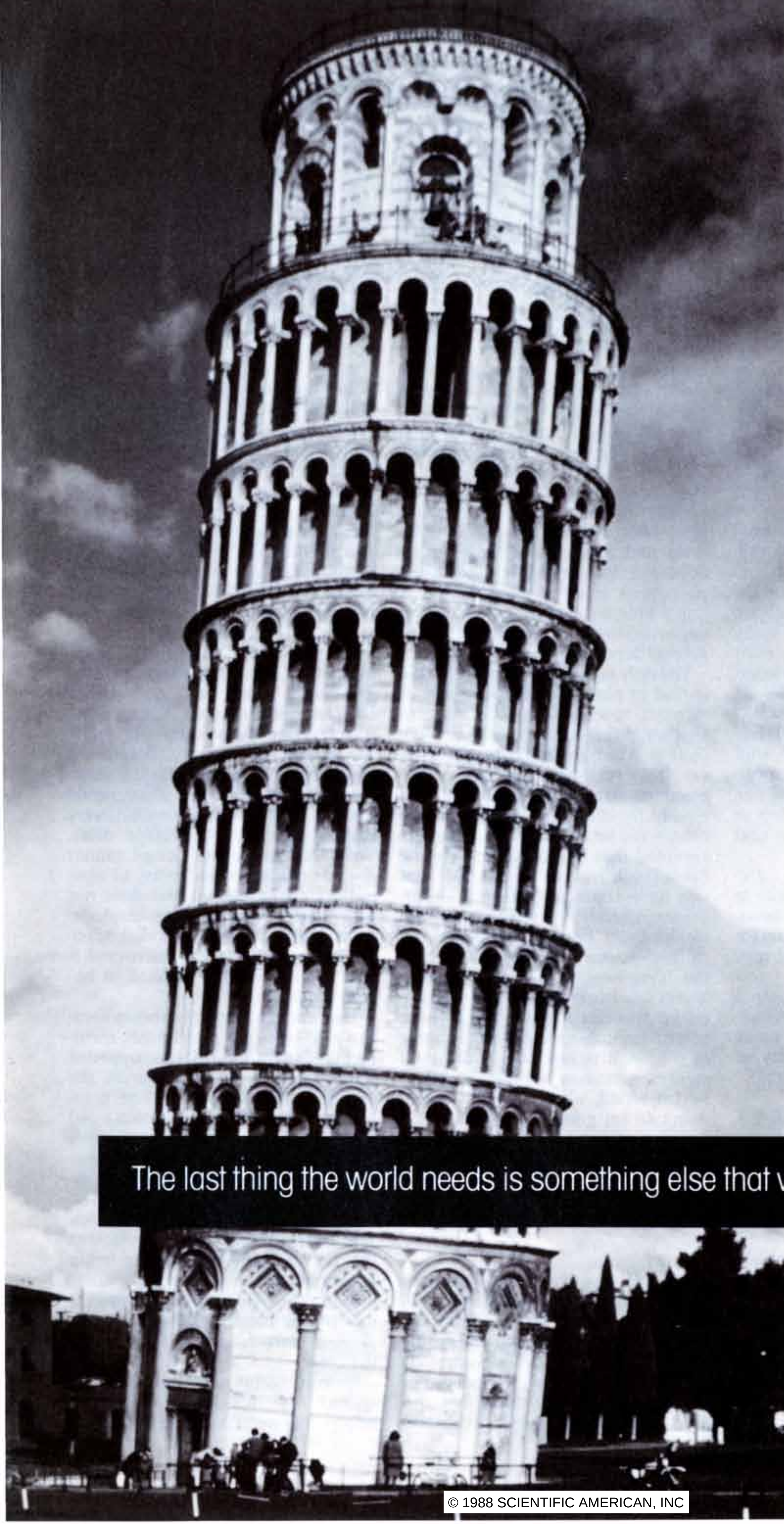
THE COVER is a photograph of the name "Einstein" seen through a lens that bends light in the same way as a gravitational lens does (see "Gravitational Lenses," by Edwin L. Turner, page 54). The yellow half-moon shapes at the center come from the inversion and mirror reversal of the letters *S* and *T*. Gravitational lensing occurs when the gravitational field of a galaxy bends light from a distant quasar. Image distortions contain information about the intervening space. Lens courtesy of Edwin L. Turner.

THE ILLUSTRATIONS

Cover photograph by S. Varnedoe

Page	Source
45	Quesada/Burke, courtesy of the Carnegie Institution of Washington
46	Jerome Kuhl
47	Ellen R. White
48	Jerome Kuhl
49	Ellen R. White
50	Jerome Kuhl
51	Ellen R. White
55	Jacqueline N. Hewitt
56-59	Joe Lertola
60	F. N. Owen and J. J. Puschell, courtesy of the National Radio Astronomy Observatory (*top and middle*); P. E. Greenfield, D. H. Roberts and B. F. Burke, courtesy of the National Radio Astronomy Observatory (*bottom*)
63	David W. Golde
64-67	Patricia J. Wynne
68	David W. Golde
69	Patricia J. Wynne
70	David W. Golde
73-78	George Retseck
81	James Kilkelly, courtesy of Paulson Casino Supplies of New Jersey, Inc.
82	Quesada/Burke, from *Hilbert,* by Constance Reid, ©1970 Springer-Verlag
83	©Arnold Newman
84	Bob Conrad (*top*), Murray Greenfield (*bottom*)
85	Bob Conrad
87	Quesada/Burke
88-90A	George V. Kelvin, Science Graphics
90B	Robin Revell, Corning Glass Works
93	Patricia N. Farnsworth, University of Medicine and Dentistry of New Jersey in Newark
94	Carol Donner
95	From *Tissues and Organs: A Text-Atlas of Scanning Electron Microscopy,* by Richard G. Kessel and Randy H. Kardon, ©1979 W. H. Freeman and Company (*top*); Carol Donner (*bottom*)
96	Carol Donner
97	Jane F. Koretz
98	Patricia N. Farnsworth, University of Medicine and Dentistry of New Jersey in Newark
99	Jane F. Koretz
101	Paul Logsdon
102	Tom Prentiss
103	Laurie Burnham
104-105	Tom Prentiss
106-107	Laurie Burnham
108	Tom Prentiss
109	Paul Logsdon
116-119	Michael Goodman
120-123	Thomas C. Moore

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Subscription rate: $24 per year in the U.S. and its possessions; $33 per year in all other countries. Postmaster: Send address changes to Scientific American. P.O. Box 953. Farmingdale. N.Y. 11737-0000.


The last thing the world needs is something else that won't hold up.

LETTERS

To the Editors:

I believe "The Bubonic Plague," by Colin McEvedy [SCIENTIFIC AMERICAN, February], underestimates the impact of public-health measures in limiting the disease. The article induced me to reread Daniel Defoe's marvelous *A Journal of the Plague Year.* Early in the book Defoe lists "Orders Conceived and Published by the Lord Mayor and Aldermen of the City of London concerning the Infection of the Plague, 1665." Among them is this directive:

"No infected Stuff to be uttered [offered for sale]. That no clothes, stuff, bedding, or garments be suffered to be carried or conveyed out of any infected houses, and that the criers and carriers abroad of bedding or old apparel to be sold or pawned be utterly prohibited and restrained, and no brokers of bedding or old apparel be permitted to make any outward show, or hang forth on their stalls, shop-boards, or windows, towards any street, lane, common way, or passage, any old bedding or apparel to be sold, upon pain of imprisonment. And if any broker or other person shall buy any bedding, apparel, or other stuff out of any infected house within two months after the infection hath been there, his house shall be shut up as infected, and so shall continue shut up twenty days at the least."

Your article states that some infection was certainly carried by fleas in furs and wool. The Lord Mayor's measure of leaving clothes and bedding for two months must have allowed time for infected fleas to die.

Perhaps the reduction in the trade in old clothes and bedding with increasing prosperity in Europe was a cause of the reduction in the severity of outbreaks of the disease with time.

R. E. DURRANT

Nottingham
England

To the Editors:

There are various reasons for thinking that public-health measures did not play much part in the disappearance of plague from early modern Europe. Perhaps the most important is that an earlier cycle of plague epidemics that began in the sixth century A.D. petered out in the course of the eighth century even though no one was taking any public-health measures of any kind. This suggests that the decline of plague was related to biological rather than social factors.

As for the contemporary belief that infected clothing was important in spreading plague, it is worth remembering that it was based on a theory now known to be false. The idea was that a toxic agent passed directly from the infected clothing to the victim, who fell ill either instantly (as described by Boccaccio in the introduction to the *Decameron*) or within a matter of hours (as occurred to the unfortunate George Viccars, the index case at the village of Eyam in 1666, who was said to have collapsed soon after opening a box of infected clothes from London). The fact that such instances were widely reported and generally believed undoubtedly encouraged the authorities to impose regulations such as the Lord Mayor's order quoted in Durrant's letter. Actually the postulated mode of transmission is impossible, because plague has a minimum incubation period amounting to several days.

The only proved mechanism for the spread of plague is the movement of infected rodents from one place to another, either actively (on their own feet) or passively (in boats and barges). It is reasonable to believe the baled-up skins of rodents that had died of the disease could be a source of infected rat fleas, and it is generally accepted that this could be how the Black Death reached Europe. But that was an exceptional, one-time occurrence, and after the initial inoculation we have no need of skins—or old clothes—to carry the plague around the Continent. Conversely, public-health measures based on misconceived theories are unlikely to have played any role in the disappearance of plague. It is just one of those instances, common enough in history, I'm afraid, where the do-gooders didn't do any good.

COLIN MCEVEDY

London

To the Editors:

In "The Reality of the Quantum World" [SCIENTIFIC AMERICAN, January] Abner Shimony appears to assess the Einstein-Podolsky-Rosen paradox (the EPR) as a very important issue. We agree. (The most typical example of the EPR is the proposition that two photons, distant from each other and independently unpolarized, show the same polarization along any arbitrary x or y axis when both of them are measured.)

We do not see, however, that quantum mechanics, in the context of the EPR or anywhere else, strongly suggests the existence of actions at a distance (aaad), which is what explaining the EPR in terms of "nonlocality" essentially amounts to.

To a limited extent, the analogue of the EPR in classical mechanics is given by two magnetized rods clinging together on a marble table, opposite poles juxtaposed. The rods can be subjected to random "kicks" and still remain parallel, but if they are separated, the kicks will disturb their parallelism. In contrast, quantum mechanics does not allow as detailed a picture, and it is possible to write formulas without taking due notice of the fact that the "rods" (the photons in this case) are at a distance from each other. The conclusions are then incorrect, and, quite naturally, they appear to require an aaad.

Such a fantastic phenomenon, however, would be nothing short of a miracle, and it would push our science back hundreds of years. Among other unacceptable features, it would violate the theory of relativity, because it would imply that *unpredictable* information, produced at one place, would have a clear effect at another, very distant place with negligible delay. The fact that human beings cannot use the present experiments to communicate is irrelevant and does not prove that future experimental designs could not make even that possible. When electricity was discovered, it was not immediately applied to human industry.

True enough, the experiments seem to agree with the presumably incorrect prediction, but that has happened often before in physics. After all, the experiments only give us one number—the degree of correlation—and because of the peculiarly imperfect knowledge allowed by quantum mechanics, the correlation is not as astonishing as it would be for the magnetized rods of classical mechanics. The fact is that in the experiments themselves nothing points the finger to an aaad; other tailor-made fantastic inventions would also do. The aaad is injected by hand to predict the EPR phenomenon, as well as to "explain" the data.

Instead, noting that quantum mechanics is a well-proved theory, we believe that we have a puzzle, and that we do not know all there is to know about it. Nature has more imagination than we do. More work is needed.

In letters to *Physics Today* (April, 1986) and *Mosaic* (winter, 1986–87) we have written a description of the EPR in less mysterious terms. Nevertheless, Shimony's article stimulates the reader in the right direction.

ORESTE PICCIONI

WERNER MEHLHOP

BRIAN WRIGHT

Department of Physics
University of California, San Diego

To the Editors:

I do not agree with the assertion that there is nothing in quantum mechanics (QM) that strongly suggests the existence of actions at a distance (aaad). Standard calculations of QM show that each photon pair in the ensemble studied by Alain Aspect and his colleagues is governed by a quantum state Ψ, which is approximately the same as the state Ψ_0 described in the article, and the correlations predicted on the basis of Ψ agree very well with the experimental results.

It is not correct to say, as Drs. Piccioni, Mehlhop and Wright do, that in view of the peculiarly imperfect knowledge allowed by QM there is nothing astonishing about this agreement; the number of events in the experiments is so large that the probable error is small, and there is negligible probability of obtaining agreement of such accuracy between predictions and observations merely by accident. Furthermore, the writers' statement that "the experiments only give us one number" may convey the wrong impression; that number is the compilation of a large amount of information, and it is a crucial number in the sense that QM predicts a sharp value for it, which differs greatly from the range of values allowed by the family of local hidden-variables theories. Thus there is strong evidence that Ψ is the objectively correct quantum state governing the photon pairs studied by Aspect and his colleagues. Finally, as the writers correctly point out, Ψ does indeed imply that "unpredictable information, produced at one place, would have a clear effect at another, very distant place." Putting all these propositions together constitutes the evidence for aaad.

Is this conclusion fantastic? No, it is not a fantasy, but it is astonishing. Would it push our science back hundreds of years? Not at all; the aaad of QM is radically different technically and conceptually from that of Newtonian mechanics, and furthermore it peacefully coexists with relativity theory in the sense that it cannot be used to send a message faster than light. Might future designs allow even that? No; the contrary has been proved rigorously on the assumption that QM is correct.

In spite of this argument that the quantum-mechanical version of aaad is not a disaster, it is interesting and important to inquire whether an alternative interpretation of the experimental evidence remains open. One much discussed possibility is that the coincidence counting rate of photon pairs is not proportional to their rate of emergence from the pairs of polarization analyzers. There is a local model by John F. Clauser and Michael A. Horne that capitalizes on the inefficiency of photodetectors; it yields counting rates that are in agreement with those predicted by Ψ. The model is not physically appealing, but logically it can be excluded only if a correlation experiment is performed with highly efficient analyzers and detectors, and for this purpose it is probably necessary to use pairs of atoms from dissociated molecules instead of pairs of photons.

Another possibility was explored by Piccioni and Mehlhop, who proposed the radical hypothesis that photons have spin higher than 1, but they report that an experiment they carried out themselves yielded results that did not confirm their hypothesis. It is desirable that further hypotheses be explored, and I am pleased that the writers are willing to make the requisite theoretical and experimental effort. Other experimenters, such as James D. Franson, are also devoting much effort to similarly motivated experiments. Of course, I have a different estimate of the probable outcomes of these experiments. I am so impressed by the correlations found by Aspect and his colleagues in a situation in which a classical relativity theorist would have expected the contrary that I think it very likely future experiments will continue to support quantum nonlocality.

ABNER SHIMONY

Departments of Physics
and Philosophy
Boston University

SCIENTIFIC AMERICAN
In Other Languages

LE SCIENZE
L. 3,500/copy L. 35,000/year L. 45,000/(abroad)
Editorial, subscription correspondence:
Le Scienze S.p.A., Via G. De Alessandri, 11
20144 Milano, Italy
Advertising correspondence:
Publietas, S.p.A., Via Cino de Duca, 5,
20122 Milano, Italy

サイエンス
Y880/copy Y9600/year Y13,000/(abroad)
Editorial, subscription, advertising correspondence:
Nikkei Science, Inc.
No. 9-5, 1-Chome, Otemachi
Chiyoda-ku, Tokyo, Japan

INVESTIGACION Y
CIENCIA
450 Ptas/copy 4950Ptas/year $35/(abroad)
Editorial, subscription, advertising correspondence:
Prensa Científica S.A.,
Calabria, 235-239
08029 Barcelona, Spain

POUR LA
SCIENCE
27FF/copy 265FF/year 315FF/year (abroad)
Editorial, subscription, advertising correspondence:
Pour la Science S.A.R.L.,
8, rue Férou,
75006 Paris, France

Spektrum DER WISSENSCHAFT
9.80 DM/copy 99 DM/year 112.20 DM/(abroad)
Editorial, subscription correspondence:
Spektrum der Wissenschaft GmbH & Co.
Moenchhofstrasse, 15
D-6900 Heidelberg,
Federal Republic of Germany
Advertising correspondence:
Gesellschaft Für Wirtschaftspublizistik
Kasernenstrasse 67
D-4000 Duesseldorf,
Federal Republic of Germany

科学
1.40RMB/copy 16RMB/year $24/(abroad)
Editorial, subscription correspondence:
ISTIC-Chongqing Branch, P.O. Box 2104,
Chongqing, People's Republic of China

В МИРЕ НАУКИ
2R/copy 24R/year $70/(abroad)
Editorial correspondence:
MIR Publishers
2, Pervy Rizhsky Pereulok
129820 Moscow U.S.S.R.
Subscription correspondence:
Victor Kamkin, Inc.
12224 Parklawn Drive,
Rockville, MD 20852, USA

TUDOMÁNY
98Ft/copy 1,176Ft/year 2,100Ft/(abroad)
Editorial correspondence:
TUDOMÁNY
H-1536 Budapest, Pf 338
Hungary
Subscription correspondence:
"KULTURA"
H-3891 Budapest, Pf. 149
Hungary

مجلة العلوم
1KD/copy 10KD/year $40/(abroad)
Editorial, subscription, advertising correspondence:
MAJALLAT AL-OLOOM
P.O. BOX 20856 Safat,
13069 – Kuwait

Advertising correspondence all editions:
SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue
New York, NY 10017
Telephone: (212) 754-0550 Telex: 236115

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1938: "The hull of the world's largest flying boat, now being built by Boeing Aircraft for Pan-American Airways, has emerged from its construction scaffolding into the open. The hull looks so large that it gives the impression more of a surface vessel than of an airplane; it measures 109 feet from bow to tail and has an inside volume equal to that of an average five-room house. A spiral staircase will connect the flight bridge or control quarters on the upper deck with the main deck, which will contain a series of spacious passenger compartments, dining room, lounge, private suites, and so on."

"Steam enters a modern turbine at a temperature hot enough to set fire to a piece of wood and .03 of a second later leaves it at a temperature too cool for a comfortable bath."

"Of particular value for educational and archive purposes is an aluminum motion-picture film developed by the Fischer Film Corporation. This film, being opaque, must be projected by means of reflection, a principle that is well known and that gives good results both for picture projection and for the reproduction of sound tracks. Among the advantages claimed for the new aluminum film is that it is not flammable."

"Why did the Russian scientists endure the danger and discomfort of establishing a weather station on the ice floes of the north polar seas? Because the whole secret of weather is to be found in the high latitudes, and in the investigation of the travel of polar air masses. What we need now is a string of weather stations along the Arctic fringe of Canada. Data from such stations, together with similar data now being gathered in northern Russia and Siberia, would in the end go far toward solving the secret of the cradle of the weather in the Northern Hemisphere."

SCIENTIFIC AMERICAN

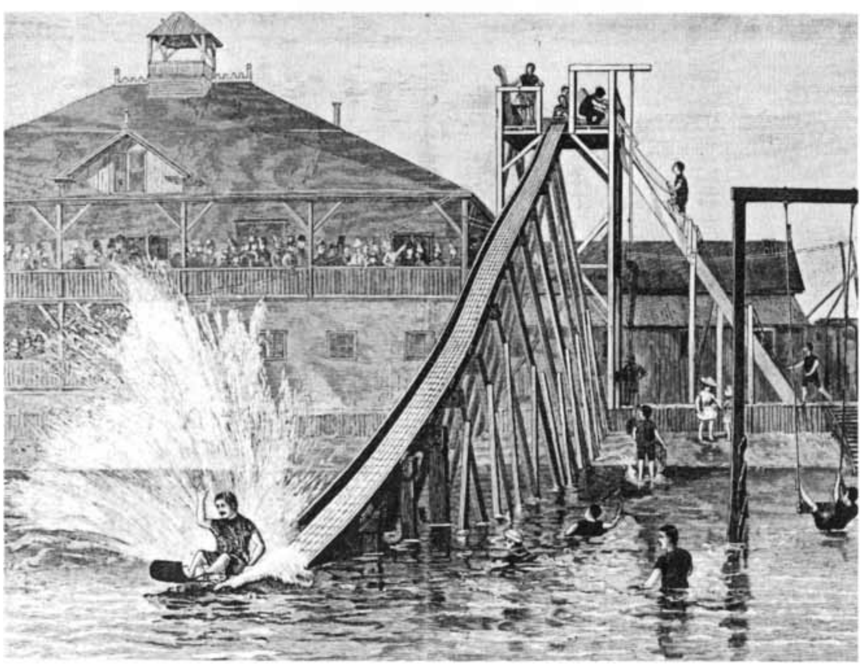
JULY, 1888: "A writer in *La Nature* remarks that the streets of American cities have been laid out with the tape-line and at right angles. This, he observes, is very fine from a geometrical standpoint, but carries with it very serious consequences from an economical point of view. In fact, if we walk along the two sides of a square instead of following a diagonal, the distance is increased in the proportion of 40 per cent; that is to say, instead of walking 100 feet, we walk 140. Hence a loss of time, strength and money."

"The women in the Sultan's seraglio, at Constantinople, have just been vaccinated, to the number of 150. The operation took place in a large hall, under the superintendence of four gigantic eunuchs. The Italian surgeon to whom the task was confided was stationed in front of a huge screen, and the women were concealed behind it. A hole had been made in the center of the screen, just large enough to allow an arm to pass through."

"The effect of complete obscurity on the normal pupil has hitherto been seen only by the light of electric discharges, which allowed of no measurements. Messrs. Miethe and Gaedicke, by their invention of the well-known explosive magnesium mixture, have furnished us with a simpler method. A photograph of the eye can be taken in a perfectly dark room, showing the pupil fully dilated, as its reaction does not begin until after exposure."

"An immense terrestrial globe, constructed on a scale of one millionth, will be shown at the Paris exhibition of 1889. The globe will measure nearly 13 meters in diameter, and will give some idea of real dimensions, since the conception of the meaning of a million is not beyond the powers of the human mind."

"Tobogganing has become such a favorite winter pastime that the idea occurred, some time ago, to Mr. C. J. Belknap, of Bridgeport, Conn., that it might also be adapted to the summer months. In 1887 he built the marine toboggan slide now described, which rises from high-water mark to the height of thirty-two feet. The chute itself, 178 feet long, is but twenty inches wide, and contains 725 wheels for the sleds to run on. The toboggans start at the signal given by a bell, only one being allowed to go at a time; and on being projected from the lower end, they ricochet across the waters of the Sound for a distance varying from 75 to 175 feet, skipping along like a flat pebble, till the force acquired in the descent is lost, after which the bather swims ashore, pulling his sled after him. The facial expression of novices taking their first adventurous slide is quite remarkable, and the sensations felt are correspondingly novel and peculiar."



A marine toboggan slide near Bridgeport, Conn.

BUILDING THE SPACE STATION

The Challenge

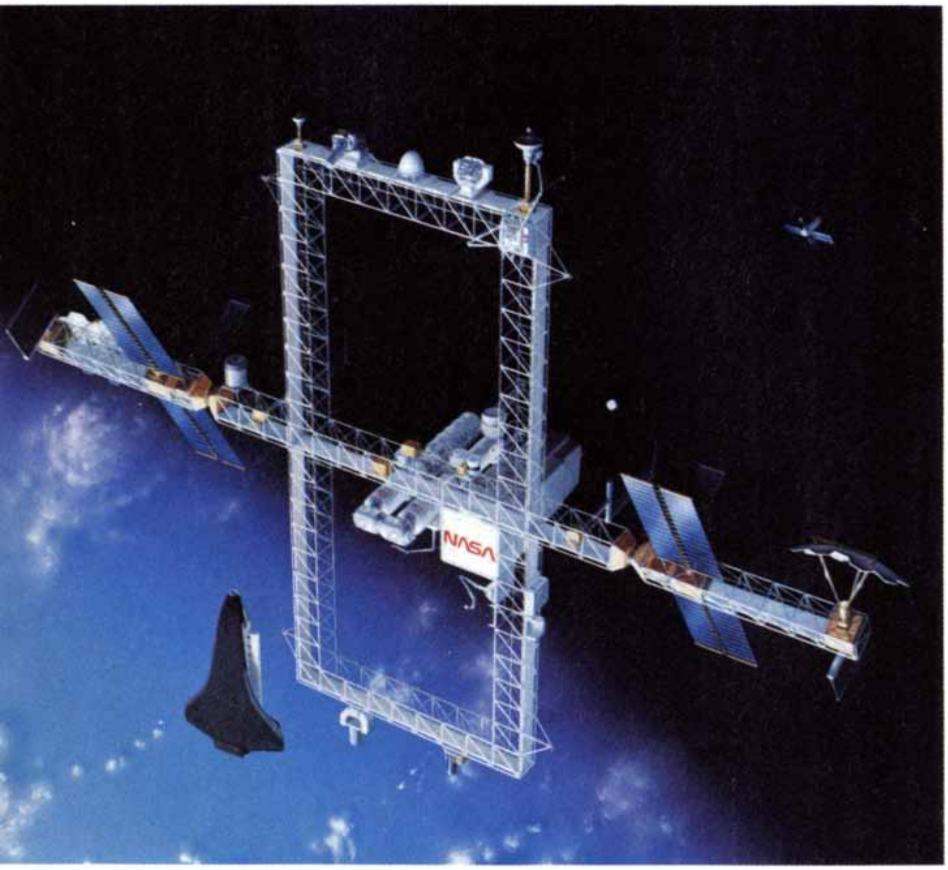
by Andrew J. Stofan

Assembly and operation of the international space station will be a major technological challenge for the United States and our international partners, Canada, Europe and Japan.

The permanent complex, envisioned to be completed in 1996, will be a research facility of unmatched versatility and practicality, occupied continuously by a team of scientists and researchers. The station will support a wide range of technological and commercial endeavors, which will lay the groundwork for leadership in civilian space activity and scientific research by America and her allies through the end of the century and beyond.

Beginning in early 1995, elements of the space station will be launched aboard the Space Shuttle into orbit 250 miles above the earth and inclined 28.5 degrees to the Equator.

The space station will consist of a single horizontal truss that stretches more than 500 feet across. Attached to the truss will be four pressurized modules: three laboratories and a habitation module. The U.S. will provide the habitation module for the eight-person crew, and a logistics module for ferrying experiments and provisions to and from the earth. The U.S., Japan and Europe will each provide a laboratory module. Canada is furnishing a mobile servicing center, which includes the space station manipulator arm. The dominant feature of the truss will be two solar power modules which will provide 75,000 watts of continuous power to the space station and subsystems for heat rejection, propulsion and altitude control. The U.S. and Europe will also provide polar orbiting platforms.



External attach points on the station will provide a valuable base for observational scientific instruments, particularly those that could benefit from servicing or resupply of consumables. And to support its ultimate use as a transportation node for expanding human presence out of the earth's orbit and into the solar system, the space station will support biological research that will not only qualify human beings for extended stays in space, but also will likely have clinical applications on the earth.

Long-duration microgravity research conducted in the station's three laboratories could make important contributions to fundamental advances in our understanding of basic physical, chemical and biological processes. The results of microgravity research will not only open up new regions of scientific inquiry, but could also reveal new processes and techniques of value to industries on the earth.

The National Aeronautics and Space Administration has awarded primary contracts to four companies to build the space station: Boeing, McDonnell Douglas, General Electric and Rockwell International.

The space station is the first of many challenges for America and her allies. It will be a staging base, and is an essential precursor to future, more ambitious missions, because it will provide the space infrastructure necessary to accomplish our goals.

Andrew J. Stofan was recently Associate Administrator for the space station at NASA. He is currently a Vice President at Martin Marietta.

The First Challenge: Assembly in Space

by David C. Wensley

Assembly of the flight elements into an operating space station is the first major challenge for America's space station team.

Construction will require as many as 20 flights of the Space Shuttle as it transfers 40,000 pounds of the building block elements on each journey to the orbital assembly point.

The initial task for the astronauts will be to connect the 15-foot-long struts that form the structural backbone and support the utility lines.

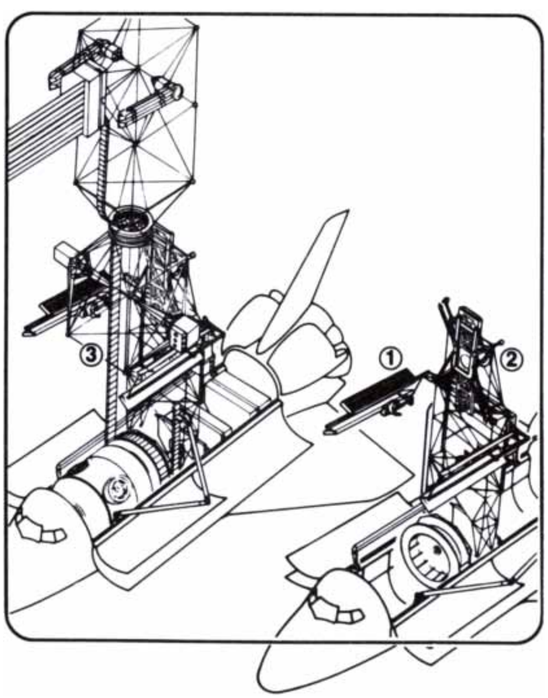
Builders on the ground will use the latest space technology—such as Litz wire cables feeding 20,000 Hertz power to each of the station's systems and payloads; optical fibres carrying scientific and operational data at 100 megabit per second rates; and pipeline networks containing ammonia, oxygen, hydrogen, nitrogen and other vital gases and fluids needed for thermal control, propulsion, life support and scientific and commercial mission needs.

Once in space, the astronauts will uncoil hinged segments of utility trays containing these lines from a 14-feet diameter spool, and guide them to prepositioned attach points inside the truss framework.

To build this initial mini-station, shuttle flight crews will use the Canadian supplied remote manipulator system to unpack the cargo bay and position each element for assembly.

After commanding release of the retention fittings that held each unit fixed as it was launched into space, a crew member will slowly maneuver each package into its mounting position on the structure, where the outside crew will make the permanent attachments.

From inside the shuttle, another member of the construction team will guide the robot arm to grapple fixtures on each of the prepackaged equipment assemblies stowed securely in the cargo bay.



Construction in space: (1) Struts and attachments on assembly work platform (2) Mobile transporter moves completed truss bays into position (3) Power-driven reel deploys utility lines as bays are completed

Then, operating once again from the control station within the orbiter's cabin, the crew will command the deployment of the station's solar power arrays and begin energizing and testing each mode of operation.

In the next sequence, connector assemblies will be joined to their assigned subsystems and the mechanical construction of the first flight element will be complete.

As early as the second mission build-up flight, the first pressurized module will be delivered by the shuttle and attached to the free-flying spacecraft. This module will later become the "Operations Center" from which the daily schedule of mission activities will be conducted.

Each arriving shuttle will bring more system building blocks and supplies. On approximately the fourth flight, the first permanent U.S. laboratory in space will be positioned on the station and the new era of scientific investigation will begin.

Meanwhile, shuttle crews will continue the build up. They will install additional solar panels and batteries, radiator sections, steerable antennas to increase communications bandwidth through the data relay satellites and propellants to maintain orbital position.

When sufficient power and other resources are available, the European and Japanese laboratories will be added to the station complex and phase one will be complete.

A successful assembly of the space station will pave the way for more ambitious assembly and checkout activities envisioned for the future.

David C. Wensley is Vice President, Project Integration at the Space Station Division of McDonnell Douglas.

The Challenge for Rockwell: ELECTRICITY

by George J. Hallinan

One of the most demanding challenges for the space station team will be the supply of electricity to the station—10 times more power than an American spacecraft has ever needed.

The permanently orbiting station will have to generate its own electricity, unlike the shuttle and earlier manned craft which could carry sufficient power in fuel cells and batteries for their short flights.

Of the space station's 75 kilowatts, 50 percent will be used for housekeeping and 50 percent for payload.

Under the direction of NASA's Lewis Research Center, Cleveland, Ohio, the Rockwell International Corporation's Rocketdyne division team—Rocketdyne, Ford Aerospace and Communications Corporation, Allied Signal Co., General Dynamics Corporation, Harris Corporation and Lockheed Corporation—will design and develop the electrical system for NASA's space station and free-flying platforms.

Electrical power will be generated by sunlight striking photovoltaic cells, commonly known as the power sources for some pocket calculators. The technological obstacles will be primarily those of scale, building 24,800 square feet of the solar panels, storing the electricity and aiming the panels at the sun. Lockheed will build these panels.

The craft will orbit the earth about every 90 minutes, with approximately 60 minutes of sunlight and 30 minutes of eclipse.

Stored energy will be used to provide power during eclipse, through use of high energy density, individual pressure vessel Ni-H_2 energy storage batteries being developed by Ford Aerospace. A fully automated computerized power management and distribution system will be used to control the generation and distribution of electricity throughout the station. General Dynamics will develop and produce the 20 khz converters for the computerized system. The system design employs maximum commonality and simple construction to minimize cost. To maximize commonality and ensure user compatibility, station and platform designs use common components (for example, power generation, energy storage and power management and control).

Rocketdyne and its partners propose to use a solar dynamic system based on a gas driven engine using a design known as the Closed Brayton Cycle. The solar dynamic system offers, by adding the two 25 kw modules, a low-drag, high-efficiency, low-life-cycle-cost electric power source that provides a cost-effective growth path.

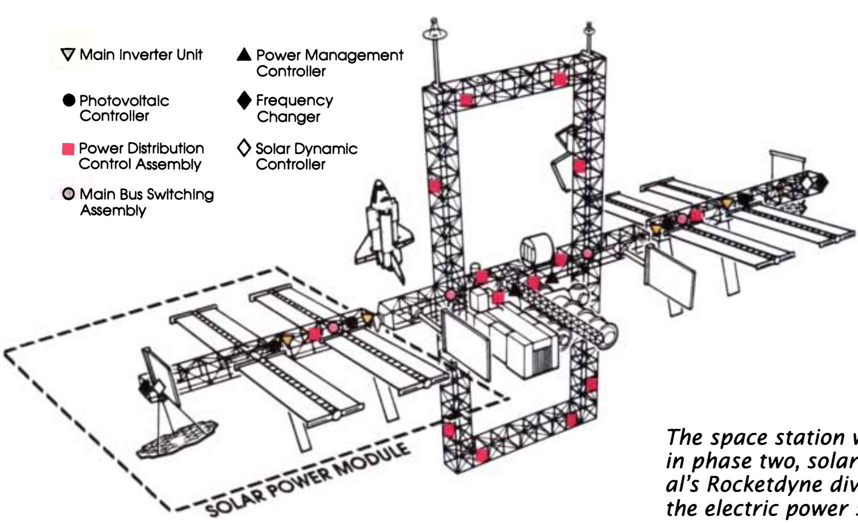
On the space station, the Brayton technology will work by using a mirror to relect sunlight into the receiver.

The receiver will contain pipes holding a eutectic saltlike substance composed of lithium fluoride-calcium difluoride and pipes carrying xenon and helium gas which will be heated by the sun. The heated gas will run through a turbine that spins a generator, creating electricity. The gas is then run back through the cycle again.

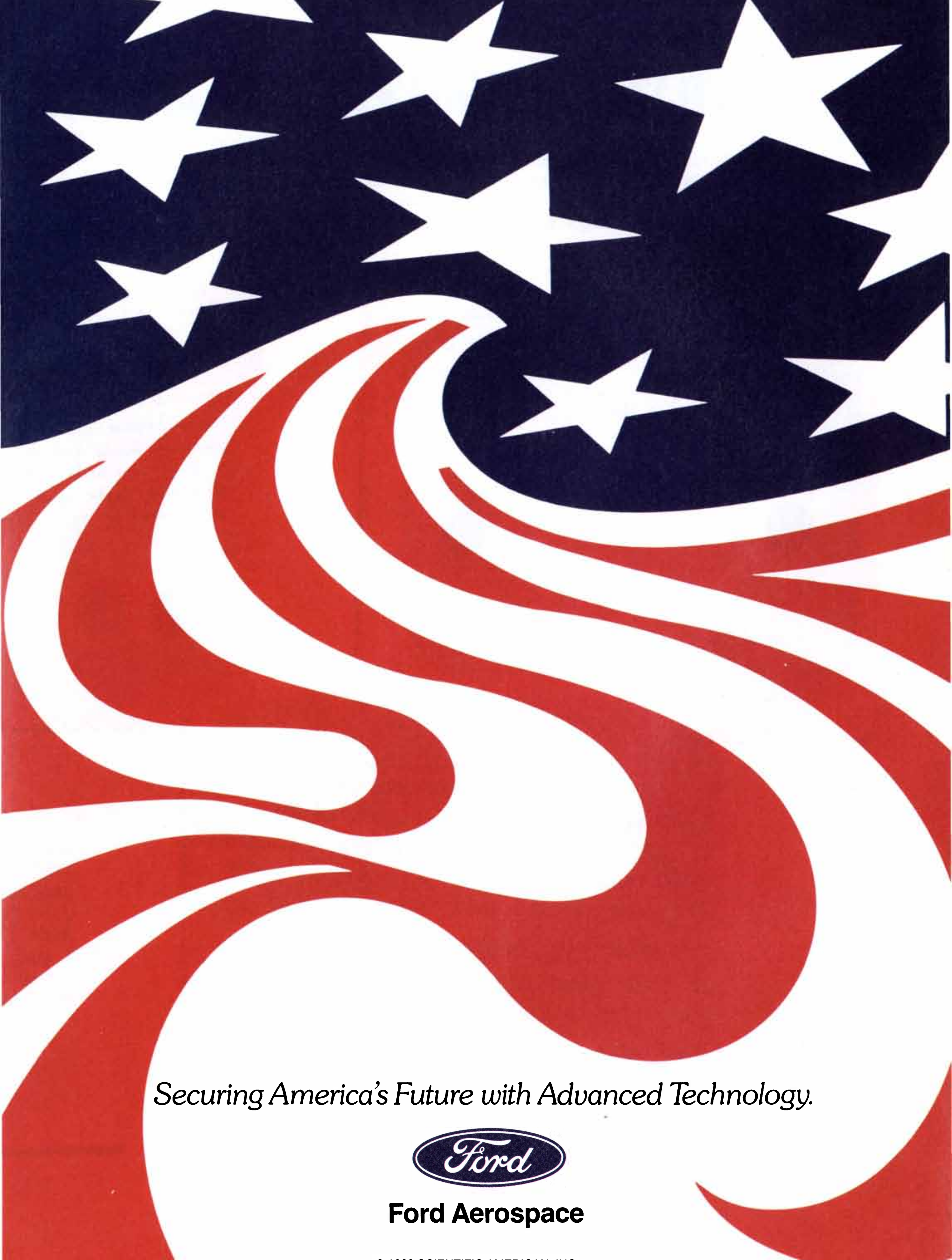
Allied Signal will develop and produce the Closed Brayton Cycle engine and receiver. Harris will provide the solar concentrator for the solar dynamic module.

To meet the needs of the station users during its 30-year life cycle, the power system has the capacity to grow from 125 kw to 275 kw by adding solar dynamic modules.

George J. Hallinan is Vice President and Program Manager—Space Station Power, Rockwell International.



The space station will first use photovoltaics for power, and in phase two, solar dynamic technology. Rockwell International's Rocketdyne division and its team members are designing the electric power system.


Securing America's Future with Advanced Technology.
Ford
Ford Aerospace

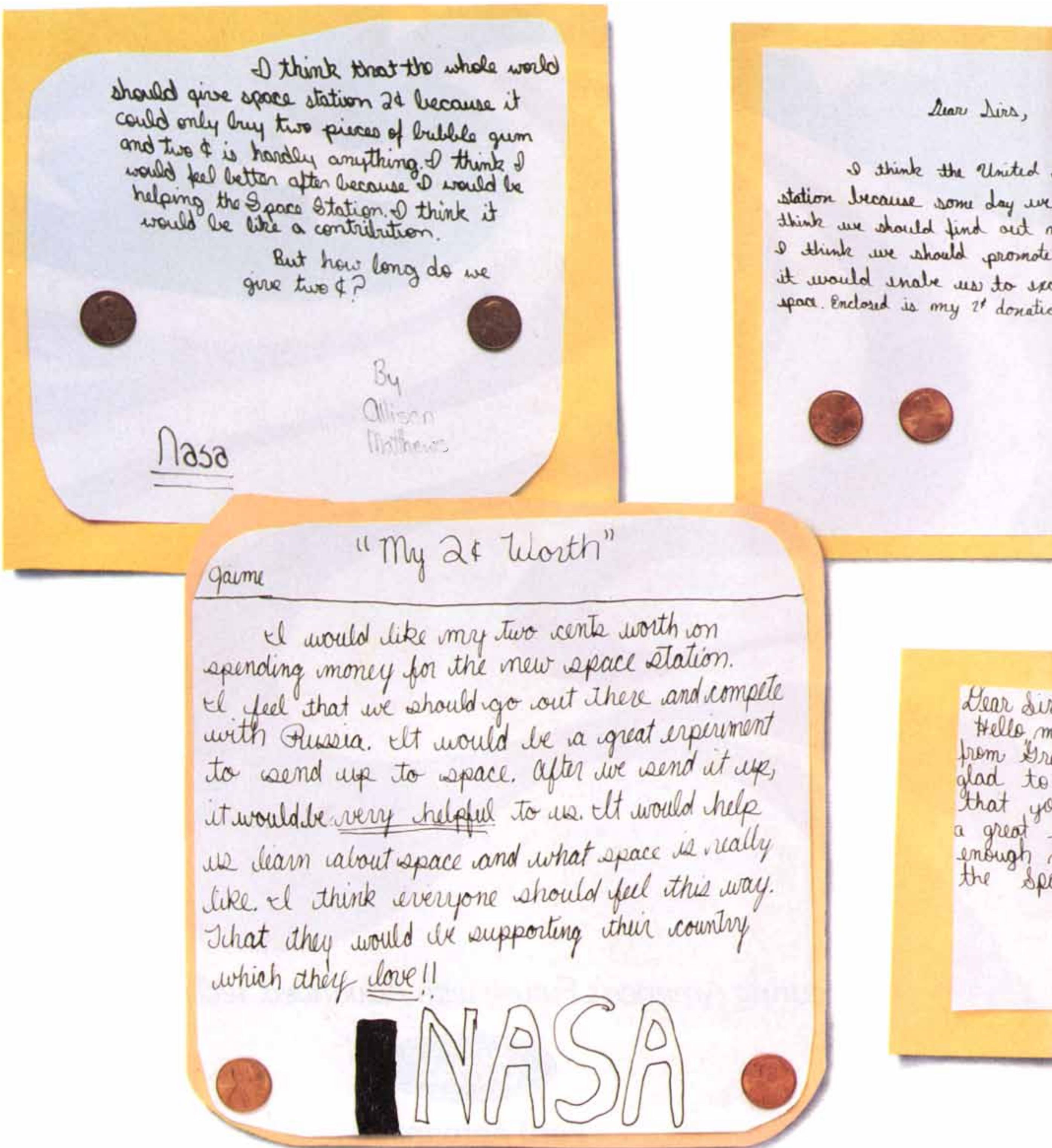
WHO'D SPEND TWO CENTS ON GUM WHEN

Not these kids.

They've learned a manned Space Station will cost each one of us just two cents a day. And they know the payback in human benefits will be enormous.

With the special power of the young to imagine what could be, children are among the first to see the value of the Space Station. So when we ran a message in the newspaper supporting it, they wrote us with their two cents' worth—not only in words, but in cash.

We're sharing some of their thoughts here as a



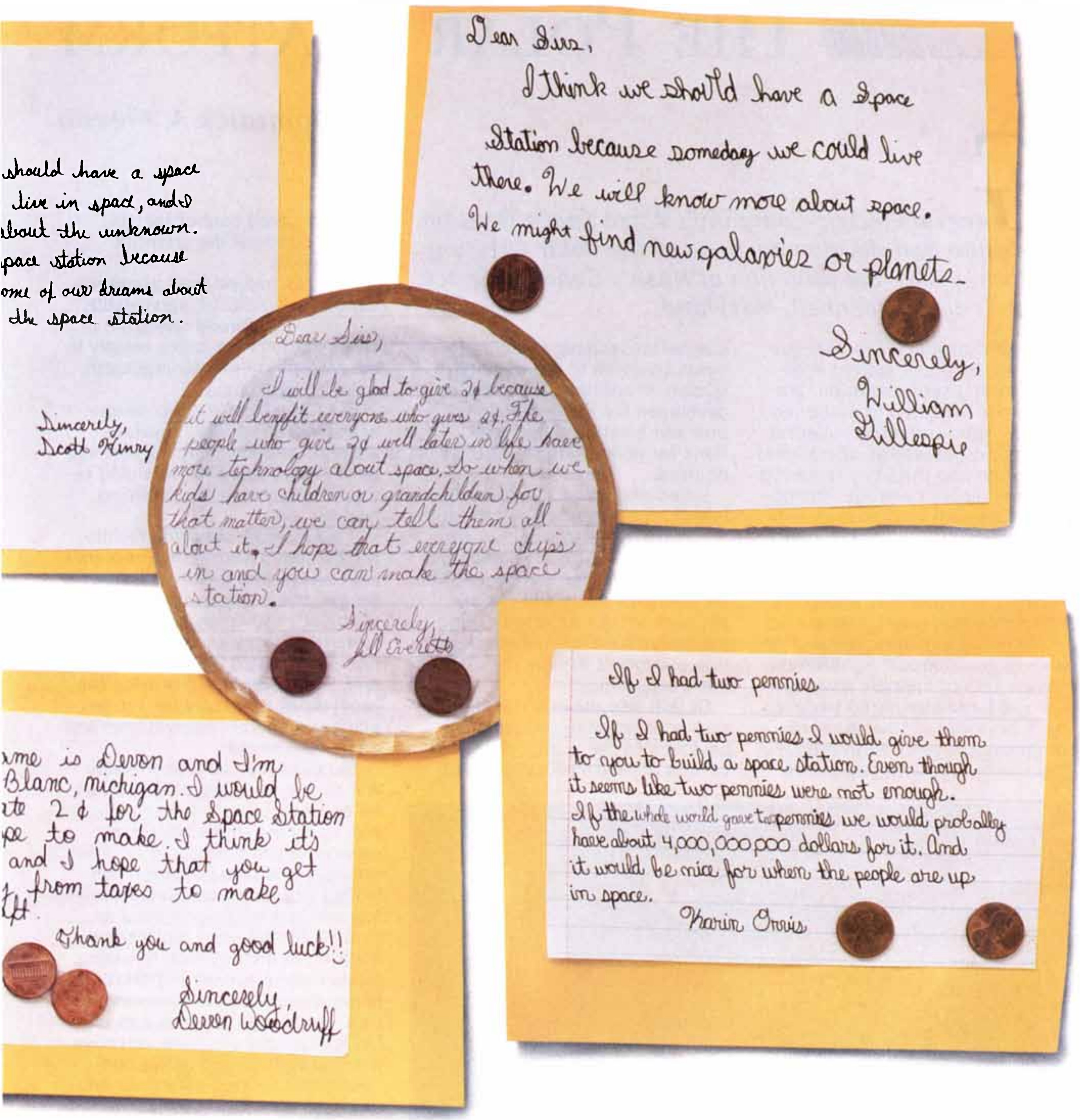
YOU CAN BUY A SPACE STATION INSTEAD?

reminder: The Space Station isn't just being built for scientists or astronauts or corporations or jobs. It is being built for America's children.

It is a legacy of American leadership, one we can leave them at a cost of only two cents a day. A better buy than bubble gum any day.

Your letter will help support the Space Station, too. Write: "Get U.S. Into Space," c/o McDonnell Douglas, P.O. Box 14526, St. Louis, MO 63178.

MCDONNELL DOUGLAS



Letters from the fifth grade at Indian Hill School, Grand Blanc, Michigan.

The Challenge for General Electric:

The Polar Platform

by Dominick A. Aievoli

General Electric Company's Astro Space Division will design and develop portions of the Polar Orbiting Platform under the direction of NASA's Goddard Space Flight Center, Greenbelt, Maryland.

GE's responsibilities include program management, system engineering and user integration. The platform is designed to accommodate scientific and meteorological experiments, as well as operational earth-observing missions requiring complete global coverage. The polar platform will carry instruments designed to study global changes in the earth's atmosphere, oceans and land surfaces.

The polar platform is a long life, serviceable spacecraft, and will be developed as a generic element capable of accommodating different mission sets of multiple payloads or single large integrated payloads. These payloads can be changed periodically or serviced in orbit. Polar platform elements will use resource subsystems and components common to the manned station. In addition, components developed for the manned station and meeting the specifications for polar platform needs will be used.

Scheduled for an October 1995 launch, the modular polar platform is compatible with launches from the Western Test Range, onboard either a Titan IV or the Space Shuttle. Launch vehicle and specific payloads will be selected during the early phases of the space station platform's design and development program.

GE will also provide the necessary equipment to integrate the payloads to the station. They will consist of instruments attached to the manned core of the station to carry out the scientific experiments.

The attached payload accommodations are modular, serviceable, and support a broad spectrum of space station customers, mostly in the astrophysics, solar and earth sciences disciplines.

The GE design features permanent reusable station interface adapters and modular support systems compatible with servicing either in place or in the servicing facility.

The customer servicing facility will be added during the program's next phase. The facility will feature progressively augmented capability, initially providing protected storage capacity and accommodation of the Flight Telerobotic system. Ultimately, it will provide the capabilities necessary for serving platforms, major observatories and attached payloads.

The customer servicing facility will enable servicing of spacecraft by properly suited astronauts as well as by the Flight Telerobotic servicer. It will provide an environmentally and contamination-controlled site for performing servicing and assembly operations including components replacements, addition of instruments, refueling, fluids replenishment and orbiting maneuvering vehicle accommodation. The design features a modular, lightweight structure, adaptive thermal control, and active contamination control, collection and disposal.

GE's Polar Orbiting Platform will give scientists valuable data on the earth's atmosphere, oceans and land surfaces.

Dominick A. Aievoli is the Space Station Program Manager, GE Astro Space.

Before he pushes the edge, IBM takes you beyond it.

From miles away and thousands of feet below, you know exactly what he's feeling.

Using an IBM 3090 with supercomputer power, you've pushed your design through relentless iterations of finite element modeling and fluid dynamics analysis. You've defined the edge.

The IBM 3090 with Vector Facility, for its part, has redefined supercomputer performance, combining up to six vector processors with one of the industry's fastest scalar processors and most efficient memory management systems, parallel processing capabilities, and full IBM System /370 compatibility.

The result is balanced supercomputer performance that can improve your job turnaround by eliminating data bottlenecks, and save the added expense of a separate

front-end processor. Plus, the IBM Vector Facility's modular design lets you start with just the computer power you need today, and make economical in-field upgrades as your needs grow. And that gives the IBM 3090VF a big edge over more expensive and less flexible vector processors.

IBM's integrated system approach combines the power of the IBM 3090VF with the advantages of a single operating environment, low entry cost, low software and maintenance costs, a wide variety of application programs, and IBM support to create a winner in total cost-of-computing.

To find out more, or to arrange to have your IBM Marketing Representative contact you, simply call 1-800-IBM-2468, ext. 20.

The Challenge for the McDonnell Douglas Team

by David C. Wensley

The McDonnell Douglas Astronautics Company's Space Station Division, located in Huntington Beach, California, will design the first five shuttle cargos used in the orbital assembly sequence.

Another of McDonnell Douglas' key assignments is to design and fabricate the intricate truss structure components that form the backbone of the orbiting station.

Made of thin tubes of lightweight graphite-epoxy composite material, this enormous gridwork in space will support the solar power generators, the habitation and laboratory modules and the huge scientific payloads that will search the skies above the station and scan the earth's surface below.

Tests of the assembly sequence and astronaut procedures have already begun in a giant water tank at the McDonnell Douglas facility in Huntington Beach.

A combination of trapped air and balance weights creates a neutral buoyancy condition simulating the weightlessness of space as the astronauts practice the skills and develop the tools that will someday be used to assemble the real craft, 220 nautical miles above the earth.

The giant spools that contain the stowed utility trays with their electrical signal, power and fluid lines, are also a product of McDonnell Douglas, as are the node elements that make up the interconnecting corner segments of the rectangular, closed loop, pressurized volume.

To maintain commonality in the design of the pressure shell, the node structures are designed by Boeing and are provided to McDonnell Douglas for outfitting.

In this step of the construction the avionics are installed in equipment racks, the viewing cupolas are attached, crew control panels and work stations are put in place, utility lines are interconnected, zero-g mobility aids and restraints are attached and surfaces are completed.

McDonnell Douglas has assem-

bled a team of subcontractors who are seasoned veterans in the space business. They will help design and build the complex elements and systems contained in this work package.

IBM has responsibility for the intricate data management system network which will be distributed throughout the station elements. Using an open architecture based on the Intel Corporation's 80386 microchip, IBM will provide a family of standard computers to operate the station's systems and payloads. This computer network will reduce the cost of development and simplify the activities of flight crews, scientific investigators, commercial users and ground operators. The instruments positioned on the space station will generate vast amounts of scientific data, all of which must be processed, reviewed, stored or transmitted to the ground for dissemination to mission sponsors and users. As these data streams are collected and coded for future analysis they will be channeled to waiting ground stations through the communications and tracking system provided by General Electric's RCA Government Communications Systems Division. This team will also provide communications links to astronauts building the craft in space and provide on-board video coverage.

Astro Aerospace will provide the mobile transporter used in construction and maintenance and as a base for the Canadian Mobile Service Center. The station's mass distribution changes and resulting motions are the concern of Honeywell, Inc., which has the responsibility for flight control.

Honeywell's task is to provide the desired flight orientation while minimizing disturbances to delicate microgravity experiments being conducted within the materials and processes laboratories. Ten Honeywell engineers will work full time along with the McDonnell Douglas team at the Houston facility. Tests of physical laws in regimes of temperature and density not attainable on the earth will be permitted in the nearly force-free environment provided by a properly isolated laboratory.

Engineers from McDonnell Douglas assess the workstation layouts inside full-scale cupola mockup.

Large momentum wheels, spinning at 6,000 r.p.m., will provide gentle control torques to counteract destabilizing forces of aerodynamic drag and residual effects of gravity. Precision attitude references, sensing direction from star trackers and ring laser gyros, will feed pointing commands to these gyros and to the attached payloads that search the universe and scan the earth's surface for scientific knowledge.

When the station has to make a major direction change, the highly sophisticated software and electronics in Honeywell's system fire the reaction jets, the 25-pound thrusters attached to the outside of the structure. The control gyros can actually store momentum during part of the orbit and then release it later, to conserve expensive fuel used by the thruster jets.

This same attitude reference will be used to compute commands to the McDonnell Douglas propulsion system during periodic thrusting maneuvers to maintain orbital altitude, change orientation or damp out motions caused by activities such as orbiter docking. Honeywell is also working on an extremely high-speed, fiber optics local area network for the space station.

Lockheed, another member of the team, will provide the externally located heat rejection system that helps maintain the space station temperatures within desired limits. Using the latest in two phase (liquid/vapor) technology, Lockheed's design offers low operating power and constant temperature mounting surfaces for attaching equipment, which improves reliability and performance by avoiding temperature fluctuations.

Lockheed will provide the rotating joint mechanisms that transfer electrical power and coolant fluid across the interface between the main truss assembly and power modules at each end of the transverse boom.

Lockheed will also build the vital outer-craft systems on which flight crews will depend when outside the station performing construction, assembly and repair tasks. The key element in this system is the Extravehicular Maneuvering Unit (EMU) which is being developed by Hamilton Standard Division of United Technologies under subcontract to Lockheed. The EMU, which is a combination space suit and life support "backpack," is a generation ahead of the system in current use. Increased metal components and less fabric will permit a higher operating pressure and eliminate the tedious pre-breathing of 100 percent oxygen that has been required up to now.

To depart the station's main modules, astronauts will simply enter an airlock chamber, don and test their space suits, secure the interior hatch and pump down the chamber to the vacuum condition of space. McDonnell Douglas will provide two such airlocks for the station, including one containing a hyperbaric chamber to be used in the event a crew member suffers from the bends.

Technicians from McDonnell Douglas testing the functional capabilities of the space station resource node.

David C. Wensley is Vice President, Project Integration at the Space Station Division of McDonnell Douglas.

New Challenges for Science

by Thomas M. Donahue

***A** wide range of scientific disciplines, such as space biology, space medicine and microgravity physical science, need the space station as a laboratory in space where their work will involve hands-on attention by astronaut scientists and visits by station personnel.*

Life scientists will examine how serious a threat to long duration space flight is posed by loss of muscle and bone mass, by cardiovascular deconditioning and by psycho-sociological deterioration in interpersonal relationships between people spending long periods of time in an environment of low gravitational forces.

A crucial element in the life sciences laboratory will be a variable "g" centrifuge capable of varying the effective gravitational force from a very small value to 1 g. Tests of physical laws in regimes of temperature and density not attainable on the earth will be permitted in the nearly force-free environment provided by a properly isolated laboratory on the space station.

Studies of fundamental importance in fluid dynamics, mass transport and combustion science will be made in this laboratory.

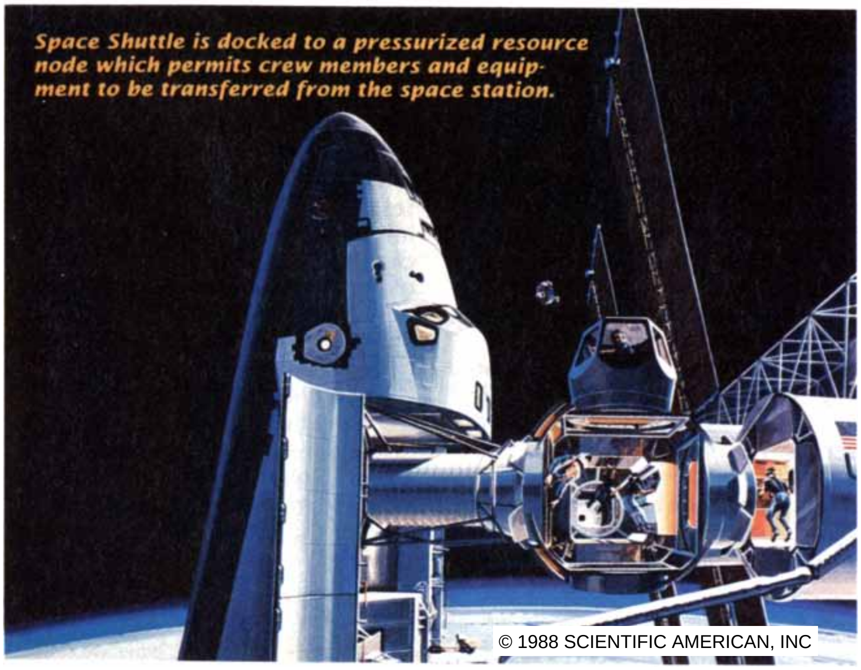
Most scientists believe that such fundamental studies must precede meaningful efforts to carry out materials processing in space.

Where extreme requirements of freedom from vibration and interference by gases do not exist, other scientific disciplines will be able to attach payloads at various places on the station. For example, Astromag, a superconducting magnet spectrometer for cosmic ray studies will be attached along with an astrometric telescope facility designed to detect planets in orbit around other stars.

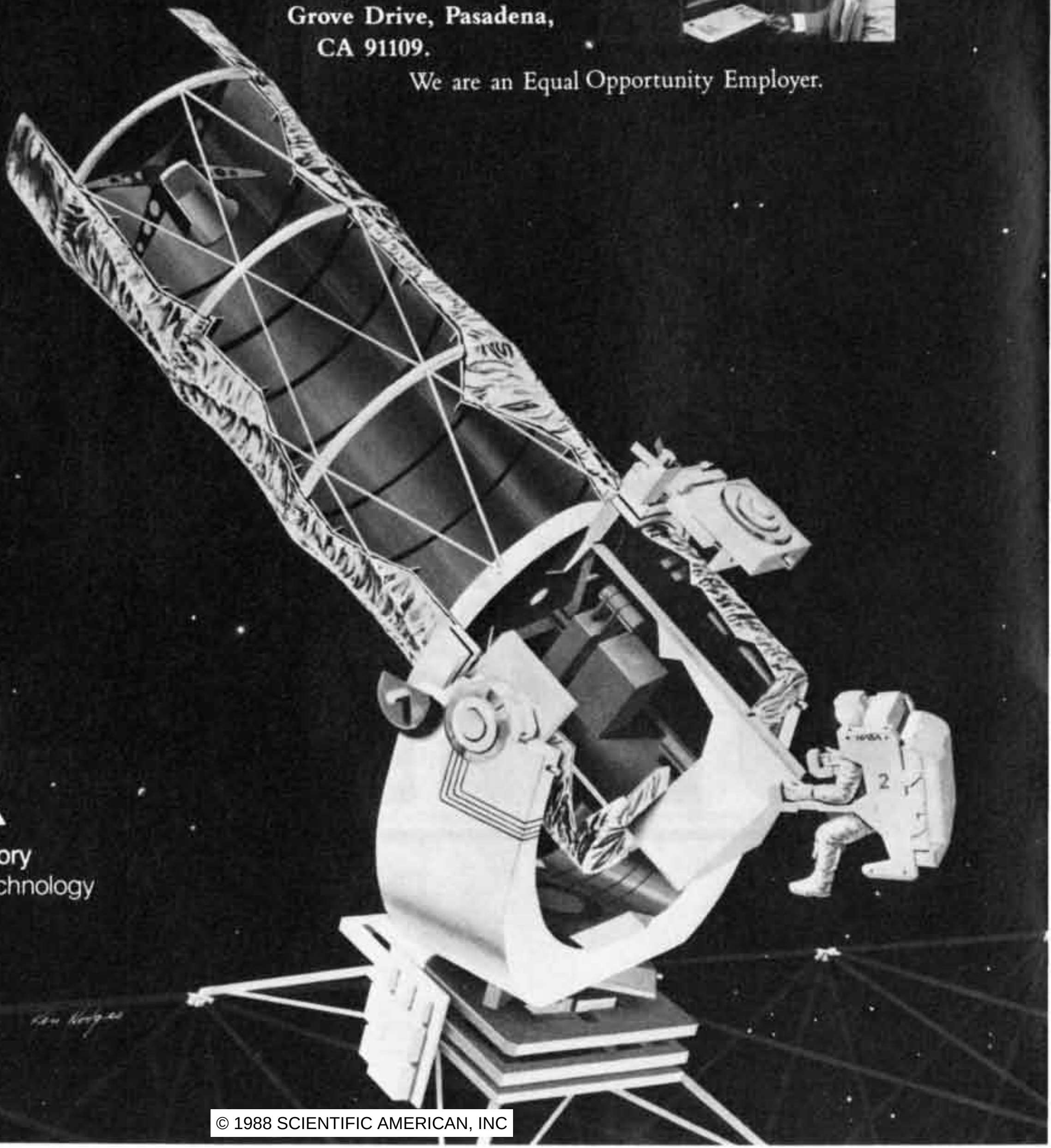
Other sciences will not mount their equipment on the space station, but plan to use it instead as a staging area. Astronomers will maintain and repair their orbiting observatories on the station. These normally would be operated at a good distance from the potential interference of the station. Further in the future, they foresee a need to use the station as a base for assembly of huge arrays of large area sensors of various kinds: multiple mirror optical interferometers, for example, might involve nine or so 1.5-meter telescopes mounted on a structure spanning 100 meters or so on a side.

For planetary scientists, the station may one day serve as a base from which large expeditions depart for other planets or where samples are returned for isolation and preliminary analysis on orbit.

Another element of the space station concept is GE's polar orbiting platform, on which an array of downward-looking remote sensing devices will be placed. This platform would be part of the Earth Observing System, a core element in the Mission to Planet Earth. This study will be designed to develop an understanding of the planet as a dynamic, interactive, evolving system in which living organisms, including human beings, interact with oceans, atmosphere and cryosphere, that in turn, interact with the solid earth, its crust and interior.


Space Shuttle is docked to a pressurized resource node which permits crew members and equipment to be transferred from the space station.

Thomas M. Donahue is the Chairman of the Space Science Board at the National Academy of Sciences.



SCIENCE AND THE CITIZEN

Snakebit

Opposition to U.S. research in biological warfare intensifies

In an interview with SCIENTIFIC AMERICAN in May, Col. David L. Huxsoll, the genial commander of the U.S. Army Medical Research Institute of Infectious Diseases in Fort Detrick, Md., recalled how NBC News once introduced a broadcast on biological weapons. First a cobra, hood extended, flashed on the screen. "Then the cobra disappears," Huxsoll said ruefully, "and there's Dave Huxsoll."

Until recently Huxsoll could afford to be sanguine about this kind of treatment. Although the Pentagon's research into biological weapons—which include microorganisms, toxins and, yes, snake venoms—has never been terribly popular, since 1980 the program has enjoyed support where it matters most: in the Administration. During the Reagan years the annual budget for biological-defense research (so named to underscore its compliance with the 1972 Biological Weapons Convention, which prohibits developing weapons) has leaped from about $15 million to more than $90 million. Administration officials justified the buildup by claiming that the Soviet Union and its surrogates had illegally stockpiled biological weapons and even used them—in the form of a lethal "yellow rain"—in Afghanistan and Southeast Asia.

Now, however, Huxsoll and other biological-defense officials face an upcoming change in administration and greatly intensified opposition. Longstanding critics are still pressing the attack. Biotechnology gadfly Jeremy Rifkin, who has sued to halt the research on the grounds that it is unsafe, is organizing protests in communities where the research is done. Matthew S. Meselson of Harvard University, who helped to discredit the Administration's case for yellow rain (it turned out to be bee excrement), has been chipping away at the so-called Sverdlovsk incident: an outbreak of anthrax in central Russia that occurred in 1979. U.S. officials have charged that the outbreak resulted from an accident at a military plant that was mass-producing the anthrax bacterium. In April, Meselson invited three Soviet physicians to the U.S. to present epidemiological data demonstrating that the outbreak had a natural origin. Military officials such as Huxsoll were not impressed, but many other physicians and epidemiologists found the data credible.

A fresh group of critics has also jumped into the fray. In May the Senate Subcommittee on Oversight of Government Management released a report accusing the biological-defense program of "lax safety enforcement." Supervision of the 100 or so contractors in industry and academia doing research is particularly poor, according to the report. In a new book called *Gene Wars,* journalist Charles Piller and Keith R. Yamamoto, a molecular biologist at the University of California at San Francisco, assert that biological-defense researchers publish seldom—and then usually in second-rate journals—compared with investigators funded by civilian organizations such as the National Institutes of Health. Either the workers are incompetent, Piller and Yamamoto conclude, or they are concealing the nature of their work in violation of the Pentagon's declared policy of openness.

The fiercest criticism focuses on the U.S. Army's Dugway Proving Ground in Utah, a stretch of mountains and desert larger than Rhode Island. In 1984 the Army announced it intended to build a containment laboratory at Dugway to determine how well gas masks, sensors and other equipment respond to airborne pathogens and toxins. Arms-control experts have deplored the plan as a provocative act, maintaining that the laboratory could help to develop and test weapons. Recently two staunchly prodefense leaders in Utah, Governor Norman H. Bangerter and Senator Orrin G. Hatch (both are Republicans), stunned the Pentagon by opposing the laboratory as unsafe.

Part of the problem for Army officials is Dugway's checkered past. Before President Richard M. Nixon banned offensive biological research in 1969, the Army tested anthrax and other lethal agents outdoors at Dugway; areas of the base are still contaminated. (The Army continues to conduct field tests with nonpathogenic microorganisms called simulants.) In 1968 an Army plane flying just north of the base, over ranchland known appropriately as Skull Valley, leaked nerve gas that killed roughly 6,000 sheep. As a result of such incidents, says Kenneth L. Alkema, director of environmental health in Utah, "the trust level is not too high over what the Army says it is going to do."

I. Gary Resnick, who supervises biological research at Dugway, seems genuinely puzzled by the uproar. He points out that the director of safety for the NIH has approved the proposed laboratory's design and that similar laboratories for studying lethal pathogens are operating at the NIH, at the Centers for Disease Control in Atlanta and at Fort Detrick. (Although it is commonly maintained that the Dugway facility would be used to test the virulence of exotic new diseases and toxins, actually this work is already done at Fort Detrick, under Huxsoll's command.) Resnick suggests, somewhat wistfully, that these laboratories are "plums" for their communities and that the Dugway facility would also be a plum for Utah. "There is a ground swell in Utah that is misunderstanding this lab," he says.

Col. Wyett H. Colclasure II, director of testing at Dugway, notes that the Pentagon is considering other locations for the laboratory, including atolls in the South Pacific. Yet he warns that for economic and other reasons the Army may build the laboratory at Dugway in spite of the local opposition. "It is not our task here to take a public-opinion poll and ask do you think we ought to do this," he says.

Notwithstanding, in the past few months military officials have mounted a vigorous public-relations campaign—issuing reports, testifying before Congress and meeting with journalists—to justify both the Dugway laboratory and the entire biological-defense program. Perhaps because relations between the superpowers are warming and because several of the most hawkish denizens of the Pentagon have recently departed, the program is being justified in a subtly different manner than in years past, according to Ivo J. Spalatin, a member of the House Foreign Affairs Committee staff. A few officials still resort to what Spalatin calls "scare tactics." In May, for example, Thomas J. Welch, a deputy assistant secretary of defense, testified that 10 countries have or are suspected of having biological weapons, up from four in 1972. (The identity of the nations, with the exception of the U.S.S.R., is classified.)

But others steer away from these military issues, dwelling on the safety and openness of the program and its benefits not only for American soldiers but also for civilians around the world. Huxsoll, for instance, likes to talk about how research done at Fort Detrick has helped the Chinese to combat hemorrhagic fever and the Malaysians to treat snakebite. "What we do here is no different from what is done at Merck and Company, or NIH or CDC," he says.

Then why can't the research be done by a civilian agency, thus eliminating the distrust that has always shadowed the military program? "I don't think the NIH or CDC can appreciate military needs," Huxsoll replies quickly. Yet some legislators are now quietly examining this option, according to Spalatin. "We don't know how realistic or responsible that alternative is yet," he says, "but we are in the process of examining it." *—John Horgan*

Getting Warmer?

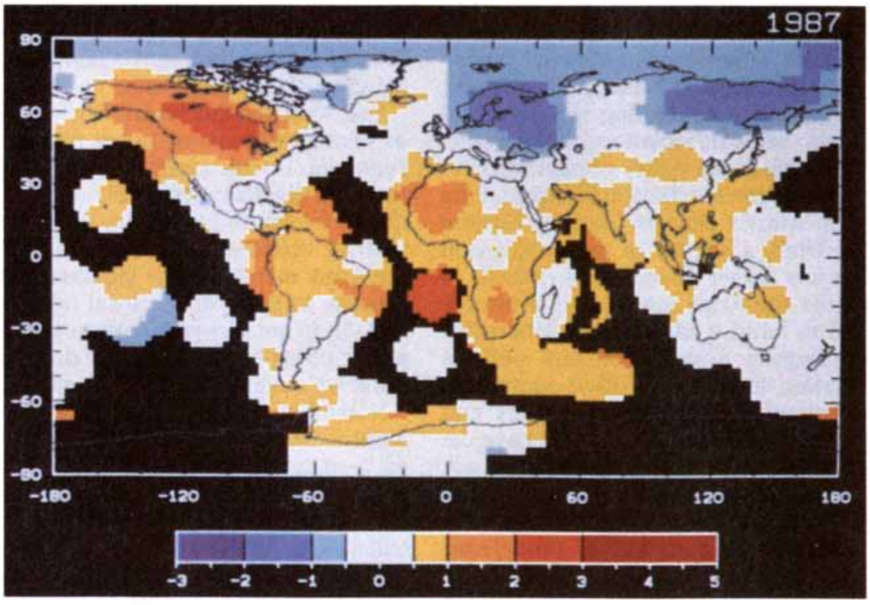
This has (so far) been the warmest decade in 127 years

Last year was the warmest on record, according to British and U.S. investigators. The global average temperature, as measured by instruments on land and at sea, was .05 degree Celsius warmer than it was during the next two warmest years, 1981 and 1983. The unusual warmth of the 1980's is most evident in the Southern Hemisphere, where seven of the eight warmest years on record have occurred within the decade.

The workers present their assessment in a letter to *Nature.* They are Philip D. Jones and Thomas M. L. Wigley of the University of East Anglia, Sergej Lebedeff and James E. Hansen of the National Aeronautics and Space Administration's Goddard Space Flight Center and others. The latest result updates an earlier study that found the average global temperature has increased by about .5 degree C. since the time of the earliest reliable records, which date back to 1861.

The investigators caution that the unusual warmth of 1987 may have been in part the result of a strong El Niño event in 1986–87. Nevertheless, they write, the persistent warmth of the 1980's "could indicate the consequences of increased concentrations of CO_2 and other radiatively active gases in the atmosphere."

The next few years should make clear whether a greenhouse warming has really started



AVERAGE TEMPERATURES in 1987, the warmest year on record, are shown as deviations above and below the average temperatures recorded over a 30-year period ending in 1980. The units are degrees Celsius.

Atmospheric scientists are in general agreement that the continuing increase in the amount of carbon dioxide, together with increases in other trace gases such as methane and nitrous oxide, is likely to cause a detectable global warming over the next few decades. The gases absorb thermal radiation emitted by the earth and prevent its escape into space, thereby warming the lower atmosphere (the greenhouse effect). The concentration of atmospheric carbon dioxide is now 23 percent higher than it was before the Industrial Revolution and is increasing by .4 percent per year.

The warmth of the 1980's might simply represent natural variability; there was, for instance, a milder unexplained warm period in about 1940. Theoretical studies do not encourage complacency, however. Veerabhadran Ramanathan of the University of Chicago wrote in *Science* recently that the predicted surface warming would be "unprecedented during an interglacial period." According to Wigley, the models suggest the global average temperature should have increased by between .4 degree and 1.1 degrees C. over the past century; the actual increase since 1861 is thus at the lower end of the range.

The models do include major uncertainties: it is not clear, for example, how much the oceans would delay warming by acting as a heat sink. On the other hand, effects that might have disguised greenhouse warming, such as aerosols from volcanic erruptions that absorb solar radiation, are not well understood either. Whether greenhouse warming is really taking place should become clear in the course of the next two decades. "All we can do is present the facts—you must draw the conclusions," Lebedeff says. *—Tim Beardsley*

Ill-starred

Can scientists themselves help to allocate funds for science?

May saw an unfavorable alignment of the science-policy stars in Washington. The conjunction of a lame-duck president and last December's agreement with Congress limiting new domestic spending meant that several "big science" projects took serious hits on Capitol Hill. One subcommittee in the House of Representatives voted to postpone all new construction starts proposed for fiscal 1989, including the planned Superconducting Supercollider and the Advanced Photon Light Source, a $456-million facility that would produce synchrotron radiation for materials-science investigations. Preliminary budget allocations in the Senate were seen as a disaster by some congressional staff members; in one view current budget allocations would require terminating the space station. The Senate's allocations also mean the National Science Foundation, which supports much of the nation's civilian basic research, will be lucky to see any significant increase, in spite of the Administration's plan to double the agency's budget over five years.

Already there have been ominous rumblings from the house of science. Research directors say facilities at several laboratories, including the Brookhaven National Laboratory and the Stanford Synchrotron Radiation Laboratory (SSRL), are working at well under capacity because of current budget shortfalls. Arthur I. Bienenstock, head of the SSRL, says he has "a machine I can't use" that could be doing important development work for the Advanced Photon Light Source. Alan Schriesheim, director of the Argonne National Laboratory, which was recently designated as a special center for commercialization of the new high-temperature superconductors, says the $2 million extra that project will bring this year "will not correct an underlying deficit." Robert J. Havlen, director of observatory services for the National Radio Astronomy Observatories, notes that inadequate maintenance has led to problems at the Very Large Array in New Mexico that are "acute enough so that we are worried we might have to restrict some observations."

What to do? At the National Academy of Sciences' annual meeting its president, Frank Press, a former science adviser to President Jimmy Carter, made the radical suggestion that scientists should decide among themselves what priority to attach to major scientific endeavors rather than leaving their fate in the hands of Congress. Press said the proposed budget's support for several mammoth science projects "tests political reality" in the face of the deficit and congressional affection for social programs.

Press went on to suggest some priorities, including a strong plea for "small science" carried out by individual scientists. He urged absolute priority for training and research grants that reach large numbers of individual scientists and engineers. National crises such as AIDS and the exploitation of opportunities such as the new high-temperature superconductors would also have first priority. On the other hand, in Press's view some major projects supported by the Administration, such as the Superconducting Supercollider and the mapping and sequencing of the human genome, should be put in a second category, to be delayed if the budget crisis demands it. The defense R&D budget, the Space Station and manned space flight he relegated to the last, "political" tier, to be judged on political merits.

Some science-policy observers recognize the academy's expertise but question the idea of having scientists decide who should be the first to be thrown out of the balloon. Alvin W. Trivelpiece, executive director of the American Association for the Advancement of Science, argues that science should always promote scientific excellence and warns against "circling the wagons and then shooting inwards." Trivelpiece, a strong advocate of the SSC, says he is "not optimistic about the chances of scientists being able to establish this kind of priority." He points out that there are already many scientifically qualified advisers in government; the problem may be that they are not listened to.

Joshua Lederberg, the president of Rockefeller University, foresees a potential trap: in the worst case, if the academy is asked to set priorities but then is excluded from taking part in the decisions, it would in effect be issuing "an open invitation to drop everything that was not of the utmost exigency." Lewis M. Branscomb of Harvard University, who is studying science-policy issues, comments that by relegating some projects to a "political" category, Press risks leaving the impression that science is being self-serving—that it is shirking its responsibility to offer advice on major projects that may be important to national defense or the economy.

Doubts or no doubts, in late May the House of Representatives tentatively agreed to go along with the Senate and accept language in a joint budget resolution that recognizes a role for the academy. No agreement was reached on precisely what that role will be, and the wording was deliberately left vague. Academy staff people have begun informal discussions with the Senate budget committee, and a joint study is planned to establish categories and criteria for the allocation of funds to science. Press thinks an interim document could be completed as early as November. *—T.M.B.*

PHYSICAL SCIENCES

Sons of STM

Scanning tunneling microscope spawns diverse applications

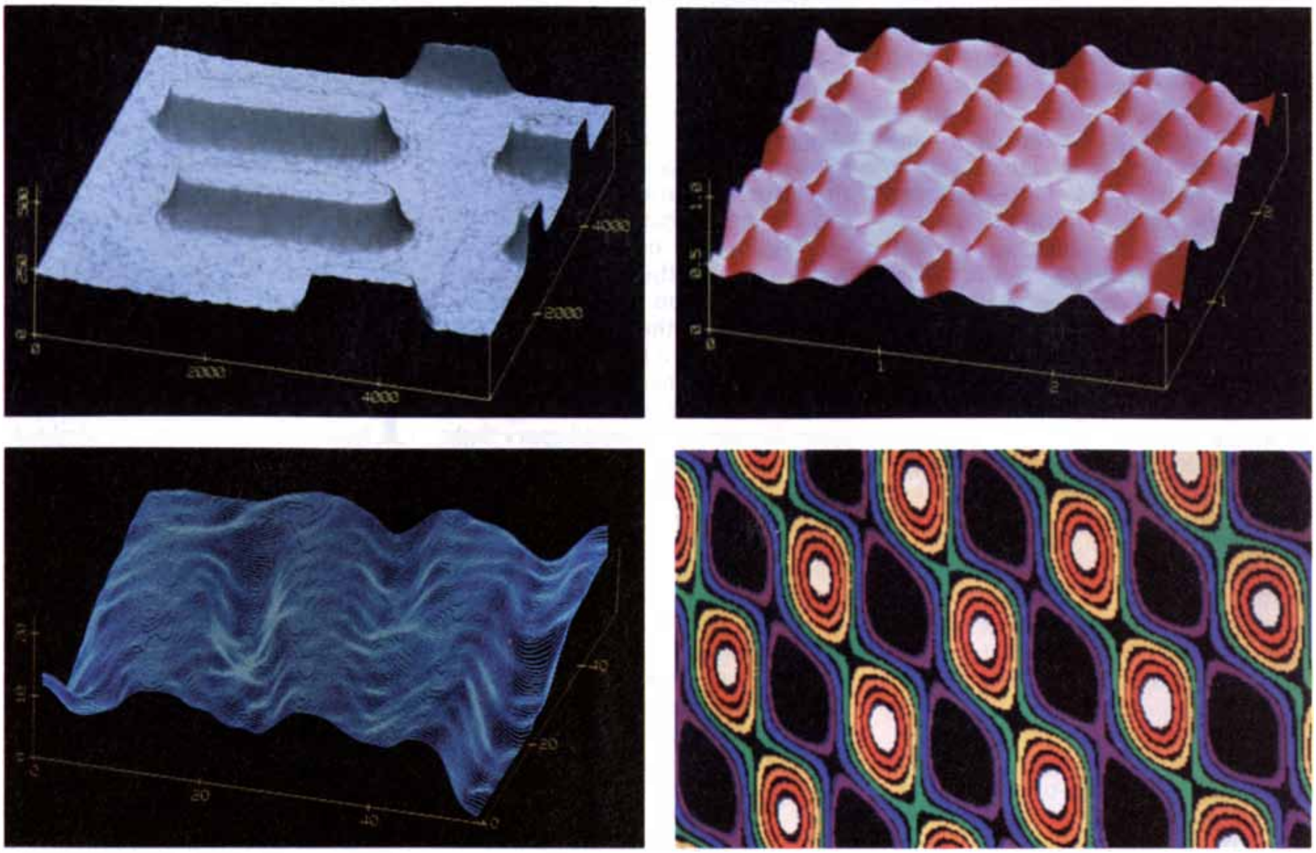
In the early 1980's Gerd Binnig and Heinrich Rohrer of the IBM Zurich Research Laboratory invented a device so sensitive to the contours of an electrically conducting surface that the machine could "see" individual atoms. The principles of the scanning tunneling microscope (STM) are now spawning a new generation of instruments for exploring the atomic-scale realm. The devices can study new surface features such as magnetic fields and temperature.

The STM maneuvers a metallic needle in three dimensions with a precision of a few angstrom units (ten-billionths of a meter). The needle has a voltage, so that when it approaches to within a few angstroms of a surface, electrons suddenly "tunnel" across the gap. This quantum-mechanical effect is exquisitely sensitive to the width of the gap: a difference of one angstrom—less than the diameter of an atom—can change the tunneling current by a factor of 10. As the tip sweeps across the surface, a feedback mechanism holds it at a fixed distance, maintaining a constant current. A computer translates the motion of the tip into a topographic map of the surface.

The STM works only on electrical conductors. Nonconductors must be covered with a metallic coat first, a technique that has yielded striking pictures of biological molecules. A team at the Swiss Federal Institute of Technology produced images of DNA complexes showing helical twists that resemble pearls in a necklace. Joseph A. N. Zasadzinski of the University of California at Santa Barbara captured a phospholipid membrane, crinkled like a piece of corrugated cardboard. He believes this "ripple phase" enables lung membranes to expand with each breath. Premature infants who lack the membrane suffer from breathing difficulties.

Another strategy for studying nonconducting specimens is to pin them

The microscopic realm is yielding a wealth of secrets to these versatile tools



SCANNING TUNNELING MICROSCOPE helps to monitor production quality of an optical-disk stamper **(top left).** ***STM reveals low bumps corresponding to sulfur atoms in the tellurium layer of a bismuth-tellurium-sulfur semimetal*** **(top right);** ***the extra atoms may explain why the substance contains more sulfur than predicted. STM captures ripple phase of a phospholipid membrane*** **(bottom left);** ***Joseph A. N. Zasadzinski made the platinum-carbon replica from which the image was taken. In an image by Scot Gould an atomic-force microscope traces the pattern of methyl groups in a crystal of the amino acid leucine*** **(bottom right).** ***STM images were made by gear from Digital Instruments, Inc.; units are nanometers.***

to a conducting surface such as a gold crystal. If the probe pushes down hard enough on a molecule, electrons can tunnel right through. "It's like looking for shells in the sand by pressing your hand on the sand's surface," says Stuart M. Lindsay of Arizona State University. In the case of DNA, tunneling occurs when the tip presses only one angstrom into the 20-angstrom-thick molecule. Lindsay recently showed that images can be made of molecules immersed in an aqueous solution. This has several advantages: the molecules retain their natural shape and can be easily controlled and kept clean in a liquid medium. The technique could reveal poorly understood biomolecular structures and perhaps even capture a reaction in progress, such as an enzyme acting on a molecule, Lindsay remarks.

The precise feedback control pioneered by the STM makes possible a variety of spin-offs. Two years ago Binnig designed one of the first, the atomic-force microscope. This instrument works on nonconducting materials because it does not rely on a tunneling current. Instead its diamond needle skims the surface like a record-player stylus, exerting a constant force of only a billionth of a pound. In the newest models a tiny mirror on the probe reflects a laser beam, which provides a more reliable measurement of tip deflection than the STM's tunneling current does.

The atomic-force microscope could probe semiconductors and other surfaces that are insulated by oxidation. It enabled Paul K. Hansma's group at Santa Barbara to find tiny pits forming in stainless steel immersed in seawater for barely 15 minutes, long before corrosion could be detected by any other method. The atomic-force microscope demands a much smaller tip than the STM does. Hansma's group selects probes by trial and error: "We bought our first diamond for six dollars at a pawn shop, smashed it with a hammer and combed through for a sharp fragment," he says.

H. Kumar Wickramasinghe of the IBM Thomas J. Watson Research Center heads a group developing scanning-probe microscopes that look not at the topography of a surface but at magnetic, electrostatic and interatomic van der Waals forces at the surface. IBM's magnetic-force microscope employs a magnetic wire that is made to wiggle at its resonant frequency by an alternating voltage applied to a piezoelectric "reed." As the wire scans, it interacts with magnetic fields at a surface, which change its resonant frequency. By maintaining the gap distance so that the wire vibrates at a constant rate, the instrument maps the magnetic profile of a surface to a resolution of 500 angstroms. It offers a new diagnostic tool for magneto-optic disks, thin-film disks and thin-film recording heads.

The IBM group has also built what Wickramasinghe calls "the world's smallest thermometer." The probe is a tungsten needle coated with an insulator except at the tip and then sprayed with a second metal. It is a thermocouple, a device that develops a voltage gap when the outer layer of metal is warmed by an outside heat source. The probe scans a surface that is being bombarded by laser pulses and pinpoints molecules that absorb the light and heat up. The elements composing the molecules can be identified from the wavelengths they absorb.

William J. Kaiser and L. Douglas Bell of the Jet Propulsion Laboratory extended the reach of the STM below a specimen's surface. They found that tunneling electrons penetrate hundreds of angstroms into a surface before they lose energy. These "ballistic" electrons are "a little searchlight for finding what's down below the surface," Kaiser says. Ballistic-electron-emission microscopy can map the Schottky barrier, an energy threshold at the critical metal-semiconductor interface in microelectronic devices. By raising the tunnel-probe voltage until the electrons have enough energy to clear the barrier, one can measure the barrier height. Changes in the height reveal defects such as the diffusion of semiconductor atoms into the metal. The method offers for the first time a way to study these vital features.

Steven B. Waltman, together with Kaiser, has also invented a clever STM spin-off: a "tunneling sensor" that could eventually be fabricated on a microchip. The sensor would resemble a diving board poised over a microscopic swimming pool. A tunneling probe affixed to the bottom of the board would detect any bending. The sensor could serve as an accelerometer 100,000 times more sensitive than conventional instruments. The JPL contemplates installing it in spacecraft and envisions uses in robotics and biomedicine. The JPL's Stephen L. Prusha is investigating whether the sensors could be implanted in a paraplegic to detect limb motion and provide feedback to electrical muscle stimulators, enabling the person to walk. —*June Kinoshita*

Rotor Rooters

Students attempt to build a human-powered helicopter

In the early 1490's Leonardo da Vinci outlined his version of a human-powered helicopter: a contraption topped by a helical propeller carved out of wood. It will not work. Nearly half a millennium later a team of budding engineers armed with computers and lightweight synthetic materials have come up with a new design. It does not work either—yet.

The students at the California Polytechnic State University hope to break this frontier in flight soon. Cal Poly's so-called da Vinci Project was born in 1981, soon after the American Helicopter Society established the $25,000 Igor I. Sikorsky Prize for builders of the first human-powered helicopter. Since then the project has consumed $100,000 in donations—as well as free computer time, materials and advice from several aerospace companies—and countless hours of students' time.

A spokesman for the helicopter society says the da Vinci Project is "unquestionably" the front-runner among several contenders (including a retired aeronautical engineer and fruit farmer in Oregon) for the Sikorsky prize. "I don't know about that," says William B. Patterson, the project's faculty adviser from the start, "but we are certainly the most dogged and determined."

The group has evolved as students graduate and as interest flares and subsides, but its hopes still rest on a design that was chosen in the first year after extensive theoretical studies. The craft consists of a stationary cycle topped by a giant rotor with a propeller on each tip. The drive mechanism is somewhat indirect. As the pilot pedals, he rotates a spool that takes up polyethylene string wound around the shafts of the propellers, causing them to spin and in turn to push the rotor around. The pilot pedals until the string breaks (a common mishap) or comes to its end, which takes about five minutes.

The students found out early that the machine cannot be tested outdoors; the slightest breeze blows it over. Tests have taken place in the Cal Poly gym and, more recently, in a

Student engineers hope to overcome broken strings and rotors and make their ungainly craft hover

HUMAN-POWERED CRAFT was built by students at the California Polytechnic State University. The cyclist, by means of a string-drive mechanism, causes propellers on opposite ends of the 140-foot-long rotor to spin. The propellers then push the rotor around. Photograph by Bernie Roddam of the Douglas Aircraft Company.

hangar at a McDonnell Douglas plant. Years of tests and modifications failed to get a hop out of a craft whose rotor was 102 feet wide. Last year the students revamped the machine and gave it a new, 140-foot-wide rotor made of carbon fiber and plastic foam. During a test in November a guy wire snapped and both propellers were smashed. The students replaced the propellers, designed a new guy-wire system and ran another test in the spring.

During the initial run the machine teetered and skidded so wildly that several participants grabbed on to its frame. Suspecting they might have kept the craft from lifting off, they decided "just to let it do what it wanted" in the next run, Scott Larwood, a senior who manages the project, recalls. As the cyclist pedaled furiously, the craft abruptly keeled over, snapping off a rotor blade.

The team hopes to redesign the rotor this summer and retest the craft in the fall. The students have apparently lowered their sights as a result of their setbacks. "The craft we have now we don't believe will win the Sikorsky prize," Larwood says, noting that to win the prize a craft must rise 10 feet into the air as well as hovering for a full minute. "We just want to hover," he says. "That would still be a world record." —*J.H.*

Upfreezing

An experiment demonstrates how buried rocks rise

Any Yankee farmer is familiar with upfreezing. Just when he thinks he has wrenched the last rock from his field, a new crop surfaces. Geologists have long blamed this upward migration of buried rocks (and other objects) on the fact that water expands as it freezes, but they have never agreed on precisely how cause leads to effect.

There are two major hypotheses: frost-push and frost-pull. Frost-push holds that during a freeze cold penetrates downward through a rock more rapidly than through the surrounding soil. Hence water directly under the rock freezes and nudges it up into the soil above. During a thaw a pedestal of ice supposedly holds the rock up long enough for thawing soil to slump around its sides and secure it in the higher position.

Frost-pull posits that in a cold snap soil freezes first around the top of the rock, gripping it tightly. As this soil expands and heaves upward it pulls the rock with it, and a cavity forms under the rock. Soil then slumps into the cavity before the rock can settle back into its former position.

Now Suzanne Prestrud Anderson of the University of Washington says she has resolved the issue. She embedded 13 thermocouples (which measure temperature) and a long thin rod in a chunk of gneiss and buried it in a wastebasket-size container filled with soil. The rod descended through the bottom of the container to a device that measures vertical displacement. Forty thermocouples and a displacement sensor also monitored the internal temperature and surface motion of the soil—not mere dirt but silt from Walla Walla, Wash., that is "known to be frost susceptible."

Anderson placed the container in a "cold room" and generated seven freeze-thaw cycles over a period of 3,000 hours (about four months), during which the gneiss rose through some five inches of silt and breached the surface. She found that cold penetrated slightly faster through the stone than through the silt but not fast enough to induce frost-push. The rock heaved upward after the temperature had dropped below freezing in and around the upper half of the rock—but not yet in or around its lower half. The findings "clearly demonstrate the mechanics of upfreezing by the frost-pull process," Anderson concludes in *Geological Society of America Bulletin.*

Does Anderson's research suggest how farmers might prevent upfreezing? "I don't see any direct way it can help," she muses. "They might try using less frost-susceptible soil, such as gravel, but of course you can't grow anything in gravel." —*J.H.*

Superduds

Making stars explode in a computer is a tricky task

Last January a group of supernova theorists gathered in Aspen, Colo., to swap ideas and perhaps gloat a little. After all, the appearance of supernova 1987A a year earlier had vindicated in spectacular fashion their basic views of how stars explode. But the conference-goers were reminded of how far they are from full understanding when James R. Wilson of the Lawrence Livermore National Laboratory announced that his computer-generated supernovas were no longer exploding.

"It was depressing," recalls Stanford E. Woosley of the University of California at Santa Cruz, who incorporates Wilson's models into his own work on stellar evolution. The failure of supernova models to explode is actually quite common, Woosley says, although it receives scant attention. "It's very complicated and a little embarrassing," he explains.

Supernova modeling first flowered in the late 1960's at Livermore, where workers have access to the sophisticated physics codes and, more important, the powerful computers used in nuclear-weapons research. "Most of the theories before computer modeling were hopelessly vague," says Wilson, who has been simulating stellar explosions since 1970, longer than anyone still active in the field.

Wilson is constantly modifying his models—complex amalgams of nuclear physics, general relativity and hydrodynamics—to reflect new and supposedly improved findings and theories. As a result of his tinkering "sometimes the models explode and sometimes they don't," he says. "That has been the pattern over the years." A new model he tested late last year was a dud, leading to his announcement at the Aspen meeting. With the help of a colleague at Livermore, Wilson says, he has recently coaxed the new model into exploding. An exploding model, of course, is not necessarily a correct one. In 1973, for example, he incorporated a new proposal on how neutrinos interact with nuclei into his model and generated beautiful explosions; one year later the proposal was proved wrong.

Wilson simulates Type II supernovas, which are thought to occur when the core of a massive star collapses and rebounds, generating a shock wave that blows off the star's outer layers. (Type I's are white dwarfs that drag matter from a companion star and then ignite in a fusion reaction.) These events, which occupy less than a second of real time, occupy one of Livermore's Cray supercomputers for up to five hours. The drama is played out not on a computer screen but on long printouts that are crammed with numbers.

Edward A. Baron of the State University of New York at Stony Brook and Jerry Cooperstein of the Brookhaven National Laboratory, Wilson's major rivals in simulating Type II's, have experienced similar ups and downs. They began modeling in 1981 and produced only duds for more than three years. In late 1984 they finally generated explosions with a model whose core is represented by a "softer" equation of state—a measure, in effect, of its springiness—than Wilson's model. "A rubber ball has a soft equation of state; a brick's is harder," Baron explains. Recently, he adds, he and Cooperstein have "put some new physics in [the model] and it's getting a little iffy."

Baron and Cooperstein's explosions are "prompt": the simulated stellar core collapses and rebounds so violently that the resulting shock wave passes all the way through the star and blows off its outer layers. Wilson's explosions are delayed: the shock wave stalls briefly in the outer region of the star and during the pause—dubbed "the pause that refreshes" by Hans A. Bethe of Cornell University—the core grows hotter and emits a pulse of neutrinos that rejuvenates the shock wave. Bethe, who won a Nobel prize in 1967 for his work in stellar physics, suggests both mechanisms may occur in nature but says he favors prompt explosions, which seem to generate more than twice the explosive energy of delayed explosions. Wilson acknowledges that "more people are probably attached to prompt explosions" than to the delayed variety and says he is searching for new equations of state that will make his model explode promptly.

The failings of supernova models, Wilson says, indicate that "after 20 years we are still not up to the problem." More observational data might help. Although Supernova 1987A confirmed several broad predictions of supernova theory, such as the emission of neutrinos, it has not provided the kind of information needed to confirm the highly detailed computer models. Modelers would also benefit, Wilson says, from better understanding of high-density nuclear physics. "The equation of state is the biggest unknown," he notes. "Of course, there may be unknowns we don't know about." —*J.H.*

Odd Couple

Astronomers spot a pulsar with a "bizarre" companion

By painstakingly deciphering a stream of rapid radio pulses, three Princeton University astronomers have found a strange and violent drama unfolding some 3,000 light-years from the earth, in the constellation Sagitta.

The path to discovery began for Andrew S. Fruchter, Daniel R. Stinebring and Joseph H. Taylor with a search for rapidly rotating pulsars. A pulsar is a neutron star, the compact kernel of collapsed matter left by a supernova. If a neutron star spins, it emits a powerful pulse of electromagnetic radiation during each rotation. Most known pulsars spin several times a second; a handful, the so-called millisecond pulsars, spin hundreds of times a second.

With a giant radio receiver in Arecibo, P.R., the workers discovered a pulsar that rotates 622 times a second. (Only one faster-spinning pulsar is known.) Tiny fluctuations in the rotation speed indicated that the pulsar was locked in a virtual embrace with an object 1.25 solar diameters away. That in itself was not remarkable; eight other binary pulsars have been spotted. What was more unusual—unique, in fact—was that the pulsar periodically disappeared behind its companion.

The investigators' computer analysis of the eclipses has revealed a great deal about the companion's character. Immediately before and after each eclipse the stream of pulses is slightly delayed, apparently by a plasma that envelops the companion and streams behind it like a comet's tail. The companion is roughly the same size as the sun and yet its mass is only about 2 percent as great, making it perhaps the least massive object observed outside the solar system. "It's a bizarre object," Stinebring says.

The workers speculate that when the pulsar formed some 100 million years ago, it was spinning much less rapidly, perhaps only a few times a second. At that time the companion probably was a normal star, much like the sun. As the companion burned up its hydrogen fuel it swelled into a red giant and matter from its outer layers began funneling toward the pulsar. The accretion of matter made the pulsar whirl ever faster and emit increasing amounts of radiation.

This radiation, including radio-spectrum photons like the ones detected on the earth and other particles accelerated in the pulsar's magnetic field, is now "blasting the hell out of the companion," Stinebring says. Heated up by the radiation, the companion is spewing even more of its mass into space in the form of a plasma wind. Eventually, perhaps after several hundred million years, the companion may evaporate entirely.

The two known millisecond pulsars that have no companions may have undergone a similar process. "The companion donates a large part of its mass to the neutron star," Stinebring observes, "and then the ungrateful neutron star turns around and kills it off." —*J.H.*

BIOLOGICAL SCIENCES

Language of Translation

Subtle chemistry helps a cell to read the genetic code

Molecular biologists are finally answering a question that has puzzled them since the genetic code was deciphered more than 20 years ago. The question has to do with how a cell interprets the code to synthesize proteins from amino acids. Biologists know that the critical "interpreters" in protein synthesis are nucleic acids called transfer RNA's, and that each tRNA specializes in binding to one of the 20 amino acids. Until now, however, no one has been able to tell what part of the tRNA molecule determines its amino acid specificity.

Recently workers in two laboratories announced that they had independently found a determinant for the tRNA that binds to the amino acid alanine. The finding has created a stir because it promises to fill a conspicuous gap in understanding of the step in protein synthesis known as translation. The nexus of translation is a messenger-RNA molecule copied from the DNA code. It describes a sequence of amino acids that makes up a particular protein. The tRNA molecules help to assemble the protein by bringing amino acids to the messenger molecule and then reading the messenger to find out where in the protein's sequence each amino acid belongs.

Investigators initially assumed that the characteristics determining tRNA identity would be gross structural features. Yet when Alexander Rich and his colleagues at the Massachusetts Institute of Technology deduced the structure of tRNA 14 years ago, they found that all tRNA's looked pretty much the same. At that point they realized the determinants could be subtler, perhaps involving just a few of the base pairs of which tRNA is made.

Work in the area languished until 1986, when John N. Abelson and his colleagues at the California Institute of Technology introduced techniques for observing the effects of mutated base pairs on tRNA specificity. A year later workers at M.I.T. applied the techniques to generate a series of 30 mutants of the alanine tRNA. Ya-Ming Hou and Paul Schimmel, a 20-year veteran


GTP ZX-TURBO
NISSAN
BRABHAM / MORTON
BRABHAM / MORTON
GTP ZX-TURBO
ELECTRAMOTIVE
83
GOODYEAR
BOSCH
GOODYEAR
NISSAN

of tRNA-identity research, found to their surprise that most of the mutations had no impact on binding—until they changed a single base pair near the end of one of the molecule's arms.

The effect was dramatic: the mutation severely curtailed the tRNA's ability to recognize alanine. Furthermore, if other tRNA's were engineered to contain the crucial base pair, their ability to bind to alanine improved. In a paper published in *Nature,* Hou and Schimmel conclude that the pair, designated 3-70 because of its position in the tRNA molecule, is a major determinant of the alanine tRNA's identity.

Schimmel had relayed his preliminary findings to William H. McClain of the University of Wisconsin at Madison, who was also studying the 3-70 pair. McClain soon obtained similar results from his own work. Writing with his colleague Kay Foss in *Science,* however, McClain downplays the significance of the 3-70 pair, saying it cannot be solely responsible for the tRNA's specificity. Hou and Schimmel maintain their results implicate 3-70 more strongly than McClain's.

Both groups predict that the determinants of the remaining 19 tRNA's will be established within the next five years, and both think the remaining determinants will be more complex, perhaps involving several base pairs throughout the molecule. The investigators also caution that, although some have heralded tRNA determinants as the second genetic code, the determinants are probably not systematic, and so disclosure of the other 19 identities will have to proceed by trial and error. *—Karen Wright*

Genetic Skeleton

What gives a cell form may also control its genes

At first glance the cytoskeleton of a mammalian cell would not appear to be closely related to gene expression. The cytoskeleton, a complex network of proteins and other molecules, endows the cell with form and enables it to move. Genes, on the other hand, embody the information needed for making proteins. Contrary to expectations, however, the cytoskeleton appears to be a crucial intervening actor in the control of gene expression.

That finding comes from work by Theodore T. Puck of the Eleanor Roosevelt Institute for Cancer Research and the University of Colorado Health Science Center in Denver and his colleagues on a line of Chinese hamster ovary cells (CHO) Puck isolated in 1958. In the laboratory the CHO cells became spontaneously transformed to the cancerous state, assuming several characteristics often seen in tumor cells. They lost their normal elongated shape, becoming compact and round. In addition many genes in the CHO cells became sequestered, or shielded from the action of substances that would ordinarily affect them.

The changes in form undergone by the CHO cells are due to disorganization of the cytoskeleton. Puck and his co-workers showed some time ago that these cells can be restored to their normal appearance by a compound called cyclic AMP (cAMP), which reorganizes the cytoskeleton. Intriguingly, they found that cAMP also exposes much of the sequestered genetic material in the CHO cells.

Is the exposure random, or does it affect specific genes and thereby serve as a form of genetic control? That was the question Puck and his co-workers set out to answer in their most recent experiments. They analyzed the shielding of particular genes before and after addition of cAMP. The indicator they employed was an enzyme called DNAse I, which breaks down the DNA chain into its subunits.

They found that the exposure to DNAse is indeed specific. Of 47 genes analyzed, 40 were resistant to the action of DNAse before the addition of cAMP. Of those 40, 15 were consistently converted to DNAse sensitivity after cAMP treatment. The other 25 were consistently resistant to digestion by DNAse after addition of cAMP. Clearly, changes in the cytoskeleton expose some genes and not others. Since it is widely thought that exposure of genes to genetic modulators influences gene activation, it follows that the cytoskeleton has a role in determining which genes are activated.

How might the cytoskeleton control gene expression? Although the answer is not known, Puck recently offered a possible explanation. At any given time parts of some chromosomes are attached to the inside of the membrane that surrounds the cell nucleus. Such attachment may expose those DNA segments to genetic modulators. The cytoskeleton extends from the nuclear membrane to the cell's outer membrane, where hormonal signals are received. Such a signal, received at the outer membrane, might be transmitted through the cytoskeleton to the nuclear membrane, changing the array of genes attached there and leading to a different pattern of gene expression, according to Puck.

In *Proceedings of the National Academy of Sciences* Puck and his coauthors conclude that the "cytoskeleton-mediated gene exposure reaction appears to be an important genetic regulatory mechanism in mammalian cells." Moreover, because the cAMP reaction seems able to restore cancerous cells to the normal state, the reaction may have particular significance for cancer therapy. *—John Benditt*

MEDICINE

Skin Saver

Researchers seek a vaccine to treat and prevent melanoma

Is it because there is less ozone? More sunworshiping? Whatever the reason, the incidence of melanoma is increasing faster than that of any other cancer. It is estimated that 1 percent of all newborns will contract this cancer of pigment-producing skin cells during their lifetime. The only effective treatment is prompt surgery of the initial tumor; once metastasis begins the disease is almost always fatal. Recently, however, some investigators have come to think melanoma may someday be treated—and even prevented from developing in the first place—with vaccines.

Oncologists have long known that the immune system can recognize and react against cancer cells just as it does against foreign pathogens. Nevertheless, the possibility that vaccines can enhance the body's natural defenses against cancer (as well as against microorganism-induced infections such as smallpox and polio) has only recently won some acceptance, according to Jean-Claude Bystryn of the New York University Medical Center, who has been studying melanoma vaccines for more than 12 years. Workers are seeking vaccines for various cancers, including those of the lung, kidney and colon. Melanoma has spurred the largest effort, Bystryn says, both because it is so difficult to treat and because its interaction with the immune system is relatively well understood.

The vaccine developed by Bystryn and his colleagues at N.Y.U. consists of substances that melanoma cells grown in culture shed from their surface. Over the past four years the workers have tested the vaccine in 55 patients who have melanoma. In the journal *Cancer* they report that the vaccine augmented or initiated an immune response—resulting in the production of antibodies and in increased reactivity of white blood cells—in about half of the patients; the disease progressed more slowly in these patients than it did in those who did not respond to the vaccine.

Malcolm S. Mitchell of the University of Southern California Cancer Center reports that he and his co-workers have achieved similar results. The U.S.C. group vaccinated 42 patients with a preparation of lysed (dead) melanoma cells; tumors were reduced in about half of the patients and in a few cases were even eliminated, according to Mitchell.

Bystryn suggests that "for the youth of the field the results are encouraging," but he stresses that much more research is needed to produce a truly effective vaccine. An important step toward this goal is determining which parts of the melanoma cell can act as antigens against melanoma and which are superfluous or even harmful. The task will be difficult. The cells in any two given melanoma tumors, even in the same patient, usually vary slightly, Bystryn points out. Such variability may explain why his vaccine and Mitchell's evoke a response in some patients but not in others.

If and when vaccines prove to be completely safe and effective in treating melanoma, they may be administered to people who have not contracted the disease but who are considered likely targets. Candidates would include those who have unusual-looking moles called dysplastic nevi and whose families have a history of melanoma, as well as those who have fair skin and are frequently exposed to sunlight. Tests with animals suggest that melanoma vaccines might work well as a purely preventive measure, according to Bystryn. He and other workers inoculated mice with the vaccine and then injected them with enough melanoma cells to kill unvaccinated mice within two months; about nine out of 10 of the vaccinated mice survived. *—J.H.*

Benevolent Bradykinins

New compounds might get to the root of pain

Inquiry into the mechanism of pain has led investigators to a group of artificial peptides that promises a new means of controlling both certain kinds of pain and the symptoms of the common cold. Drugs based on the peptides might avoid some serious problems posed by existing pain-relieving drugs.

Most analgesics are salicylates (such as aspirin) or opiates (such as morphine). Salicylates are not effective against severe pain and can irritate the gut. Opiates have profound effects on many organ systems that often make their prescription medically undesirable—and they are addictive. Acetaminophen and some newer pain relievers, which (like salicylates) interfere with the synthesis of the pain-inducing prostaglandins, also have limited effect on severe pain.

The new peptides are modified forms of bradykinin. Bradykinin, first identified in the 1940's by the slow contraction it causes in smooth muscle, is a chain of nine amino acids. The suggestion of its involvement in pain arose because it is found in elevated quantities in injured tissue and because when injected it is a potent inducer of pain, as well as of inflammation. More recently experiments have shown that it also seems to be largely responsible for precipitating the symptoms of the common cold.

Four years ago John M. Stewart of the University of Colorado School of Medicine found that modified bradykinins in which the amino acid in the seventh place is replaced with a different one can suppress the effects of real bradykinin, including the ability to cause pain. Presumably such an altered bradykinin binds to bradykinin receptors on nerve cells but fails (because of the changed amino acid) to initiate a nerve impulse, thereby jamming the receptor. Yet it remained unclear how central a role bradykinin plays in causing pain.

That question has now been answered by Larry R. Steranka of the Nova Pharmaceutical Corporation in Baltimore, Solomon H. Snyder of the Johns Hopkins University School of Medicine and others. They report in *Proceedings of the National Academy of Sciences* that rats injected with altered bradykinins took longer to withdraw a paw to which increasing pressure was applied; moreover, they were less sensitive to injections of urate crystals, which cause gout.

The workers also showed that bradykinin receptors are found specifically on nerves known to carry pain signals. Snyder says the two sets of findings together constitute "the first definitive evidence that bradykinin is the normal initial stimulus to pain." The prevalence of its receptors in the central nervous system suggests that bradykinin may also act as a neurotransmitter in pain fibers, according to Salvatore J. Enna, senior vice-president for research and development at Nova Pharmaceutical.

Nova, which Snyder cofounded in 1982, has bradykinin blockers (licensed from Stewart) in clinical trials for possible treatment of the common cold. The company also hopes later this year to conduct trials of bradykinin blockers as topical analgesics for burns, which cause the secretion of large amounts of bradykinin. Peptides cannot be taken orally, which limits their potential, but Enna notes that Nova is working on nonpeptide bradykinin blockers. *—T.M.B.*

OVERVIEW

In Vino Scientia

Science is improving California's premium wines

At the J. Lohr winery in San Jose, Barry Gnekow has pioneered the use of ultrafiltration to remove the protein molecules that sometimes cause a haze in white wine; he thinks the approach preserves flavor better than the traditional techniques of "fining," in which the protein is captured with earth or egg white. At Sterling Vineyards in Calistoga, Daniel S. Roberts is experimenting with a portable instrument that combines heat and humidity measurements to derive an index of water loss from the leaves of a vine; the instrument might provide useful information about the plant's physiological state and so inform decisions about irrigation. At those California wineries and at many others, growing sophistication about the chemistry of grapes and wine and the factors that affect it is changing the vintner's art.

Some of the technological fixes common in winemaking during the 1970's, such as the widespread use of centrifuges to speed the settling of solids, have fallen out of favor. The makeup of wine is complex and subtle: the specific compounds that give a wine its flavor are for the most part unknown. Tastes vary too, and what one judge considers good may be thought excessive by another. Nevertheless, "one of the real trends is toward being more analytical," says Janice C. Morrison of the University of California at Davis.

In many instances a better understanding has led winemakers back to traditional approaches, according to

Thomas A. Ferrell, president of Sterling. Many premium vineyards, for example, are turning away from the vigorous, bushy vines that have been dominant in California to smaller plants similar to those found in most European vineyards. Such plants produce fewer grapes but make for better wine. Morrison has shown that shading caused by luxuriant foliage affects grapes not only directly, by preventing sunlight from reaching them, but also indirectly, by shading other leaves. In both cases the result is a vegetal taste in wine—usually considered unattractive—due to the presence of compounds called pyrazines.

Many vineyards are now experimenting with special trellises that encourage the vines to spread out, exposing the fruit to more sunlight. Simi Winery, Inc., near Healdsburg, and Stag's Leap Wine Cellars in the Napa valley are using wires that are raised as the vine grows. Closer spacing of individual plants can also decrease shading. According to Rex Geitner of Stag's Leap, this encourages competition among the roots, which reduces leaf cover. Stag's Leap is cooperating with U.C. Davis in another approach. Vines are usually grafted onto a rootstock from a different species that thrives under local conditions, but some roots encourage less growth than others. The collaboration aims to find good "devigorating" rootstocks.

Another way to discourage vegetative growth is to stress the plant by limiting the availability of water. Mark A. Matthews of Davis says that moderate water stress before the grapes begin to ripen also decreases the amount of malate, an organic acid, in the fruit. Low malate content is often considered advantageous because it limits the extent of a bacterial fermentation that can follow yeast fermentation and produce undesirable flavors.

According to Matthews, some water stress also stimulates the production of phenolic compounds, which are concentrated in the skin of the grape and can give a wine astringency. To that end, several wineries, including the Robert Mondavi Winery in Oakville, are deliberately reducing irrigation or improving drainage. Water stress also boosts the production of anthocyanins, the pigments in black grapes.

At harvesttime, water stress has a cost: overall yield is reduced. Traditionally, decisions about when to harvest were made on the basis of the grapes' sugar content. Tim Mondavi, winemaker at Robert Mondavi, is now also focusing on the amounts of various phenolics, which change as grapes ripen. By determining which of them, if any, are correlated with critical approval of a wine, he hopes to devise a new way to judge when grapes have reached perfection. Simi has experimented with picking grapes at night, when they are cooler. According to Paul Hobbs, assistant winemaker at Simi, the strategy reduces the amount of potassium extracted from the skins during pressing, leading to a slightly more acidic juice that holds a better fruit flavor and does not require supplements of tartaric acid.

At the winery itself, practices have also changed markedly in recent years. Usually grapes are crushed before they are pressed, but William Bonetti of Sonoma Cutrer in the Russian River valley omits that stage, believing that the resulting juice contains fewer of the less pleasant-tasting phenolics. Mondavi likewise avoids overmacerating grape skins. Like most vintners, he uses a tank press, in which a bladder slowly inflated with compressed air gently presses the grapes against the sides of a tank and thereby avoids crushing the seeds.

Traditional wisdom holds that the extracted juice should be protected from oxidizing, which can destroy compounds that may be important for flavor. The usual strategy is to add sulfur dioxide when the grapes are pressed as an antioxidant and an antimicrobial agent. Now many premium wineries are using sulfur dioxide in much smaller quantities, partly because some asthmatics are sensitive to the compound but also to minimize the extraction of bitter-tasting tannins. "The concept is to retain the originality of the vineyard," according to Bonetti.

Unlike some vintners, Bonetti still adds a small amount of sulfur dioxide to the pressed grapes to reduce oxidation, and he is rigorous about keeping oxygen away from the juice and the finished wine. He has invented a machine that bottles the wine by way of a tube that extends to the bottom of the bottle. In what is now a common practice, the bottle is filled with nitrogen before the wine is added.

Others are abandoning the traditional injunction against oxygen. At Domaine Chandon, a maker of sparkling wine in Yountville, Dawnine Sample-Dyer adds oxygen to the juice to oxidize phenolics. The oxidation products, which can discolor sparkling wine, then precipitate out. Mondavi and Hobbs also allow oxidation for some wines: "we can oxidize juice and not lose flavor," Hobbs says.

For the in-the-bottle fermentation that causes sparkling wine to sparkle, Domaine Chandon is experimenting with yeasts trapped in tiny spheres of calcium alginate, a protein compound derived from seaweed. The strategy makes the yeast debris easier to remove after fermentation. In the traditional *méthode champenoise* the debris is encouraged to accumulate in the bottle neck by "riddling": inverting the bottle and rotating it a fraction of an inch at a time over a period of several months. The bottle neck is then frozen and the cap is removed, expelling a plug of ice containing the debris. The alginate spheres, in contrast, fall into the neck immediately when the bottle is inverted.

The public's demand for perfectly clear wines has led most makers to pass their product through filters to remove bacteria. Although Gnekow at J. Lohr is using even finer filters to remove protein as well, others are taking different approaches. In pursuit of a more natural product, Tim Mondavi has begun to leave premium Pinot Noirs unfiltered—at some slight cost in clarity. Linda F. Bisson of Davis hopes to produce a genetically engineered yeast that could be added toward the end of fermentation to destroy haze-producing protein.

Some vintners hope genetic engineering will also help to dispel a cloud on the wine industry's horizon. The Food and Drug Administration has proposed strict limits on wine's content of ethyl carbamate, a suspected human carcinogen. The compound is a breakdown product of the urea produced by yeast, and much effort at Davis' Department of Viticulture and Enology aims at reducing it. Cornelius S. Ough, chairman of the department, has found that some yeast strains make less urea than others, opening the possibility of transferring that trait to a good wine yeast.

Good wine need not contain alcohol, according to Gnekow. In addition to ultrafiltration he has pioneered reverse osmosis of wine, a process in which alcohol and water diffuse out of wine through a membrane so fine that it will not pass any of the larger organic molecules that give wine its flavor. By adding distilled water to the resulting concentrate, he produced an alcohol-free wine that, entered incognito, won a gold medal at the Los Angeles County Fair and a bronze medal in a competition in Dallas. Gnekow asserts that in 30 years alcohol will be optional in wines. Time will tell whether Gnekow's marketing acumen is a match for his technological skill. *—Tim Beardsley*

SCIENTIFIC AMERICAN July 1988 Volume 259 Number 1

The Geography of U.S. Presidential Elections

Enduring geographic cleavages divide the electorate. They weigh heavily in the Electoral College system and demand that a winning candidate build a geographic coalition

by J. Clark Archer, Fred M. Shelley, Peter J. Taylor and Ellen R. White

The presidential election of 1988 falls on an unusual and largely forgotten centennial. The election of 1888 marked the last time a candidate who held a plurality of popular votes failed to win the presidency. The hapless candidate was the incumbent Democratic president, Grover Cleveland, who polled 5,540,365 votes to the 5,445,269 votes of his Republican challenger, Benjamin Harrison. Cleveland's popular plurality, which gave him a margin of nearly 100,000 votes, counted for nothing in the Electoral College, where Harrison was elected president by 233 votes to 168.

Quite simply, Cleveland's votes were in the wrong places. He was the victim of the peculiar mathematics of the Electoral College system, designed a century earlier to span deep regional divisions in the young nation by forcing presidential candidates to seek geographically broad constituencies. Yet conditions like those to which the Electoral College was a response persisted in Cleveland's time, and they may still be present today. Our statistical analyses of historical and recent presidential elections reveal enduring geographic patterns of partisan support and volatility—patterns that are certain to reassert themselves in this election regardless of its specific candidates and issues.

The Electoral College, in which each state is given a number of votes equal to its combined representation in the two houses of Congress, is one of the solutions devised by the framers of the Constitution to their formidable task of welding disparate and often disputatious sections into a single nation. The 13 colonies had long been divided along economic lines into three regions. New England's economy was based on trade, fisheries and shipbuilding; the South specialized in plantation production of cotton, tobacco and other crops, and the Middle Atlantic colonies combined agriculture with manufacturing and commerce. These three regions of colonial America acted as distinctive "culture hearths": they differed not only in their economy but also in ethnicity, religion, dialect, settlement patterns and land use.

Political conflict among the sections had developed long before Independence, and within each section divisions had opened between seaboard commercial interests and settlers on the frontier, who often resented the growing economic power of the coastal cities. To produce an acceptable framework for nationhood, the Founding Fathers tried to devise a system of government that would prevent any one region from dominating presidential elections. By guaranteeing every state a distinct voice in the election process, the Electoral College was one response to that imperative.

During the early 19th century an additional feature of the Electoral College developed informally: the winner-take-all tradition, in which the candidate getting a plurality of the popular vote in any state takes all that state's electoral votes. As a result overwhelming popular majorities in a few states cannot compensate for poor showings in many other states. A candidate winning slight popular pluralities in two states, for example, would garner more electoral votes than a candi-

J. CLARK ARCHER is assistant professor of geography at the University of Nebraska—Lincoln. He holds a 1974 Ph.D. from the University of Iowa. FRED M. SHELLEY is assistant professor of geography at the University of Southern California; his Ph.D., awarded in 1981, is also from the University of Iowa. PETER J. TAYLOR is reader in political geography at the University of Newcastle. He received his doctorate in 1970 from the University of Liverpool. He is editor of the *Political Geography Quarterly* and series editor of *Geography of the World-Economy,* published by the Cambridge University Press. ELLEN R. WHITE is director of the Center for Cartographic Research and Spatial Analysis at Michigan State University. She has an M.S. in geography from Virginia Polytechnic Institute and State University; her maps and graphics have appeared in a number of publications in addition to the present article.

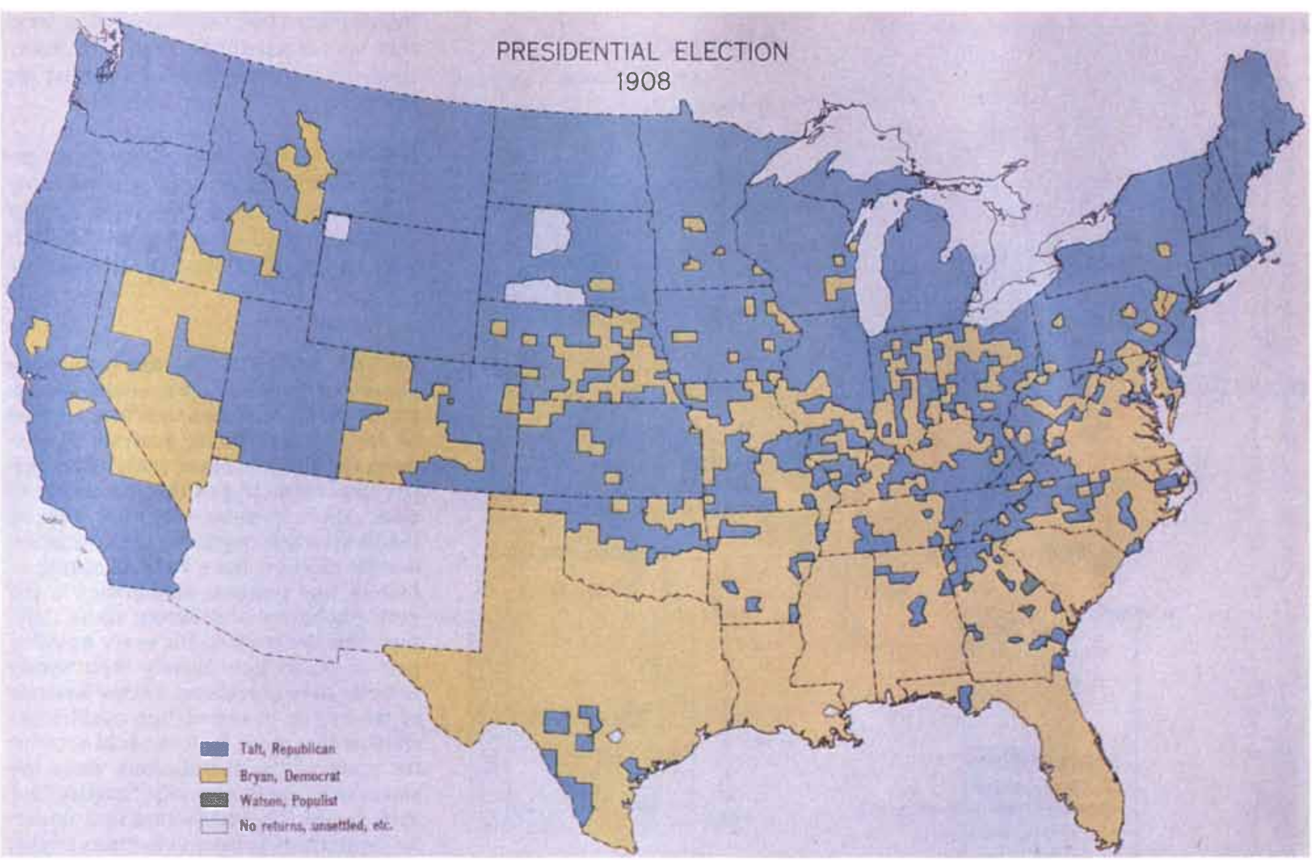
date who won an overwhelming popular majority in one state but fell just short of a plurality in the other—even though the second candidate's popular votes would outnumber those of the first. This, in essence, was Cleveland's downfall in 1888: he lost New York's 36 electoral votes by only a small margin in the popular vote while gaining the electoral votes of several states, mainly Southern ones, by popular majorities of more than two-thirds. A geographic coalition rather than a simple popular majority wins presidential elections.

Thus the Electoral College system encourages candidates to woo voters state by state. It amplifies the influence of sectional divisions in the strategy of presidential campaigns and, by giving prominence to regional issues, perhaps even helps to perpetuate such divisions. How have sectional cleavages been manifested in past presidential elections? In what form and to what degree do they persist this year, and what should they mean to each party's strategists as they try to draw a winning map of the U.S.? These are the questions that prompted our statistical research on the geography of presidential elections.

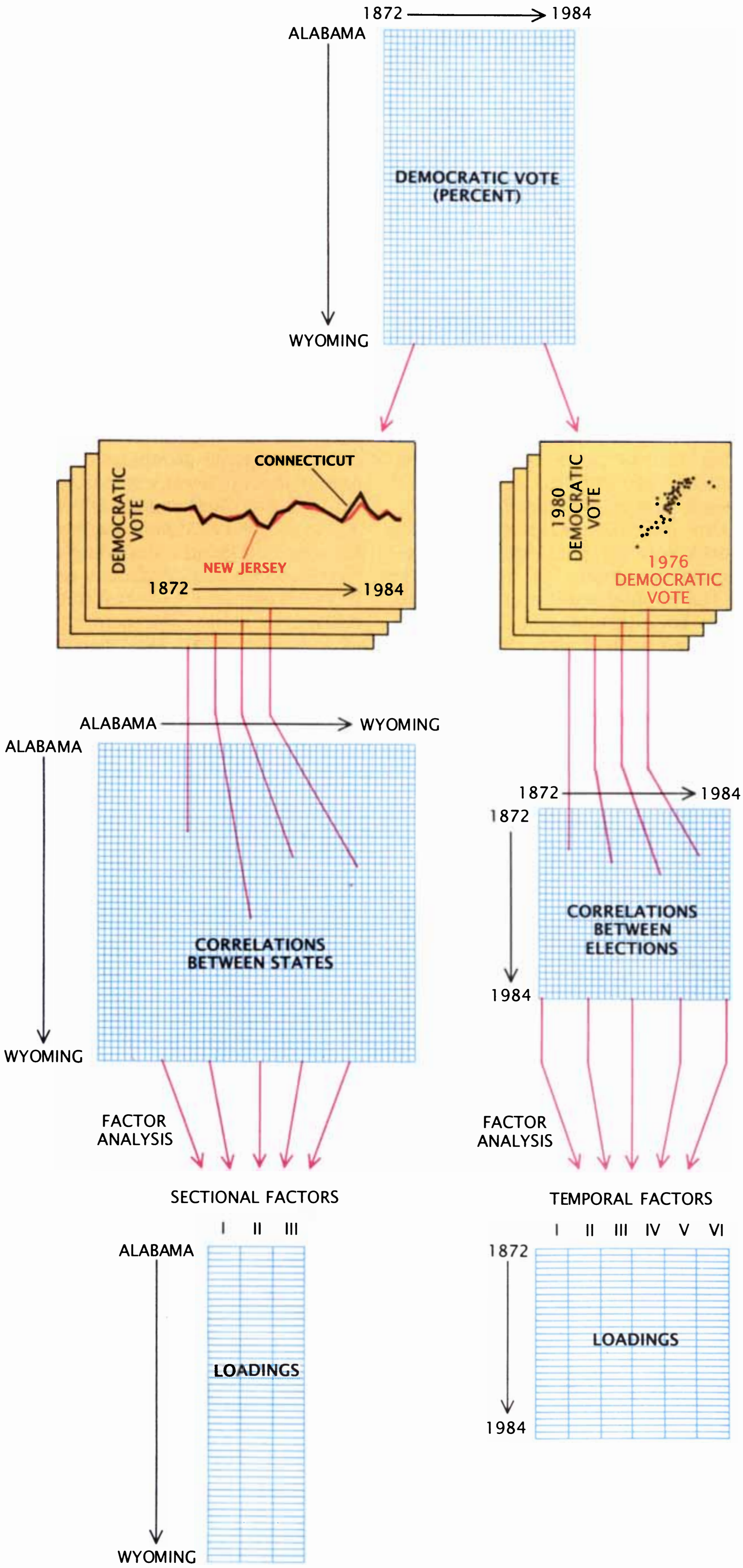
Our first task was to identify the dominant sectional divisions in presidential elections since the Civil War. We did so by employing factor-analytic techniques to detect states that have had similar voting histories. The statistical procedures called factor analysis were originally developed by psychologists trying to discern distinct components of intelligence underlying scores on large batteries of intelligence-test tasks. Factor analysis is also valuable in other fields, including archaeology, climatology, economics and geography. It serves to reveal the number of underlying "factors" needed to describe statistically the patterns of correlation seen in a large data matrix.

Our own data consisted of each state's vote for the Democratic presidential candidate as a percentage of the popular vote in every election from 1872 through 1984. (For states admitted to the Union after 1872 the data set began later, and we excluded Alaska and Hawaii altogether.) For each possible pair of states we determined a correlation coefficient, indicating how strongly their Democratic voting percentages had tended to be correlated over the series of elections. We then subjected the matrix of correlation coefficients to factor analysis.

We discovered that most of the correlations could be statistically "explained" by invoking just three underlying factors. Each state received a specific "loading" for each factor, representing the degree to which the factor was reflected in the state's voting behavior. The factors, we found, corresponded to geographic divisions. When states were distinguished on a map according to the factor for which they had the highest loading, three distinct regional groupings were apparent: the Northeast, extending as far south as Maryland and the Ohio River and as far west as Minnesota, Iowa and Missouri; the South, extending as far west as Texas and Oklahoma, and the West, encompassing the rest of the 48 contiguous states. The factors testify to a persistent electoral geography, prompted by sectional variations in culture and economy.



ELECTORAL GEOGRAPHY emerges from a map of counties won by each candidate for president in 1908. Republican William Howard Taft defeated Democrat William Jennings Bryan and several third-party candidates. In keeping with a long-standing geographic pattern, the Democrats had solid support in the South and the Republicans in the Northeast; the outcome was mixed in the West. The map is from the 1932 *Atlas of the Historical Geography of the United States,* by Charles O. Paullin.

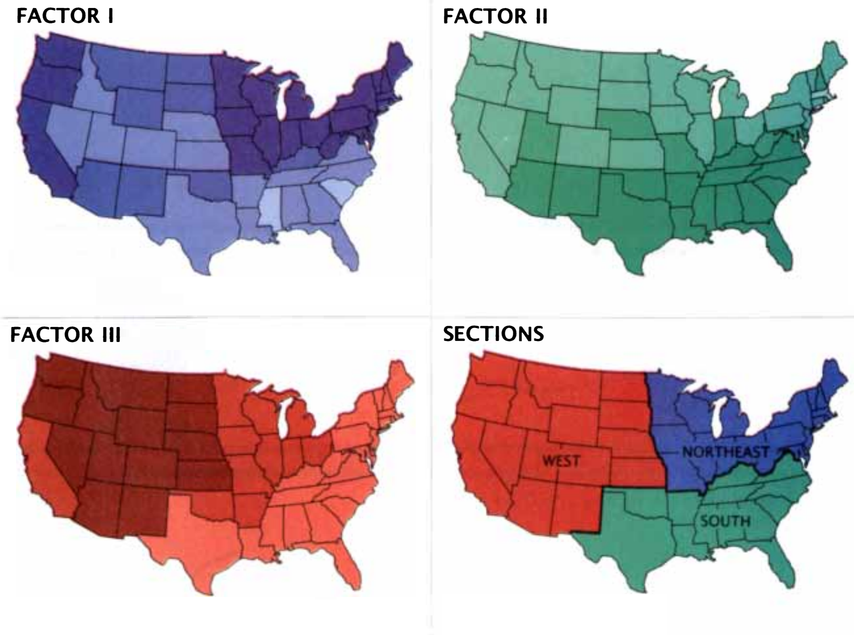


What roles have these three sections played in elections over the past century? To answer this question we sought temporal rather than spatial correlations: correlations not between specific states, calculated over the entire series of elections, but between specific elections, based on the Democratic vote in all the states. The correlation coefficient for each pair of elections expressed the degree to which the pattern of Democratic support across all 48 states in one election resembled the pattern in the other. We constructed a matrix that included correlation coefficients for every possible pair of elections from 1872 through 1984 and subjected it to factor analysis. The analysis pointed to three major factors and several minor ones underlying the correlations.

When we calculated the loadings of every election for each of the factors, we found that successive elections often had high loadings for the same factor. The factors identified electoral epochs: periods when, for several elections in a row, individual states tended to vote the same way with respect to the national vote. During each electoral epoch a section's Democratic or Republican vote fluctuated at a level that was consistently higher or lower than the national level of support for the party.

The first epoch was the most enduring, beginning after Reconstruction and continuing almost unbroken until after World War II. It was an era of consistent Democratic

FACTOR ANALYSIS reveals how many statistical "factors" are necessary to describe the correlations seen in a large set of data. In their factor analysis of elections the authors began with a data matrix (*top*) consisting of the percentage of each state's popular vote that went to the Democratic candidate in every presidential election since 1872. In search of factors that underlie similarities in the voting behavior of different states (*left*), they first determined for every possible pair of states how closely their voting records were correlated. Factor analysis of the matrix of correlation coefficients showed that three factors could account for most of the correlations when the states were given a specific "loading" for each factor. To find factors that underlie similarities between elections (*right*), the authors determined correlation coefficients for paired elections from every state's Democratic vote. Factor analysis then revealed three major temporal factors and three minor ones at work.



GEOGRAPHIC DIVISIONS are evident in maps on which states are colored according to their loading for each of the three factors revealed by factor analysis of states' voting records (*see illustration on opposite page*). A darker color shows a higher loading. Each section, defined by a single factor, groups states having similar voting behavior.

strength in the South, more than offset by Republican dominance of the industrializing Northeast. Hence the West and the states in the former Northwest Territories—Ohio, Indiana and Illinois—emerged as crucial electoral battlegrounds. The fact that eight of the nine Republican presidents elected between 1860 and 1920 came from one of those three states attests to the region's pivotal role.

We call this most consistent geographic pattern in American electoral history the Sectional Normal Vote. It was interrupted only in 1896, when William Jennings Bryan's candidacy allied Southern Republicans and Western Democrats with the Populists, a third party based in the West, and challenged the Republican-dominated Northeast. A new sectional cleavage—dividing the Northeast from the combined South and West—briefly emerged. In the end, however, the Republican candidate, William McKinley, was elected with the help of victories in the old Northwest, and by the election of 1900, when McKinley won a second term, the West and the South were again playing their traditional electoral roles. Although the epoch of the Sectional Normal Vote saw long series of victories by one party or the other (Franklin Delano Roosevelt's string of Democratic wins in the 1930's and 1940's, for example), the parties' relative strengths among the sections remained quite consistent.

The Sectional Normal Vote came to an abrupt end in 1948, as the adoption of a more liberal position on civil rights by the national Democratic party fractured the hitherto solid Democratic support in the South. The so-called Dixiecrats revolted and supported an alternative candidate, Strom Thurmond, instead of the national candidate, Harry Truman. Democratic support in the South would never again be as unified as it had been during the Sectional Normal Vote.

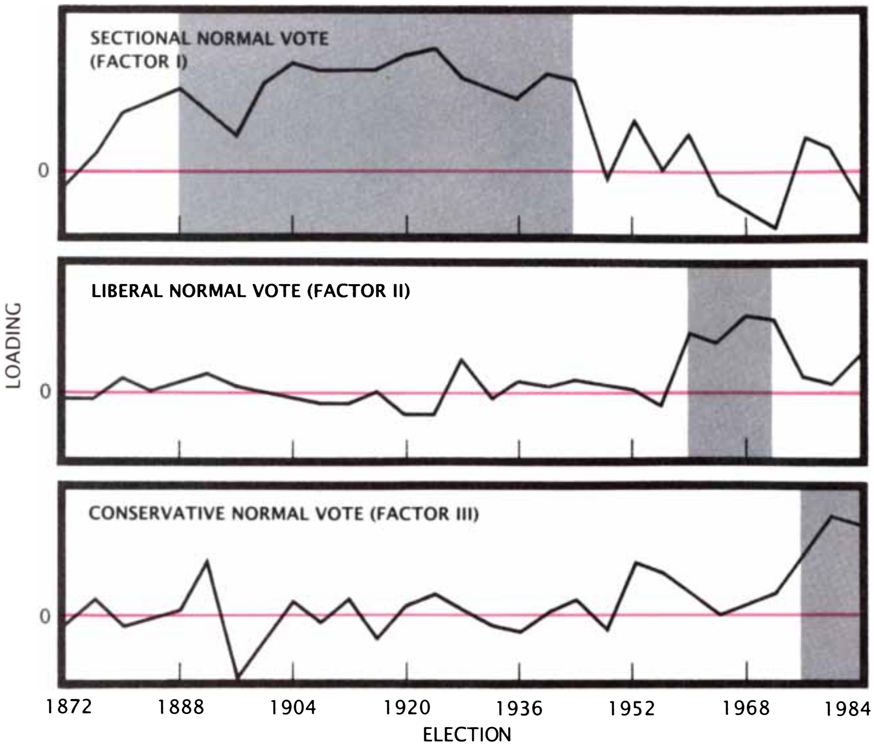
Two contrasting epochs of relative sectional support for Democratic and Republican candidates emerged from the series of elections after 1948. In the first, encompassing the elections of 1964, 1968 and 1972, the Democratic candidates, including Lyndon Johnson, Hubert Humphrey and George McGovern, tended to promote an expansion of Federal involvement in the affairs of the nation. Hence we call the distinctive electoral geography of that epoch the Liberal Normal Vote. In the 1976, 1980 and 1984 elections the rallying point became a call for a limited or even a reduced role for the Federal Government—a call voiced by Gerald Ford and Ronald Reagan on the Republican side and Jimmy Carter among the Democrats. We refer to the geographic pattern of those elections as the Conservative Normal Vote.

To explore in detail the contrast between the electoral geography of the Liberal Normal and Conservative Normal votes, we focused on the more than 3,000 counties and parishes in the U.S. We made maps of the country showing, county by county, the average percentage of the popular vote that went to Democratic candidates during each electoral epoch. The map of the Liberal Normal Vote [*see top illustration on page 49*] reveals a major contrast with the Sectional Normal Vote. Throughout the Deep South, Democratic candidates averaged less than one-third of the vote in many counties—counties that two decades earlier had cast large majorities for Roosevelt. Other areas of low Democratic support during the period of the Liberal Normal Vote included the central Great Plains and the Rocky Mountain region.

At the same time the Democrats consolidated a new stronghold in the Northeast, which the Republicans had dominated before World War II. The county-scale map also shows smaller pockets of high average Democratic support in the upper Middle West and along the Pacific coast.

The Conservative Normal Vote succeeded the Liberal Normal pattern in the election of 1976. Whereas Southern Democrats had often deserted their traditional party ties during the Liberal Normal Vote, many of them supported Jimmy Carter in the elections of 1976 and 1980, with the result that the map shows renewed Democratic strength in the South [*see bottom illustration on page 49*]. Yet the Democratic resurgence there was matched by a return of Republican strength in some (if by no means all) of the Northeast. Democratic support persisted in the upper Middle West, in central cities and coal-mining areas and in the "rust belt" counties that are dependent on declining industries such as steelmaking and automobile manufacturing.

The broadest expanses of low to very low Democratic support during the Conservative Normal Vote are found in the West—in the Great Plains, the Rocky Mountain region and California. With the exception of counties containing major universities or large Native American or Hispanic populations, the solid West has assumed an electoral role for the Republican party similar to the role the South played for the Democrats before 1948. Carter, the only Democratic nominee to cap-



ELECTORAL EPOCHS—periods when the voting behavior of each geographic section stayed fairly consistent—are evident in graphs showing each election's loading for the three major temporal factors that emerge from factor analysis (*see illustration on page 46*). A series of elections that all have a high loading for a single factor defines an electoral epoch. In successive epochs the sectional groupings stay mostly unchanged, but the voting behavior of each section as a whole may shift dramatically.

ture the White House since 1964, was borne into office by a coalition of states in the South and the Northeast; west of the Mississippi he won only Texas (which we classify as Southern) and Hawaii.

The county-level maps chart complex patterns. Islands of Democratic support appear in heavily Republican regions; Republicans may dominate suburban or rural counties of an otherwise Democratic state. How important do sectional cleavages remain in the face of smaller-scale variations in voting preference, which reflect patterns of urbanization and other local cultural, social or economic variations? After all, broad regional distinctions in American culture are often said to be blurring as a result of increased mobility and improved communications, the growth of the national media and the emergence of a postindustrial economy.

To find out whether this assumed homogenization has lessened the importance of sectional cleavages, we compared the influence of those divisions on county-scale voting patterns with the influence of urbanization: residence in the central city, a suburb or a rural area. We compiled a data set consisting of the Democratic percentage of the popular vote in each county for every presidential election from 1940 through 1984. For the purpose of our analysis, we distinguished the counties in two ways. We classified them as central-city, suburban or rural, based on U.S. Bureau of the Census information from 1954 (for the elections from 1940 through 1964) and 1982 (for the elections from 1968 through 1984). We also distinguished them by their location, in keeping with our delineation of Northeast, South and West.

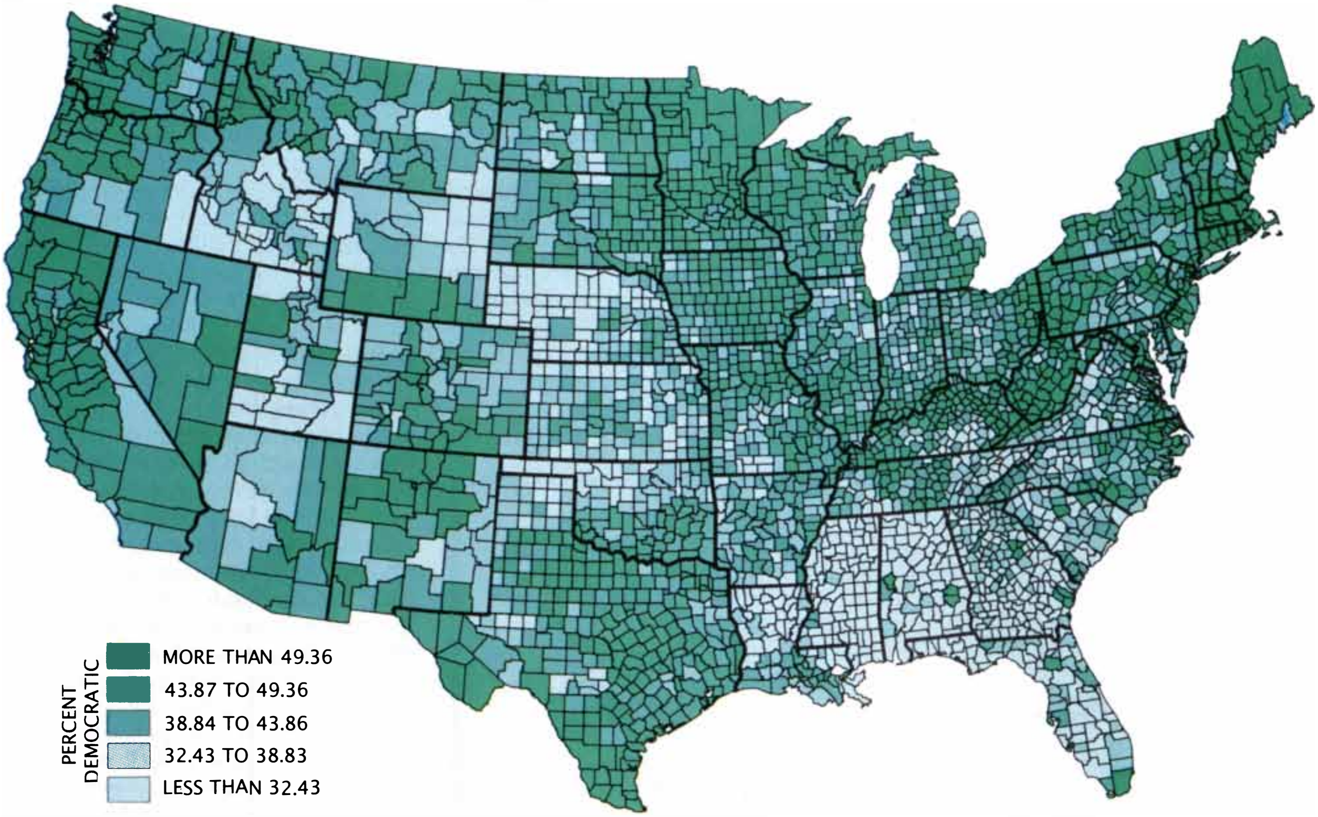
The comparative importance of urbanization and sectionalism was examined by multiple-regression analysis, a technique that can serve for assessing the statistical dependence of one variable on two or more "independent" variables. Here the county-level Democratic vote was the "dependent" variable; each county's status as predominantly urban, suburban or rural on the one hand and as Northeastern, Southern or Western on the other served as the independent variables.

We found that variations in urbanization account for surprisingly little of the nationwide pattern of voting behavior. The proportion of the county-level variation in the Democratic vote that could be explained statistically by distinguishing urban, suburban and rural counties ranged from .03 percent for the 1952 and 1980 elections to 2.4 percent for the 1968 election; for all 12 elections in the data base the proportion averaged only .8 percent.

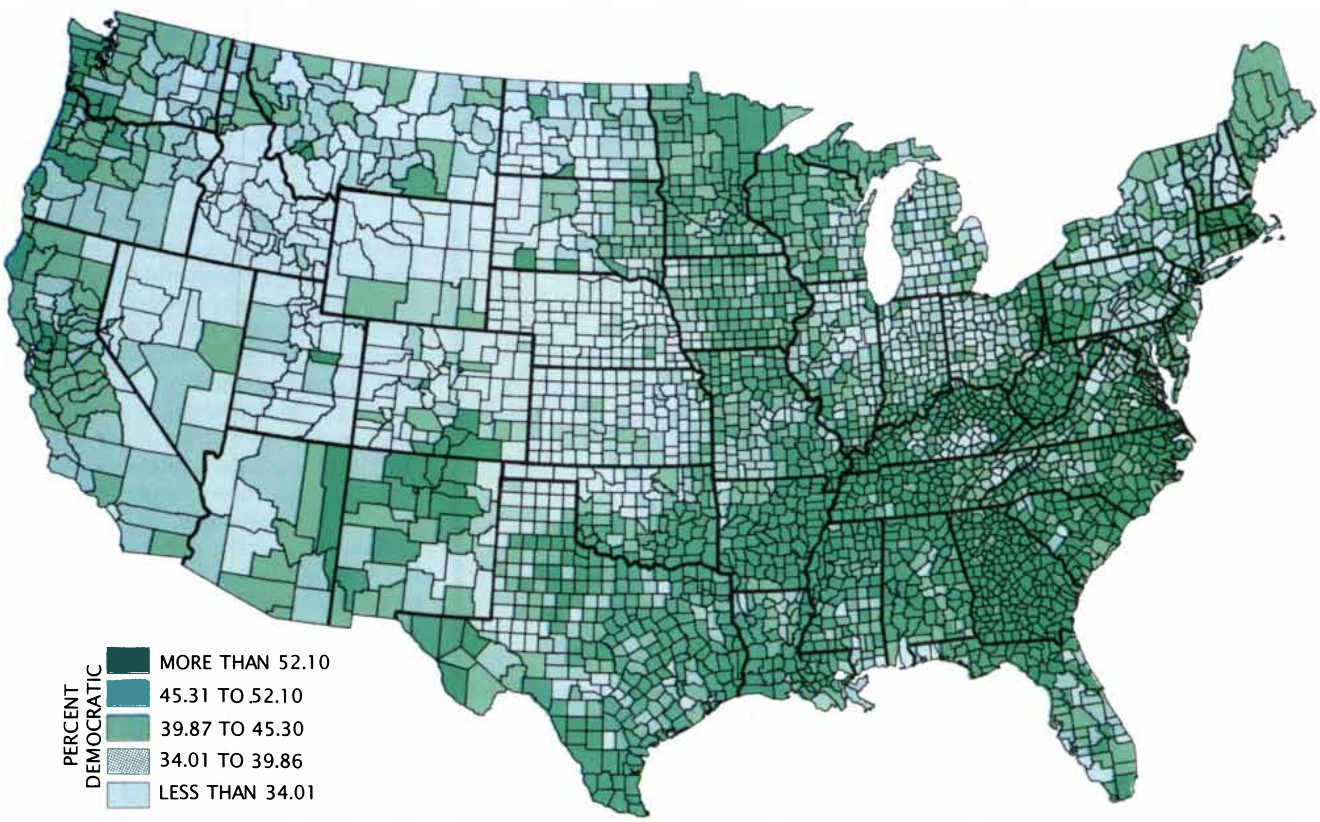
In contrast, distinguishing counties as Northeastern, Southern or Western accounted for an average of 24 percent of the county-level variation. The percentage fell as low as 2.7 percent for the election of 1948, when the Dixiecrat opposition to civil rights fragmented the Democratic party in the South, but it rose above 35 percent for the elections of 1940, 1944, 1952 and 1980. By subdividing the sections—splitting the Northeast into New England, the Middle Atlantic and the Middle West; the South into the Core South, the Rim South and the Southwest, and the West into the Great Plains, the Rocky Mountains and the Pacific coast—we amplified the explanatory power of sectionalism. The proportion of the voting pattern that could be explained by reference to sectionalism then rose to an average of 36 percent, and it reached fully two-thirds for the election of 1940.

These statistical results would probably gratify the framers of the Constitution. Our results indicate that the Electoral College system, devised in part as a response to geographic cleavages existing in the electorate at the time of Independence, still fills a crucial need: the sectional divisions the Founding Fathers confronted persist today, albeit much transformed. In shaping political behavior, regional culture—a common history, dialect and religious denomination, among other things—still outweighs other fundamental divisions prevailing in American society.

What does this electoral geography mean for a presidential candidate trying to build a winning geographic coalition? Although the sections, grouping states that tend to vote together, are quite stable, the nature of the sectional vote is not. It changes most dramatically in "realigning" elections, which mark the transition from one electoral epoch to the next, but certain areas of the country are highly volatile even within an epoch. Although areas of predictable partisan support often serve as foun-



LIBERAL NORMAL VOTE, the geographic pattern that prevailed during the elections of 1964, 1968 and 1972, is mapped at the level of counties. Each county's average level of support for Democratic candidates is indicated by the intensity of color. A clear split is evident between the heavily Democratic Northeast and the South, which tended to support Republican candidates. Democratic strength was also low in the West, in the Great Plains and the Rocky Mountain region in particular.



CONSERVATIVE NORMAL VOTE, mapped county by county, has a geographic pattern that contrasts with the pattern of the Liberal Normal Vote, which it succeeded in 1976. Democratic strength returned in the South while ebbing in much of the Northeast. The West became the most dependably Republican part of the country, with a few exceptions—notably the heavily Hispanic counties of northern New Mexico and southern Colorado and the Native American areas of northeastern Arizona.

dations in a party's campaign strategy, the party is well advised to compete most intensely in the volatile areas.

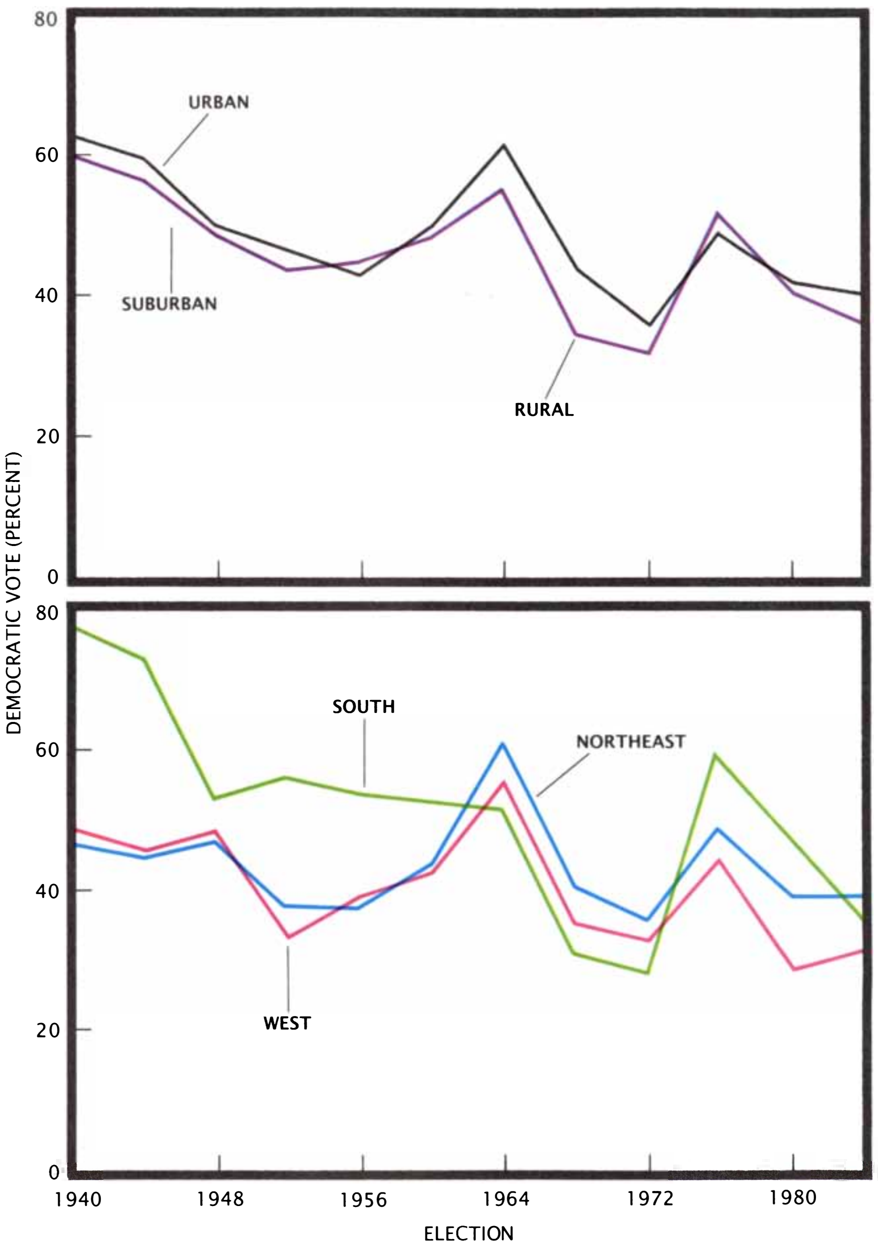
Which areas of the country have been the most variable in partisan support during recent elections, and which have been the most consistent? We continued our county-level analysis of political geography by devising an index of electoral volatility. The index takes national voting trends over the presidential elections from 1940 through 1984 as a point of comparison. Counties whose Democratic voting percentage maintained a similar relation to the national percentage in successive elections got low scores on the index, whereas counties whose Democratic proportion showed a changing relation to the national vote got higher scores. The scores do not reflect absolute levels of partisan support; a county that maintained a Republican majority through an election won by Democrats would not be considered volatile if, proportionally, its level of Democratic support had followed the national trend.

Mapping the volatility index county by county reveals that volatility, like partisan support, is a function of geography [*see illustration on opposite page*]. New England, upstate New York and the upper Middle West emerge as highly volatile regions. The Deep South also stands out, with the highest levels of volatility in the country appearing in Alabama, Georgia and Mississippi. These loci of volatility in part reflect recent electoral history: the reversal of partisan loyalty in the Northeast and South in 1976, as the Liberal Normal Vote gave way to the Conservative Normal Vote. More broadly, the volatility map distinguishes regions whose voters are often discontented with the candidates or positions of the national parties. Such areas may offer third-party candidates unusually strong support, such as George Wallace had in the South in 1968 and John Anderson had in New England in 1980.

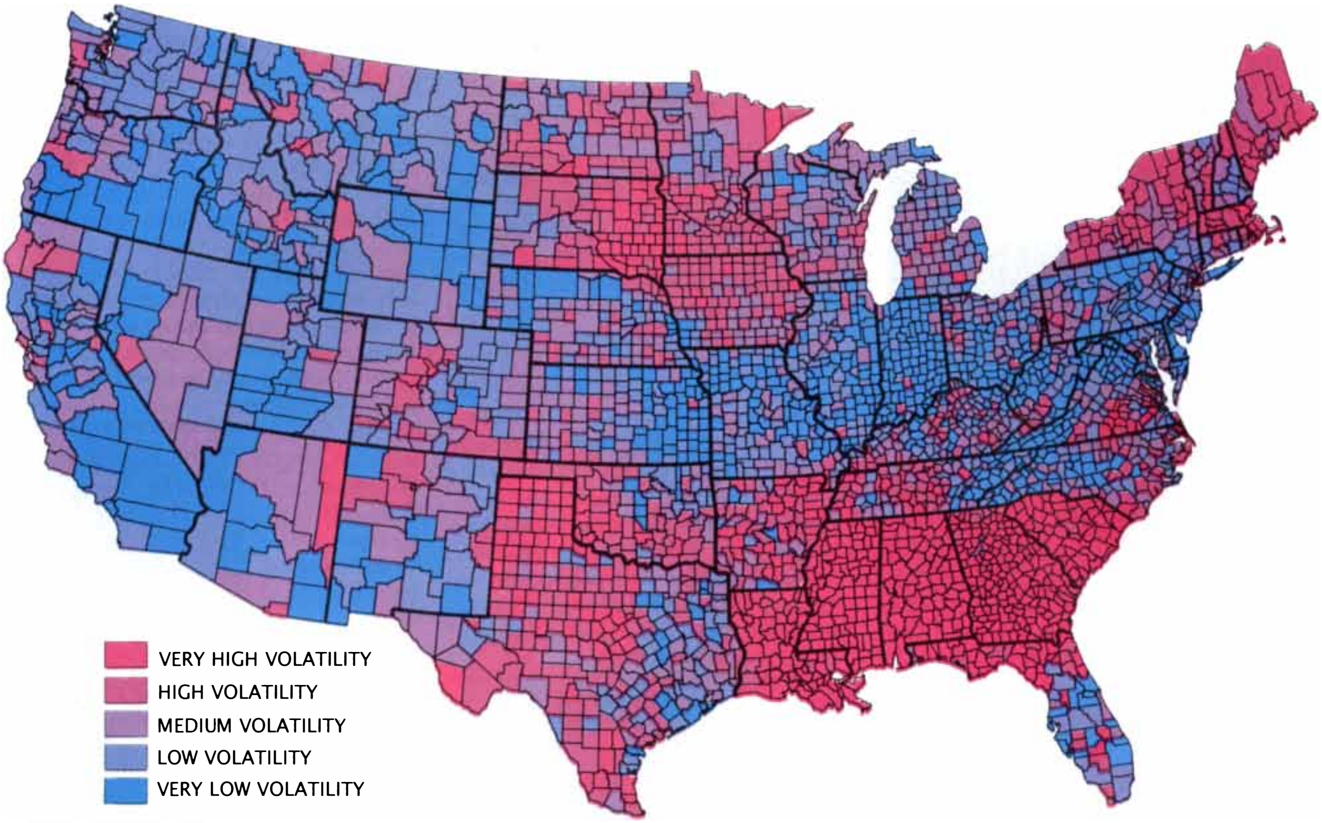
By the same token, lack of volatility also has a distinctive geography. Scores on the volatility index are lowest in a broad belt of counties extending from Pennsylvania and the Middle Atlantic states west across Ohio, Indiana and Illinois to Missouri, then broadening north and south into Nebraska and Kansas. Along the Rocky Mountain front the belt widens still further to embrace much of the interior West and the Pacific coast.

This geography—three bands, two of high volatility and one of low volatility, spanning the eastern half of the U.S.—illustrates a link between the sectional divisions that confronted the Founding Fathers and the electoral geography of today. Settlers migrating westward during the 18th and 19th centuries generally followed routes that paralleled lines of latitude. The bands of similar volatility evident today are one relic of the common culture and political attitudes carried west from the three culture hearths of the Atlantic seaboard.

This, then, is the geography that is likely to guide presidential-campaign strategy in the foreseeable future. Because of its predictable (Republican) orientation, the interior West is likely to receive less attention than the volatile areas of the Northeast and South. In the present election an awareness of regional distinctions was one motivation for the coordinated presidential primary held in 14 Southern states on "Super Tuesday," March 8. The coordinated primary, largely devised by the Democratic party, was intended



SECTIONALISM OUTWEIGHS SOCIOECONOMICS in explaining county-level voting patterns over the past 12 elections. Little difference is seen in the average Democratic vote of counties distinguished as urban, suburban or rural, a distinction that parallels many socioeconomic cleavages (*top*). Counties grouped by geographic section, however, often differ dramatically in their average Democratic vote (*bottom*).



ELECTORAL VOLATILITY, like party preference, has a persistent geography. Volatility was determined for each county from the frequency with which its Democratic vote in successive elections varied in a sense opposite to the national trend. For example, if a county's Democratic vote declined from one election to the next while national support for the party rose, the county might qualify as volatile (whatever its absolute Democratic vote). The highly volatile areas evident on the map—the Northeast, the South and the upper Middle West—are critical campaign battlegrounds for candidates of both parties.

to give the region a greater voice in the choice of a nominee and thereby reduce the chance of defections by Southern Democratic voters in the general election. Once the nominees are chosen, both parties' candidates can be expected to campaign vigorously in states such as Florida, New York and Texas, where a peak in volatility coincides with a large prize of electoral votes. It is a measure of those states' influence that each of them has cast electoral votes for the winner in at least 13 of the last 15 elections.

As the early returns in this year's presidential election are reported on the evening of November 8, what geographic clues might presage the outcome? Assuming the Conservative Normal Vote persists, and in particular its feature of solid Republican support in the West, the Democratic party probably can win the election only by a strong showing in both the Northeast and the South, such as it had in 1976. Republican wins or close contests in the eastern third of the country—and in particular in such volatile large states as Florida and New York—will probably signal an overall Republican victory.

Indeed, the Democratic party is at a disadvantage from the point of view of electoral geography. Until 1980 the Northeastern electoral section, as we have defined it, had a bare but nonetheless absolute majority in the Electoral College, holding 270 votes to the South's 147 and the West's 121. Reapportionment after the 1980 census reduced the Northeast's electoral vote to 254 and increased the vote of the South and West respectively to 155 and 129. As a result the 1984 election was the first one in history that the electoral vote of the Northeast alone could not at least in principle have decided. Both parties now need support in at least two electoral sections to win the presidency. Republican candidates, however, enjoy the advantage of having to target only one of the two more volatile sections in addition to their apparently certain base in the West; Democrats must win them both.

From a geographic perspective the most striking outcome for this year's election (a decidedly unlikely one, to be sure) would be a strong Democratic showing in the western half of the country. A resurgence of Democratic strength in the West might signal the eruption of a distinctive but transient sectional structure, such as appeared in 1896. It might also mark the start of a new electoral epoch, in which the lasting geographic divisions in the U.S. electorate would assert themselves in a new way.

FURTHER READING

GEOGRAPHY OF ELECTIONS. Peter J. Taylor and Ronald J. Johnston. Penguin Books, 1979.

SECTION AND PARTY: A POLITICAL GEOGRAPHY OF AMERICAN PRESIDENTIAL ELECTIONS FROM ANDREW JACKSON TO RONALD REAGAN. John C. Archer and Peter J. Taylor. John Wiley & Sons, Inc., 1981.

AMERICAN FEDERALISM: A VIEW FROM THE STATES. Daniel J. Elazar. Harper and Row, 1984.

AMERICAN ELECTORAL MOSAICS. J. Clark Archer and Fred M. Shelley. Association of American Geographers, 1986.

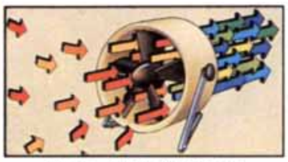
THE SHAPING OF AMERICA: A GEOPOLITICAL PERSPECTIVE ON 500 YEARS OF HISTORY. Donald W. Meinig. Yale University Press, 1986.

THE LIFESTYLE RESOURCE™

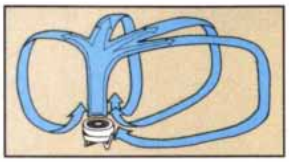
At The Lifestyle Resource we give you all the facts and details necessary to make an informed purchase. Your satisfaction is our primary concern. If your purchase doesn't meet your expectations, return it in original condition within 30 days for prompt refund.

AERODYNAMIC COOLING

The Turbo-Aire fan moves more *air and provides more cooling than ordinary fans* — you get more than twice the air delivered by oscillating room fans. Patented and computer-designed to maximize efficiency and minimize noise and vibration, this fan creates an exceptionally strong, smooth jetstream column that sets a room stirring with secondary air currents. Refreshes better than the intermittent blast of air from an oscillating fan. Aerodynamic housing increases blade-tip efficiency over conventional fans. Adjusts to any angle. Floor, table or wall mount. Set in the hassock position, the five bladed, 12″ fan redistributes air in an entire room, makes it as useful in winter as in summer. Can reduce air-conditioning costs. 300% more efficient than ordinary fans with the same type and size motor. 3 speed fan control, the most advanced 12″portable house fan. Weighs less than a case of cold soft drinks — and more easily moved! 2-yr. limited warranty. UL listed. **$69.95** #2160.



Turbo-Aire's jet stream



Hassock position—venturi effect

Conventional fan

FINGERTIP PHONE DIALER

The new 8K Sharp Dial Master stores and dials up to 400 separate phone numbers. To dial just hold it up to a telephone mouth piece and press the dial button. Fast-search ability by name or company. Full two-line, 24-character, LCD display. Two speed dialing adapts to any local phone system. One-touch redial. Credit card number and long distance computer access keys. Secret key requires password for access to protected information. Two separate memo files. 10-digit calculator with memory. Each listing holds a combination of up to 80 characters and numerals. Easy-touch raised keys. Backup battery. 1-yr. warranty. Even with everything on it the 8K Sharp Dial Master is wallet-sized — in dimension and price. **$89.95** #2150.

ZONE OF CALM

Noise produces stress in the human organism. Today the volume of civilization nearly everywhere seriously interferes with our abilities to relax, read, sleep, concentrate or otherwise function at optimum efficiency. But you needn't be victimized by noise pollution. The *new* Marsona 1200A Sound Conditioner electrically synthesizes a variety of pleasing natural sounds that help mask out and reduce the annoyance of unwanted noise. You can simulate sounds of ocean surf, summer rain, mountain waterfalls. You control not only the volume but also wave pattern, wave or rain rhythm, the seeming nearness or distance of the source. The new Marsona is the finest instrument of its kind, through a 5 inch speaker it creates sounds as close to nature as you will find. UL listed. **$149.95** #2200.

BRING MOUNTAIN TOP FRESHNESS INDOORS

The new Bionaire 700, no bigger than a table-model radio, will clean and recharge your stale indoor air to virtually mountain top freshness. Get relief from breathing allergy causing dust, pollen, tobacco smoke, animal hair and dander, cooking odors, soot, and mold spores. The Bionaire 700 will clean and rejuvenate the air in a 12x12x8 ft. room 4 times an hour, while the filtering system removes up to 99% of all particulate pollutants as small as .01 microns in the process. The filtering process begins with an activated carbon pre-filter that helps remove odors and large particles; next, with the patented electret main filter, the Bionaire removes particles as small as 1/10,000th the diameter of a human hair. Finally, Bionaire's unique negative ion generator—which not only precipitates any remaining particles, but also generates millions of negative ions—reproduces the stimulating effect of fresh mountain air. Two year limited warranty, UL listed, weighs 5.2 lbs., **$149.95** #2070. One Set of two replacement filters **$19.95** #2071.

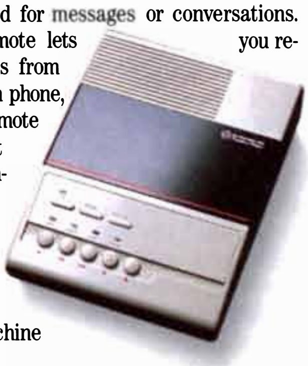
WHICH WATCH?

You choose. You get this year's most sought-after chrono style and a totally professional competition sports watch, powered by a precision Citizen quartz movement. Dive with it: it can go an incredible 660 feet deep. Monitor elapsed time for running, swimming, or cycling: the deep-knurled, uni-directional bezel can't accidentally reverse, and you can set it with gloves on. All metal case with scratch resistant mineral glass crystal seals out dust and water. Shock resistant 3 hand movement with date read out window. Luminous hands let the action go on even in the dark. You've seen watches just like this for $145. Now, enjoy the same exceptional quality and look for a fraction of the cost. Adjustable PVC band can be trimmed for perfect fit. Yellow with black, #1941; Black with red, #1951; Black with blue, #1961. **$69.95.**

THE LIFESTYLE RESOURCE™

CUTTING THE CORD

A leading consumer magazine writer likens a person's first conversation walking around talking on a cordless telephone to the exhilaration of that first bike ride minus training wheels. Later, that article rates Southwestern Bell's FF-1700 Cordless Phone tops for range in controlled tests among 21 brands and models. By handling incoming and outgoing calls to range of 1000′ (the article rated a maximum of 1500 ft.), with outstanding speech quality and convenience features, the FF-1700 ended on top of the consumer magazine ratings reports. Base unit serves as freestanding speakerphone with dialpad, so you get two phones in one. Plus intercom, paging and 10-channel selection. Digital security code protects line from outside access. To take your freedom a step further, Southwestern Bell's FA-450 Telephone Answering Machine gives you the latest technology and newest features at a most attractive price. Single cassette operation, call screening, household memo function, voice-activated record, one-touch playback. Two-way record for messages or conversations. Beeperless remote lets you retrieve messages from any pushbutton phone, also allows remote announcement changes. Southwestern Bell's FF-1700 Cordless Phone **$179.95** #2130; FA-450 Answering Machine **$99.95** #2140.



IF YOU FORGET THE WHO, WHAT, WHERE OR WHEN... IT WON'T

There it is, spread before you in all its impossibly complicated, completely unmanageable glory. Just one short week which might consist of 22 meetings, 134 messages, 297 calculations, 5 lunch dates, 1 business trip, your Aunt Sally's 80th (81st?) birthday, your child's 1st grade recital, and a shopping list of 10 things you need to pick up at the supermarket. How can you possibly keep track of it all? The Psion uses the latest technology to give you the tool that science fiction has long imagined — a completely portable hand-held device that organizes and carries all the information that's important in your life. There is nothing else like it with anywhere near the powers of the Psion. It enables you to carry your desk in a pocket. Or to put all the important information in your personal life where it can be easily retrieved. Or do both at the same time. Plus much more. You can ask the Psion to remember your mother's birthday — and remind you ahead of time. To keep a diary of your appointments, a log of your expenses, a record of the batting averages of your Little League team. The rental agents you met in Maine last summer, the couple you met in Atlanta, the places you like to eat in San Francisco, and the shipping dates you promised to clients. You can even find your plumber without knowing his name! Doctors carry Psion to have with them a complete record of what medication each patient is receiving. Executives carry Psions to have company data to work with when they are traveling. Mothers carry Psion to keep track and remind them of who's supposed to be where, when. Salesmen carry Psions to keep all their client records right at hand. The Psion is simply the strongest, smallest tool of information management yet devised. Operates on a single 9V battery. Weighs only 9 oz. Measures 5½″ x 3″ x 1″. Ready to use right out of the box. Can be equipped to communicate with personal computers, printers, telecommunications. Two-way data transfer. Additional plug-in memory packs give you up to a total of 320K of "on-board" memory. If your life is complicated, Psion can help you streamline it. No more wishing you had with you that name or number or fact or piece of information that spells the difference between frustration and satisfaction. The Psion puts the power of information where it will always do the most good — right in the palm of your hand. Model CM 8K Organiser II **$179.95** #2030; Model XP 32K Organiser II **$249.95** #2040; Additional 64K Plug-in Memory Packs **$89.95** #2050; Optional AC Adaptor **$19.95** #2060.

Here's a sample of what your Psion can hold

	Name Address & Phone #	Appointments with Reminder Alarms	Notes	Birthdays and/or Anniversaries	Expense Log	Number of Alarms	Memories for Saving Calculations
8K MODEL CM	110	25	25	20	30	8	9
32K MODEL XP	350	75	75	75	75	8	9
64k ADDITIONAL MEMORY PACKS ADD SPACE FOR	1,000 more	200 more	200 more	200 more	200 more	N/A	N/A

64K Plug-in Memory Packs

Memory is used interchangeably for Appointments, Notes, Names and Addresses etc. If you have fe er Names and Addresses, you can have more Appointments; if you have fe er Notes, you can have more Expense Logs; if you have fewer Appointments, you can have more Names and Addresses etc.

THE LIFESTYLE RESOURCE™

CREDIT CARD ORDERS CALL TOLL-FREE 24 HOURS **800-872-5200**

DEPT. SFAG08; 921 EASTWIND DR. SUITE 114: WESTERVILLE, OH 43081

SEND TO (PLEASE PRINT)

ADDRESS

CITY　STATE　ZIP

☐ CHECK OR MONEY ORDER　☐ MASTERCARD　☐ VISA　☐ AMEX　SIGNATURE

ACCT. #　EXP. DATE

ITEM NO.	QTY.	DESCRIPTION	ITEM PRICE	TOTAL PRICE
			SUB TOTAL	
			SHIPPING (see table at left)	
			TOTAL	

Shipping Charge covers UPS, handling and insurance for guaranteed delivery.

Up to	$20 . . . . $ 3.95		$50.01 to $60 . . . . $ 7.95
$20.01 to	$30 . . . . $ 4.95		$60.01 to $70 . . . . $ 8.95
$30.01 to	$40 . . . . $ 5.95		$70.01 to $100 . . . . $10.95
$40.01 to	$50 . . . . $ 6.95		Over $100 . . . . $12.95

UPS 2ND DAY AIR

UPS Second Day available for an additional $7.50 per order.

ORDER WITH CONFIDENCE

- *Most orders ship within 48 hours of receipt.*
- *Credit Cards aren't billed until after shipment.*
- *No risk 30-day return privilege.*

THE MEDIA DEVELOPMENT GROUP　©1988 G46 ADVERTISEMENT

Gravitational Lenses

These cosmic flukes offer a unique window on the secrets of the universe. Systematic searches now under way are designed to realize the scientific promise of the objects

by Edwin L. Turner

Statistical flukes, occasional occurrences of improbable events, are usually a worrisome source of confusion and error in science. Yet for several years astronomers have devoted considerable effort to finding and studying a class of celestial phenomena whose very existence depends on rare cosmic accidents. These are gravitational-lens events, which occur when two or more objects at different distances from the earth happen to lie along the same line of sight and so coincide in the sky. The radiation from the more distant object, typically a quasar, is bent by the gravitational field of the foreground object. The bending creates a cosmic mirage: distorted or multiple images of the background object.

Gravitational-lens images can exhibit a panoply of distortions: they may be displaced, magnified, shrunken, rotated, inverted, multiplied, deformed or spectrally aberrant. The effects depend on the precise geometry of the alignment, the spatial features of the background source and the strength and shape of the gravitational field of the foreground object. The phenomenon may therefore reveal many otherwise undetectable features of the image source, of the foreground object and of the space lying between them. Such observations could help to resolve several fundamental questions in cosmology. In the hope of someday realizing the scientific promise of gravitational lenses, several groups, including one of which I am a member, are carrying out systematic searches for lenses. At present, however, the mere discovery of a lens system represents significant progress.

EDWIN L. TURNER is professor of astrophysical sciences at Princeton University. He studied physics as an undergraduate at the Massachusetts Institute of Technology and received his Ph.D. in astronomy in 1975 from the California Institute of Technology. After brief stints as a postdoctoral fellow at the Institute for Advanced Study and assistant professor at Harvard University he was appointed to the Princeton faculty in 1978. His research has ranged over a number of topics in extragalactic astronomy and cosmology, including the dynamics of binary galaxies, galaxy clustering, quasar evolution, the cosmic X-ray background and gravitational lensing. Turner has a strong interest in advanced statistical techniques.

The history of gravitational-lens studies began, grandly enough, with Einstein's landmark 1915 paper on the theory of general relativity. The paper proposed three empirical tests of the new theory, the most famous one being the deflection of starlight grazing the limb of the sun. Classical Newtonian theory predicts that light will fall by a certain amount in the solar gravitational field; general relativity asserts that the deflection will be twice as large. Sir Arthur Eddington's confirmation of Einstein's prediction during a solar eclipse in 1919 won wide acceptance for general relativity.

Aside from this instance, gravitational lensing received scant attention until 1936, when Einstein published a short calculation showing that if two stars at different distances were exactly coincident in the sky, the image of the more distant one would form a ring. He dismissed such an alignment as being too improbable to be of practical interest. The following year Fritz Zwicky of the California Institute of Technology and Henry Norris Russel of Princeton University suggested other, more probable scenarios. In particular Zwicky, with his customary prescience, pointed out that lensing events involving extragalactic objects such as galaxies and clusters of galaxies were likely to be both observable and of serious scientific interest.

The modern history of gravitational-lens studies began in the early 1960's with the theoretical work of Sjur Refsdal, now at the University of Hamburg, and others, who considered how a gravitational-lens image might be analyzed—if one were found. They also suggested that gravitational lenses could reveal important cosmological information, such as the velocity of expansion of the universe and the density of dark matter. But these ideas were relegated to the playground of the theorist's imagination, because no one had ever observed a real gravitational-lens system.

Then in 1979 interest in the subject leaped with the discovery of a lens system (0957+561) in the constellation Ursa Major [see "The Discovery of a Gravitational Lens," by Frederic H. Chaffee, Jr.; SCIENTIFIC AMERICAN, November, 1980]. True to Zwicky's prediction, both this lens system and those systems subsequently discovered have involved the alignment of two or more extragalactic objects, in most cases an extremely distant quasar as the source and a moderately distant galaxy or galactic cluster as the lensing object in the foreground.

In the past decade theoretical and observational research on gravitational lenses has grown rapidly and steadily. At this writing at least 17 candidate lens systems have been discussed in the literature, including recently discovered giant luminous arcs and what appears to be an almost perfect "Einstein ring." The world's most powerful radio and optical telescopes are now devoting substantial amounts of time to finding and studying these objects.

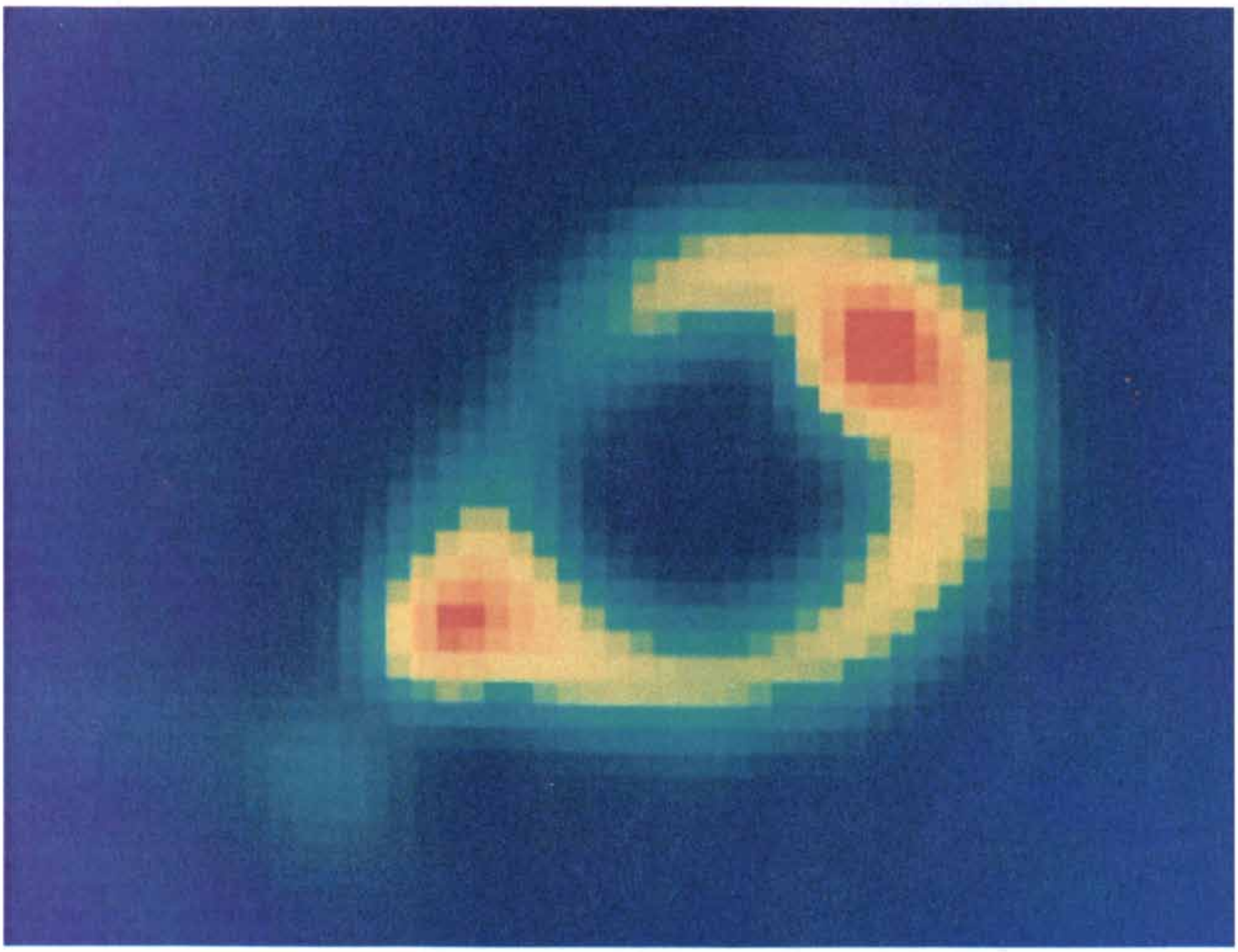
At first thought gravitational lenses appear to be not much more than amusing curiosities. They provide little experimental information about general relativity and gravitation that cannot be obtained more easily and reliably from studies within our own solar system. Why then should astronomers be so interested in these cosmic flukes? To answer this question, one must consider a typical gravitational-lensing event in greater detail.

Radiation that was emitted by a quasar early in the history of the universe travels toward our galaxy. As time passes, the expansion of the universe stretches the wavelength of the radiation—the well-known phenomenon of red shift. The expansion also causes the beam of radiation to spread as it propagates through space. Somewhere along its trajectory the radiation passes near a galaxy and bends in the latter's gravitational field. The amount of bending depends on which part of the field the radiation passes through, just as the angle of deflection of light passing through an optical lens depends on exactly where the light strikes the lens. Consequently the beam can cleave into separate pathways and give rise to multiple images, typically three or five (the original image plus extra images, which theory holds should occur in pairs). Each image appears in a slightly different part of the sky and may be magnified or distorted to varying degrees. Some of the images may also be reversed, as in a mirror, or inverted.

A gravitational-lens image arises, then, from a complex interplay of light paths with the lensing galaxy's gravitational field, which in turn depends on the spatial distribution of the gravitating material. By the dictates of general relativity this material encompasses all forms of matter and energy. Hence the lens object may be luminous or dark, ordinary matter or exotic and even unknown matter, matter condensed into stars or spread thinly as a gas of elementary particles. Moreover, all irregularities in the distribution of matter lying near the trajectory of radiation will contribute additional small deflections and thereby perturb the observed images.

Gravitational-lens images, then, are unique among astronomical objects. They are imprinted not only with the properties of the original source but also with information about the large-scale geometry and evolution of spacetime and about inhomogeneities in the universe. It is the hope of deciphering some of this information that motivates astronomers to study them.

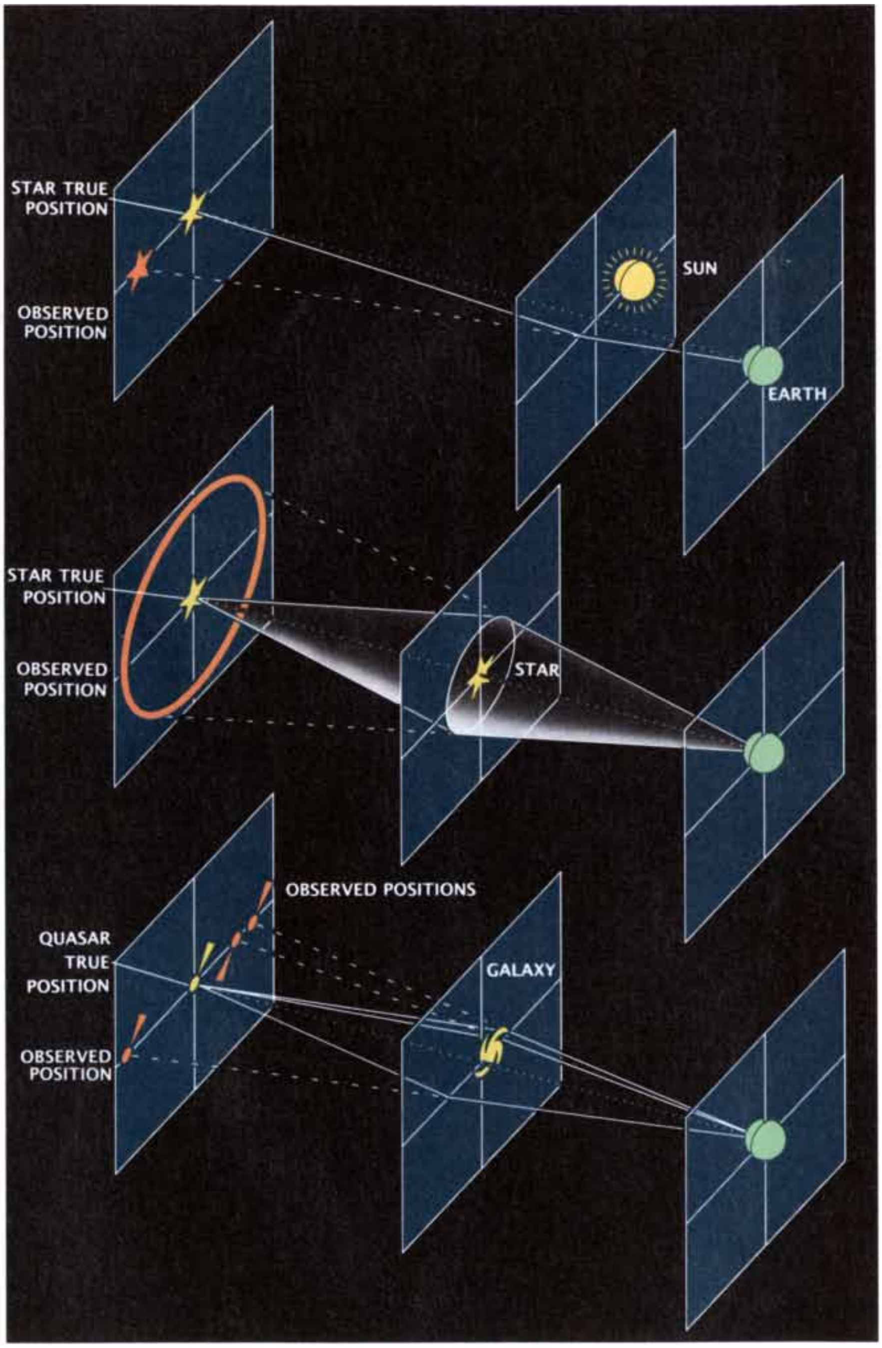
There are three general categories of information that astronomers hope to glean. First, as Zwicky originally proposed, a gravitational lens can operate as a natural telescope of cosmic dimensions: the lens may magnify the source image, revealing structural information that would otherwise be too small to resolve. Second,



MYSTERIOUS COSMIC LOOP (MG 1131+0456) found by Jacqueline N. Hewitt and her colleagues (including the author) may be the first sighting ever of an Einstein ring, a gravitational-lens image arising from a perfect alignment of objects (*see illustration on next page*). The object could be a typical radio source called a core jet, whose image is being distorted by an intervening elliptical galaxy. In this false-color radio image the two orange points are probably a double image of the source's intense core. The yellow circular form would then be a ring image of the jet, which extends directly behind the galaxy's center. The source, in the constellation Leo, was catalogued in the 1950's, but its unusual ringlike shape was discovered only last year in the gravitational-lens survey being carried out with the Very Large Array (VLA) radio-interferometric telescope.

the lensed image would yield information about the average large-scale properties of the universe, including the value of the cosmological constant. Third, the image would enable investigators to detect the presence of inhomogeneities in the universe, particularly those in the so-called missing mass or dark matter that is thought to account for most of the universe's total density. The last two applications offer the prospect of exploiting gravitational lenses as new tools for attacking classic problems.

As an example of the second category, gravitational-lens observations may enable astronomers to determine the size and age of the universe directly. This application derives from big-bang cosmology, which is based on the observation that, at large scales of distance, objects recede from one another at velocities proportional to the separation between them. One can explain this empirical fact if one assumes the universe is expanding uniformly. In an expanding universe the velocity of receding objects is proportional to the distance to the objects.



GRAVITATIONAL LENSING occurs when the gravitational field of an object in the foreground bends the paths of photons emitted by a more distant source. The deflection distorts the apparent position, size and shape of the source. Einstein proposed the deflection of starlight by the sun (*top*) and the ring that would appear if stars were aligned perfectly (*middle*). Lens systems found to date result from the alignment of extragalactic quasars and galaxies (*bottom*). Note the inverted image.

In the nearly 60 years since the expansion was discovered astronomers have tried to measure the "proportionality constant" relating velocity to distance. The constant would give a direct measure of both the universe's size and its age: the time elapsed since the big bang. To determine the constant, however, one needs to know the absolute distance to remote extragalactic objects. That task is formidable; traditional techniques for measuring astronomical distance have failed to yield an uncontroversial value in spite of decades of effort. The best results disagree by a factor of two.

Gravitational lenses offer a fundamentally new way of measuring astronomical distance and thus the proportionality constant as well. The new surveying technique, unlike all others, will work equally well for very remote objects and nearby ones. It is possible to calculate the geometry of the lensing event [*see bottom illustration on opposite page*]. Such an analysis might determine that the path of one image is, say, one-billionth longer than that of the other one. Light traveling along the two paths would therefore take different amounts of time to reach the earth. If the quasar suddenly flared in brightness (as quasars often do), one would observe the flare first in one image and slightly later in the other image. The path-length difference is then simply the observed time difference multiplied by the speed of light. Since this distance is one-billionth of the total distance, one can deduce the absolute distance to both the galaxy and the quasar.

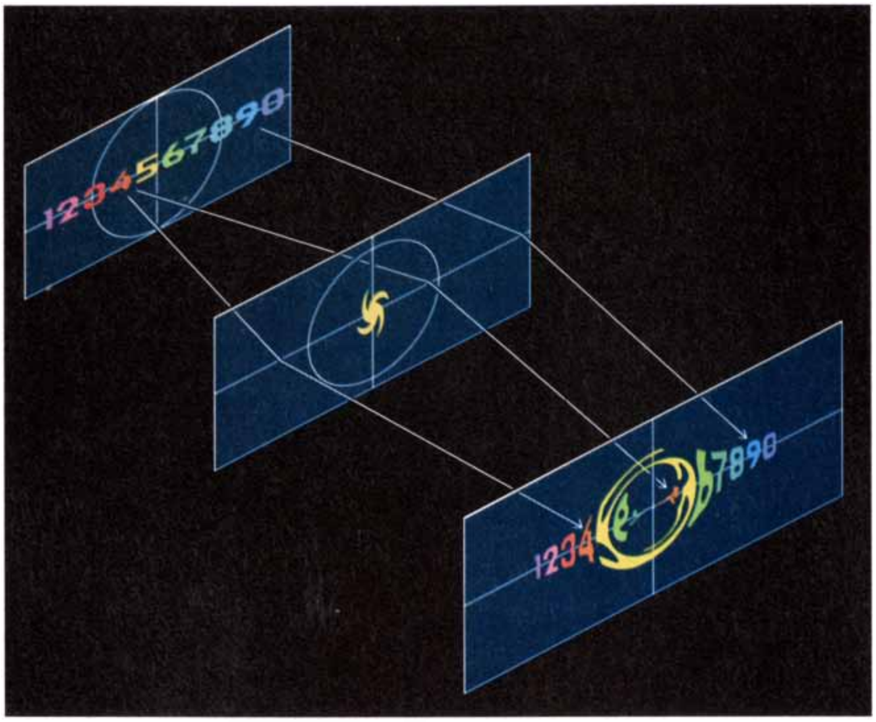
Gravitational lenses are particularly well suited to the third category of problems: the study of dark matter in the universe. Beginning with the pioneering work of Zwicky in the 1930's, investigators have steadily accumulated evidence for large-scale gravitational fields that are far stronger than can be explained by the observed (that is, luminous) stars and interstellar material. Most astronomers interpret this to mean that from 90 to 99 percent of the universe's total mass is made up of some undetected component—the postulated dark matter. Yet searches for low-level emission or absorption of radiation over a wide range of wavelengths have failed to reveal any direct indication of this dark matter.

Barring direct detection, the properties of dark matter must be studied entirely through its gravitational effects—hence the immediate connection to lens studies. Perhaps the simplest relevant observation is the frequency of lens events. Purely on statistical grounds the number of lens systems among objects at a particular distance should be proportional to the density of gravitating objects in the universe [*see illustration on next page*]. In practice, observational limits, particularly the resolving power of existing telescopes, restrict the types of lensing events that can be detected: one can study only systems in which the lensing objects are sufficiently massive to produce distinct, resolvable images. On the other hand, even a failure to detect lens events can provide valuable information by limiting the density of certain hypothesized classes of lenses. For example, the observed frequency of lens systems rules out the possibility that there are very large numbers of massive dark galaxies and black holes.

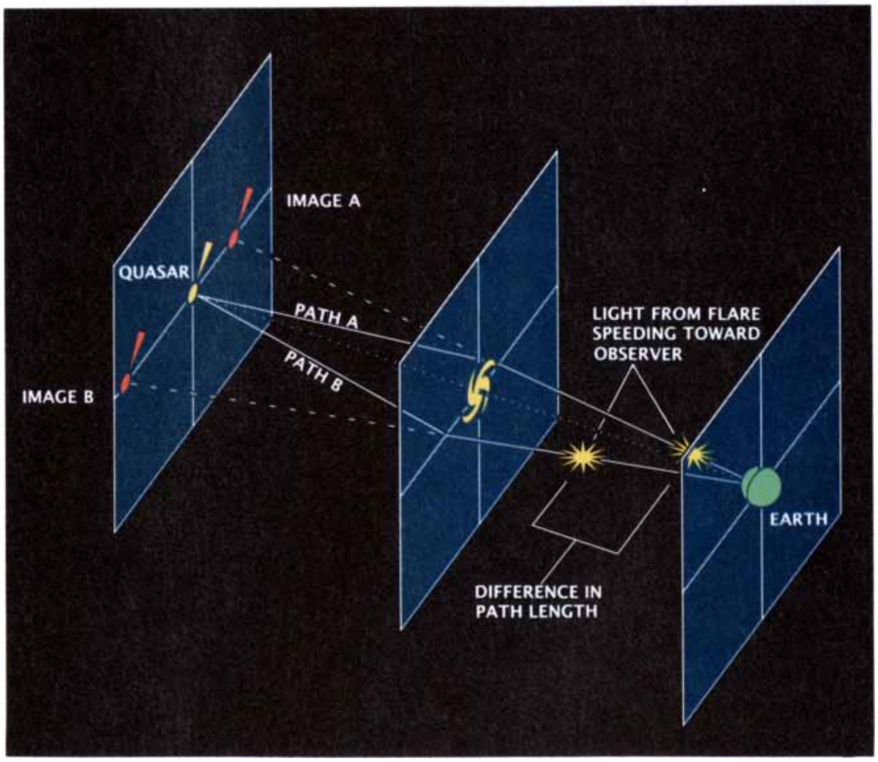
These examples only begin to list the potential applications of gravitational lenses. Observations of gravitational lensing could in principle also allow an independent measurement of galaxy masses and facilitate studies of the spatial structure of the intergalactic medium. Magnification by gravitational lensing could explain why an unexpectedly large number of high-red-shift quasars are found near low-red-shift galaxies; the galaxy may be amplifying the image of the distant quasar, making it easier to observe. Image amplification could also account for quasars that appear unusually bright or variable, quasarlike objects that have peculiar emission spectra and "superluminal" components of quasars that appear to exceed the speed of light. The possibilities are limited only by the theorist's ingenuity.

Yet these will remain merely tantalizing prospects as long as astronomers lack a statistically relevant sample of lens events. Investigators have just begun to search the sky systematically for lens systems; most of those known to date were discovered by chance and are statistically as meaningless as a public-opinion poll of people who happen to be encountered in the course of an afternoon walk.

Recognizing the potential benefits of gravitational-lens studies, several groups have undertaken ambitious surveys they hope will yield statistically relevant samples as well as indi-



GIANT NUMERALS in the sky (*upper left*) would appear distorted to an observer on the earth (*lower right*) because of the gravitational field of an intervening galaxy. Light striking within a critical impact zone (*broken circle*) would form a multiple image. Numerals closest to the line of sight to the galaxy undergo the greatest distortions.

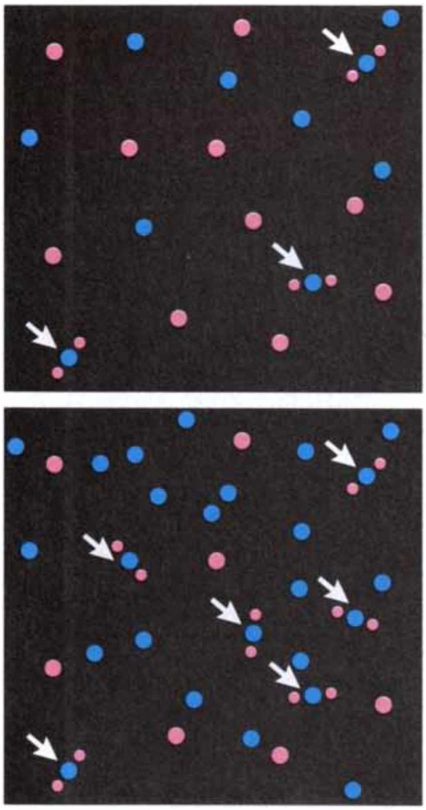


SURVEYING TECHNIQUES exploiting gravitational lenses could enable astronomers to measure absolute distances to remote quasars. An observer can deduce the geometry of the lens system from the angular separation of the images, the mass distribution of the galaxy and the ratio of the distances to the galaxy and to the quasar (based on their red-shift velocities, which are proportional to distance). Suppose this analysis finds path *B* is one-billionth longer than path *A*. If image *B* is seen to flare two years later than image *A*, the absolute difference between paths *A* and *B* is two light-years. Therefore path *A* is two billion light-years long.

vidual events of particular interest. Some groups have focused on careful screening of known quasars, with special attention paid to the most distant objects, which are most likely to be lensed by intervening galaxies. Other groups are extending the search to previously unknown quasars or other remote objects. I am collaborating in a project of the latter kind being carried out by astronomers at the California Institute of Technology, the Haystack Observatory, the Institute for Advanced Study, the Massachusetts Institute of Technology and Princeton University.

The first major difficulty faced by these surveys is the fact that lensing events are extremely uncommon to begin with. Among known classes of astronomical objects only quasars are typically far enough away to have a significant chance of being aligned with a foreground object. Even among quasars, lens systems are rare: roughly 2,000 quasars had been catalogued before the first chance discovery of a lensed one in 1979. Careful screening of known quasars can improve the yield but probably not to more than a fraction of a percent.

To add to the problem, quasars themselves are quite difficult to cull from the vastly more numerous stars in our galaxy, which they superficially resemble. In fact, during the nearly 25 years since quasars were discovered, much research has been focused on improving techniques for finding and identifying them. The initial step of a gravitational-lens survey, then, requires major observation programs aimed at producing useful samples of ordinary quasars. Hence the study of gravitational lenses is burdened from the outset by a kind of cosmic four-leaf-clover hunt.



DARK MATTER in the universe may reveal itself in a gravitational-lens survey. Quasars (*pink*) are lensed by dark objects (*blue*) in proportion to the density of the dark objects. If there are few dark objects (*top*), there should be fewer lens events (*arrows*) than there would be if there were many dark objects (*bottom*). Telescopes can resolve only those lensing events caused by sufficiently massive objects; hence a large population of small-mass objects could go undetected.

Our search for lens systems begins at radio wavelengths. One reason is that ordinary stars are generally weak sources of radio emission, whereas a large fraction of the known bright radio sources are associated with extragalactic objects, including quasars. A second reason is that radio astronomers currently have at their disposal many of the most powerful instruments and techniques in astronomy. The high angular resolution of radio telescopes reveals the very small-scale structure of an object, which is of great importance for understanding a lens system. Concentrating on radio sources does carry a price, however: only a few percent of all quasars are detectable radio sources. Because sources that do not emit radio waves have the same chance of lining up to form a lens event, the radio survey filters out all but a few percent of the lens systems in the sky.

The first step in the radio survey is being carried out by an M.I.T. group under the direction of Bernard Burke with the 300-foot telescope in Green Bank, W.Va., which is operated by the National Radio Astronomy Observatory (NRAO). The investigators scan large areas of the sky for sources of six-centimeter-wavelength radiation and determine their positions and brightness. Many thousands of such sources have been catalogued.

We then choose a subset of the sources for "snapshot" observations with the NRAO's Very Large Array. The VLA is an interferometric radio telescope consisting of 27 transportable dish antennas distributed over an area some 40 kilometers in diameter on a high plateau in New Mexico. For many purposes the VLA is the world's most powerful radio telescope: it has the amazing ability to form good images of a distant radio source in only a few minutes of observation (typically with a resolution of .4 second of arc and sensitive to a fiftyfold variation in brightness). Four of the 17 published lens candidates were discovered in the VLA-based survey. Extrapolation suggests that this project will ultimately yield from 10 to 30 good candidates.

Once the VLA has taken "snapshots" of selected sources, the radio images are inspected. Even ordinary radio sources can assume many shapes, ranging from simple starlike points to jets of plasma streaming from a central point source to chaotic structures not easily characterized. The problem is to winnow the best gravitational-lens candidates from this bewildering cosmic zoo. The trick is to exploit the empirical fact that most ordinary radio sources show no more than a single bright, unresolved component. If such a source is subject to strong gravitational lensing, two or more images of the single point will appear. Particularly good candidates show more than two such components separated by no more than 10 arc-seconds. (Larger separations are considered implausible because of the huge lens mass that would be required.) Sources in which components are stretched out along an arc rather than a straight line are good suspects as well, because lensing can produce curved distortions—the most perfect example being the Einstein ring.

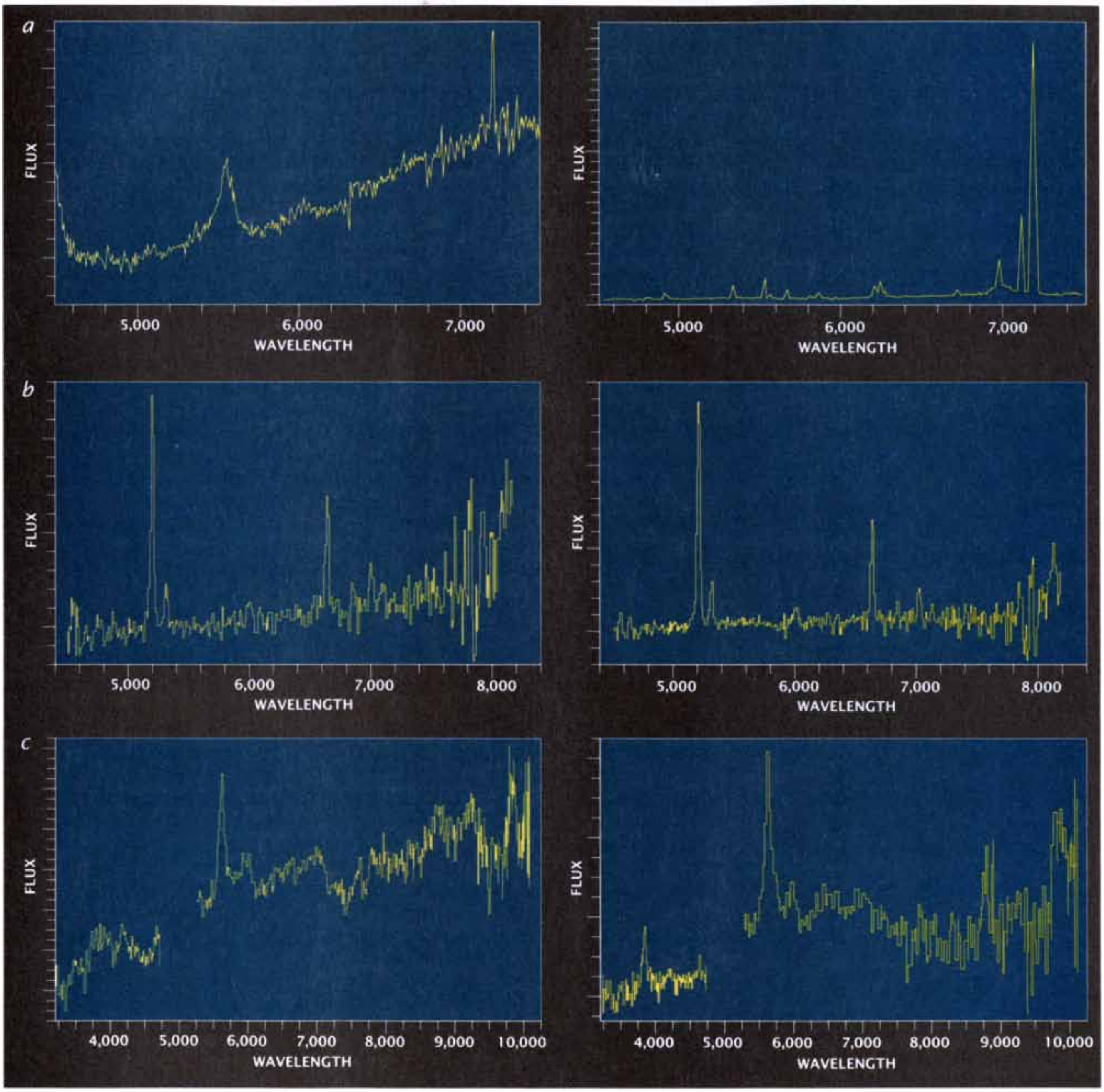
At this point the search turns to optical observations. The general theory of relativity predicts that the gravitational deflection of radiation is completely independent of its wavelength, and so if a lensing event produces multiple images of a source at radio wavelengths, it will also produce analogous multiple images at optical wavelengths. Hence a high-sensitivity optical detector is trained on the radio sources to obtain an optical image for comparison with the radio image.

These observations are being carried out primarily with an extremely sensitive detector, called a charge-coupled device (CCD), mounted at the prime focus of the National Optical Astronomy Observatories' four-meter Mayall telescope on Kitt Peak. The CCD enables this telescope—the third-largest in the U.S.—to detect, in about 30 minutes of exposure per source, objects at least 16 million times dimmer than the faintest star visible to the unaided eye.

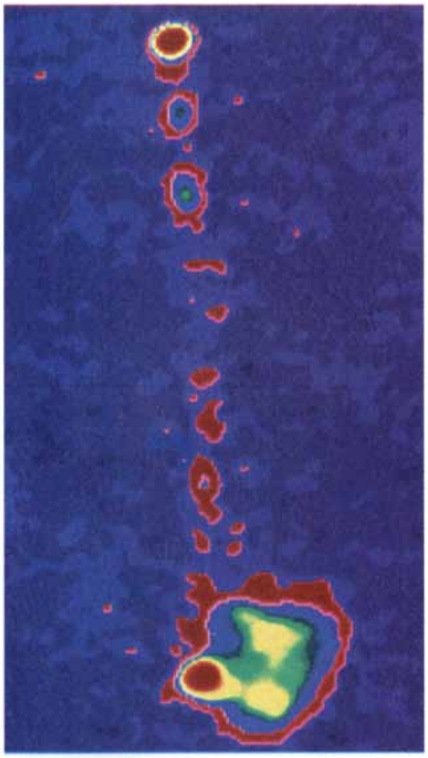
After an optical image is made it is compared with the radio image. In a few cases no optical counterpart of the radio source is visible, but in most cases there is one optical object associated with the source. Occasionally, however, two or more of the radio components have optical counterparts; this is the possible signature of lensing. Optical observations, then, filter the list of VLA lens candidates to produce a roster of prime suspects.

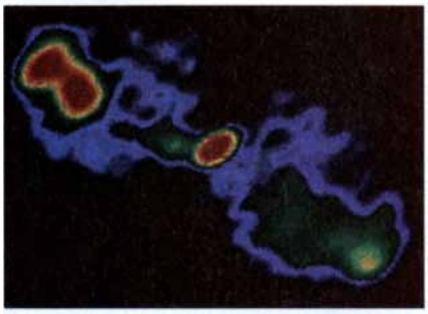
The sources that survive the screening process to this point are subjected to a critical final test. Each component in a multiple image is "fingerprinted" by taking detailed optical spectra showing the apparent brightness at a large number of wavelengths. Details of the optical spectrum are determined by a complex combination of many physical properties—the source's distance, motion through space, temperature, density, chemical composition and so forth. The spectrographic test should therefore distinguish real multiple images caused by lensing (which should have identical spectra) from images of similar but physically separate objects masquerading as a lens system.

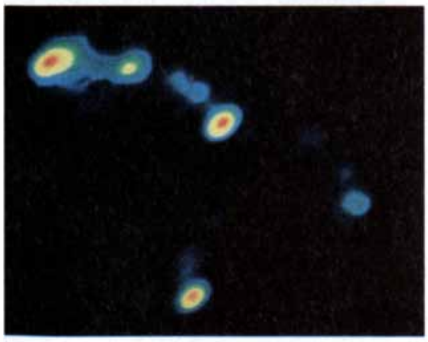
In principle this is a virtually ironclad test, because it seems highly unlikely that two independent objects would be identical in all their physical properties. In practice, unfortunately, the situation is less clear-cut. First, it is often difficult or impossible to obtain good spectra of faint objects sep-



SPECTRAL "FINGERPRINTS" test whether pairs of celestial objects are in fact multiple gravitational-lens images. Pairs with clearly dissimilar spectra (*a*) are almost certainly physically distinct. Pairs with nearly identical spectra (*b*) are likely to be lensed images. In many cases, such as pair 1146+111 (*c*), the results are ambiguous. Measurements made at wavelengths of between 5,000 and 7,000 angstrom units were strikingly similar and the object appeared to be a strong lens candidate. But later measurements made at other wavelengths revealed significant differences, casting doubt on the earlier conclusion.







WHICH of these three radio images is the result of gravitational lensing? The top two (1007+417 and 2355+490) are ordinary, unlensed core jets, pointlike objects ejecting gigantic streams of plasma. At the bottom is the gravitational-lens system 0957+561, discovered in 1979. One part of a quasar has been multiply imaged, resulting in two pointlike components aligned vertically at the center. The lensing galaxy, which also happens to be a radio source, appears as a faint blue blur just above the lower image; in optical observations the galaxy shows up clearly. All three brightness contour maps were made by the VLA.

arated by only a few arc-seconds. In some cases the task has consumed nearly a full night of observations at the Kitt Peak or the Mount Palomar telescope. Second, different quasars can have quite similar spectra by chance or because the objects are related in some way. Worse yet, certain effects, such as nonuniform magnification by the lens, can produce small aberrations in the spectra of lensed images. Spectral analysis is perhaps more like a comparison of smudged signatures than a comparison of clean fingerprints. Still, if the spectra of two or more optical components are consistent or very similar, the object is usually regarded as a good enough candidate to be published and discussed in professional journals.

Once a lens candidate has been identified by one of the surveys, it generally merits substantial further study to test whether it is actually a lens system and, if it is, to understand it and model it in detail. One of the first priorities of such follow-up studies is to search for the lensing object itself. In addition one generally wants to obtain the highest-resolution images over the widest possible range of wavelengths in order to constrain certain theoretical models. Between hunting for the lensing object and obtaining further data of various kinds, a single good gravitational-lens candidate can consume perhaps a tenth as much observing time as was consumed in making the entire survey from which it was drawn.

The first generation of lens surveys is not yet complete, and new classes of lenses continue to appear. Still, the obvious question arises: How many lenses have been found so far? The difficulty in answering this question highlights a fundamental problem in lens studies. Of the 17 gravitational-lens candidates reported so far in the professional literature, only five are considered to have been reliably established by subsequent observations. Another three are generally regarded as weak or speculative cases with less—perhaps much less—than a 50 percent chance of actually being lens systems. In the remaining nine cases the evidence is mixed or is sparse enough so that the final judgment could swing either way.

As might be concluded, little of the scientific promise of gravitational lenses has yet been realized. The work has not yielded a clear value for the proportionality constant or any of the other fundamental cosmological parameters. There have been interesting results on dark matter—its distribution in specific objects and limits on its total abundance in specific forms. For example, it has been established that there are not many black holes of galactic mass. (These results are useful but so far have only confirmed preexisting conclusions or assumptions. Information on missing-mass distributions in a galaxy can be inferred from its rotational behavior, for example.) And in at least one system (2016+112) Zwicky's dream of a cosmic telescope has come true, providing a glimpse of a peculiar quasar that would otherwise be too faint to detect. The present list of scientific achievements is modest but, considering the immature state of the field, perhaps not discouraging.

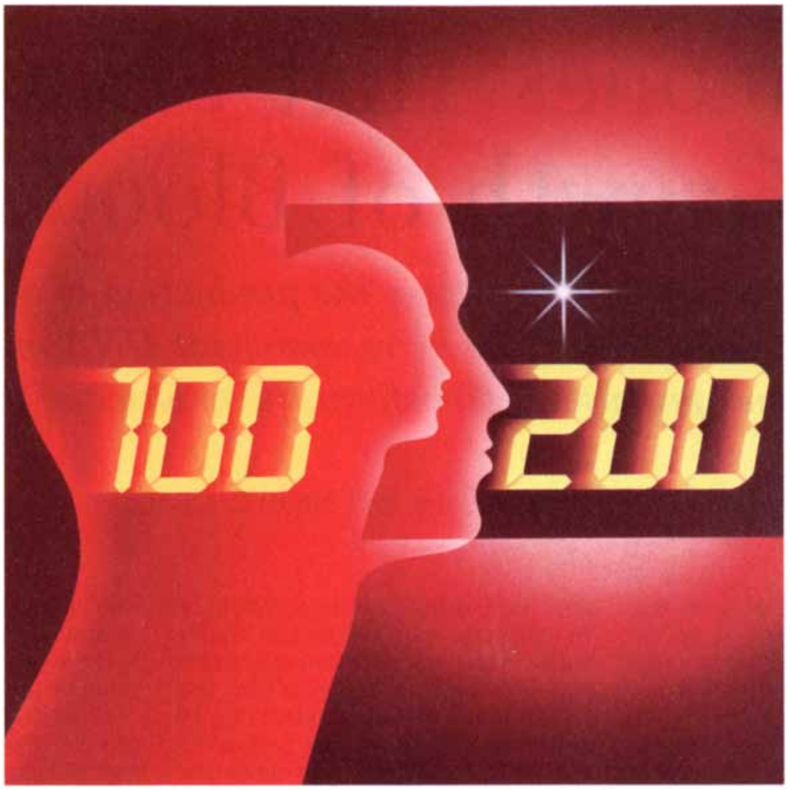
The ultimate value of gravitational-lens studies remains a debated question. In the idealized world of astrophysical theorizing it is possible to dream up many elegant and powerful applications of lenses to fundamental questions. In the real world of complex actual cases and limited information about them, not all the potential applications can ever be achieved in practice. To some extent the difficulties are inherent in the nature of cosmic gravitational lenses. Indeed, they arise from the very effects that make lenses potentially informative. As an example, lensed images contain information about both the total mass distribution of the lensing galaxy and the mean large-scale properties of the universe. One wants to learn about both, but it is often difficult to say anything about one without already knowing something about the other—a classic Catch-22 situation. In spite of such formidable obstacles, gravitational-lens applications seem worth pursuing, if for no reason other than that they offer a new tool for tackling fundamental problems that so far have largely eluded the traditional tools of the trade.

FURTHER READING

QUASARS AND GRAVITATIONAL LENSES: PROCEEDINGS OF THE 24TH LIÈGE ASTROPHYSICAL COLLOQUIUM. Edited by Jean P. Swings. Presses universitaires de Liège, 1983.

QUASARS AND GRAVITATIONAL LENSES. Edwin L. Turner in *Science*, Vol. 223, No. 4642, pages 1255–1259; March 23, 1984.

GRAVITATIONAL LENSES AS TOOLS IN OBSERVATIONAL COSMOLOGY. Claude R. Canizares in *Observational Cosmology*, edited by Adelaide Hewitt, Geoffrey Burbidge and Li Zhi Fang. D. Reidel Publishing Co., 1987.



ALCOA'S SECOND CENTURY IS BEGINNING MUCH AS OUR FIRST...

This month Alcoa is 100 years old. It was founded on the promise of science and engineering and an abiding faith in the future. Much has changed over the years, but these tenets remain.

As we begin our second century we continue to regard science, engineering and faith in the future as fundamental forces guiding Alcoa into a tomorrow of new, technically-rich products and processes.

In the lead are imaginative and intelligent people. Annually, we honor some of them for stimulating and capturing the benefits of technological excellence. The 1988 awards and their recipients are:

The 1988 Francis C. Frary Award for a career of outstanding individual contribution to Alcoa technology:

James T. Staley
Senior Technical Consultant
Aerospace Products

Jim Staley, whose scientific accomplishments have often altered conventional wisdom, has achieved prominence in the field of physical metallurgy, especially as it is applied to high-strength, heat-treatable aluminum alloy development for the aerospace industry. Most notable among many scientific accomplishments were Jim's contributions to the development of a breakthrough aerospace aluminum alloy that possessed a unique combination of properties necessary for today's aircraft.

The 1988 Arthur Vining Davis Award for outstanding group achievement in Alcoa technology:

Advancements in aluminum alloy and aerospace technology have progressed hand-in-hand from Kitty Hawk to Cape Canaveral. Alcoa has continued its leadership tradition with the introduction of a series of new aerospace alloys for aluminum sheet, plate, forgings, extrusions and rivets. The following Alcoans developed a new process specifically designed to manufacture aluminum products in a special temper and contributed to the development and commercialization of these alloys:

Bruce E. Anderson
William A. Anderson
Lee E. Aughinbaugh
Daniel C. Boley
Charles E. Brooks
Melvin H. Brown
Robert H. Brown
Stanley J. Cieslak
Walter D. Coker
Robert E. Davies
James L. Eriksson
Edmund C. Franz
Robert L. Garrett
Luther Greenhill
Ray M. Hart
John E. Hatch
Harold Y. Hunsicker
John E. Jacoby
R. Steve James
Ronald J. Kegarise
Franklin L. King
Paul W. Kroger
Sootae Lee
Bernard W. Lifka
John Liu
Paul L. Mehr
Larry L. Mueller
Norman W. Nielsen
Glenn E. Nordmark
Christopher R. Owen
Grant G. Owen
Robert C. Pahl
Basil M. Ponchel
Chester S. Recko
Raymond T. Richter
Ralph R. Sawtell
Richard Schmidt
Donald E. Scott
Sylvester Scott
Eugene D. Seaton
Samuel L. Shelby
Dell F. Skluzak
Arvid H. Sorenson
Donald O. Sprowls
Edwin H. Spuhler
James T. Staley
Richard J. Stokwisz
James D. Walsh
Robert W. Westerlund

The 1988 Chairman's Award for Significant contributions to the development and implementation of materials, processes and systems technologies:

William H. Dunlap

Known as the "Father of Alcoa Precision Forgings," Bill recognized in the late 1960's that an emerging technology intended for producing small parts could be applied to larger, precision forgings. This innovation led to the production of complex contours for aluminum and titanium precision parts.

Daniel H. Hugh

Dan Hugh's expertise in rolling lubricants, process shielding and roll technology has allowed Alcoa to maintain world leadership in the production of rolled aluminum products.

Ambrose G. (Rusty) Rust

Rusty Rust has been a major contributor in the evolutionary improvement of primary aluminum production by the Hall Smelting Process. His expertise has helped Alcoa reduce the energy intensity of this process, a critical factor in efficiency.

ALCOA CENTENNIAL 1988

ALCOA

Hormones That Stimulate the Growth of Blood Cells

Each hemopoietin regulates the production of a specific set of blood cells. Now made by recombinant-DNA methods, these hormones promise to transform the practice of medicine

by David W. Golde and Judith C. Gasson

In Leviticus the Bible declares that "the life of the flesh is in the blood." Metaphorical significance aside, the statement is literally true: each type of blood cell is required for life. Erythrocytes (red blood cells) carry oxygen to the tissues. Leukocytes (white blood cells) defend the body from pathogens and tumors. Until recently only a few methods were available to bolster the functions of blood cells: vaccines and better nutrition boosted immune defenses; transfusions might make up for lost red-cell function. Within the last few years a group of hormones have been discovered that may change all that. These protein growth factors are known as hematopoietic hormones or hemopoietins, from the Greek words *haima* (blood) and *poiein* (to make).

Indeed, making blood is just what these hormones do. All the various types of blood cells develop from a single progenitor called a stem cell. Each hormone in the array of hemopoietins causes specific classes of blood cells to be made and "primes" them, enhancing their function. Because genes for several hemopoietins have recently been cloned, these hormones can now be made in quantity, and physicians may soon be able to elicit production of specific types of blood cells as a routine form of therapy. As a result the need for blood transfusions may be greatly diminished, bone-marrow transplantation may be simplified and rendered less risky and the immune system may be bolstered to fight against pathogens, AIDS or tumors. In short, the hemopoietins may effect a revolution in medicine as profound as the introduction of antibiotics half a century ago.

DAVID W. GOLDE and JUDITH C. GASSON are colleagues at the University of California at Los Angeles School of Medicine who collaborate in research on the colony-stimulating factors. Golde, professor of medicine, serves as chief of the Division of Hematology-Oncology, director of the U.C.L.A. AIDS Center and director of the U.C.L.A. General Clinical Research Center. Gasson, assistant professor of medicine, is associate director of the division Golde heads.

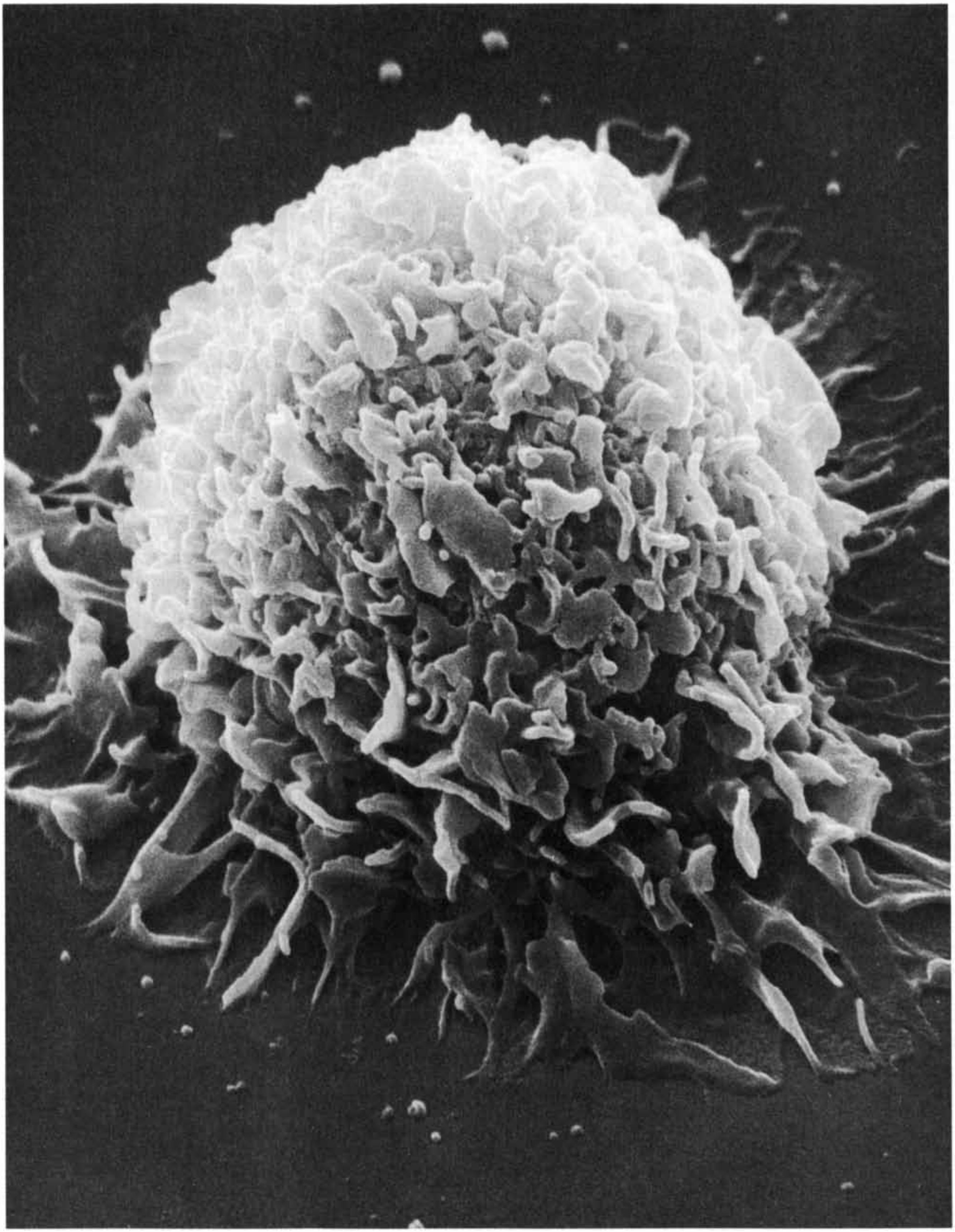
Although red blood cells come in only one form, the white cells of the immune system include three different lineages that carry out specialized functions: granulocytes, monocytes and lymphocytes. The granulocytes are in turn subdivided into three groups called neutrophils, eosinophils and basophils—names that describe the type of stain for which each cell has an affinity. The neutrophil is essential in the host's defense against bacteria and some fungi; the eosinophil has a role in defending against parasites such as worms and protozoans; the function of the basophil is less well understood. Monocytes (and related cells called macrophages) are crucial in the defense against intracellular parasites such as viruses and certain bacteria. Lymphocytes help in recognizing and destroying many types of pathogens.

How does this panoply of cell types develop from a single precursor? The process resembles the elaboration of a family tree, with a series of steps taking each descendant farther from the undifferentiated stem cell. Most of this development takes place in the bone marrow, where the stem cells reside. When a stem cell divides, it can replicate itself as a stem cell or become committed to a particular developmental pathway. It is not yet fully understood what governs the "decision" a stem cell makes in becoming committed to a specific lineage. Clearly the process involves expression of some of the cell's genes, but whether commitment is primarily a random process or one dependent on environmental factors is not clear.

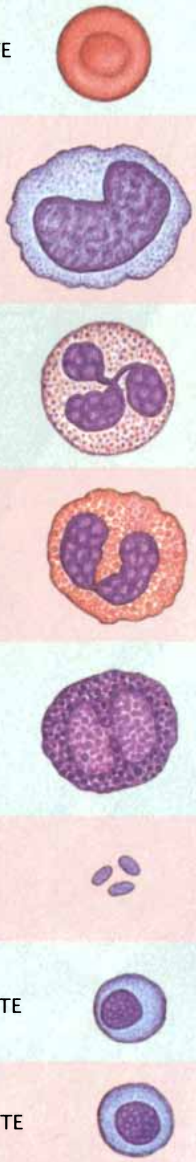
In any event, the result of commitment is that the stem cell produces receptors on its surface that respond to specific hormonal signals. Those signals in turn push the cell farther down the pathway toward ultimate specialization. The first major branching of the tree divides the precursors of lymphocytes from those of all other types. At this stage the committed stem cells are not distinguishable on the basis of their form. As differentiation proceeds, the first recognizable precursors appear, including erythroblasts (red-cell precursors) and myeloblasts (precursors of granulocytes and monocytes).

Although it is not difficult to sketch the family tree of the cells of the blood, it is a more complex task to explain how the process is regulated. The pathway leading to mature red blood cells was investigated first and is now fairly well understood. The existence in the blood of a substance regulating the production of red blood cells was postulated early in this century. Since the function of erythrocytes is to transport oxygen, it seemed logical that red-cell production should be tied to the need for oxygen-carrying capacity. Direct evidence for a circulating "factor" that stimulated erythropoiesis (production of red blood cells) in response to changes in the level of oxygen in the environment and in the tissues was provided by Kurt R. Reissmann of the United States Air Force School of Aviation Medicine.

Reissmann connected the vascular systems of two laboratory rats and



MACROPHAGE is a white blood cell that has a central role in the immune response. A motile cell found in many body tissues, the macrophage phagocytoses (ingests) both pathogens and wastes. The "ruffles" on the surface of the cell serve in locomotion, cell spreading and phagocytosis. The macrophage is among the cells that release hormones called colony-stimulating factors (CSF's), which induce white blood cells to proliferate and mature. The macrophage is shown enlarged some 9,000 diameters in this scanning electron micrograph made by Shirley G. Quan in the laboratory of the authors.



BLOOD CELL	LIFE SPAN IN BLOOD	FUNCTION
ERYTHROCYTE	120 DAYS	OXYGEN TRANSPORT
MONOCYTE	3 DAYS	HOST DEFENSE, IMMUNE SURVEILLANCE, (PRECURSOR OF TISSUE MACROPHAGE)
NEUTROPHIL	7 HOURS	HOST DEFENSE
EOSINOPHIL	?	HOST DEFENSE AGAINST PARASITES, ALLERGIES
BASOPHIL	?	INFLAMMATION AND ALLERGY
PLATELETS	7-8 DAYS	BLOOD CLOTTING
T-LYMPHOCYTE	?	CELLULAR IMMUNITY
B-LYMPHOCYTE	?	ANTIBODY DEFENSES (PRECURSOR OF PLASMA CELL)

BLOOD CELLS are of many types, each of which has specific functions. The erythrocyte is the oxygen-carrying red blood cell; all the others are leukocytes (white blood cells). Some monocytes leave the bloodstream and mature into macrophages in the tissues. The neutrophil, eosinophil and basophil are subtypes of the granulocyte.

exposed one partner to low oxygen levels; there was a stimulation of red-cell production in the other partner as well, implying that a blood-borne substance had passed between them. The observations of Reissmann and others paved the way for definitive studies by Allan J. Erslev of Jefferson Medical College. Erslev and his colleagues rendered rabbits anemic by bleeding them. Plasma (the noncellular, liquid part of the blood) from the anemic animals was then injected into normal rabbits, where it stimulated erythropoiesis.

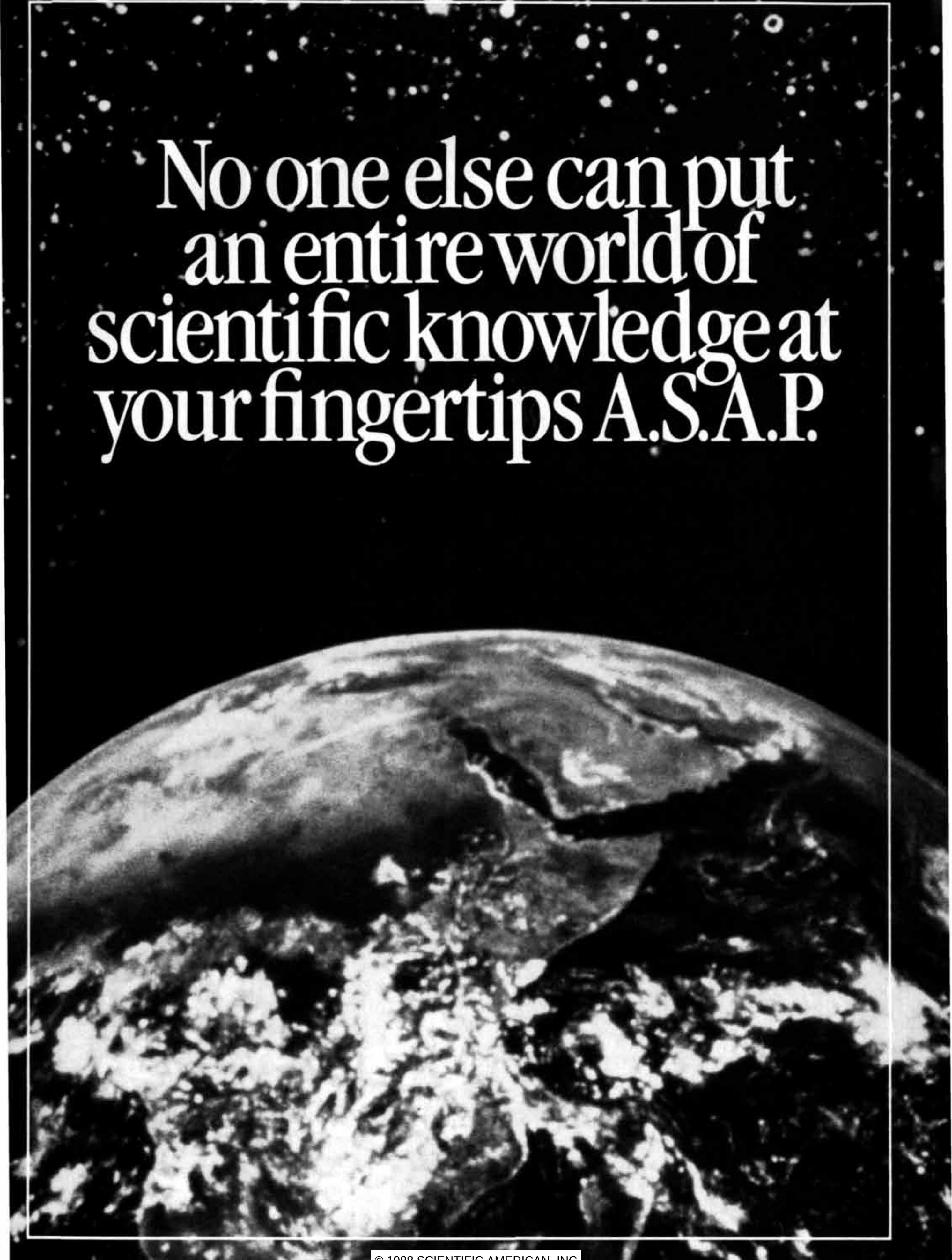
In the years after Erslev's work much was learned about the red-cell growth substance. It was found that the factor is made in the kidney, circulates in the plasma and is excreted in the urine. In laboratory culture its effect is to stimulate incorporation of iron into developing red-blood-cell precursors and to stimulate the growth of colonies of red blood cells. Although much knowledge of the effects of the factor was obtained, the postulated molecule itself was not easy to find—largely because its concentration in body fluids is low. It was not until 1977 that Takaji Miyake and Eugene Goldwasser of the University of Chicago succeeded in isolating erythropoietin, as the factor came to be called; their efforts yielded only a few milligrams from 2,500 liters of human urine.

Once the protein molecule was in hand, the work moved quickly. Investigators at Amgen and Genetics Institute, Inc., quickly cloned the erythropoietin gene. The purified hormone has a molecular weight of about 34,000 daltons (a hydrogen atom has a weight of one dalton), and the man-made molecule has all the effects observed with the natural one. Those effects have recently been put to use in clinical trials of biosynthetic erythropoietin. Joseph W. Eschbach and John W. Adamson of the University of Washington School of Medicine administered the hormone to patients with severe kidney disease who were dependent on dialysis to free their blood from wastes and toxins.

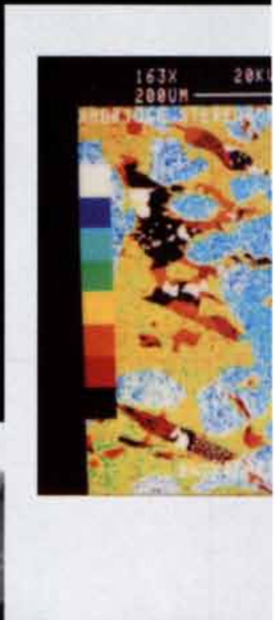
Dialysis corrects most of the problems caused by severely impaired kidney function except for the accompanying anemia (decrease in the number of circulating erythrocytes). The anemia is due to the low level of erythropoietin, but it may be exacerbated by blood-borne inhibitors and toxins that interfere with the effect of the hormone. The work done by Eschbach and Adamson shows, however, that biosynthetic erythropoietin can overcome the inhibitors, returning the red-cell count to normal. More than 300 dialysis patients have been experimentally treated, and clinically meaningful increases in red-cell production have been seen in almost all of them.

Renal disease is not the only area where erythropoietin will ultimately have profound consequences. Introduction of the hormone will dramatically alter the way blood banks function. Because erythropoietin can theoretically increase red-cell production tenfold, the need for transfusions will decrease greatly. Surgical patients will need fewer transfusions because of erythropoietin's capacity to stimulate their own red-cell production; they may store their own blood and receive erythropoietin before, during and after surgery. In the future the hormone may make it possible for blood banks to grow red cells, thereby transforming the blood bank into a production facility as well as a storage facility. Erythropoietin will also be widely employed to restore red-cell levels in patients suffering from blood-cell cancers or undergoing chemotherapy for many forms of cancer.

Although elucidating the effects of erythropoietin was clearly not a simple job, it was somewhat simpler

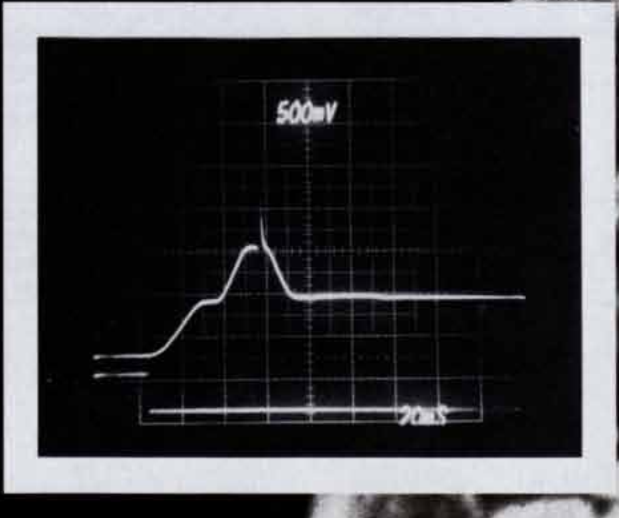

No one else can put an entire world of scientific knowledge at your fingertips A.S.A.P.

Particle
A



S.E.M. Image of a Heart Valve –
Type 55

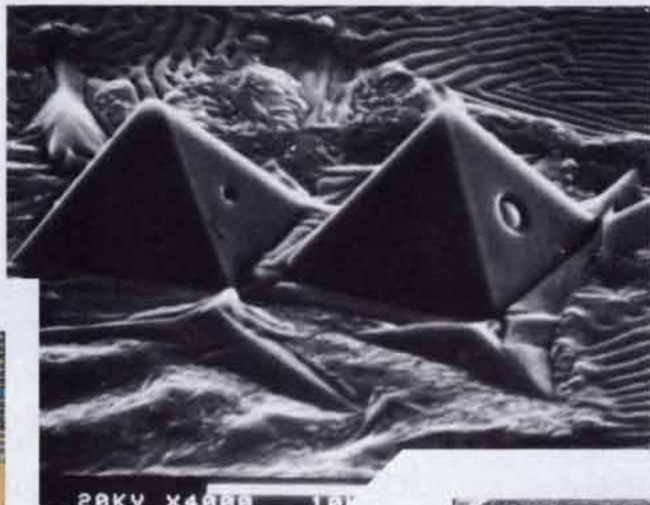
Oscillograph Image –
Type 667



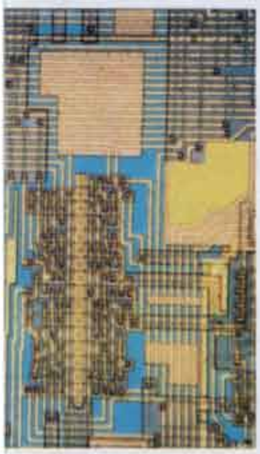
Electrophoresis Pat
Enzyme Digestion of
Polacolor ER Typ


alysis of a Metal Alloy –
Film Type 339
CAD/CAM Image of a Ball Bearing –
PolaChrome Instant 35mm Film
Polaroid
Instant Slide
Molecular Structure of Galactosamine –
PolaBlue Instant 35mm Film
Polaroid
CH2OH
HO
O
H
OH
NH2
PolaBlue Instant Slide
Photomicrograph of Integr
AutoFilm Type 3
X-ray of Pens –
Type 53
n from an
smid DNA –
69 Film
COST OF
% OF MANU
Percent
20
15
10
5
0
1st Qtr
External F
Prevention

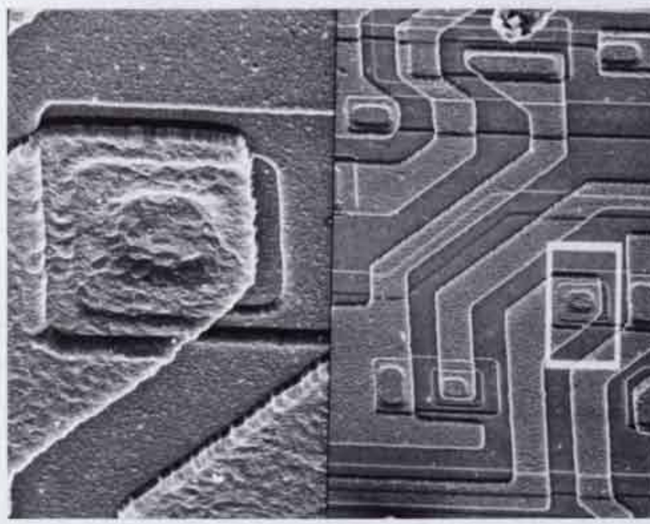
S.E.M. Image of Inner Surface Cavity
of a Nickel-Tin Eutectic Alloy Sphere –
Type 55



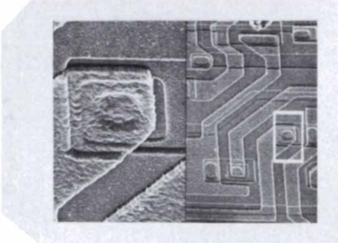
ed Circuit –



S.E.M. Image of an Integrated Circuit –
PolaPan Type 52



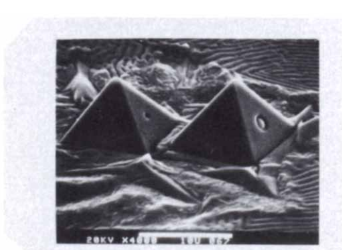
Cost Analysis Graph –
Colorgraph Type 691



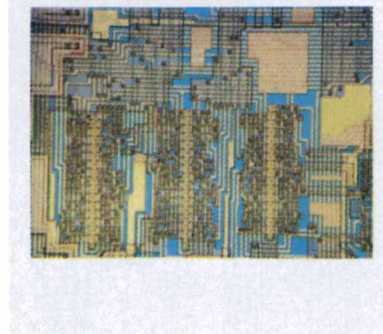
Electronics. PolaPan Type 52 sheet film was used to capture this split-field S.E.M. of an integrated circuit. This black and white, fine grain print film offers a wide tonal range and provides superb detail.

Quality Assurance. This particle analysis of a metal alloy, shown on Polaroid High Speed AutoFilm Type 339, displays consistent saturated color and requires no pulling, timing, or peeling. It was made with the Polaroid FreezeFrame Video Image Recorder.

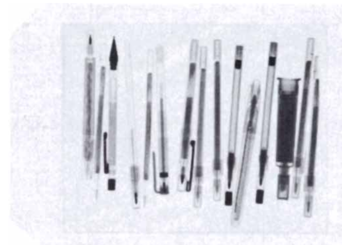
Failure Analysis. This S.E.M. image of a heart valve, magnified 18x, is on Type 55 instant film. This black and white sheet film provides you with a positive and negative to make high quality, professional images.



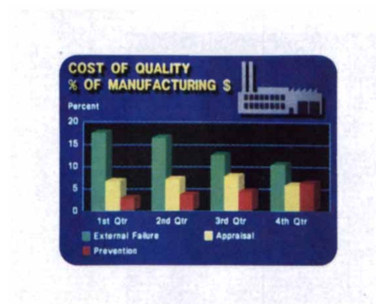
Metallurgy. This S.E.M. image magnified 4000x is of the inner surface shrinkage cavity formed on solidification of a nickel-tin eutectic alloy sphere levitated in microgravity aboard the space shuttle. To capture the fine detail, Polaroid Type 55 black and white instant film was used.



Clean Room Microscopy. Polaroid AutoFilm Type 339 was used to produce this photomicrograph of an integrated circuit viewed through an optical microscope.

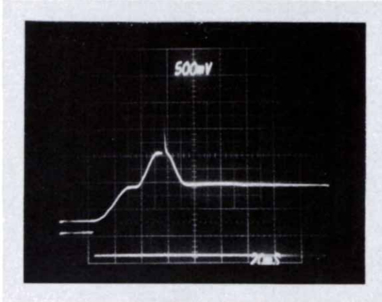


Non-Destructive Testing. This X-ray of a representative group of pens was made on Type 53 to view the alignment of the parts. This general purpose high speed film requires no print coating.

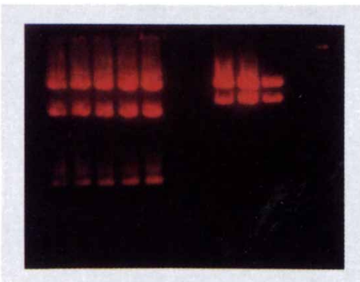


Presentations. An instant color overhead transparency, such as this cost analysis graph, can be made using Polaroid Colorgraph Type 691 film. This full color film creates small format overhead transparencies so the latest findings can be presented instantly. The actual graph was generated on a personal computer using the PalettePlus Computer Image Recorder.

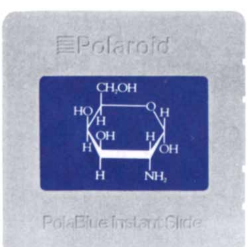
Graphic Design. To capture the subtle color differentiations in this CAD/CAM image of a ball bearing, new Polaroid High Contrast PolaChrome Instant 35MM Slide Film was used. New High Contrast PolaChrome provides bright, high quality, color slides in minutes for presentation or documentation.



Test and Measurement. Polaroid Type 667 black and white high speed print film and Polaroid DS-34 Direct Screen Camera were used to document this oscillograph.



Biotechnology. Polaroid Polacolor ER Type 669 film and the Polaroid MP-4 Multipurpose Camera were used to produce this image of an electrophoresis pattern from an enzyme digestion of Plasmid DNA. Type 667 and Type 53 black and white films are also used for electrophoresis documentation.



Physical Science. This white on blue slide of the molecular structure of galactosamine was made with new Polaroid PolaBlue Instant 35MM Slide Film. New PolaBlue provides bright, high quality, white on blue slides in minutes for presentation or documentation. It's more cost effective than traditional methods of making blue slides.

Polaroid offers you more than forty different instant film types – an entire world of solutions – for general purpose or highly specific applications.

With Polaroid instant color or black and white print films, you can actually speed up the process of scientific inquiry by providing faster documentation. With records available in seconds, you can be certain you have what you need, so you can quickly move on to the next stage of your experimentation or research.

You'll also be able to analyze your work more quickly and thoroughly than ever before.

As for presentations and meetings, there's simply no faster way to get the images you want than Polaroid instant 35mm slides and overhead transparencies.

Polaroid also offers you a complete range of instant imaging systems, including close-up and direct screen cameras, copy stands, and electronic imaging systems for recording from video sources and personal computers.

In many cases, our instant imaging equipment has been developed by working hand-in-hand with the leading scientific equipment manufacturers. The results are instant imaging systems that deliver fast, easy, and lasting photographic records from computers, video systems, microscopes, oscilloscopes, and many other scientific instruments.

And you get it all A.S.A.P. – As Soon As Polaroid.

Polaroid instant images. If you're not using them, you're missing instant access to an entire world of scientific knowledge.

For a complete brochure (please specify whether you want information about the Quality Assurance field or the Research field) about Polaroid instant imaging solutions for the professional scientific community, or for a free consultation, call **1-800-343-5000**. Or write, Polaroid Corporation, Dept. 687, P.O. Box 5011, Clifton, N.J. 07015.

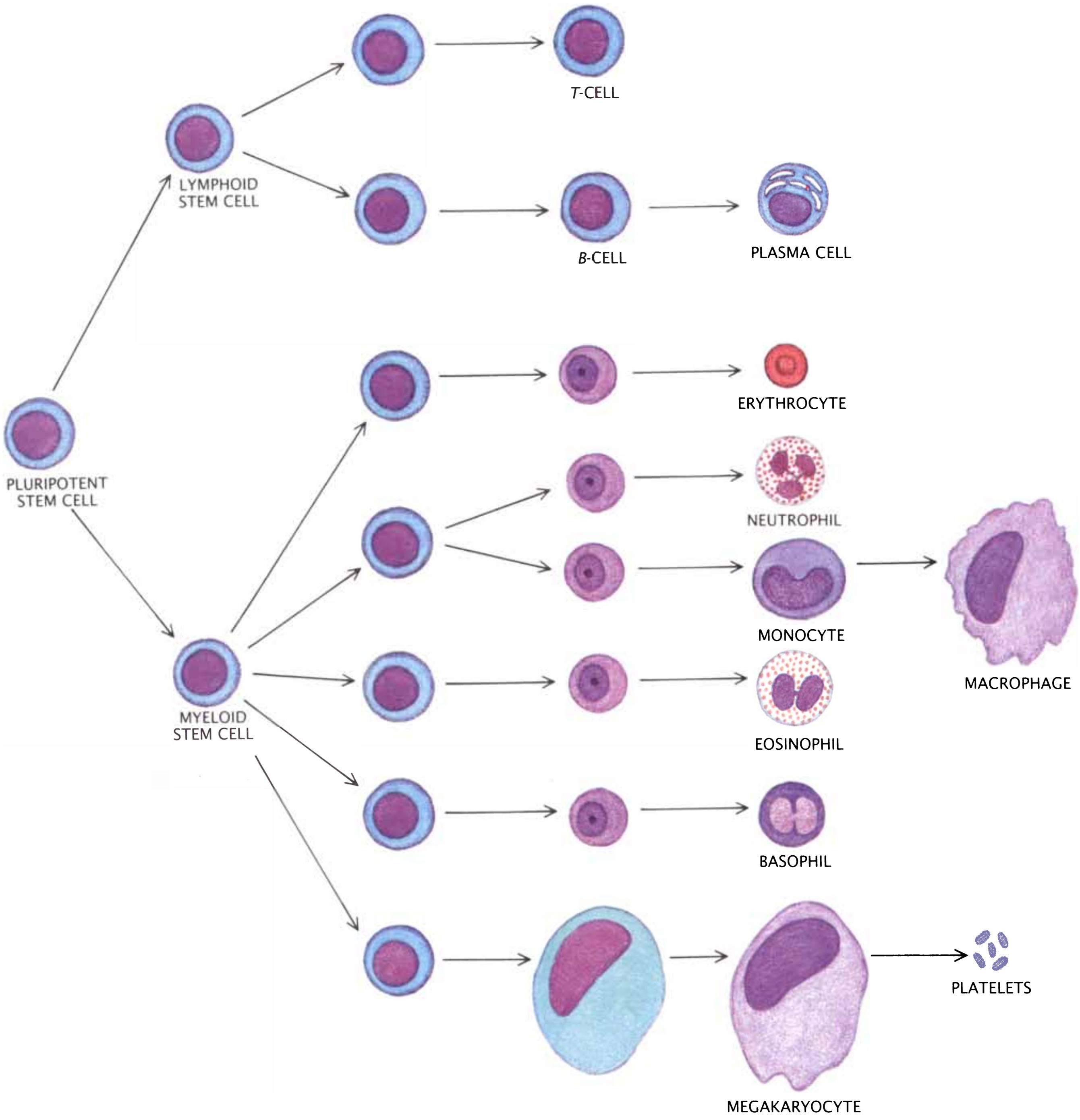
Polaroid

Result	Product	Format/Size	Speed ISO & DIN equivalents	Normal Development Time at 75°	Special Characteristics
Instant Color Prints	Type 779	Integral Pack 3½x4¼ in.	ISO 640/29°	Self-timing	Balanced for daylight and electronic flash.
	Time Zero Type 778	Integral Pack 3½x4¼ in.	ISO 150/23°	Self-timing	Balanced for daylight. Image visible in 10-15 seconds.
	Type 909	Integral Pack 4x4 in.	ISO 640/29°	Self-timing	Balanced for daylight and electronic flash.
	Type 339	Integral Pack 4.5x4.2 in.	ISO 640/29°	Self-timing	General purpose, high speed, medium contrast.
	Polacolor ER Type 669	Pack 3¼x4¼ in.	ISO 80/20°	60 seconds	Medium contrast; extended dynamic range. Balanced for daylight and electronic flash.
	Polacolor ER Type 559	Pack 4x5 in.	ISO 80/20°	60 seconds	Medium contrast; extended dynamic range. Balanced for daylight and electronic flash.
	Polacolor ER Type 59	Sheet 4x5 in.	ISO 80/20°	60 seconds	Medium contrast; extended dynamic range. Balanced for daylight and electronic flash.
	Polacolor ER Type 809	Sheet 8x10 in.	ISO 80/20°	60 seconds	Medium contrast; extended dynamic range. Balanced for daylight and electronic flash.
Instant B&W Prints	Type 331	Integral Pack 4.5x4.2 in.	ISO 400/27°	Self-timing	Medium contrast, automatic development.
	Type612	Pack 3¼x4¼ in.	ISO 20000/44°	30 seconds	Ultra high speed for oscilloscope recording.
	Type 667	Pack 3¼x4¼ in.	ISO 3000/36°	30 seconds	General purpose, high speed film. No coating required.
	Type 52	Sheet 4x5 in.	ISO 400/27°	15 seconds	Wide tonal range, superb detail.
	Type 53	Sheet 4x5 in.	ISO 800/30°	30 seconds	Medium contrast film. No coating required.
	Type 57	Sheet 4x5 in.	ISO 3000/36°	15 seconds	General purpose, high speed film.
	Type 553	Pack 4x5 in.	ISO 800/30°	30 seconds	General purpose, high speed film. No coating required.
	Type 803	Sheet 8x10 in.	ISO 800/30°	30 seconds	Fine grain medium contrast film. No coating required.
Print & Negative	Type 55	Sheet 4x5 in.	ISO 50/18°	20 seconds	Negative requires brief clearing in sodium sulfite solution, washing and drying before use.
	Type 665	Pack 3¼x4¼ in.	ISO 80/20°	30 seconds	Negative requires brief clearing in sodium sulfite solution, washing and drying before use.
Instant Color Overhead Transparency	Colorgraph Type 691	Pack 3¼x4¼ in.	ISO 80/20°	4 minutes	Small format overhead projection use.
	Colorgraph Type 891	Sheet 8x10 in.	ISO 80/20°	4 minutes	For overhead projection use.
Instant B&W Radiographic Transparency	Type TPX	Sheet 8x10 in.	N/A	60 seconds	Medium to high contrast. High speed Orthochromatic sensitivity. No darkroom needed.
Instant 35mm Color Slides	PolaChrome CS	35mm Roll (12 & 36 Exposures)	ISO 40/17°	60 seconds	General purpose, continuous tone color.
	High Contrast PolaChrome	35mm Roll (12 Exposures)	ISO 40/17°	2 minutes	High contrast color for text and graphic imaging.
	PolaBlue BN	35mm Roll (12 Exposures)	ISO 4/7° (Tungsten) ISO 8/10° (Electronic Flash)	4 minutes	White-on-blue negative film for word slides, charts and graphs.
Instant 35mm B&W Slides	PolaPan CT	35mm Roll (12 & 36 Exposures)	ISO 125/22°	60 seconds	General purpose, continuous tone black and white.
	PolaGraph HC	35mm Roll (12 Exposures)	ISO 400/27°	2 minutes	High contrast black and white. Ideal for line copy.

We also have many other films to meet additional applications.

than clearing up the regulatory mechanisms that govern white-blood-cell production. By the early 1960's much work had been done on the rates at which leukocytes develop from their common ancestor. The factors that act on each set of cells, however, were not known. One reason was that until a little more than 20 years ago cells of the human bone marrow could not be grown effectively in laboratory culture. In 1966 that situation was changed by a discovery made independently and at about the same time by Dov H. Pluznik and Leo Sachs at the Weizmann Institute of Science in Israel and by Thomas R. Bradley of the University of Melbourne and Donald Metcalf of the Walter and Eliza Hall Institute of Medical Research in Australia.

What these two teams found was that suspensions containing individual cells from mouse bone marrow could be made to yield colonies of mature white blood cells. Each of the colonies was a clone, that is, it was made up of genetically identical descendants of a single ancestor. The system's operation depended on the presence in the culture vessel of



BLOOD CELLS MATURE in a pattern resembling a family tree. The pluripotent stem cell, precursor of all mature types, is found in the bone marrow. The first step in the maturation of the blood cells is the division into two main lineages: the lymphoid (consisting of the lymphocytes) and the myeloid (consisting of the erythrocyte and the rest of the leukocytes). Thereafter, under the influence of protein signals, each precursor cell develops step by step into a mature cell type.

"feeder" layers to induce and sustain the colonies. At first the feeder layers contained suspensions of various types of cells. Later it was found that the cells were superfluous: it was enough to supply the "conditioned" medium in which the feeder layers had been grown. This was a dramatic step, since it made clear that the conditioned medium contained substances with the capacity to make white blood cells divide and mature; ultimately such hormones came to be known as colony-stimulating factors, or CSF's.

What were the CSF's and where did they come from? The second question was tackled first. Although most of the early work had been carried out with mouse bone marrow, by 1970 the refinement of the colony-culture system made it possible to grow human leukocytes as a matter of routine. Much of that refinement was due to the work of William Robinson of the University of Colorado, Norman N. Iscove of the University of Toronto and Paul A. Chervenick of the University of Pittsburgh School of Medicine, who developed techniques for growing colonies of human myeloid cells (the lineage that includes erythrocytes, monocytes and granulocytes). Their system entailed feeder layers consisting of human leukocytes or medium conditioned by exposure to the same types of cells.

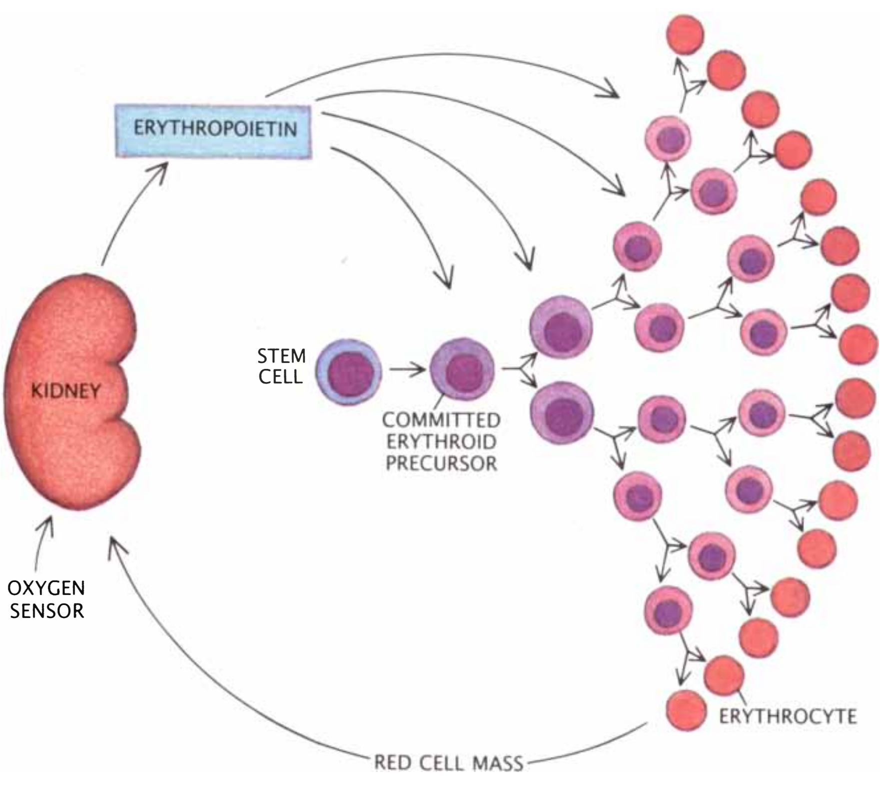
The finding that human leukocytes released CSF's initiated an intense search for the "colony-stimulating cell" in the peripheral blood that was presumed to be the source of the hormones. The first fruit of that search was the identification of the monocyte as the blood leukocyte primarily responsible for the release of the CSF's; that finding was made by Chervenick and Al F. LoBuglio of Ohio State University and independently by Martin J. Cline of the University of California School of Medicine at San Francisco and one of us (Golde). Later we found that the macrophage (a descendant of the monocyte that is present in the tissues rather than in the blood) also releases CSF's. Even more recently it has been found that cells of the monocyte-macrophage lineage make substances that can cause the release of CSF's from other cells; examples of these substances are interleukin-1 (IL-1) and tumor necrosis factor (TNF). Such findings have led to the idea that the macrophage is a sentinel, responding to microbial invasion and sending signals to trigger increased synthesis of white blood cells.

The monocyte and the macrophage are not the only white blood cells that release CSF's. Further work showed that when lymphocytes are activated, they also release a potent CSF. Lymphocytes come in at least two forms, known as *T* and *B* lymphocytes, each with a variety of functions. Some *B* lymphocytes ultimately differentiate into plasma cells, which make large quantities of antibodies. *T* lymphocytes have a wide range of roles, including serving as a kind of "master" cell, precisely orchestrating many aspects of the immune response. It is in this capacity that the *T* lymphocytes make and give off colony-stimulating factors. When they are exposed to specific antigens, they release CSF's that marshal the production of white blood cells.

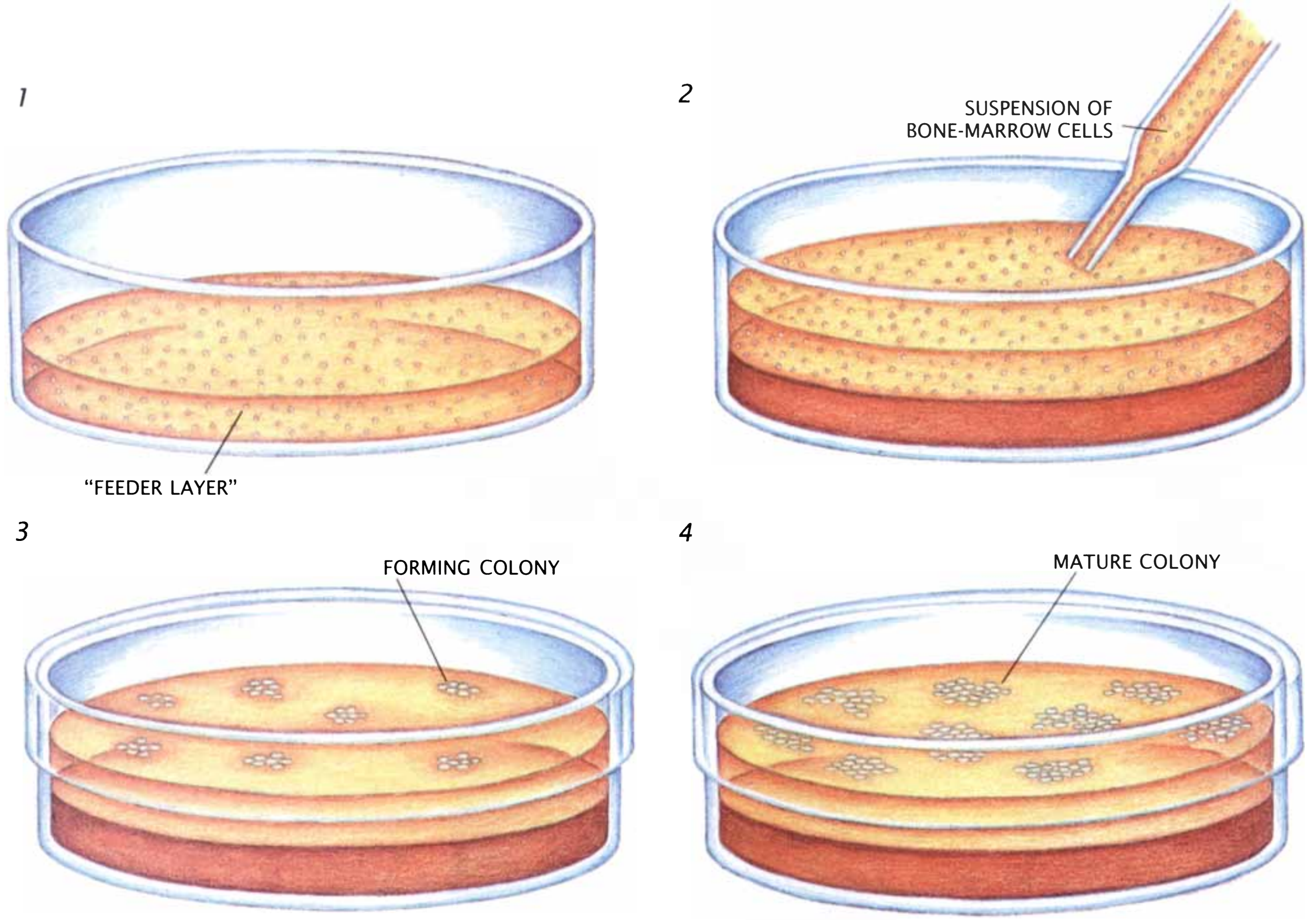
As the 1970's passed, some of the sources of the colony-stimulating factors were becoming clear. Yet the question of what these molecules actually were had not been answered. The first substantial evidence came from the mouse-bone-marrow colony-assay system described above. In that system four different CSF's were identified. The first, identified by E. Richard Stanley of the Walter and Eliza Hall Institute and Metcalf, only induced colonies of the monocyte-macrophage lineage, and so it was named macrophage CSF (M-CSF). Three other CSF's were identified after that. Of those, the one that stimulated granulocyte and monocyte colonies was called GM-CSF, and the one that stimulated only granulocyte colonies, G-CSF. The third, which resulted in colonies containing a mixture of cell types, was called multi-CSF, or interleukin-3 (IL-3).

Having identified a set of CSF's by their activity in culture, investigators wanted very much to be able to obtain them in pure form. One advantage of having the purified protein in hand was that it could lead to cloning the gene for the protein. Once cloned, the gene could be inserted into mammalian or bacterial cells. There the protein could be made in quantities large enough for research purposes and for therapeutic trials. As it happened, macrophage CSF was the first to be obtained in highly purified form (by Metcalf's group), but it was not the first CSF whose gene was cloned.

That honor went to the murine (mouse) form of GM-CSF. Antony Burgess of the Walter and Eliza Hall Institute purified the protein from medi-



ERYTHROPOIETIN is the protein hormone that causes the precursors of red blood cells to proliferate and mature. It is secreted by the kidney in response to the blood's oxygen content. Under the influence of erythropoietin, erythrocyte precursors become smaller, make more hemoglobin (the red substance that binds the oxygen molecule) and lose their nuclei as they become specialized for carrying oxygen.



COLONY ASSAY led to the identification of the colony-stimulating factors beginning in the late 1960's. "Feeder" layers, containing various types of white blood cells in a semisolid medium, were placed in a small laboratory dish (*1*). Bone-marrow cells (including stem cells) were added to form a second layer (*2*). When the dish was incubated, colonies of white blood cells formed in the second layer (*3*). The colonies were counted and the cells identified (*4*). When the contents of the first layer were varied, different types of colonies formed, implying the existence of a range of colony-stimulating factors.

um conditioned by lung tissue from mice that had been injected with endotoxin (a substance found in the cell walls of certain bacteria that triggers a potent immune response). Nicholas M. Gough and Ashley R. Dunn of the Walter and Eliza Hall Institute then employed a partial amino acid sequence of the protein to construct complementary DNA (cDNA) probes. The probes in turn were exploited to pick the GM-CSF gene out of a "library" of mouse DNA sequences. Work on the human hormone lagged, but ultimately our group purified it from medium conditioned by a line of human *T* lymphocytes that had been transformed by HTLV-II (a human virus that can cause leukemia). Gordon G. Wong and Steven C. Clark of Genetics Institute, collaborating with us, then exploited a novel laboratory system to retrieve the GM-CSF gene from a line of monkey cells.

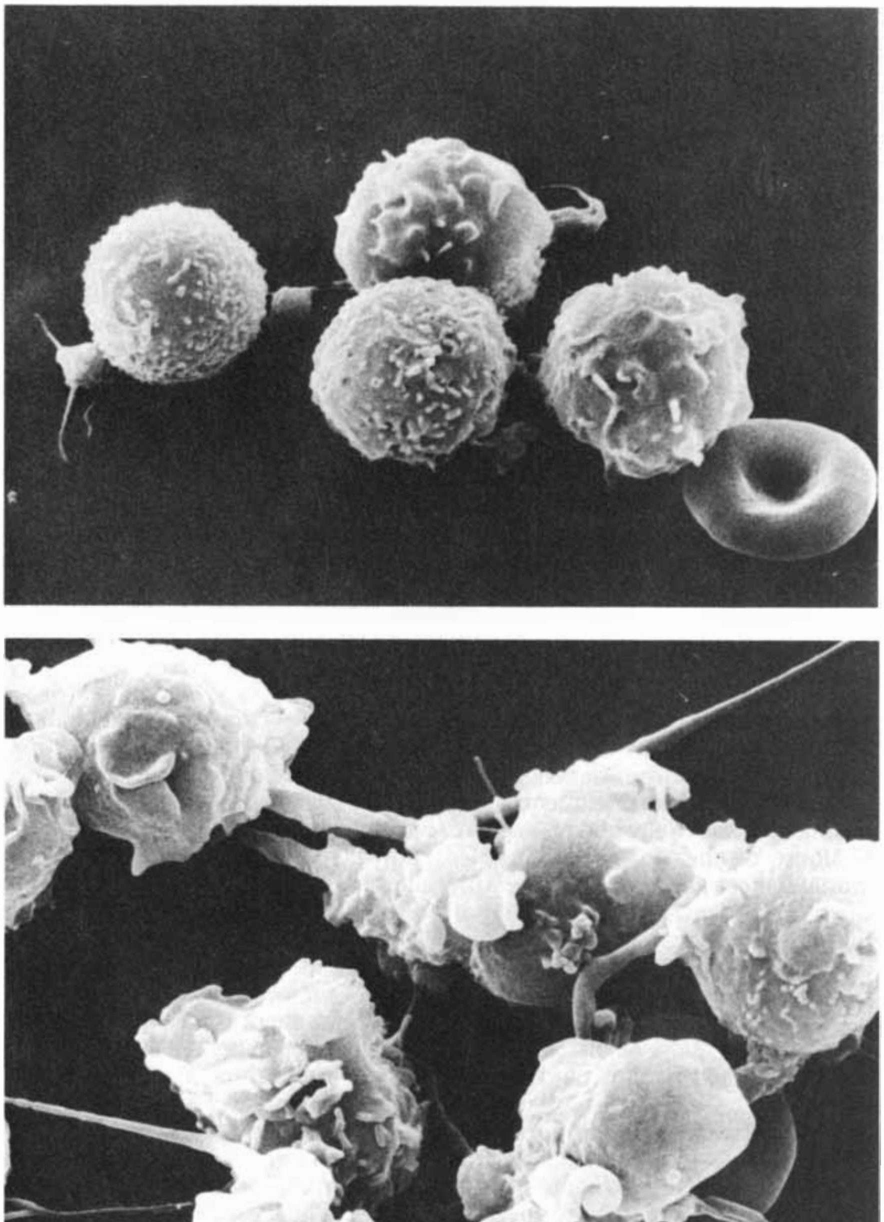
Thus GM-CSF was the first human hematopoietic hormone whose gene was molecularly cloned; it was also the first to be made by recombinant-DNA methods. Yet others were not far behind. Yu-Chang Yang of Genetics Institute and Clark cloned sequences encoding IL-3; Karl Welte and Malcolm A. S. Moore of the Sloan-Kettering Memorial Cancer Research Institute with Lawrence M. Souza of Amgen (and at the same time Shigekazu Nagata of the Univerity of Tokyo and his colleagues) did the same for G-CSF. While that work was going on, Ernest S. Kawasaki and David F. Mark of the Cetus Corporation cloned part of the DNA for the macrophage CSF. Once the genes had been cloned and the hormones had been made by recombinant methods, the biosynthetic hormones could be tested for their capacity to grow white blood cells. Each of the recombinant hormones has shown the specific activity expected of it on the basis of observations from the colony-assay system.

Although cloning the genes for the human CSF's was an important step, it did not provide direct information about the function in the human body either of the genes or of the corresponding hormones. Using the cloned DNA as probes, however, the human genes that encode the hormones have now been located on specific human chromosomes. Each gene is apparently present in a single copy. (Genes may be present in multiple copies; the purpose of such amplification may be to enable the cell to make more of the gene's product.) The genes for GM-CSF and IL-3 are quite close to each other on chromosome 5, which suggests they may have had a common ancestor. Interestingly, the genes for M-CSF (a protein unrelated to the other two) and its receptor are also on chromosome 5. The G-CSF gene, however, is on chromosome 17.

When are these genes turned on or off in hematopoiesis? Work on that question is proceeding rapidly and is beginning to yield an outline of how the process as a whole is regulated. Some of the CSF genes are not expressed under ordinary conditions; they remain quiescent until the cell receives a specific signal. GM-CSF, for example, is released by lymphocytes when they are activated by specific antigens (foreign proteins). It is also made and released by fibroblasts (connective-tissue cells that have a role in wound healing) and endothelial cells (cells that line the blood vessels) when these cells are exposed to substances from monocytes and macrophages, such as tumor necrosis factor and IL-1. Macrophages can also synthesize and release GM-CSF when stimulated.

Similarly, the G-CSF gene can be turned on in a number of cell types in the presence of activating signals. It is made by cells of the monocyte-macrophage lineage when they are exposed to endotoxin and by fibroblasts in response to substances released by monocytes and macrophages. M-CSF is also made by many cells, including the macrophage itself, in response to endotoxin, GM-CSF or IL-3. The IL-3 gene is turned on in activated lymphocytes. Thus, in response to a network of intercellular signals, genes for the CSF's are aroused from their quiescent state and begin to give rise to their products.

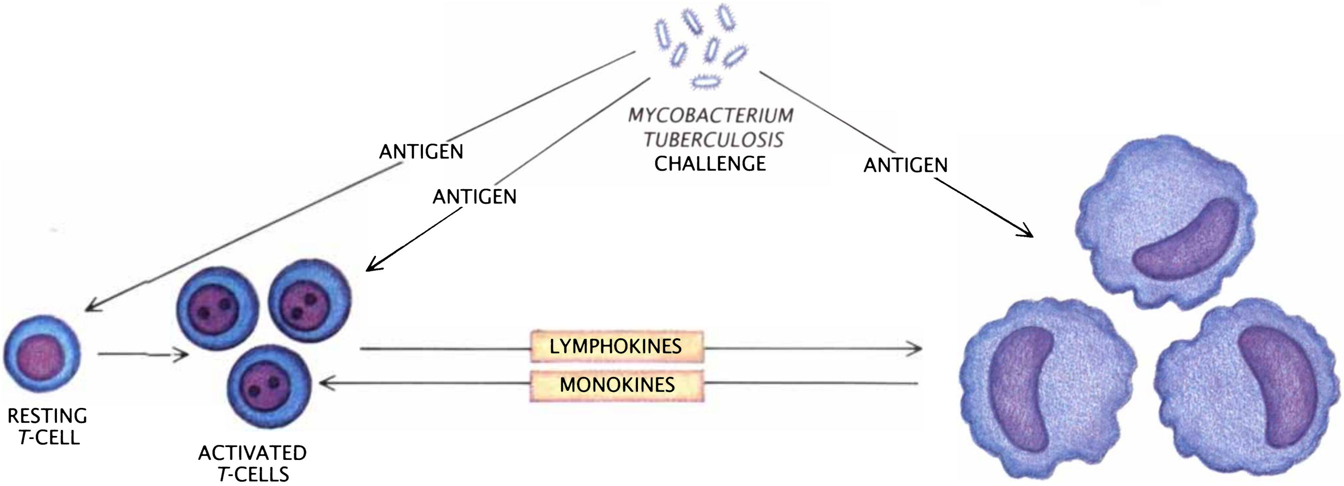
This network of interactions may appear complex. It is. What is more, the network is not fully understood. What is already known, however, provides the basis for a preliminary outline of how the immune system responds to a pathogen in terms of CSF's. In that picture the *T* lymphocyte and the macrophage have a central role. In response to a particular antigen a subset of *T* cells are activated, releasing GM-CSF and IL-3. At the same time macrophages responding to their specific stimuli synthesize GM-CSF, G-CSF and M-CSF. The release of IL-1 and TNF from the macrophages triggers production of GM-CSF, G-CSF and M-CSF in local populations of endothelial and mesenchymal cells (a type found in muscle, bone and connective tissue). As the result of the release of this array of CSF's, the subpopulations of leukocytes that are required in the immune response begin to proliferate and mature.

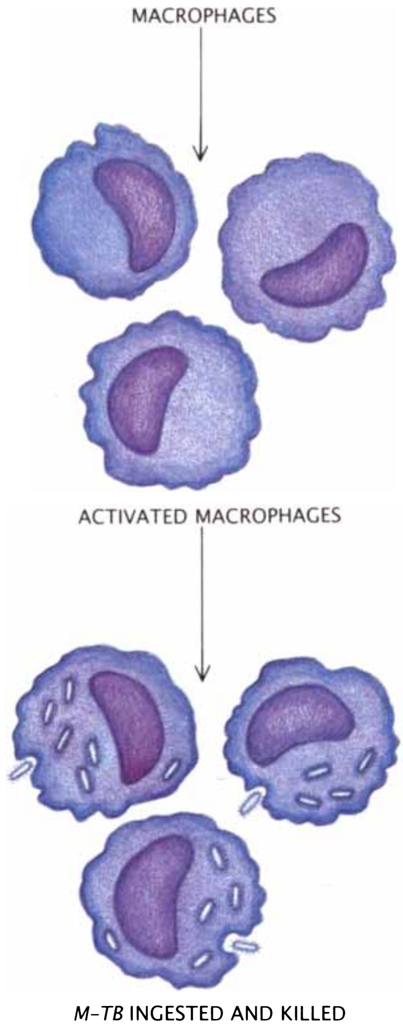


GRANULOCYTES ARE ACTIVATED by granulocyte-macrophage colony-stimulating factor (GM-CSF). In the upper panel the granulocytes are in the resting phase (the disklike object at the lower right is an erythrocyte). In the lower panel the granulocytes have been stimulated by GM-CSF. The long projections—called filopodia—enable the granulocyte to adhere to surfaces and to make contact with pathogens. In these micrographs made by Gerald E. Garner and Leonard Hancock, Jr., of the Baxter Healthcare Corporation the cells are enlarged some 3,000 diameters.

Yet the role of the CSF's in immunity does not end there. Although these factors were originally identified by their capacity to stimulate the growth and maturation of stem cells and precursor cells, it seems they also have profound effects on mature white blood cells. Neutrophils activated by invading microorganisms synthesize and release highly toxic oxygen derivatives that can kill the invaders. Richard H. Weisbart of the University of California at Los Angeles (in collaboration with our group) showed that GM-CSF primes the neutrophil, thereby making the response more potent. GM-CSF did not trigger the oxidative burst directly, but it did lead to a markedly increased release of oxidants when the neutrophils were exposed to known triggering agents such as bacterial proteins.

The neutrophil's mission is to "search and destroy": if its weapons are to have an effect on the invader, the enemy must first be found. GM-CSF and G-CSF both increase the directed movement of neutrophils toward triggering agents. The CSF's also augment a neutrophil's capacity to ingest microorganisms. How this happens is not fully understood. Studies by Weisbart and his co-workers, however, have shown that GM-CSF regulates the number and affinity of cell-surface receptors on the neutrophil that recognize bacterial products. Thus as the





MACROPHAGES ARE ACTIVATED in a process that entails interacting with *T* lymphocytes. When *Mycobacterium tuberculosis* (the tuberculosis agent) enters the body, its proteins activate a specific subset of *T* cells. Those *T* cells release substances called lymphokines that prime macrophages to ingest and kill bacteria. Among the lymphokines are GM-CSF, interleukin 3 (IL-3) and interferon. Macrophages in turn release monokines, including interleukin 1 (IL-1), which stimulate the lymphocytes.

granulocytes survey the environment within the body, CSF's may render them better able to detect pathogens.

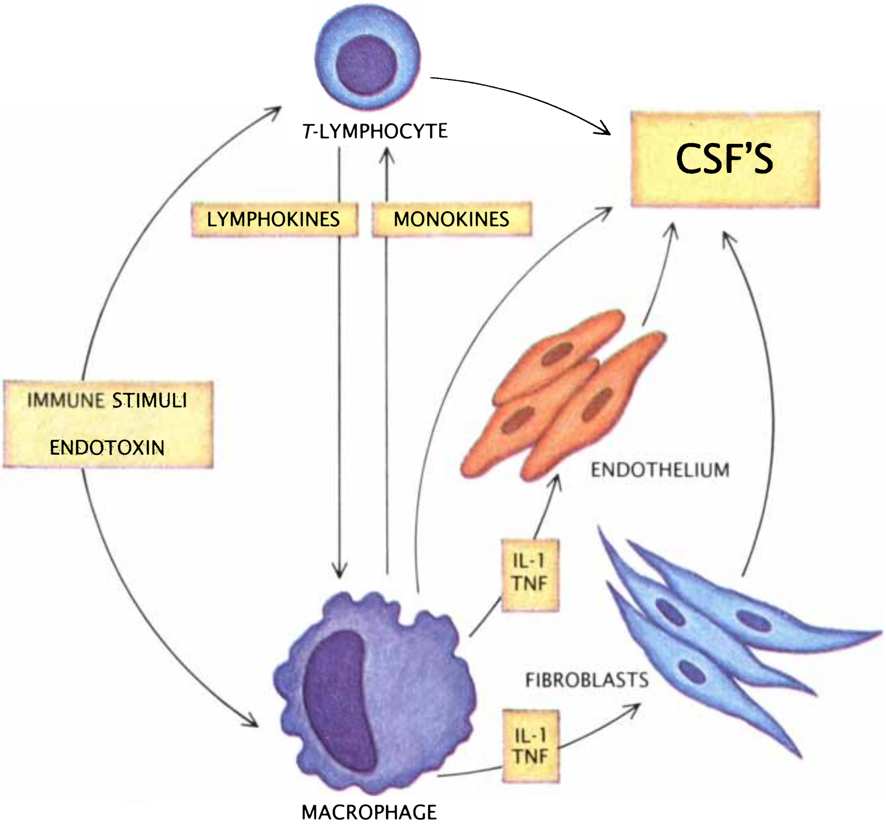
The capacity of the neutrophil for searching and destroying is central to the "host defense" mounted by the immune system. In the end the host defense rests on the ability of mature effector cells—such as neutrophils and macrophages—to kill invading organisms. When the bone marrow cannot make enough neutrophils or monocytes, host defense is impaired. The response of the effector cells also depends on the collaboration of *T* lymphocytes. When there are too few *T* cells or the *T* cells function poorly (as in AIDS or advanced tuberculosis), the macrophage is incapable of containing the intracellular pathogen. Such deficiencies are clearly catastrophic from the clinical point of view; the CSF's, which can help to correct them, have great therapeutic potential.

Some of that potential is already being realized. Studies of several CSF's made by recombinant methods have been carried out in laboratory animals. Such trials indicate that GM-CSF and G-CSF are relatively nontoxic and are effective in stimulating the proliferation of host-defense cells. Human G-CSF causes increased neutrophil production in both mice and monkeys. In monkeys the increase was to a level 50 times the normal one, and there were few side effects. Human GM-CSF is not effective in mice, but in monkeys it yields increases in neutrophils, eosinophils and monocytes. The administration of IL-3 alone to monkeys results in only a modest increase in circulating white cells, but in combination with GM-CSF it has a potent stimulating effect on bone marrow and leukocyte production.

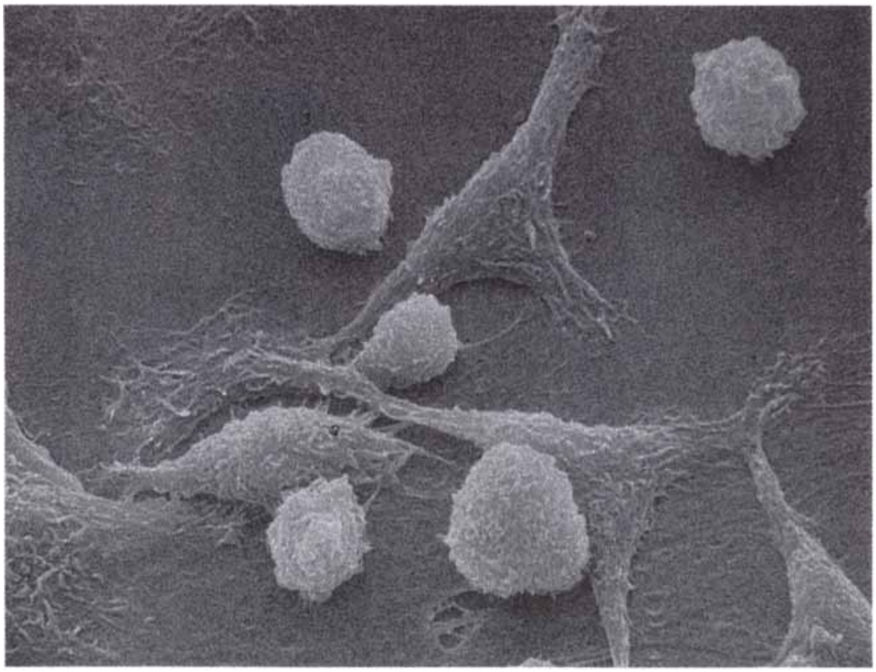
There have so far been only a few trials in human beings of recombinant CSF's, but those that have been carried out tend to confirm the exciting prospects offered by the animal work. Ronald Mitsuyasu, working with us and with Jerome Groopman of New England Deaconess Hospital, conducted the first human phase I and II (safety and efficacy) trials of GM-CSF in patients suffering from AIDS, with its characteristic decrease in white-blood-cell count. This initial study showed that GM-CSF was well tolerated and that it increased levels of neutrophils, monocytes and eosinophils circulating in the blood. GM-CSF is also being investigated as a form of therapy in cancer, preleukemic conditions and aplastic anemia (anemia due to malfunction of the bone marrow).

Studies of the other CSF's are not as advanced as those of GM-CSF, but they are gaining momentum. Janice Gabrilove and her colleagues at Sloan-Kettering carried out a study of G-CSF in bladder-cancer patients being treated by combination chemotherapy, which often leads to suppression of bone-marrow function and reduced white-cell counts. The hormone had few side effects, and it clearly stimulated production of neutrophils in the bone marrow and reduced bone-marrow suppression. M-CSF also appears to be relatively nontoxic in monkeys and human beings, but it has not undergone rigorous clinical trials. IL-3 has not been assessed in clinical trials.

The clinical results available so far suggest that the CSF's will one day prove their worth in treating AIDS patients and in overcoming the leukopenia (low white-cell counts) associated with cancer chemotherapy. If one accepts the capacity of these factors to increase host-defense cell numbers and activity, then it is clear that the CSF's will be of value in other areas as well. For example, bone-marrow transplantation is a difficult procedure that



COLONY-STIMULATING FACTORS ARE RELEASED by an array of cells in response to the presence of pathogens. Immune stimuli (antigens) cause *T* cells to release CSF's; endotoxins (molecules found in the cell wall of certain bacteria) cause macrophages to do the same. In addition, the macrophage releases substances that evoke release of CSF's from two other cell types: endothelial cells (which line the walls of blood vessels) and fibroblasts (connective-tissue cells that have a role in wound healing).



MACROPHAGES ATTACK TUMOR CELLS in a micrograph made by Quan. Both the macrophages (which are flat and irregularly shaped) and the rounded tumor cells come from a mouse. Rather than ingesting tumor cells (as they do bacteria), macrophages kill the tumor cells by releasing toxins such as tumor necrosis factor and by methods (which are still poorly understood) that entail surface contact.

can require weeks of in-hospital care for the recipient and a shorter period for the donor. Administration of CSF's might allow the recipient to recover in as little as a week. Treating donors with CSF's may make it possible to remove much smaller quantities of marrow, allowing donation to be done without hospitalization.

Most of the applications we have discussed entail strengthening a weakened immune system—as in AIDS, cancer chemotherapy or aplastic anemia. More radical applications of the CSF's, however, may come not in such situations but in those where the aim is to bolster a normal immune system. In the future a multitude of infections may be treated by increasing the number and potency of host-defense cells. Part of the therapy for parasite-caused diseases may be to exploit CSF's in regulating the number of eosinophils. In addition, some experimental cancer therapies involve marking tumor cells with antibodies so that they can be destroyed by monocytes and neutrophils. Such strategies may be enhanced by CSF's that increase the level and activity of these effector cells.

The uniqueness of CSF's in medicine lies in their capacity to make the patient a more formidable defensive entity. In the past the only means available to medicine for improving the host-defense mechanism were indirect (such as improved nutrition) or were specific to a single disease (such as immunization). A variety of means have been developed to subvert invading organisms, antibiotics being the most dramatic examples. Future research will undoubtedly offer new and more potent ways of interfering with the metabolism of the invader. Thanks to the CSF's, however, clinicians will also have a new strategy based on giving the patient a stronger defense against microbes and even against cancer.

FURTHER READING

HEMATOPOIETIC STEM CELLS. Edited by David W. Golde and Takaku Fumimaro. Marcel Dekker, Inc., 1985.

THE GRANULOCYTE-MACROPHAGE COLONY-STIMULATING FACTORS. Donald Metcalf in *Science*, Vol. 229, No. 4708, pages 16-22; July 5, 1985.

THE HUMAN HEMATOPOIETIC COLONY-STIMULATING FACTORS. Steven C. Clark and Robert Kamen in *Science*, Vol. 236, No. 4805, pages 1229-1237; June 5, 1987.

MYELOID GROWTH FACTORS. D. W. Golde and J. C. Gasson in *Inflammation: Basic Principles and Clinical Correlates*, edited by J. I. Gallin, I. M. Goldstein and R. Synderman. Raven Press, 1988.

The Supercontinent Cycle

Several times in earth history the continents have joined to form one body, which later broke apart. The process seems to be cyclic; it may shape geology and climate and thereby influence biological evolution

by R. Damian Nance, Thomas R. Worsley and Judith B. Moody

Is plate tectonics a random process or is it orderly? According to the theory of plate tectonics, the earth's rigid outer layer, called the lithosphere, is a mosaic of slablike plates that move with respect to one another at speeds averaging a few centimeters per year. The plates float on a hot, plastic layer of the earth's mantle called the asthenosphere. Most of the plate movements are driven by a process known as sea-floor spreading, in which molten material from the asthenosphere rises through the lithosphere at high ridges on the ocean floor, where it cools to become the crust that makes up the ocean bottom. Newly created oceanic crust moves steadily away from the mid-ocean ridges toward the continents. If the sea floor and the adjacent continent are on the same lithospheric plate, the continent is carried along by the conveyor belt of oceanic crust. Alternatively, the oceanic crust may sink under the continent to rejoin the mantle, in a process known as subduction.

The continents are generally viewed as passive objects that are ferried about by sea-floor spreading. They are not entirely unchanged by the processes of plate tectonics, however. Separate blocks of continental crust can collide and merge, forming new, larger continents. Conversely, continents can be torn apart by deep rifts that eventually become the centers of new ocean basins. Indeed, there is evidence that several times in the history of the earth the continents have undergone these processes on a grand scale: several times most or all of the continents have gathered to form a single supercontinent, which has later split into many smaller continents only to rejoin and form a supercontinent again.

What governs the formation and destruction of supercontinents? Do they appear and disappear simply by chance, because of the random shifting of continental plates? Various regularities in the geologic record have led the three of us to believe that a much more orderly, even cyclic, process must be at work. Drawing on the ideas of Don L. Anderson of the California Institute of Technology and on the prescient observations of the Dutch geologist J. Umgrove (set out in his 1947 book *The Pulse of the Earth*), we have devised a theoretical framework that describes what may be the underlying mechanisms of such a "supercontinent cycle."

In our theory the dominant force comes from heat. It is generally understood that tectonic plates are driven by convective motions in the underlying mantle, which are powered by heat from the decay of radioactive elements. The radioactive decay (and the resulting production of heat) is a continuous process whose rate has declined smoothly with time, and so the production of heat cannot in itself account for the episodicity inherent in an alternation between continental assembly and continental breakup.

The key phenomenon, we think, is not the production of heat but rather its conduction and loss through the earth's crust. Continental crust is only half as efficient as oceanic crust at conducting heat. Consequently, as Anderson has pointed out, if a stationary supercontinent covers some part of the earth's surface, heat from the mantle should accumulate under the supercontinent, causing it to dome upward and eventually break apart. As fragments of the supercontinent disperse, heat can be transferred through the new ocean basins created between them. After a certain amount of heat has escaped, the continental fragments may be driven back together.

In other words, we think the surface of the earth is like a coffee percolator. As in a coffee percolator, the input of heat is essentially continuous. Because of poor conduction through the continents, however, the heat is released in relatively sudden bursts.

This theoretical framework and its corollaries make it possible to tie together a number of observations in widely disparate fields. They make it possible, for example, to understand the timing of the extreme changes in sea level that have taken place in the past 570 million years. The framework also helps to explain and link many other events of the past 2,500 million years, such as periods of intense mountain building, episodes of glaciation and changes in the nature of life on the earth. The supercontinent cycle, in our view, is a major driving process that has provided the impetus for many of the most important developments in the earth's history.

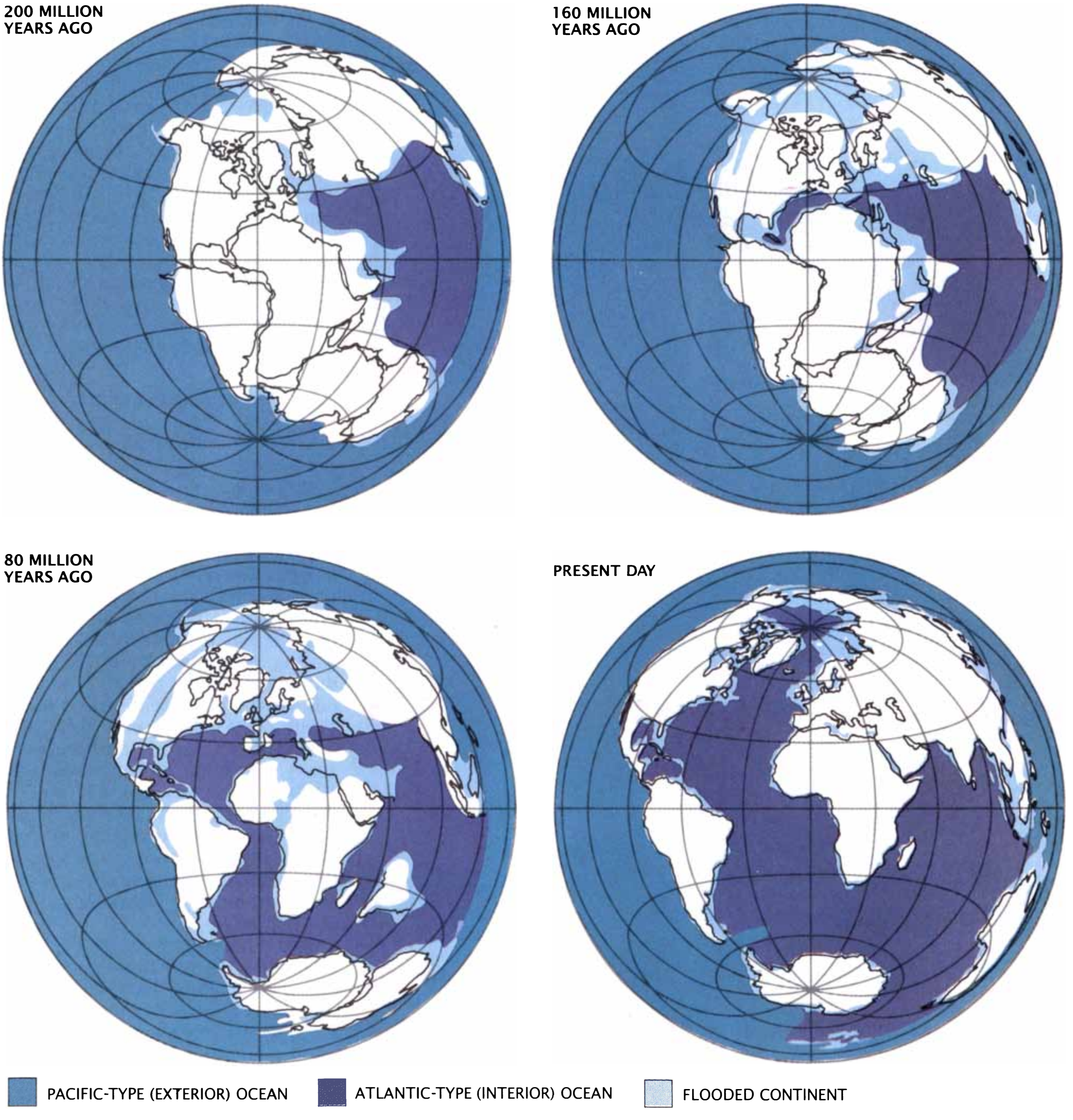
The Opening of Oceans

Our model builds on an earlier description of episodic plate motions known as the Wilson cycle. Named for J. Tuzo Wilson of the Ontario Science Center, the Wilson cycle is the process by which continents rift to form ocean basins and the ocean basins later close to reassemble the continents. In the first stage of the Wilson cycle volcanic

R. DAMIAN NANCE, THOMAS R. WORSLEY and JUDITH B. MOODY have combined their respective specialties of tectonics, oceanography and geochemistry in a particularly close partnership. During the week Worsley stays with Nance's family in Athens, Ohio, where they both teach at Ohio University, and on weekends Worsley travels to Columbus, Ohio, where Moody is president of J. B. Moody and Associates; Moody and Worsley have been married for nine years. Nance got his Ph.D. in 1978 from the University of Cambridge, and he taught at St. Francis Xavier University in Nova Scotia before going to Ohio in 1980. Worsley's Ph.D., granted in 1970, is from the University of Illinois; he went to Ohio in 1977 after teaching at the University of Washington. Moody has a Ph.D. (1974) from McGill University. From 1981 until this year she worked at the Battelle Memorial Institute.

"hot spots" form in a continent's interior; the hot spots are then connected by rift valleys, along which the continent eventually splits. When the continent fragments, the rift valleys grow to become a new ocean as hot mantle material wells up through the rifts to form the sea floor. The continental fragments move apart, sliding away from these elevated "spreading centers" as mantle material wells up.

As the material making up the ocean floor ages, it cools, becomes denser and subsides, increasing the depth of the ocean. Eventually, about 200 million years after the first rift formed, the oldest part of the new ocean floor (the part directly adjacent to the continental fragments) becomes so dense that it sinks under the continental crust: it is subducted. The processes of subduction then close the ocean, bringing the continents back together. Eventually the continents collide and



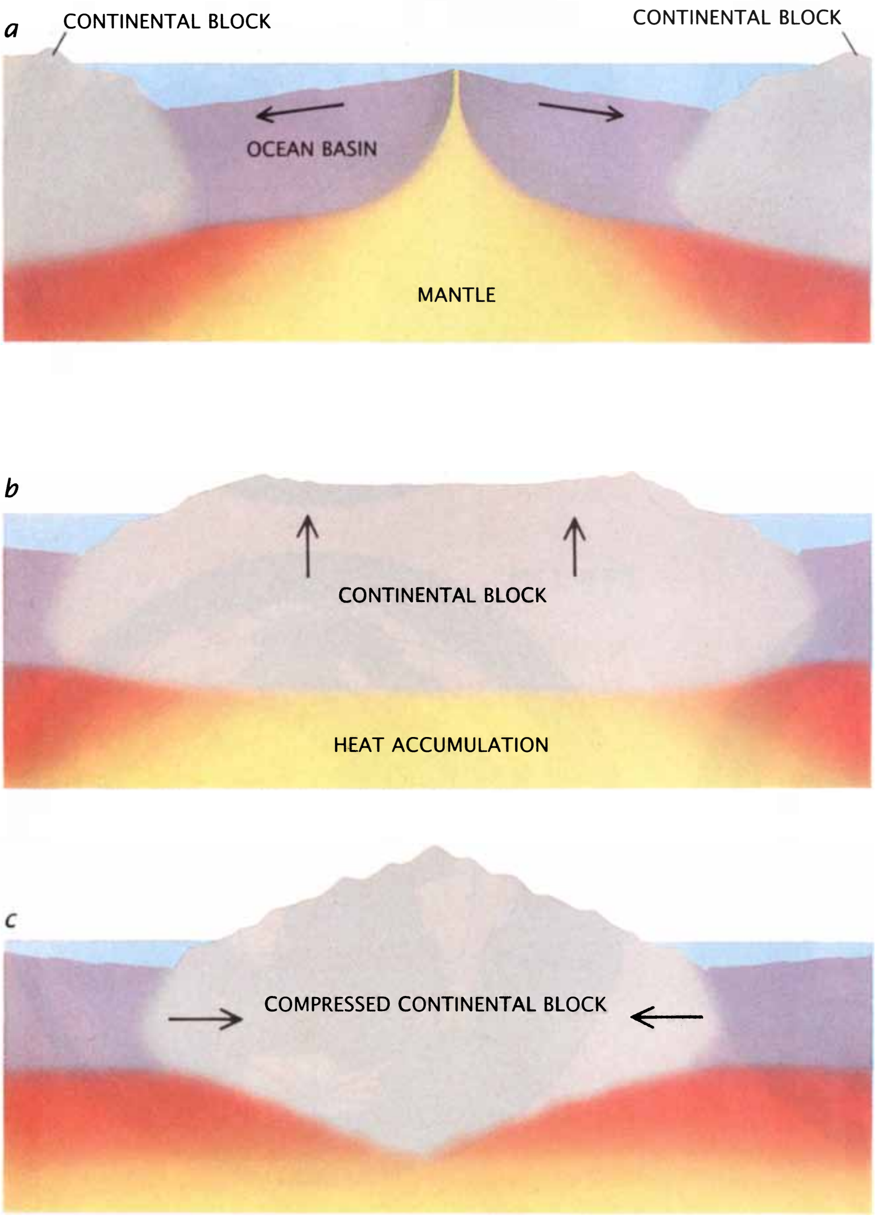
BREAKUP of Pangaea, a supercontinent that formed some 300 million years ago, has dominated later geologic history. About 200 million years ago heat accumulating under the supercontinent broke through in rifts that eventually became oceans. The growth of these shallow oceans at the expense of the older and deeper superocean raised the sea level, partially drowning the continents. Sea level rose to a maximum about 80 million years ago; then it fell, as the new oceans became older and deeper, and the world's present geography was established. According to the authors' hypothesis, Pangaea was only the most recent of a series of supercontinents that have broken up and reassembled during the past 2,600 million years; this supercontinent cycle has shaped the geology and climate of the earth and provided a force for biological evolution. These maps are based on work by A. G. Smith of the University of Cambridge and J. C. Briden of the University of Leeds.

rejoin, and the compressive forces of collision create mountain belts.

When viewed in terms of Wilson cycles, there is a striking contrast between the evolution of the continental margins surrounding the North Atlantic and the margins of the Pacific. The margins of the North Atlantic have undergone a series of Wilson cycles during the past billion years; the regions bordering the Pacific have apparently undergone none. In other words, oceans have repeatedly opened and closed in the vicinity of the present-day North Atlantic, while a single ocean has been maintained continuously in the vicinity of the Pacific.

In our model, then, the Pacific is the descendant of the oceanic hemisphere that has surrounded each incarnation of the supercontinent; each of the Wilson cycles that took place in what is now the North Atlantic region occurred as part of the breakup and reassembly of a supercontinent. The Atlantic should therefore be expected to close again, once more reuniting the continents in a supercontinent surrounded by a single superocean.

At present the sea-floor crust of the Pacific is being subducted under all the continents that surround it, whereas the floor of the Atlantic generally butts up against surrounding continental blocks. In our framework this means that the continents are still in the process of dispersing after the breakup about 200 million years ago of the most recent supercontinent, which Alfred Wegener, the father of the theory of continental drift, christened Pangaea, or "all earth" [see "The Breakup of Pangaea," by Robert S. Dietz and John C. Holden; SCIENTIFIC AMERICAN, October, 1970]. The continents are now approaching their maximum dispersal. Soon (on a geologic time scale) the crust of the Atlantic will become old and dense enough to sink under the surrounding continents, beginning the process that will close the Atlantic ocean basin.



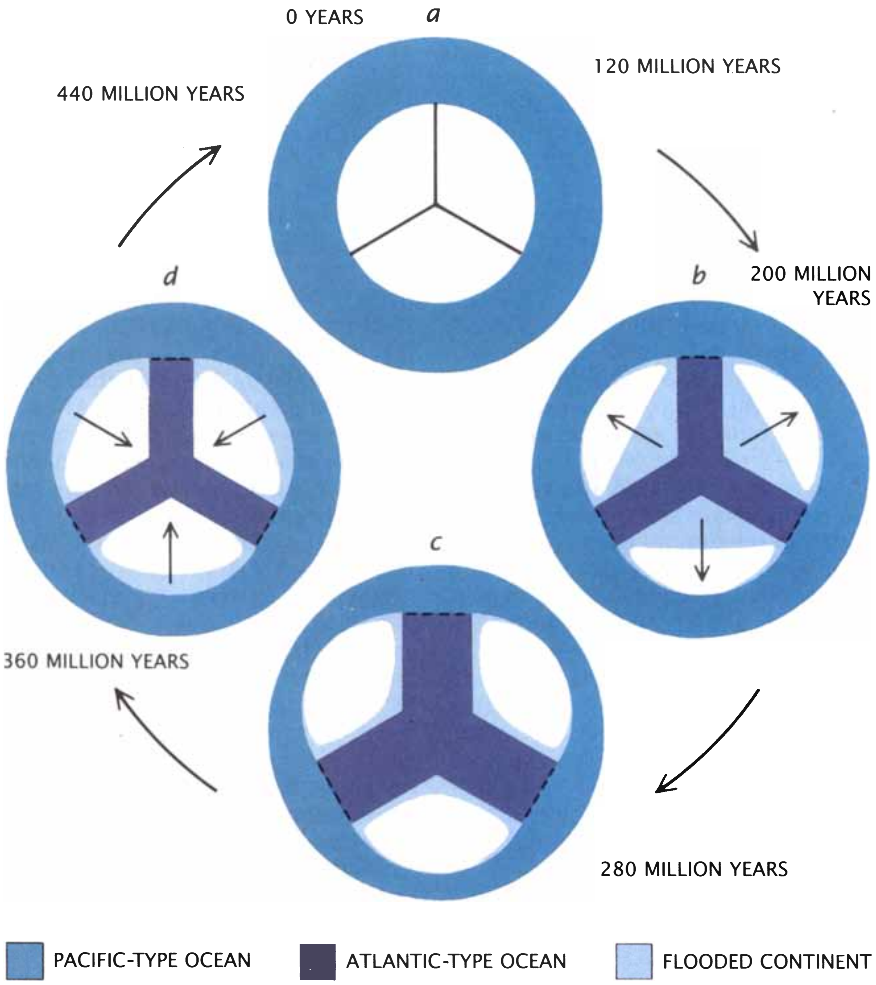
SEA LEVEL with respect to the continents is controlled by several tectonic factors. One is the age of the sea floor, which is created by the upwelling of hot material from the earth's mantle at mid-ocean "spreading centers" (*a*). As the sea floor spreads it cools, becomes denser and sinks; older ocean is therefore deeper, and so sea level becomes lower when the average age of the world's oceans increases. The accumulation of heat under stationary continental crust (*b*) alters sea level by buoying the continent upward. Sea level is also affected by compression (*c*) or extension (not shown) of continents. When continents are compressed, the total area of the world ocean increases while the volume of water remains constant: sea level is lowered.

Surprising Regularity

A second underpinning of our supercontinent-cycle hypothesis is the timing of various episodes of mountain building and episodes of rifting. The ages of mountain ranges that could have been produced by the compressive forces that accompany continental collisions reveal a surprising regularity. This kind of mountain building was particularly intense, occurring in several parts of the world, during six distinct periods. The periods were broadly centered on dates about 2,600 million years ago, 2,100 million years ago, a time between 1,800 and 1,600 million years ago, 1,100 million years ago, 650 million years ago and 250 million years ago. The timing shows a certain periodicity: the interval between any two of these periods of intense compressive mountain building was about 400 to 500 million years.

What is more, about 100 million years after each of these periods of mountain building there appears to have been a period of rifting. Large numbers of mantle-derived rocks—rocks that may have been produced when magma welled up into cracks created by rifting—date from times broadly centered on 2,500 million years ago, 2,000 million years ago, a time between 1,700 and 1,500 million years ago, 1,000 million years ago and 600 million years ago. The mountain building of 250 million years ago, of course, was followed by the rifting and eventual breakup of Pangaea.

These regularities indicate to us that supercontinents are created in a cyclic process, in which one complete cycle takes about 500 million years. By examining these geologic records and others, and by taking into account such factors as the rate at which sea-floor spreading takes place in present-day oceans, we have calculated a more precise timetable for the supercontinent cycle. After the fragments of a supercontinent first separate—probably some 40 million years or so after rifting begins—we estimate that it should take about 160 million years for the fragments to reach their greatest dispersal and for subduction to begin in the new oceans. After the continents begin to move back together, another 160 million years or so should elapse before they re-form a supercontinent. The supercontinent should survive for about 80 million years before enough heat accumulates under it to cause rifts to form. Forty million years later that rifting will lead to another breakup, 440 million years after the previous one.



SUPERCONTINENT CYCLE is depicted schematically. A supercontinent (*a*) can survive for about 80 million years before the accumulation of heat causes rifts to form, and for another 40 million years or so before it is torn apart into separate continents. The continents drift apart (*b*) until they reach their greatest degree of dispersal, about 160 million years later (*c*). Then they move back together (*d*), eventually re-forming the supercontinent (*a*). The entire cycle takes about 440 million years.

Effects on Sea Level

How can one test whether the supercontinent cycle proceeds as we have described? The cycle is likely to have striking effects on sea level, for which there are clear geologic records covering the past 570 million years. Assuming a constant amount of water in the world oceans, sea level (in relation to continental mass) is largely determined by two factors: the total volume of the world's ocean basins, which depends in part on the average depth of the sea floor, and the relative elevation of continents. The supercontinent cycle would involve the creation and destruction of ocean basins and the thermal uplifting of continents, and so it should have a profound influence on both factors.

As the material making up the ocean floor moves away from mid-ocean ridges during sea-floor spreading, it cools and subsides, its depth increasing as the square root of its age. Wolfgang H. Berger and Edward L. Winterer of the Scripps Institution of Oceanography have calculated how the average age of the world ocean floor should change during the breakup of a supercontinent. Before the breakup the average age of the ocean should remain constant, because in the superocean surrounding the supercontinent new sea floor is created at about the same rate as old sea floor is destroyed by subduction under the landmass. During the breakup, the subducting, "Pacific type" superocean will be replaced by an increasing proportion of nonsubducting, "Atlantic type" oceans. Later, when the supercontinent begins to reunite, these "interior" oceans will be destroyed by subduction and replaced by Pacific-type ocean again. These processes affect the average age of the world ocean floor.

Immediately after the breakup of a supercontinent the world ocean floor should, on the average, become progressively younger and shallower as young, Atlantic-type oceans begin to replace the older, Pacific-type ocean. When the Atlantic-type oceans reach the same average age as the Pacific-type ocean, the trend should reverse: the growth of increasingly old Atlantic-type oceans should cause the world ocean floor to age and deepen. The maximum average depth should occur when Atlantic-type oceans reach their greatest average age, just before they begin to be subducted. Then, as the oldest areas of the Atlantic-type oceans are subducted and the oceans close, the world ocean floor should become younger and shallower again.

Calculations of sea level based on these parameters alone suggest that a supercontinent's continental shelves should be flooded, because the ocean basin surrounding a supercontinent is younger and shallower than, for example, the floor of today's world ocean. A second factor, however, must be added to the analysis: the degree to which a supercontinent would be uplifted by the heat that would accumulate under it. If the supercontinent is lifted high enough, sea level in relation to the continental mass could still be low even if the sea floor is comparatively shallow.

One way to estimate how much a supercontinent might be uplifted thermally is to consider present-day

Africa. Africa has remained essentially stationary for at least the past 200 million years, during which time a good deal of heat from the mantle has accumulated under it. (Some of that heat is being released in the rift valleys now forming in various areas of the continent.) By comparing the height relative to sea level of Africa's shelf break (the true edge of the continent) with the height of the shelf breaks of other continents, we can estimate that thermal uplifting has buoyed Africa by about 400 meters. As a lower limit, then, one can expect that a supercontinent would be thermally uplifted by at least 400 meters.

Other factors should also cause the supercontinent to be emergent (elevated in relation to sea level). For example, the collisions that take place during the assembly of the supercontinent should compress and thicken continental crust, decreasing the earth's total land area. This would add to the total area of the ocean basins and thereby lower sea level. Conversely, the stretching and extension of crust that accompanies the breakup of a supercontinent should lower the total area of the world ocean basin, thereby raising sea level.

By adding together all these components, it is possible to determine how sea level should change in the course of every phase of the supercontinent cycle. As we have noted, during the existence of a supercontinent sea level should be relatively low. As the supercontinent breaks up, sea level should rise, both because the continental fragments will stretch and subside thermally and because the breakup will replace old, Pacific-type ocean with young, Atlantic-type ocean. Sea level should continue to rise for about 80 million years, as the younger oceans make up a greater fraction of the world ocean. Then, as the Atlantic-type oceans age and expand, sea level should decline for another 80 million years or so, until the Atlantic-type oceans begin to be subducted.

When the continents begin to come together, sea level should rise, as older Atlantic-type crust is subducted. That rise in sea level should continue for another 80 million years, until the supercontinent begins to be reassembled. Then, as continents collide and the growing supercontinent is uplifted thermally, sea level should decline for about 80 million years. Once the supercontinent has been formed, sea level should remain static for another 120 million years, until the supercontinent breaks up again.

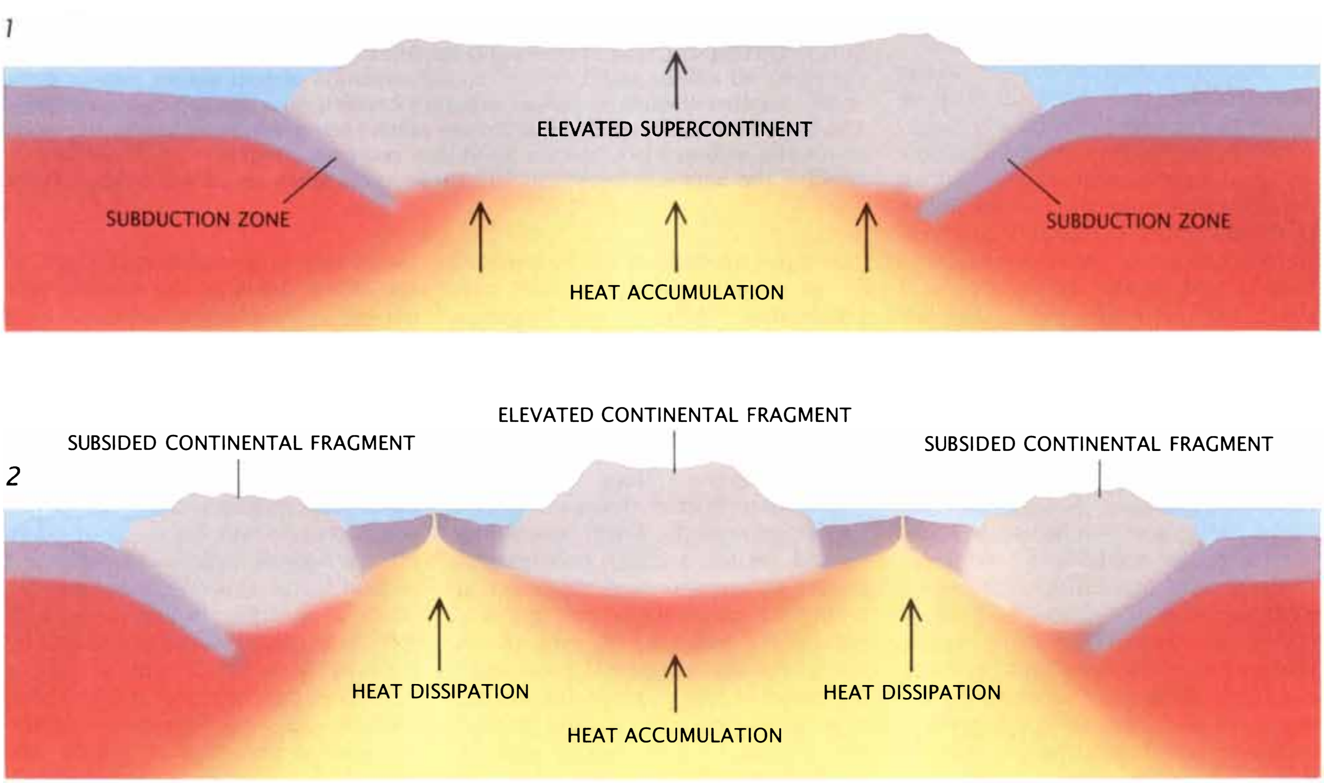
These predictions of changes in sea level match the geologic record of the past 570 million years, which is as far back as sea level can be determined with any reliability. In particular, the timing and the relative magnitudes of sea-level changes predicted by the model match the preserved record. The absolute values, of course, always vary from continent to continent; the model gives only global averages.

Testing the Model

Evidence for the supercontinent cycle can also be found by examining the isotopes of sulfur and carbon found in certain marine sediments. (Isotopes are atoms of the same element that have different atomic weights.)

During the early stages of breakup, a supercontinent is likely to include a number of marine rifts that, like the modern Red Sea, are weakly connected to the world ocean. These rifts can undergo a continuous process of partial evaporation, in which certain elements, such as sulfur, precipitate out of the seawater to form minerals. When seawater containing sulfur is evaporated, heavy sulfur (which has an atomic weight of 34) precipitates out more readily than light sulfur (which has an atomic weight of 32).

If an evaporating marine rift contin-



HEAT controls the elevation of a supercontinent and its fragments. A supercontinent (*1*), surrounded by subduction zones (where sea-floor material sinks under the continent), remains stationary in relation to the underlying mantle. Heat accumulates under it, buoying it upward. After the supercontinent breaks up (*2*), fragments subside as they drift away. A fragment that stays in place (*center*) remains elevated. Present-day Africa is one example of such a stationary fragment.

ues to mix with the world ocean, it should act as a sink for heavy sulfur: it should tend to pull heavy sulfur from the ocean as a whole and bury it in evaporitic sediments. Then the world ocean should become relatively depleted in heavy sulfur and enriched in light sulfur. Hence in sediments derived from the world ocean as a whole during the period of the supercontinent, we should expect to see relatively high levels of light sulfur and low levels of heavy sulfur. This is indeed what is found in open marine sediments formed about 200 and 600 million years ago—that is, during the past two instances of the supercontinent.

Carbon isotopes also give evidence of supercontinents. The lighter isotope of carbon (carbon 12) diffuses in solution faster than a heavier isotope (carbon 13). As a result light carbon is more likely to be taken up by organisms and incorporated into their biomass. Organisms are therefore a sink for light carbon. During periods of low sea level the rate of organic productivity in the world ocean should be high, because greater amounts of nutrients such as phosphorus and nitrogen—eroded from continental crust and carried to the sea by rivers—will be available when more continental crust is exposed.

Thus when sea level is low, more carbon (particularly light carbon) will be incorporated into organisms, and the water of the world ocean will be comparatively depleted in light carbon and enriched in heavy carbon. When examining such seawater-derived sediments as limestone (calcium carbonate), then, we should expect to find relatively high levels of heavy carbon and low levels of light carbon if the sediments were produced during a period of low sea level. In similar sediments produced during a period of high sea level, we should expect to find relatively more light carbon and less heavy carbon. And indeed, the ratio of heavy carbon to light carbon in such sediments closely matches the predictions of our model for the past 600 million years.

Climate and Life

Perhaps the most important effects of the supercontinent cycle are its influences on climate and life. What should those effects be? Most of the climatological effects of the supercontinent cycle will be driven by the changes in sea level that are caused by the processes of continental breakup, dispersal and reassembly.

When sea level is low—that is, when the world is dominated by a single emergent supercontinent or when individual continents are widely dispersed (as they are now) and the world ocean floor is at its oldest—large amounts of silicate minerals in the continental crust, such as calcium silicates, are exposed to weathering and erosion; they are dissolved in rivers and carried into the world ocean. When these dissolved silicates are mixed into seawater, they combine with dissolved carbon dioxide to produce solid precipitates [see "How Climate Evolved on the Terrestrial Planets," by James F. Kasting, Owen B. Toon and James B. Pollack; SCIENTIFIC AMERICAN, February]. For example, calcium silicates may combine with carbon dioxide to produce calcite (limestone) and quartz. This process draws down carbon dioxide from the atmosphere.

Carbon dioxide in the atmosphere helps the earth to retain the heat it gains from solar radiation. When carbon dioxide is drawn down into oceanic deposits, this "greenhouse" warming effect is diminished, and the world climate becomes colder. If there is an emergent continental landmass near enough to the pole, glaciers will form (as they have on modern-day Antarctica and Greenland).

Glaciation has several important effects on the global climate. For one thing, it removes water from the ocean basins, causing sea level to drop still lower. Glaciation also amplifies circulation and mixing in the world ocean. Much of the ocean circulation of the present-day earth, for example, is driven by a global "heat engine," in which warm, salty water from the Tropics and the subtropics flows toward the pole, where it gives up its heat, sinks to the bottom and flows back toward the Equator [see "Polynyas in the Southern Ocean," by Arnold L. Gordon and Josefino C. Comiso; SCIENTIFIC AMERICAN, June]. By mixing surface water into the deep water, the heat engine distributes oxygen and other nutrients throughout the ocean. Ice at the pole keeps the polar waters cold, helping to maintain the temperature difference that drives the heat engine.

Such vertical circulation combines with the increased supply of continent-derived nutrients to raise the level of biological productivity at times of low sea level. High productivity should trap still more carbon in organic matter, further lowering levels of atmospheric carbon dioxide.

The environments that are most hospitable to marine life are the continental shelves, where continent-derived nutrients are abundant and shallow depths allow sunlight to penetrate to the sea floor. When biological productivity is high, however, these shelf environments are unavailable, exposed by the low sea level. As a result many established species will become extinct, and new, innovative species will be favored. This does not mean, however, that life will be particularly diverse. On the contrary, such conditions—high nutrient levels but few available environmental niches—lead to ecosystems in which a large amount of biomass is concentrated in relatively few successful species.

Climate after Breakup

Thus the effects of low sea level include a propensity toward glaciation, strong vertical circulation in the world ocean, high biological productivity, biological innovation and a low degree of biological diversity. What should be the effects of high sea level, which would be expected just after a supercontinent breaks up or just before it is reassembled?

When continents are drowned, a relatively small amount of silicates will be available to sequester atmospheric carbon dioxide in sea-floor deposits. Meanwhile carbon dioxide will be released into the ocean and from there into the atmosphere by the hot mantle material that wells up at sea-floor spreading centers. Also, the subduction of oceanic crust and the resulting melting of limestone deposits in the subducted sediments will release still more carbon dioxide into the atmosphere at the volcanoes that mark subduction zones.

Atmospheric levels of carbon dioxide will therefore rise and the earth's climate will become warmer. Polar ice caps should melt, raising sea level still higher and further drowning the continents. The absence of polar ice will reduce vertical and horizontal circulation of the world ocean, causing it to begin to stagnate: oxygen and nutrient levels in ocean waters will decline, and with them biological productivity. On the other hand, the drowned continental crust will provide a large area for the shallow seas that are most hospitable to life. The resulting ecosystems will resemble those of the present-day Tropics, where the climate is warm, nutrient levels are low and a relatively large number of environmental niches are available. Like today's Tropics, these ecosystems would be characterized by low productivity and great species diversity.

How well does the record of past climates and life forms bear out these predictions? One of the most impressive confirmations of our model is found in records of glaciation. All the known episodes of glaciation in the history of the earth took place at times when according to our model sea levels should have been low. The converse is not true. That is, not every period when sea levels should have been low had an episode of glaciation: probability dictates that in some periods there would not have been an emergent continent near the pole.

Biological Evidence

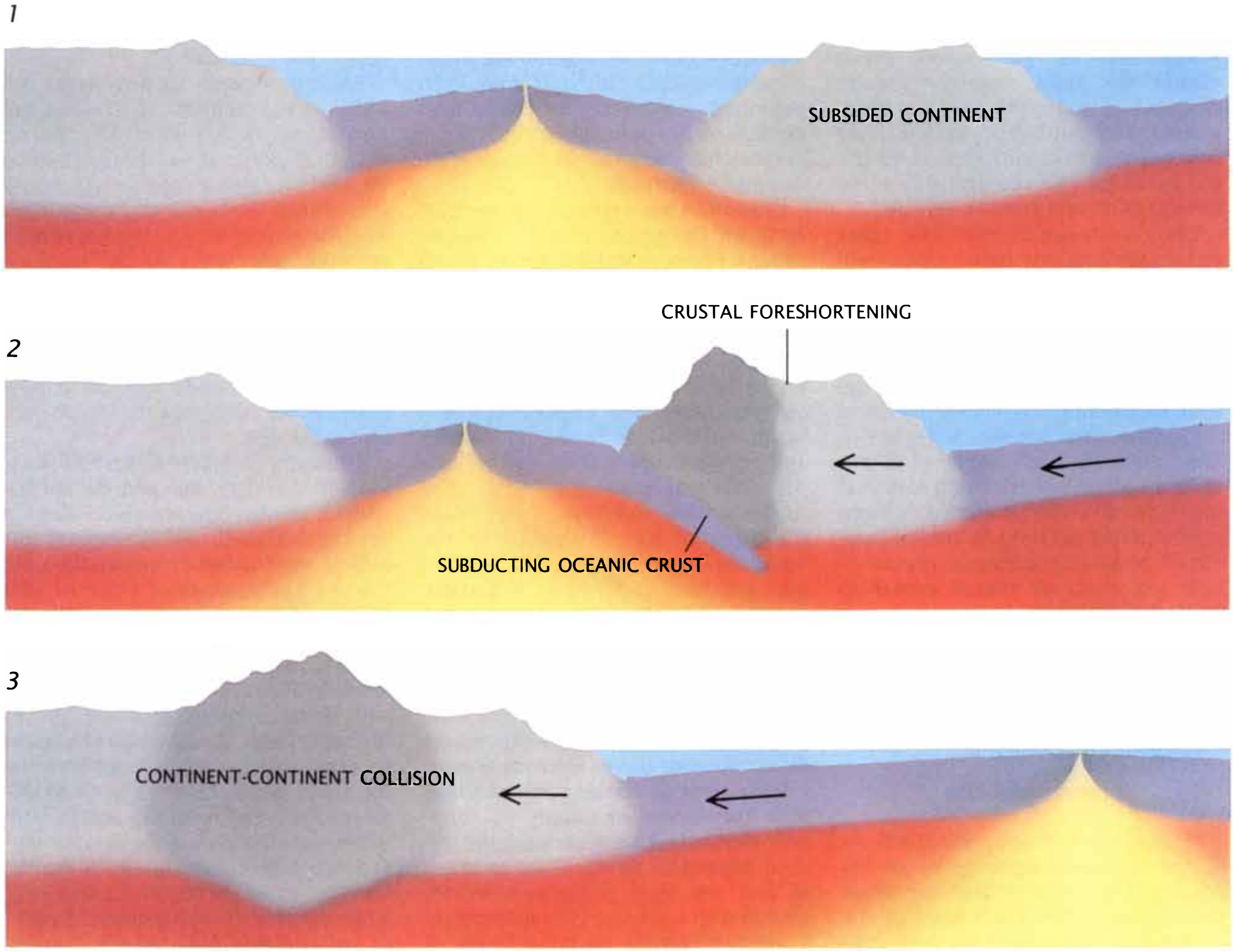
The biological record is a little more ambiguous, for a variety of reasons. Perhaps most important, the fossil record is not uniform throughout time. The record is based mainly on deposits buried on continental crust—buried when sea levels were high. When sea levels are low, marine organisms will generally live offshore, beyond the exposed continental shelves. Deposits recording these periods are rare: such deposits are likely to have been destroyed later, by subduction of the ocean floor. Nevertheless, the available evidence tends to confirm our hypothesis.

For example, the few geologic records of marine life during the existence of the most recent supercontinent, Pangaea, indicate low diversity of species, as we would expect when sea levels are low. In addition, the period of drowning that followed the breakup of Pangaea was characterized by high levels of species diversity.

Looking further back in time, the breakup of the previous supercontinent about 600 million years ago was marked by the first recorded appearance of shelled animals. During the period following this breakup there was what has been called an explosion of diversity [see "The Emergence of Animals," by Mark A. S. McMenamin; SCIENTIFIC AMERICAN, April, 1987]. In particular, shelled animals radiated into a highly diverse array.

Looking back still further, the first recorded multicellular animals are found in marine sediments that are about a billion years old. These sediments would have been deposited, according to our model, right after the breakup of a supercontinent. It is quite possible that this biological innovation occurred during the existence of the supercontinent but was not recorded until the continent broke up and sea levels rose, drowning the continental shelves.

A still older innovation may also be linked to the supercontinent cycle. About 2,100 million years ago, just prior to an assumed supercontinental breakup, blue-green algae first developed heterocysts, the organelles that



FORESHORTENING of continental blocks increases the total area of the world ocean and thereby lowers the global sea level. Just after the breakup of a supercontinent (*1*) the new sea floor (*left center*) butts up against continental fragments. Later, as the new ocean basin closes (*2*), the continent moves over the sea floor, subducting it; the resulting compressive forces foreshorten the continent and raise mountains, in the process lowering sea level. Still later, when continents collide during reassembly of the supercontinent (*3*), the continental crust is foreshortened even further, again lowering sea level.

make it possible to fix nitrogen (to crack apart nitrogen molecules and bind the constituent atoms to carbon in organic matter) even in the presence of atmospheric oxygen. Without heterocysts or similar organelles, the chemical reactions of nitrogen fixation can be interrupted by oxygen atoms that bind to the nitrogen atoms. The atmosphere was then just beginning to contain oxygen; the innovation made it possible for many later organisms—the predecessors of today's photosynthetic plants—to survive in the new, oxygen-bearing atmosphere, which otherwise could have been poisonous to them.

A New Framework

The supercontinent-cycle hypothesis represents a new framework, a new way to understand the geologic history of the earth. It suggests that the processes of plate tectonics on the largest scale are primarily governed not by chance but by a regular, cyclic process.

The supercontinent cycle also represents a new way of understanding the history of life on the earth. The large-scale climatological effects brought about by various phases of the supercontinent cycle—continental drowning or emergence, glaciation and ocean circulation, stagnation in the world ocean and other effects—drove many of the important biological innovations that have directed the later course of evolution. In a sense, then, the supercontinent cycle is indeed the pulse of the earth: with every beat the earth's climate, geology and population of living organisms are advanced and renewed.

FURTHER READING

THE TWO PHANEROZOIC SUPERCYCLES. Alfred G. Fisher in *Catastrophes and Earth History*, edited by W. A. Berggren and John Van Couvering. Princeton University Press, 1983.

GLOBAL TECTONICS AND EUSTASY FOR THE PAST 2 BILLION YEARS. Thomas R. Worsley, Damian Nance and Judith B. Moody in *Marine Geology*, Vol. 58, No. 3/4, pages 373-400; July 25, 1984.

POST-ARCHEAN BIOCHEMICAL CYCLES AND LONG-TERM EPISODICITY IN TECTONIC PROCESSES. R. Damian Nance, Thomas R. Worsley and Judith B. Moody in *Geology*, Vol. 14, No. 6, pages 514-518; June, 1986.

TECTONIC CYCLES AND THE HISTORY OF THE EARTH'S BIOCHEMICAL AND PALEOCEANOGRAPHIC RECORD. Thomas R. Worsley, R. Damian Nance and Judith B. Moody in *Paleoceanography*, Vol. 1, No. 3, pages 233-263; September, 1986.

Randomness in Arithmetic

It is impossible to prove whether each member of a family of algebraic equations has a finite or an infinite number of solutions: the answers vary randomly and therefore elude mathematical reasoning

by Gregory J. Chaitin

What could be more certain than the fact that 2 plus 2 equals 4? Since the time of the ancient Greeks mathematicians have believed there is little—if anything—as unequivocal as a proved theorem. In fact, mathematical statements that can be proved true have often been regarded as a more solid foundation for a system of thought than any maxim about morals or even physical objects. The 17th-century German mathematician and philosopher Gottfried Wilhelm Leibniz even envisioned a "calculus" of reasoning such that all disputes could one day be settled with the words "Gentlemen, let us compute!" By the beginning of this century symbolic logic had progressed to such an extent that the German mathematician David Hilbert declared that all mathematical questions are in principle decidable, and he confidently set out to codify once and for all the methods of mathematical reasoning.

Such blissful optimism was shattered by the astonishing and profound discoveries of Kurt Gödel and Alan M. Turing in the 1930's. Gödel showed that no finite set of axioms and methods of reasoning could encompass all the mathematical properties of the positive integers. Turing later couched Gödel's ingenious and complicated proof in a more accessible form. He showed that Gödel's incompleteness theorem is equivalent to the assertion that there can be no general method for systematically deciding whether a computer program will ever halt, that is, whether it will ever cause the computer to stop running. Of course, if a particular program does cause the computer to halt, that fact can be easily proved by running the program. The difficulty lies in proving that an arbitrary program never halts.

I have recently been able to take a further step along the path laid out by Gödel and Turing. By translating a particular computer program into an algebraic equation of a type that was familiar even to the ancient Greeks, I have shown that there is randomness in the branch of pure mathematics known as number theory. My work indicates that—to borrow Einstein's metaphor—God sometimes plays dice with whole numbers!

This result, which is part of a body of work called algorithmic information theory, is not a cause for pessimism; it does not portend anarchy or lawlessness in mathematics. (Indeed, most mathematicians continue working on problems as before.) What it means is that mathematical laws of a different kind might have to apply in certain situations: statistical laws. In the same way that it is impossible to predict the exact moment at which an individual atom undergoes radioactive decay, mathematics is sometimes powerless to answer particular questions. Nevertheless, physicists can still make reliable predictions about averages over large ensembles of atoms. Mathematicians may in some cases be limited to a similar approach.

My work is a natural extension of Turing's, but whereas Turing considered whether or not an arbitrary program would ever halt, I consider the probability that any general-purpose computer will stop running if its program is chosen completely at random. What do I mean when I say "chosen completely at random"? Since at the most fundamental level any program can be reduced to a sequence of bits (each of which can take on the value 0 or 1) that are "read" and "interpreted" by the computer hardware, I mean that a completely random program consisting of n bits could just as well be the result of flipping a coin n times (in which a "heads" represents a 0 and a "tails" represents 1, or vice versa).

The probability that such a completely random program will halt, which I have named omega (Ω), can be expressed in terms of a real number between 0 and 1. (The statement $\Omega = 0$ would mean that no random program will ever halt, and $\Omega = 1$ would mean that every random program halts. For a general-purpose computer neither of these extremes is actually possible.) Because Ω is a real number, it can be fully expressed only as an unending sequence of digits. In base 2 such a sequence would amount to an infinite string of 0's and 1's.

Perhaps the most interesting characteristic of Ω is that it is algorithmically random: it cannot be compressed into a program (considered as a string of bits) shorter than itself. This definition of randomness, which has a central role in algorithmic information theory, was independently formulated in the mid-1960's by the late A. N. Kolmogorov and me. (I have since had to correct the definition.)

The basic idea behind the definition is a simple one. Some sequences of bits can be compressed into programs much shorter than they are, because they follow a pattern or rule. For example, a 200-bit sequence of the form 0101010101... can be greatly compressed by describing it as "100 repetitions of 01." Such sequences certainly are not random. A 200-bit sequence generated by tossing a coin, on the other hand, cannot be compressed, since in general there is no pattern to the succession of 0's and 1's: it is a completely random sequence.

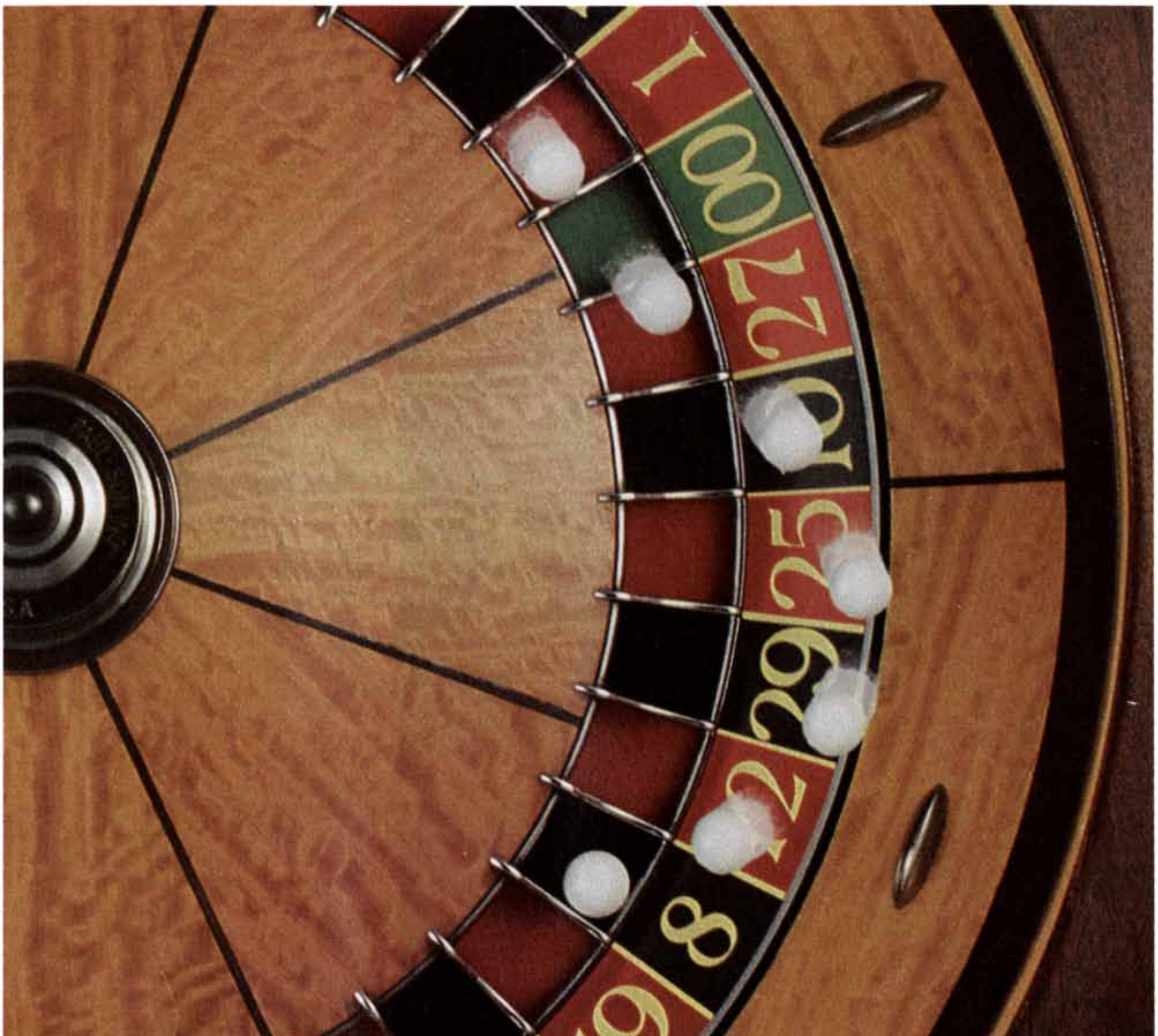
Of all the possible sequences of bits, most are incompressible and therefore random. Since a sequence of bits can be considered to be a base-2 rep-

GREGORY J. CHAITIN is on the staff of the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y. He is the principal architect of algorithmic information theory and has just published two books in which the theory's concepts are applied to elucidate the nature of randomness and the limitations of mathematics. This is Chaitin's second article for SCIENTIFIC AMERICAN.

resentation of any real number (if one allows infinite sequences), it follows that most real numbers are in fact random. It is not difficult to show that an algorithmically random number, such as Ω, exhibits the usual statistical properties one associates with randomness. One such property is normality: every possible digit appears with equal frequency in the number. In a base-2 representation this means that as the number of digits of Ω approaches infinity, 0 and 1 respectively account for exactly 50 percent of Ω's digits.

A key technical point that must be stipulated in order for Ω to make sense is that an input program must be self-delimiting: its total length (in bits) must be given within the program itself. (This seemingly minor point, which paralyzed progress in the field for nearly a decade, is what entailed the redefinition of algorithmic randomness.) Real programming languages are self-delimiting, because they provide constructs for beginning and ending a program. Such constructs allow a program to contain well-defined subprograms, which may also have other subprograms nested in them. Because a self-delimiting program is built up by concatenating and nesting self-delimiting subprograms, a program is syntactically complete only when the last open subprogram is closed. In essence the beginning and ending constructs for programs and subprograms function respectively like left and right parentheses in mathematical expressions.

If programs were not self-delimiting, they could not be constructed from subprograms, and summing the halting probabilities for all programs would yield an infinite number. If one considers only self-delimiting programs, not only is Ω limited to the range between 0 and 1 but also it can be explicitly calculated "in the limit from below." That is to say, it is possible to calculate an infinite sequence of rational numbers (which can be expressed in terms of a finite sequence of bits) each of which is closer to the true value of Ω than the preceding number.



UNPREDICTABILITY is a concept familiar to casinos: it is what enables them to make a profit on games such as roulette. The author argues that mathematicians, like gamblers, may also have to reconcile themselves to an inherent unpredictability in their occupation. It arises from the randomness that exists in the solutions to rather simple problems in number theory.

One way to do this is to systematically calculate Ω_n for increasing values of n; Ω_n is the probability that a completely random program up to n bits in size will halt within n seconds if the program is run on a given computer. Since there are 2^k possible programs that are k bits long, Ω_n can in principle be calculated by determining for every value of k between 1 and n how many of the possible programs actually halt within n seconds, multiplying that number by 2^{-k} and then summing all the products. In other words, each k-bit program that halts contributes 2^{-k} to Ω_n; programs that do not halt contribute 0.

If one were miraculously given the value of Ω with k bits of precision, one could calculate a sequence of Ω_n's until one reached a value that equaled the given value of Ω. At this point one would know all programs of a size less than k bits that halt; in essence one would have solved Turing's halting problem for all programs of a size less than k bits. Of course, the time required for the calculation would be enormous for reasonable values of k.

DAVID HILBERT (1900):
"Every definite mathematical problem must necessarily be susceptible of an exact settlement, either in the form of an actual answer to the question asked, or by the proof of the impossibility of its solution and therewith the necessary failure of all attempts.... However unapproachable these problems may seem to us and however helpless we stand before them, we have, nevertheless, the firm conviction that their solution must follow by a finite number of purely logical processes....
"We hear within us the perpetual call: There is the problem. Seek its solution. You can find it by pure reason...."

FUNDAMENTAL DECIDABILITY of all mathematical questions was espoused by David Hilbert, who was about 50 years old when this photograph was made. He believed a finite set of axioms and rules of reasoning sufficed to prove or disprove all theorems.

So far I have been referring exclusively to computers and their programs in discussing the halting problem, but it took on a new dimension in light of the work of J. P. Jones of the University of Calgary and Y. V. Matijasevič of the V. A. Steklov Institute of Mathematics in Leningrad. Their work provides a method for casting the problem as assertions about particular diophantine equations. These algebraic equations, which involve only multiplication, addition and exponentiation of whole numbers, are named after the third-century Greek mathematician Diophantus of Alexandria.

To be more specific, by applying the method of Jones and Matijasevič one can equate the statement that a particular program does not halt with the assertion that one of a particular family of diophantine equations has no solution in whole numbers. As with the original version of the halting problem for computers, it is easy to prove a solution exists: all one has to do is plug in the correct numbers and verify that the resulting numbers on the left and right sides of the equal sign are in fact equal. The much more difficult problem is to prove that there are absolutely no solutions when this is the case.

The family of equations is constructed from a basic equation that contains a particular variable k, called the parameter, which takes on the values 1, 2, 3 and so on [*see top illustration on page 84*]. Hence there is an infinitely large family of equations (one for each value of k) that can be generated from one basic equation for each of a "family" of programs. The mathematical assertion that the diophantine equation with parameter k has no solution encodes the assertion that the kth computer program never halts. On the other hand, if the kth program does halt, then the equation has exactly one solution. In a sense the truth or falsehood of assertions of this type is mathematically uncertain, since it varies unpredictably as the parameter k takes on different values.

My approach to the question of unpredictability in mathematics is similar, but it achieves a much greater degree of randomness. Instead of "arithmetizing" computer programs that may or may not halt as a family of diophantine equations, I apply the method of Jones and Matijasevič to arithmetize a single program to calculate the kth bit in Ω_n.

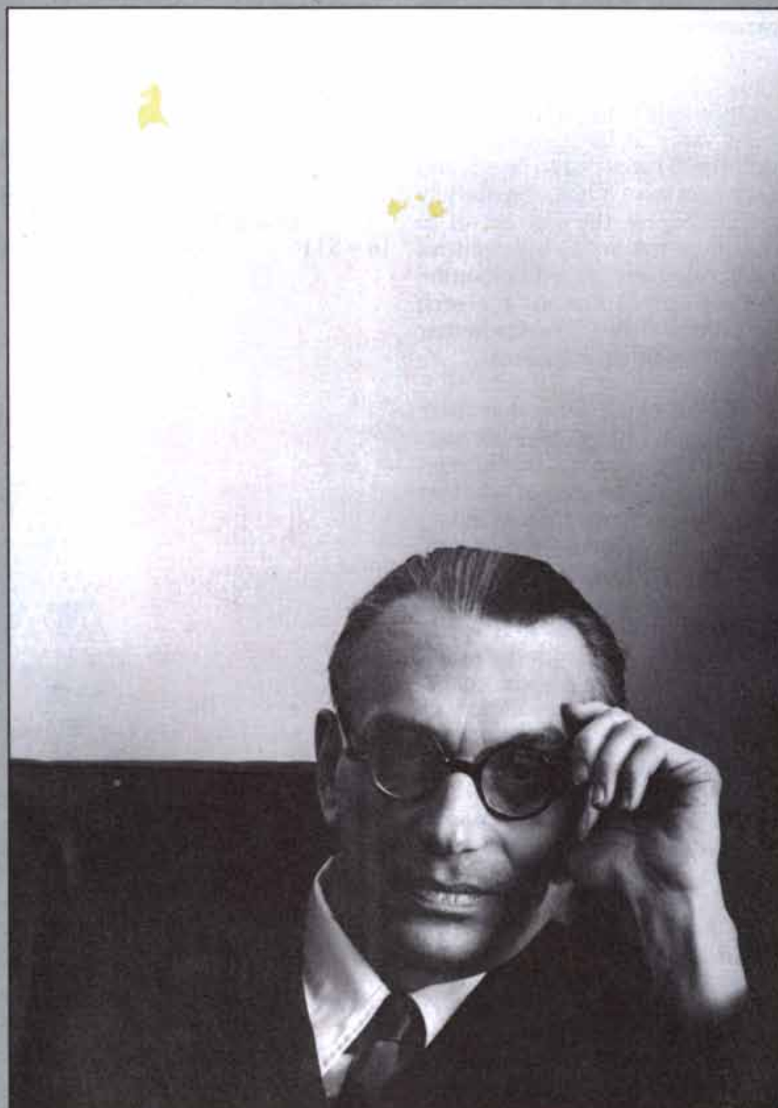
The method is based on a curious property of the parity of binomial coefficients (whether they are even or odd numbers) that was noted by Édouard A. Lucas a century ago but was not properly appreciated until now. Binomial coefficients are the multiplicands of the powers of x that arise when one expands expressions of the type $(x+1)^n$. These coefficients can easily be computed by constructing what is known as Pascal's triangle [*see bottom illustration on next page*].

Lucas's theorem asserts that the coefficient of x^k in the expansion of $(x+1)^n$ is odd only if each digit in the base-2 representation of the number k is less than or equal to the corresponding digit in the base-2 representation of n (starting from the right and reading left). To put it a little more simply, the coefficient for x^k in an expansion of $(x+1)^n$ is odd if for every bit of k that is a 1 the corresponding bit of n is also a 1, otherwise the coefficient is even. For example, the coefficient of x^2 in the binomial expansion of $(x+1)^4$ is 6, which is even. Hence the 1 in the base-2 representation of 2 (10) is not matched with a 1 in the same position in the base-2 representation of 4 (100).

Although the arithmetization is conceptually simple and elegant, carrying it out is a substantial programming task. Nevertheless, I thought it would be fun to do it. I therefore developed a "compiler" program for producing equations from programs for a register machine. A register machine is a computer that consists of a small set of registers for storing arbitrarily large numbers. It is an abstraction, of course, since any real computer has registers with a limited capacity.

Feeding a register-machine program that executes instructions in the LISP computer language, as input, into a real computer programmed with the compiler yields within a few minutes, as output, an equation about 200 pages long containing about 17,000 nonnegative integer variables. I can thus derive a diophantine equation having a parameter k that encodes the kth bit of Ω_n merely by plugging a LISP program (in binary form) for calculating the kth bit of Ω_n into the 200-page equation. For any given pair of values of k and n, the diophantine equation has exactly one solution if the kth bit of Ω_n is a 1, and it has no solution if the kth bit of Ω_n is a 0.

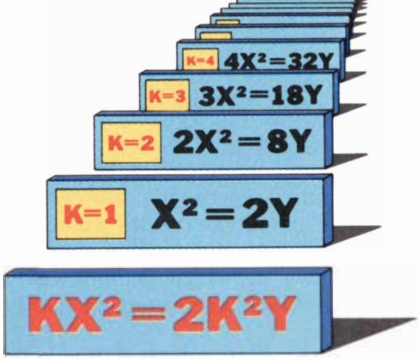
Because this applies for any pair of values for k and n, one can in principle keep k fixed and systematically increase the value of n without limit, calculating the kth bit of Ω_n for each value of n. For small values of n the kth bit of Ω_n will fluctuate erratically between 0 and 1. Eventually, however, it will settle on either a 0 or a 1, since for very large values of n it will be equal to the kth bit of Ω, which is immutable. Hence the diophantine equation actu-



KURT GÖDEL (1944):

"It has turned out that the solution of certain arithmetical problems requires the use of assumptions essentially transcending arithmetic.... Of course, under these circumstances mathematics may lose a good deal of its 'absolute certainty,' but, under the influence of the modern criticism of the foundations, this has already happened to a large extent...."

FUNDAMENTAL UNDECIDABILITY of mathematical questions was proved by Kurt Gödel, here shown also at about the age of 50 in a photograph made by ©Arnold Newman. Gödel published his proof in 1931, when he was 25 and Hilbert was 70.



"FAMILY" OF EQUATIONS can be generated by assigning whole-number values to a parameter, *K*, of a basic equation.

ally has infinitely many solutions for a particular value of its parameter k if the kth bit of Ω turns out to be a 1, and for similar reasons it has only finitely many solutions if the kth bit of Ω turns out to be a 0. In this way, instead of considering whether a diophantine equation has any solutions for each value of its parameter k, I ask whether it has infinitely many solutions.

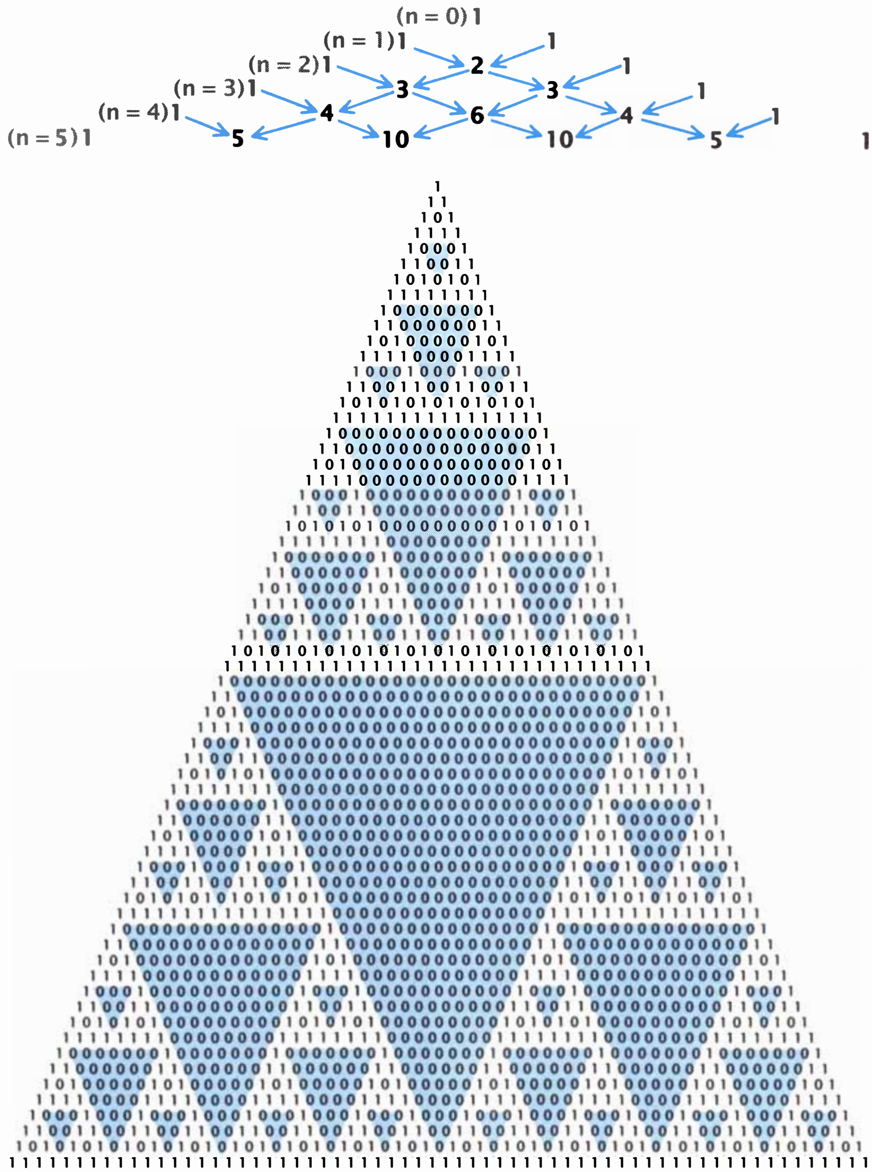
Although it might seem that there is little to be gained by asking whether there are infinitely many solutions instead of whether there are any solutions, there is in fact a critical distinction: the answers to my question are logically independent. Two mathematical assertions are logically independent if it is impossible to derive one from the other, that is, if neither is a logical consequence of the other. This notion of independence can usually be distinguished from that applied in statistics. There two chance events are said to be independent if the outcome of one has no bearing on the outcome of the other. For example, the result of tossing a coin in no way affects the result of the next toss: the results are statistically independent.

In my approach I bring both notions of independence to bear. The answer to my question for one value of k is logically independent of the answer for another value of k. The reason is that the individual bits of Ω, which determine the answers, are statistically independent.

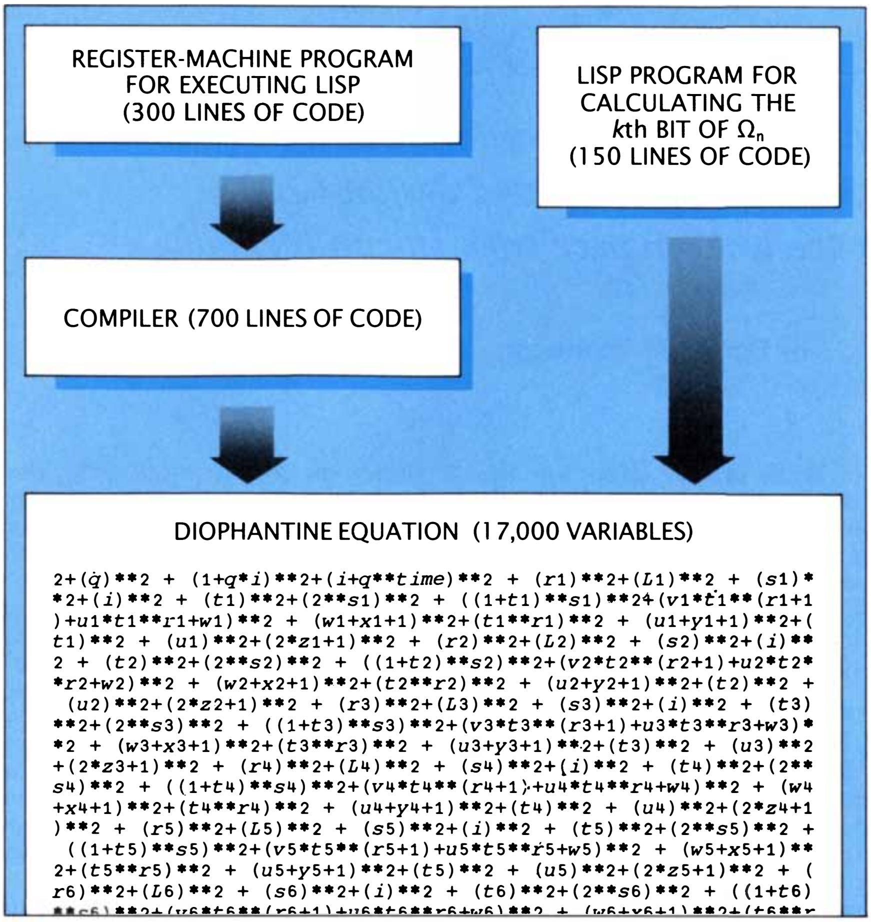
Although it is easy to show that for about half of the values of k the number of solutions is finite and for the other half the number of solutions is infinite, there is no possible way to compress the answers in a formula or set of rules; they mimic the results of coin tosses. Because Ω is algorithmically random, even knowing the answers for 1,000 values of k would not help one to give the correct answer for another value of k. A mathematician could do no better than a gambler tossing a coin in deciding whether a particular equation had a finite or an infinite number of solutions. Whatever axioms and proofs one could apply to find the answer for the diophantine equation with one value of k, they would be inapplicable for the same equation with another value of k.

Mathematical reasoning is therefore essentially helpless in such a case, since there are no logical interconnections between the diophantine equations generated in this way. No matter how bright one is or how long the proofs and how complicated the mathematical axioms are, the infinite series of propositions stating whether the number of solutions of the diophantine equations is finite or infinite will quickly defeat one as k increases. Randomness, uncertainty and unpredictability occur even in the elementary branches of number theory that deal with diophantine equations.

How have the incompleteness theorem of Gödel, the halting problem of Turing and my own work affected mathematics? The fact



PASCAL'S TRIANGLE (*top*) is a method by which to calculate the coefficients in an expansion of expressions of the form $(x + 1)^n$. One begins with a triangle of 1's and computes the values in each successive line by adding adjacent numbers and placing 1's at each end of the line. In this way one can determine, for example, that $(x + 1)^4 = 1x^4 + 4x^3 + 6x^2 + 4x^1 + 1x^0$. The triangle can be converted into an eye-catching fractal pattern if one replaces odd coefficients with a 1 and even coefficients with a 0 (*bottom*). The pattern displays in graphic form a property of the coefficients that is applied in the "arithmetization" of computer programs into algebraic equations.



ARITHMETIZATION OF Ω is accomplished by substituting the binary representation of a specific program to calculate the *k*th digit of Ω_n (expressed in base 2) for a variable in the equation derived from a general computer program. Ω_n is an "*n*th order" approximation of Ω, which is the probability that a computer will stop running if the bits that constitute its program are determined randomly, say, by flipping a coin.

is that most mathematicians have shrugged off the results. Of course, they agree in principle that any finite set of axioms is incomplete, but in practice they dismiss the fact as not applying directly to their work. Unfortunately, however, it may sometimes apply. Although Gödel's original theorem seemed to apply only to unusual mathematical propositions that were not likely to be of interest in practice, algorithmic information theory has shown that incompleteness and randomness are natural and pervasive. This suggests to me that the possibility of searching for new axioms applying to the whole numbers should perhaps be taken more seriously.

Indeed, the fact that many mathematical problems have remained unsolved for hundreds and even thousands of years tends to support my contention. Mathematicians steadfastly assume that the failure to solve these problems lies strictly within themselves, but could the fault not lie in the incompleteness of their axioms? For example, the question of whether there are any perfect odd numbers has defied an answer since the time of the ancient Greeks. (A perfect number is a number that is exactly the sum of its divisors, excluding itself. Hence 6 is a perfect number, since 6 equals 1 plus 2 plus 3.) Could it be that the statement "There are no odd perfect numbers" is unprovable? If it is, perhaps mathematicians had better accept it as an axiom.

This may seem like a ridiculous suggestion to most mathematicians, but to a physicist or a biologist it may not seem so absurd. To those who work in the empirical sciences the usefulness of a hypothesis, and not necessarily its "self-evident truth," is the key criterion by which to judge whether it should be regarded as the basis for a theory. If there are many conjectures that can be settled by invoking a hypothesis, empirical scientists take the hypothesis seriously. (The nonexistence of odd perfect numbers does not appear to have significant implications and would therefore not be a useful axiom by this criterion.)

Actually in a few cases mathematicians have already taken unproved but useful conjectures as a basis for their work. The so-called Riemann hypothesis, for instance, is often accepted as being true, even though it has never been proved, because many other important theorems are based on it. Moreover, the hypothesis has been tested empirically by means of the most powerful computers, and none has come up with a single counterexample. Indeed, computer programs (which, as I have indicated, are equivalent to mathematical statements) are also tested in this way—by verifying a number of test cases rather than by rigorous mathematical proof.

Are there other problems in other fields of science that can benefit from these insights into the foundations of mathematics? I believe algorithmic information theory may have relevance to biology. The regulatory genes of a developing embryo are in effect a computer program for constructing an organism. The "complexity" of this biochemical computer program could conceivably be measured in terms analogous to those I have developed in quantifying the information content of Ω.

Although Ω is completely random (or infinitely complex) and cannot ever be computed exactly, it can be approximated with arbitrary precision given an infinite amount of time. The complexity of living organisms, it seems to me, could be approximated in a similar way. A sequence of Ω_n's, which approach Ω, can be regarded as a metaphor for evolution and perhaps could contain the germ of a mathematical model for the evolution of biological complexity.

At the end of his life John von Neumann challenged mathematicians to find an abstract mathematical theory for the origin and evolution of life. This fundamental problem, like most fundamental problems, is magnificently difficult. Perhaps algorithmic information theory can help to suggest a way to proceed.

FURTHER READING

ALGORITHMIC INFORMATION THEORY. Gregory J. Chaitin. Cambridge University Press, 1987.

INFORMATION, RANDOMNESS & INCOMPLETENESS. Gregory J. Chaitin. World Scientific Publishing Co. Pte. Ltd., 1987.

THE ULTIMATE IN UNDECIDABILITY. Ian Stewart in *Nature*, Vol. 232, No. 6160, pages 115–116; March 10, 1988.

Capacitors

They are the electronic guardians of integrated circuits. Innovative manufacturing, materials and design have enabled a 200-year-old device to keep pace with silicon technology

by Donald M. Trotter, Jr.

The silicon chip may be the heart of modern electronics, but to do its job properly an integrated circuit still needs help from an old-fashioned component: the capacitor. Capacitors act as electrical buffers, diverting spurious electrical signals and storing surges of charge that could damage circuits or disrupt their operation. They cost just a few cents each but are the foundation of a billion-dollar international business. In televisions and radios, in calculators and computers—wherever there are integrated circuits, capacitors can also be found.

The principles of capacitor design were introduced almost 250 years ago with the invention of a device called a Leyden jar. Those principles have not changed, but advances in materials and manufacturing technologies have produced a staggering increase in capacitor efficiency. The charge that could be stored in a "one quart" Leyden jar in the 18th century can now be squeezed into a device not much larger than the head of a pin. Indeed, in the past 30 years capacitors have undergone size reductions that rival those achieved by chip technology.

Miniaturization is an urgent priority of the capacitor industry because the advantages of shrinking integrated circuits cannot be realized unless capacitors shrink with them. The two most popular capacitor designs are engineering marvels: one makes use of brittle ceramic layers less than a thousandth of an inch thick, and the other condenses the surface area of a sheet of newspaper into a device smaller than a sugar cube. In order to understand the theoretical basis for these designs, it helps to look at the capacitor's origins.

DONALD M. TROTTER, JR., is a senior research scientist at the Corning Glass Works, where he spent several years investigating experimental capacitors. He has bachelor's and master's degrees from Kansas State University and a Ph.D. from the University of Texas at Austin, which was awarded in 1977. He joined Corning in 1980 after a postdoctoral appointment at Cornell University. Trotter is a science-fiction enthusiast and has published several short stories in that genre.

Leyden jars were the first capacitors, perfected by the English astronomer and physician John Bevis in 1746. A Leyden jar is a glass jar lined with metal foil and covered on the outside with a second piece of foil [*see illustration on page 88*]. A metal rod pushed through a rubber stopper in the neck of the jar makes contact with the inner foil. Both the inner foil and the outer foil can act as electrodes, but they have no net charge other than what is supplied by an external source.

The source of charge can be a battery, a generator or merely a rubber rod that has been stroked with fur. If the rubber rod, which carries mobile electrons, is touched to the metal rod in the Leyden jar, electrons flow freely from the rod into the jar. A net negative charge is thereby transferred to the inner electrode. Because the storage capacity of the jar is limited by the mutual repulsion of electrons, the flow of electrons eventually stops. Similar principles apply to the storage of a net positive charge, generated by the removal of electrons from the inner electrode. The ability to store charge is known as capacitance.

In a Leyden jar capacitance is enhanced by the presence of the second electrode on the outside of the jar. When this electrode is grounded, the charge stored on the inner electrode attracts an equal charge of the opposite sign from the ground. Since the supply of opposite charge in the ground is for practical purposes unlimited, the ground can supply as much neutralizing charge as is needed. As this charge builds up on the outer electrode, it attracts the electrons in the inner electrode, canceling some of the repulsive forces that curtail electron storage. Consequently the jar's capacitance increases.

It is not, however, unlimited; the balance between attractive and repulsive forces is always tilted toward repulsion, because the repulsive forces originate closer, by the thickness of the insulator, to the electron source. Hence as more charge is added to the inner electrode the neutralizing effect of the ground is eventually overwhelmed. At that point no more charge can be stored in the jar.

There are two obvious ways to increase the capacitance of a Leyden jar. One way is to increase the surface area of the electrodes, giving the charge more room to spread out and reducing the repulsion between electrons. The other way is to reduce the thickness of the glass separating the stored inner charge from the neutralizing outer charge. (If the glass becomes too thin, however, electrons can be pulled through it, creating a spark that dissipates the charge.) Although they are difficult to implement in a Leyden jar, these are two of the three classic strategies by which scientists and engineers seek to improve the performance of modern capacitors.

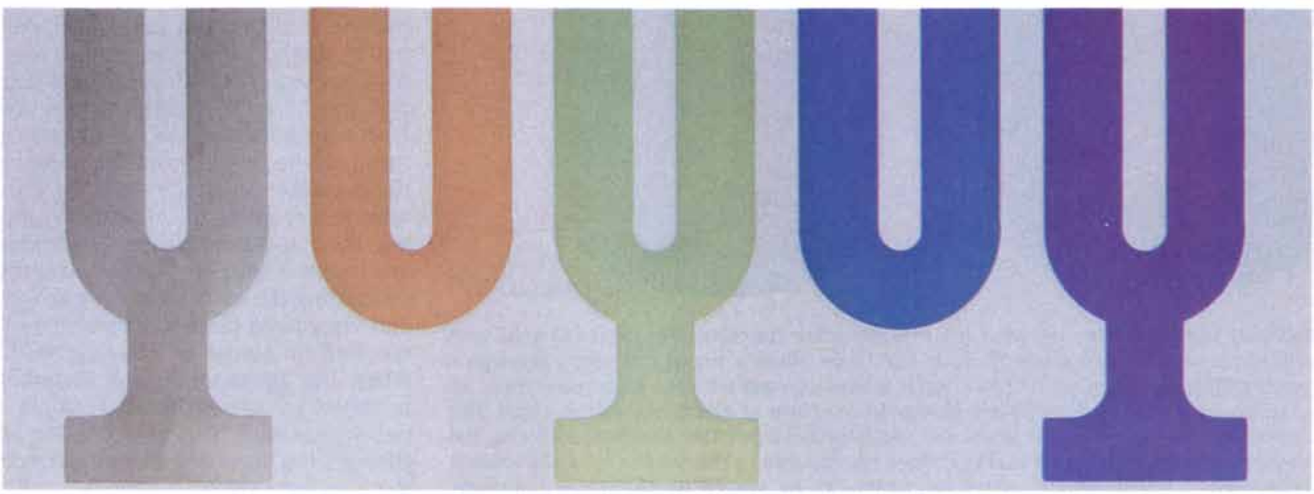
The electrical behavior of the electrons in the insulating material offers a third means of increasing capacitance. Even though the electrons in the insulator are not mobile, they can shift slightly under the influence of the attractive and repulsive forces from the electrodes on each side. Electrons on one side of the insulator "bulge" out from the surface, creating a negative charge; on the other side they retreat from the surface, giving rise to a positive charge.

The charges thus generated on the insulator help to neutralize the charges on the electrodes, and some insulators can bear charges that are nearly as large as those on the electrodes themselves. Neutralization reduces repulsive forces and allows more charge to reside on the electrodes, increasing capacitance. The degree to which this phenomenon occurs is reflected in a property of the insulator called the dielectric constant. The dielectric constant of a material indicates how many times the capacitance of a device increases when that material, rather than a vacuum, is placed between the electrodes. The glass used in a Leyden jar has a dielectric constant of about five. New insulators in capacitors now coming on the market have dielectric constants of close to 20,000.

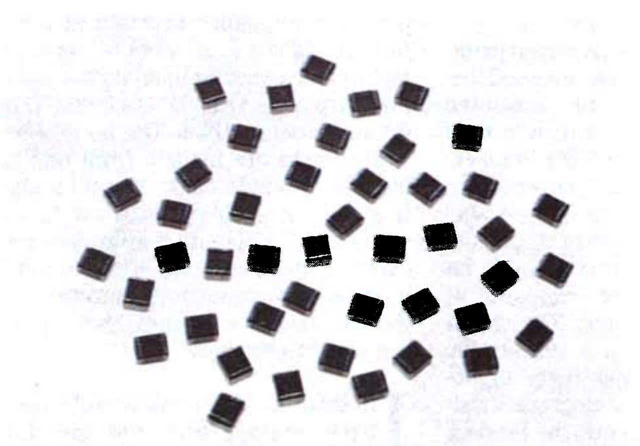
Such materials account for the efficacy of the multilayer ceramic capacitor, one of the two standard capacitor types that together represent 95 percent of all capacitors sold. The other type, the electrolytic capacitor, achieves even higher capacitance per unit volume without exploiting insulators that have high dielectric constants. Thus the two kinds of capacitors rely on different strategies to maximize their efficiency. Their method of construction differs as much as their design.

The multilayer ceramic capacitor (MLCC) is a compact version of the Leyden jar [*see top illustration on page 90*]. Imagine a Leyden jar slit along its length and unrolled, yielding a large, flat sheet of insulator with electrodes on each side. The sheet could be cut into small squares and the squares stacked on top of one another, inside electrode against inside electrode and outside electrode against outside. If a small strip were removed from all the inside electrodes on one end of the stack, then all the outside electrodes could be connected by covering that end with a metal termination. Similarly, all the inside electrodes could be connected on the other end by removing strips from the outside electrodes and applying a termination there. Each termination would unite many interleaved electrodes in what is effectively one large electrode.

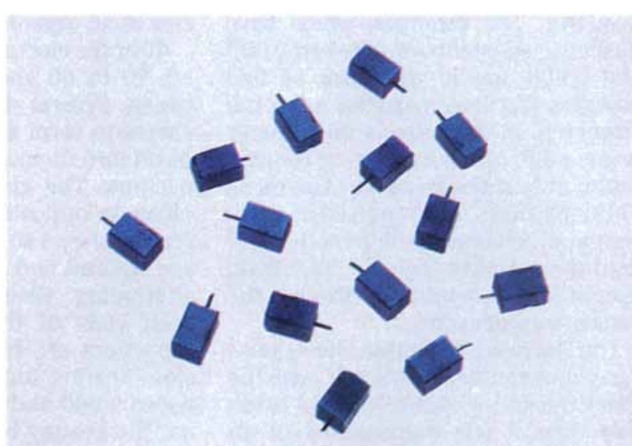
Of course, that is not how MLCC's are constructed, but it is the rationale of



EXPERIMENTAL INSULATOR for capacitors is made from an oxide of niobium metal. The bare metal (*left*) is shown next to samples whose surfaces have been oxidized. The oxide film is so thin that it gives rise to interference effects in the reflected light; hence the thin oxide layers look purple, blue and green, whereas thicker films are red and orange. Although niobium oxide is an excellent insulator, it can be unreliable because the glassy surface crystallizes when a voltage is applied to it.



MULTILAYER CERAMIC CAPACITORS, more commonly called MLCC's, are not much bigger than a BB. The silvery knobs at each end are terminations; the brown in between is a barium-titanate-based ceramic insulator. MLCC's are often coated with plastic before they are soldered to a circuit board, where they serve to divert power surges and spurious electrical signals.



TANTALUM SLUGS are intermediates in the manufacture of tantalum capacitors. The metal slugs are coated with layers of tantalum oxide that, like the niobium oxide films in the top photograph, reflect different colors depending on their thickness. The blue films are thicker than the purple ones. These capacitors and MLCC's are the two main classes of capacitors.

LEYDEN JAR illustrates the principles of capacitor function. The jar is covered with foil electrodes on the inside (*red*) and outside (*blue*); a metal rod stuck through a rubber stopper in the neck of the bottle makes contact with the inner electrode. An external source supplies electric charge in the form of electrons (*minus signs*). The charge can be stored on the inner electrode because positive charges collect on the outer one (*left*). More neutralizing charge is available to the outer electrode when it is grounded (*right*), and so more electrons can be stored on the inner electrode.

the design. In actual manufacturing practice, ceramics consisting of barium titanate and small amounts of other oxides are generally used as the insulator. The ceramics, which have dielectric constants of between 2,000 and 6,000, are in the form of fine powders that have particles a few micrometers in diameter (a micrometer is one millionth of a meter, or roughly four hundred-thousandths of an inch). These particles are dispersed in a solvent in which organic binders that will hold the particles together have been dissolved. The resulting slurry has the consistency of paint.

The slurry is cast in thin sheets onto a paper or stainless-steel belt, with the sheet thickness controlled by a blade that rides a few thousandths of an inch above the belt. The slurry dries as the solvent evaporates, leaving a cohesive "green" (unfired) tape that is smooth and limp like fine cloth. The tape is cut into sheets from six to eight inches square, and thousands of electrodes are printed on each sheet through a thin screen that delineates the electrode patterns. The electrodes are made of silver-palladium ink, a mixture of finely divided metal particles in an organic binder.

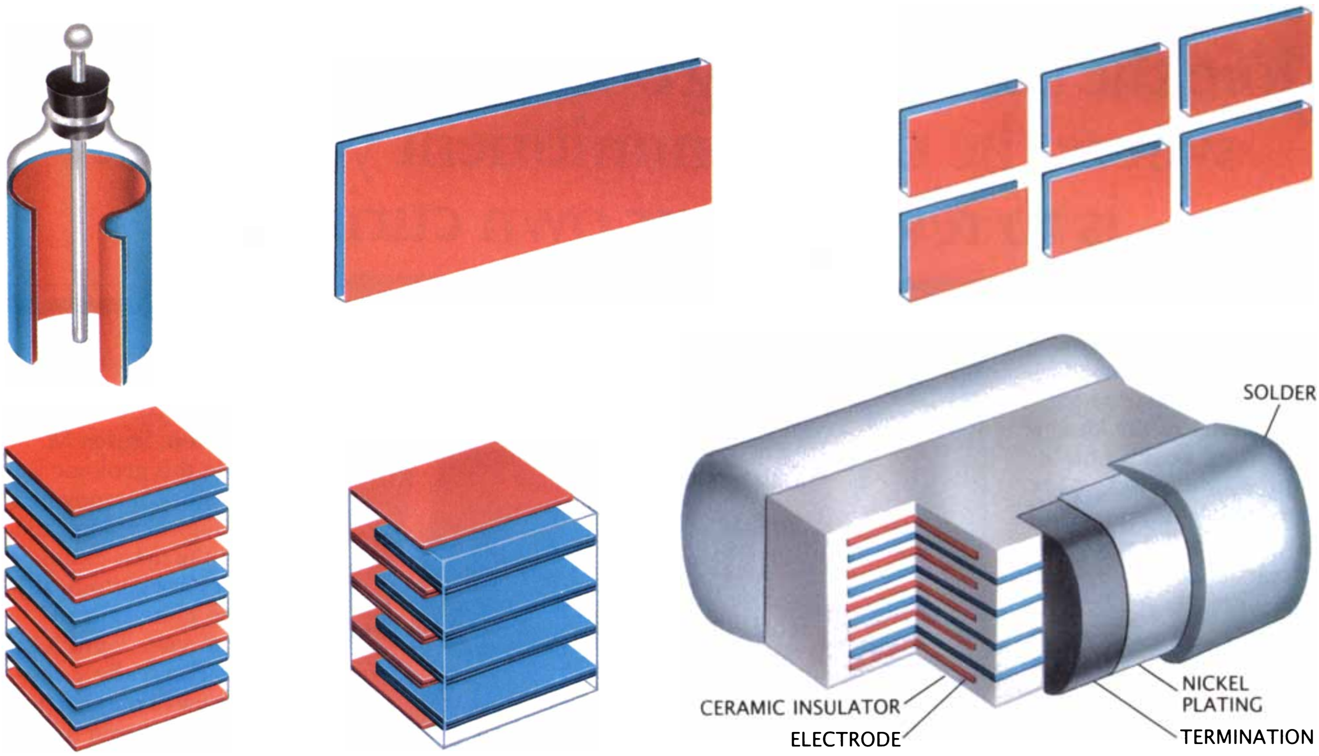
After the electrodes have been printed, 30 to 60 sheets are pressed between several layers of unprinted sheets to form a stack, which is then diced into thousands of individual capacitors. The electrode patterns are offset in opposite directions on successive sheets so that when the sheets are stacked and diced, electrodes on alternating sheets are exposed at both ends of the stack. The green capacitors are fired in a furnace by slow heating to temperatures of between 1,000 and 1,400 degrees Celsius. The heating burns out the binders in the ceramic layers and the inks and causes the ceramic layers to "sinter": the individual particles bind to one another but do not melt together. At the end of the firing the ceramic layers make up a fine-grained, coherent whole that is brittle but strong. The metal particles in the electrodes also sinter during firing, forming continuous metal sheets.

Terminations are applied by dipping the ends of each capacitor in another silver-palladium ink that is formulated to adhere to the ceramic surface. The ink is fused by a second firing. The terminations are then plated with nickel to prevent the silver from leaching out, and wire leads can be attached to the terminations for subsequent soldering into a circuit. Frequently the finished MLCC is embedded in plastic to seal out moisture.

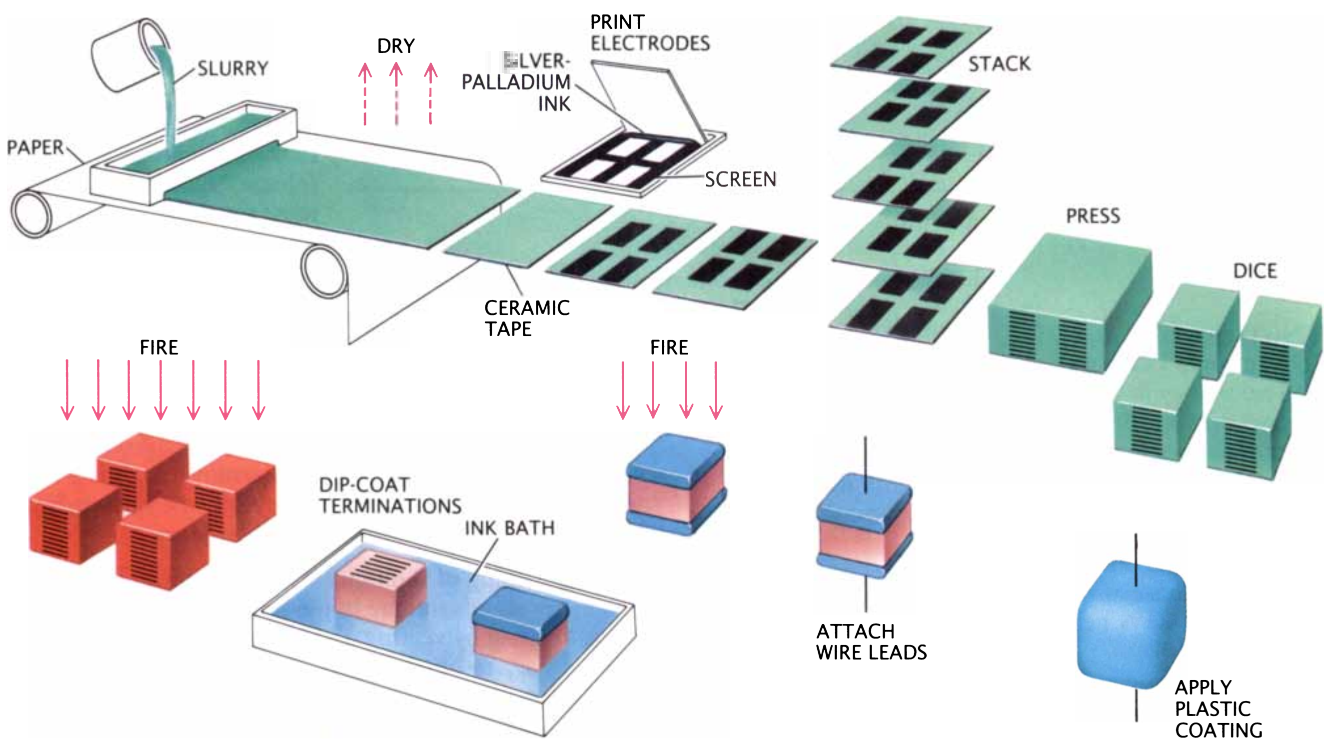
A slight variation in this procedure makes it possible to manufacture capacitors with lead-tin electrodes. The electrode patterns are printed on the green ceramic tape with an ink containing carbon and ceramic powders instead of silver and palladium; during firing the carbon burns out, leaving empty spaces with occasional supporting pillars of ceramic where the electrode patterns had been. Porous terminations are applied by dipping the capacitor ends in a silver ink containing a foaming agent and then firing the capacitors again. Electrodes are injected into the vacant sites by evacuating the air from the capacitors and immersing them in a bath of molten lead-tin solder at 320 degrees C. When the pressure in the chamber is raised to several hundred pounds per square inch, the solder is forced through the porous terminations into the electrode sites.

Lead-tin capacitors are somewhat more reliable and less expensive than silver-palladium devices, but because thicker ceramic layers are necessary to withstand the pressure of electrode injection, the capacitance per unit volume is also smaller. In either case the final product is a capacitor between an eighth and a quarter of an inch square, made up of ceramic layers from 20 to 40 micrometers thick. The metal electrode layers are usually from one to five micrometers thick. The total surface area of such a device can be as much as 100 square centimeters per cubic centimeter and, with ceramic layers in new capacitors becoming as thin as 10 micrometers, that figure promises to improve.

In spite of the considerable electrode surface area and the thin insulating layers of MLCC's, the main reason for their high capacitance per unit volume is the extraordinarily high dielectric constants of barium titanate and the other materials employed as insulators. Ceramics based on lead magnesium niobate, now com-



DESIGN OF AN MLCC is derived from the Leyden jar. If the jar were slit down the side and spread out flat, then cut into sections stacked inner electrode against inner electrode (*red*) and outer electrode against outer (*blue*), the jar would be transformed into interleaved layers of glass and electrode. The electrode layers could be reunited into two distinct electrodes by removing strips from all the inner electrodes on one side of the stack and from all the outer electrodes on the other, and then connecting the electrodes protruding at each end with terminations. This structure is apparent in an MLCC cross section (*bottom right*): a multilayer ceramic-electrode sandwich whose terminations are wrapped in nickel and solder.



MANUFACTURE OF MLCC'S with insulating layers less than a thousandth of an inch thick begins with a ceramic powder (mainly barium titanate) mixed with a solvent and a binder. The slurry is spread and dried to form a ceramic tape, which is then cut into sheets from six to eight inches square and printed with thousands of electrodes in silver-palladium ink. The electrodes are offset so that they are exposed at both ends on alternate sheets. From 30 to 60 of the sheets are stacked and diced, then fired at 1,000 to 1,400 degrees Celsius. The ends are dipped in a silver-palladium ink bath and the terminations are fired at a lower temperature. Finally wire leads are attached to the terminations and the capacitor is coated with plastic.

ing on the market, will exceed the dielectric constant of barium titanate by a factor of two or three while retaining the temperature stability of the current ceramic. Thus the tradition of optimizing dielectrics in MLCC's seems to be going strong.

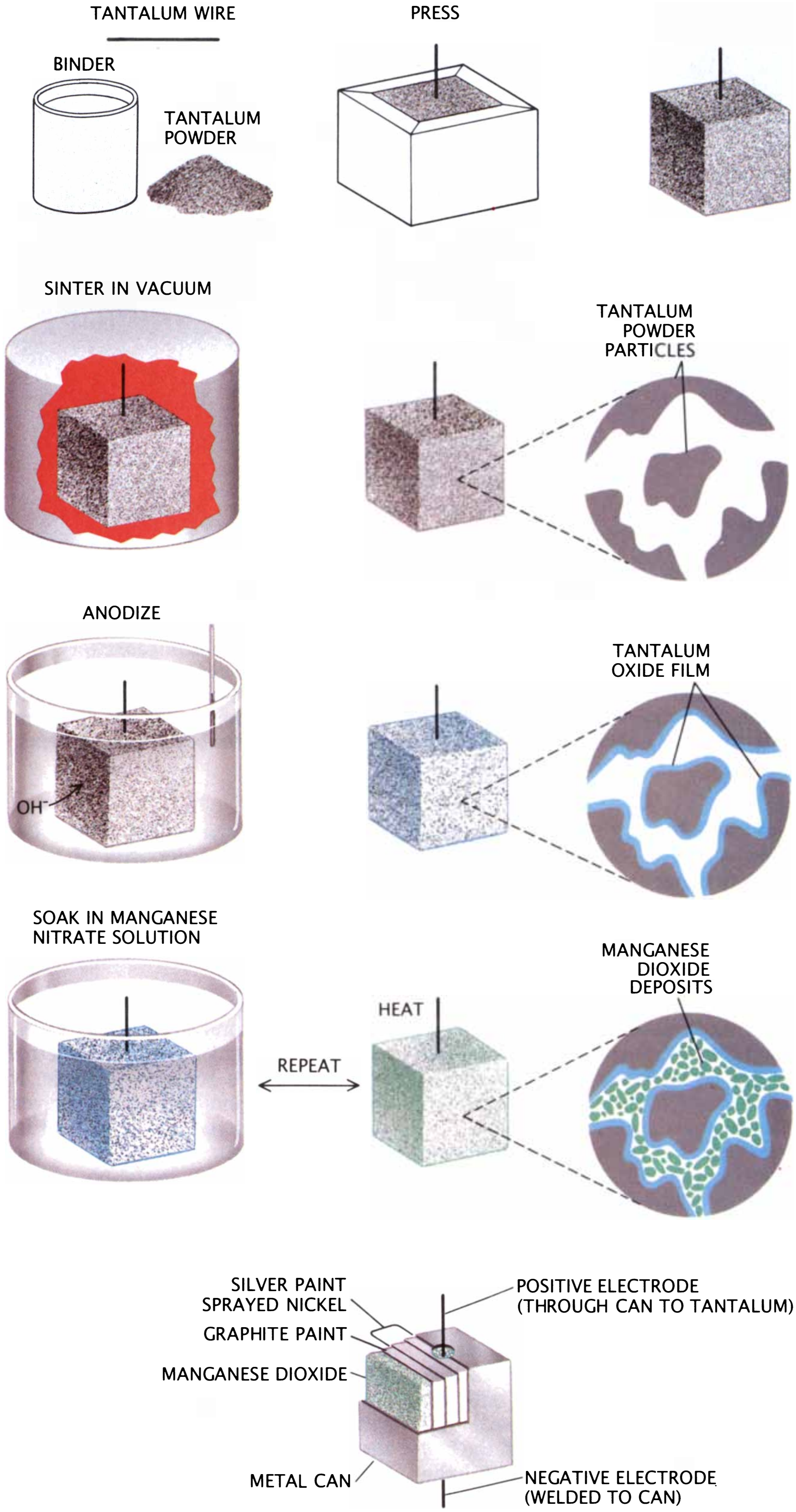
In contrast, the insulating layers in electrolytic capacitors have dielectric constants of between eight and 27—yet such devices exhibit capacitances larger than those of ceramic capacitors having the same volume. They owe their advantage to a combination of extremely large surface areas and extremely thin insulating layers. A typical electrolytic capacitor has a surface area of one square meter per cubic centimeter, giving it a surface-to-volume ratio 10 times that of an MLCC, and its insulating layers are 100 times thinner.

One can conceive of an electrolytic capacitor as a Leyden jar with very thin glass that has been crumpled into a very small cube. It is built up from a slug of metal that is 60 percent porous. In most modern electrolytic capacitors the slug is made from pulverized tantalum, a hard, gray metal. The tantalum powder is compressed into a slug and heated in a vacuum to temperatures near 2,000 C. for several hours, causing the tantalum particles to adhere strongly to one another. Nooks and crannies in the powder increase the surface area of the slug, which will serve as one electrode of the capacitor.

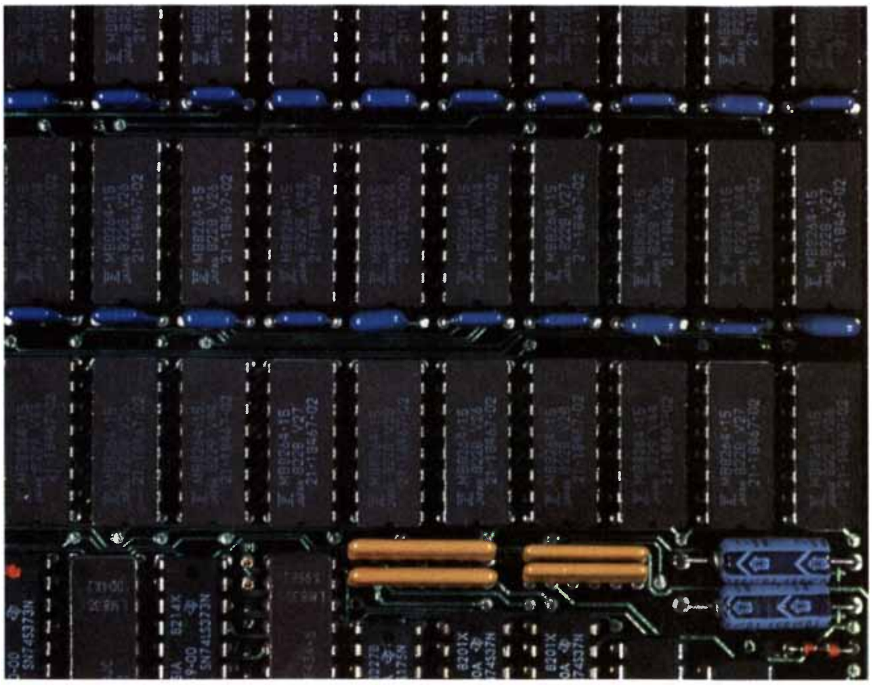
The easiest way to apply a continuous insulating layer to the irregular contours of the tantalum electrode is simply to alter the surface of the metal so that it becomes nonconducting. This can be done by oxidizing the slug surface to yield an insulating layer of tantalum oxide. In fact, such a layer will develop during the formation of the slug if the tantalum is heated in air rather than in a vacuum, but the layer is quite grainy and therefore would be prone to charge leaks that could destroy the device.

Instead, a smooth, glassy insulator can be formed by a process known as anodization. The tantalum slug is made to serve as the anode in an electrolytic cell: it is submerged in a dilute, conducting solution of acid or base together with a second electrode, and a battery is connected to both. The circuit is configured so that electrons are removed from the tantalum electrode and added to the other electrode. The conducting solution completes the circuit.

The solution soaks into the pores in the tantalum slug and comes in con-



MANUFACTURE OF TANTALUM CAPACITORS starts with a porous slug pressed out of tantalum powder and a binder. The slug is sintered to make the metal particles cohere, then anodized to produce an insulating film of tantalum oxide. Next the slug is immersed in manganese nitrate solution and heated, leaving deposits of semiconducting manganese dioxide in its pores. The manganese dioxide serves as one electrode and the underlying tantalum serves as the other. The capacitor is coated with graphite paint, silver paint and sprayed nickel, then packaged in a can.



COMPUTER MEMORY BOARD incorporates the two major classes of capacitors. The familiar dark brown rectangles are packages of integrated circuit chips; at the end of each package MLCC's coated with bright blue plastic provide diversion paths for spurious electrical signals. The striped blue devices in the corner are tantalum capacitors that buffer fluctuations in the power supplied to the board's components.

tact with its entire surface. As electrons are removed from the slug, it attracts oxygen-containing ions bearing a negative charge from the solution. The ions combine with the metal to form tantalum oxide. The oxidized layer grows thicker until it reaches a limiting thickness of between one-tenth and one-fifth of a micrometer. The layer is transparent, but because its optical thickness is comparable to half a wavelength of light, it imparts a rich color to the slug owing to interference effects [*see illustration at bottom right on page 87*]. The particular hue depends on the thickness of the layer, and so color is often used to gauge whether the tantalum oxide insulator has reached the desired thickness.

Next the insulator must be coated with a material that can function as the second electrode. Early investigators filled the slug's pores with a conducting liquid, but capacitors with fluid electrodes are hard to seal and vulnerable to freezing. In 1956 the Bell Telephone Laboratories invented the technique that is now generally used in making modern tantalum capacitors. The technique involves soaking the anodized slug in a concentrated manganese nitrate solution—a pale pink syrup. When the slug is subsequently heated to 300 degrees C., the water in the absorbed solution evaporates and the manganese nitrate decomposes, leaving deposits of manganese dioxide throughout the slug's interior. The process is repeated several times until the manganese dioxide layer attains sufficient thickness.

Manganese dioxide is a ceramic, but it conducts electricity well enough to serve as the electrolytic capacitor's second electrode. It has another desirable feature: it gives the capacitor some power to heal itself in the event of a break in the insulating layer of tantalum oxide. Such a failure would allow electrons to leak through the barrier between electrodes. If the break is small enough, the heat produced by the flow of electrons can raise the temperature in its vicinity and cause the manganese dioxide to decompose, freeing oxygen and a lower oxide of manganese. The free oxygen reacts with the exposed tantalum to form more tantalum oxide, patching the flaw in the insulator. Furthermore, the lower oxide of manganese produced in the vicinity of the break is not a good conductor and thus helps to isolate the flaw from the rest of the electrode.

Once the layers of insulator and electrode are complete an electrolytic capacitor must be packaged. The slug is sprayed with nickel and painted with graphite and silver to protect and smooth its fragile, irregular contours. It is then wrapped in a snug metal can. A wire lead connected to the can makes contact with the manganese dioxide electrode by means of the nickel, graphite and silver layers, and another wire embedded in the slug is connected to the tantalum electrode. The capacitor is ready to be soldered into a circuit.

In spite of their superior capacitance per unit volume, electrolytic devices are not suitable for every application. For one thing, the polarity of the charge fed to an electrolytic capacitor cannot be reversed. Whereas the electrodes in an MLCC can store charge of either polarity, in the electrolytic capacitor the positive charge must always be stored on the metal electrode, since that is the polarity in which it was anodized. The practical upshot of this inflexibility is that electrolytic capacitors are generally used only with direct current, whereas MLCC's can be used with alternating current. Electrolytic capacitors are also more subject to sparking than ceramic capacitors because their insulating layers are so thin. Consequently MLCC's are preferred in applications involving relatively high voltages.

The two kinds of capacitors differ in other details of their electrical responses, and the circuit designer must take all of them into consideration. But both kinds are shining examples of the way elegant materials and processes have been combined to achieve a high degree of miniaturization in a device that is still, fundamentally, a Leyden jar. Driven by the demands of the integrated-circuit industry, capacitor manufacturers are bound to improve dielectric constants, shave insulator thicknesses and expand electrode areas in order to reduce the size of capacitors still further. Tomorrow's inventions are likely to make today's capacitors seem as cumbersome as the Leyden jar.

FURTHER READING

DIELECTRIC MATERIALS AND APPLICATIONS. Edited by Arthur R. von Hippel. John Wiley & Sons, Inc., 1954.

DIELECTRICS AND WAVES. Arthur R. von Hippel. John Wiley & Sons, Inc., 1954.

PRINCIPLES AND APPLICATIONS OF FERROELECTRIC AND RELATED MATERIALS. Malcolm E. Lines and Alastair M. Glass. Oxford University Press, 1977.

DEVELOPMENTS IN ELECTROLYTIC CAPACITORS. Walter J. Bernard in *Journal of the Electrochemical Society*, Vol. 124, No. 12, pages 403C–409C; December, 1977.

How the Human Eye Focuses

As people age, their ability to focus on nearby objects gradually declines. Several probable causes have been identified, including changes in the eye's geometry and biochemistry

by Jane F. Koretz and George H. Handelman

It happens to almost everyone: by about the age of 45 it becomes impossible to read without the help of glasses. How does the healthy young eye focus on a nearby object? Why does near vision fade? The answer to the first question has long been incomplete, and the answer to the second remains a matter of conjecture. By means of photographic studies of the lens of the eye and mathematical modeling, we have recently gained new insight into both problems. We have shown that several processes conspire to progressively limit one's ability to focus on close objects; other processes counteract the decline for a while, but these ultimately fail—typically in the fifth decade of life.

When a person looks at something, light reflected from the object passes through the cornea (a transparent sheath across the front of the eye) and a fluid known as the aqueous humor, on through the pupil of the iris and into the lens, which is normally clear and is shaped and oriented something like the lens of a camera. From there the light travels through the gel-like vitreous body to the retina, the part of the eye that converts light into electrical signals that are transmitted to the brain for interpretation.

In order for the image to come into focus, the light must be bent so that the rays converge at the fovea, the center of the retina. The nearer something is to the eye, the more the light must be bent if the object is to be seen clearly. The cornea, aqueous humor and vitreous body each have a fixed refractive power, or ability to bend light, but the lens can accommodate: it can sharpen the curvature of its front and back surfaces, thereby increasing its focusing power.

JANE F. KORETZ AND GEORGE H. HANDELMAN are colleagues at the Rensselaer Polytechnic Institute. Koretz, who earned her Ph.D. in biophysics from the University of Chicago, is associate professor of biology and a member of the Center for Biophysics. Handelman holds the Amos Eaton Chair at Rensselaer. He received his doctorate in applied mathematics from Brown University and served as chairman of the department of mathematics and dean of the School of Science at Rensselaer for 18 years. The authors dedicate this article to the memory of Henry N. Fukui of the National Eye Institute.

What, then, enables the lens to accommodate? A broad explanation, which our work and that of others generally supports, was put forward in the mid-19th century by the German physicist Hermann von Helmholtz in his *Treatise on Physiological Optics.* Helmholtz noted that the lens is suspended by filaments that project from the so-called ciliary muscle, which encircles the equator, or rim, of the lens like a collar but does not come in direct contact with it. These nonelastic filaments, called the zonules of Zinn, are now known to form three rings of hairlike "spokes" around the lens; one ring attaches to the equator and the other two attach somewhat in front and in back of it.

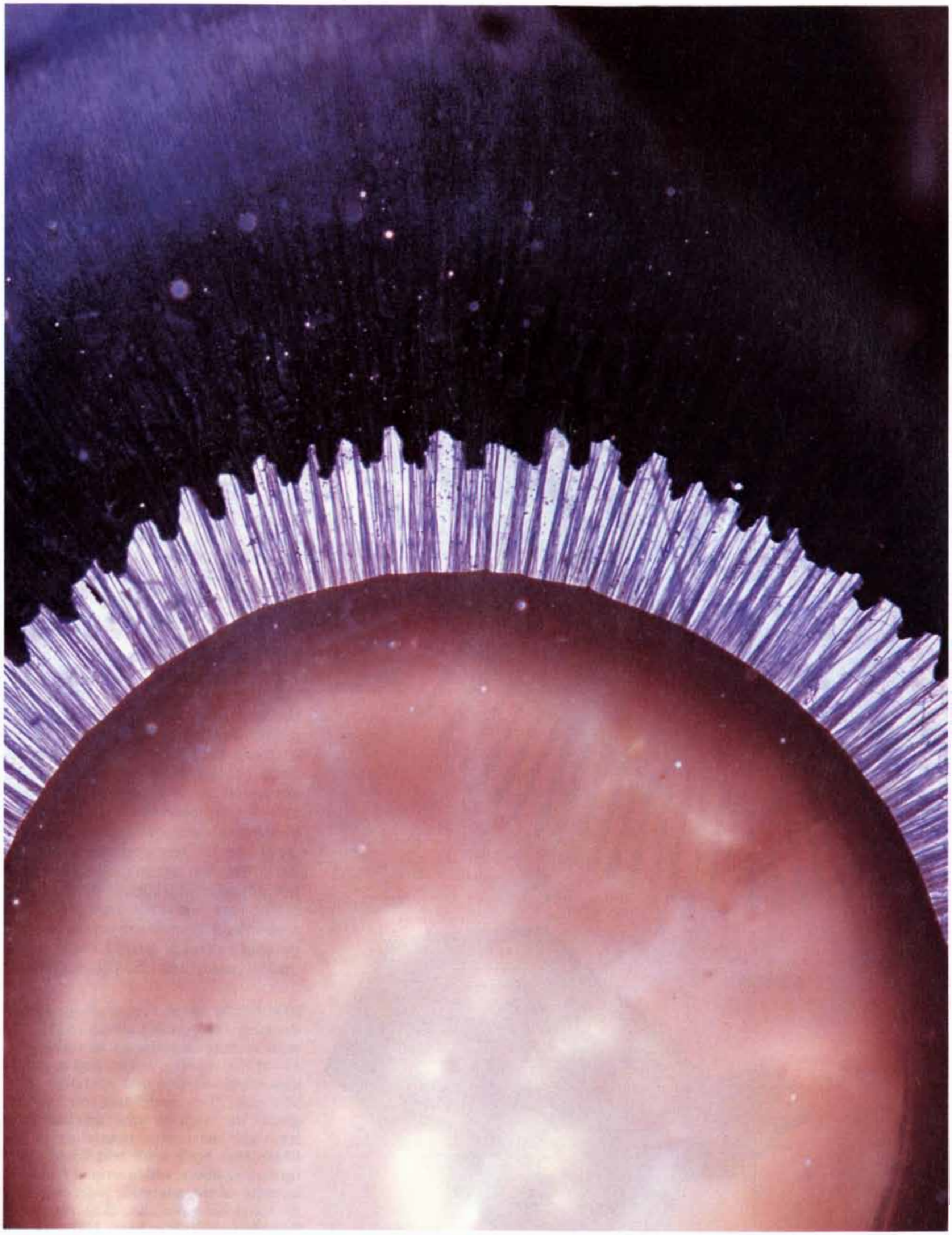
Helmholtz proposed that when the eye is focused on infinity (which for human beings begins about 20 feet away), the sphincterlike ciliary muscle relaxes and therefore expands; the diameter of the circular muscle reaches a maximum. As the muscle expands it pulls the zonules taut, causing them in turn to pull on the lens. The pulling flattens the front and back of the lens and increases the diameter of its equator. In this condition—called the unaccommodated state—the ability of the lens to bend light is at a minimum. The combined refractivity of the cornea, aqueous humor, unaccommodated lens and vitreous is just right for focusing an image of a distant object on the fovea.

When the eye attempts to focus on a point closer than 20 feet away, the ciliary muscle contracts, reducing the diameter of its opening and also causing the muscle to move slightly forward. Both changes reduce the stress on the zonules and thereby lessen the stress exerted by the zonules on the lens. The lens thereupon undergoes elastic recovery: much as a foam-rubber ball expands after it has been compressed, the lens rebounds to a more relaxed state. As the lens focuses on progressively closer objects, it becomes thicker from front to back, its surfaces become more sharply curved and the diameter of the equator shrinks. This relaxation process is precisely controlled to provide the exact degree of extra refractive power needed for focusing on objects closer than 20 feet away.

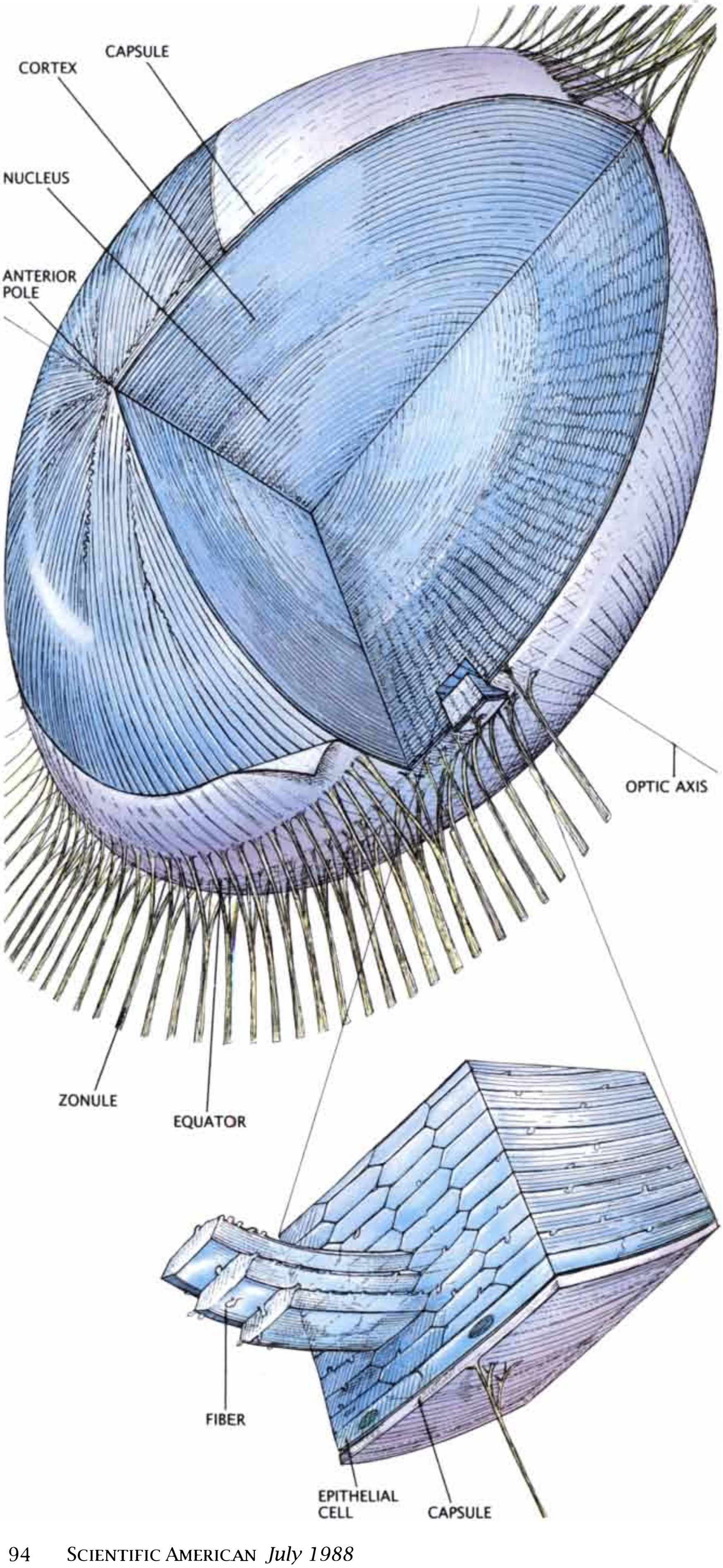
The lens, then, is unaccommodated—flattest and the least refractive—when it is under maximum stress: when the eye focuses on infinity and the ciliary muscle is totally relaxed. The lens is maximally accommodated—most sharply curved and most refractive—when it is under the least stress: when the eye focuses on the closest discernible object and the ciliary muscle is fully contracted.

Helmholtz' accommodative model is widely accepted today, but it leaves many questions unanswered. For instance, what effects do small changes in the pattern of stress exerted by the zonules have on lens shape? How much relaxation of the force exerted by the zonules is needed to produce enough lens curvature for, say, reading, and at what angle must the zonules meet the lens? Does the vitreous, to which Helmholtz paid little attention, have a role in accommodation?

Moreover, Helmholtz conceived of the lens as a readily deformable bag of fluid. Actually the material within the "bag" consists of long, ribbonlike fibers that interlock and also nest within one another like the layers of an onion. The bag itself, the lens capsule,



LENS OF HUMAN EYE is suspended by filaments known as the zonules of Zinn (*purple "threads"*) from the ciliary muscle, a ring of tissue essentially concentric with the lens. In this frontal view the muscle is not visible; it is embedded in the ciliary body (*dark, ruffled region*). The ciliary muscle and the zonules exert major control over the ability of the lens to accommodate: to change shape so as to increase its focusing power. The accommodative facility usually is lost by the age of about 45. Patricia N. Farnsworth of the University of Medicine and Dentistry of New Jersey in Newark made the photograph.



is composed of fibers of a different kind, oriented parallel to the lens surface. Both kinds of fiber, which are rich in protein, resist stretching but not bending. These findings raise an additional question: What effect does the structural organization of the lens have on accommodation?

Someone searching for answers to such questions would ideally examine the zonules and the vitreous directly in the living eye, measuring the magnitude and direction of the forces they exert on the lens capsule and the lens body. The investigator would also measure the distribution of forces throughout the lens to determine the effect of the internal structure on the response to external stresses. In reality, of course, it is impossible to make such direct measurements.

It is possible, however, to describe changes in the shape of the living lens as it accommodates and to calculate the magnitude and direction of the forces that would have to act on the lens in order to produce the described changes. This information can then be correlated with what is known about the architecture of the eye to determine which structures are capable of exerting the calculated forces.

That is the approach we adopted. We began by creating a mathematical model of the lens body—the lens minus the capsule. For simplicity we concentrated on the lens of a young person about 10 years old. Such a lens can be represented quite accurately as a somewhat distorted sphere composed of two hemispheres of differing radii, one representing the front of the lens and the other the back. We then made certain assumptions in our model about the elasticity of the lens body, for instance that it responds differently to stress exerted parallel to and at a right angle to the optical axis of the eye. The model also

STRUCTURE of the lens is shown in this drawing. The zonules attach to the capsule, or outer membrane, in three places—at the equator, or rim, and at points somewhat anterior and posterior to it. The inner "lens body" has two main regions: the nucleus (the original fetal lens) and the cortex, consisting of the fibers that have been laid down since birth. The fibers, which extend from the anterior to the posterior pole, originate as epithelial cells (*detail*) at the outer boundary of the lens body. With time the cells elongate into ribbons, lose their nuclei and are covered by newer cells; as a consequence the lens grows thicker. The lens fibers nest within one another much the way the layers of an onion do.

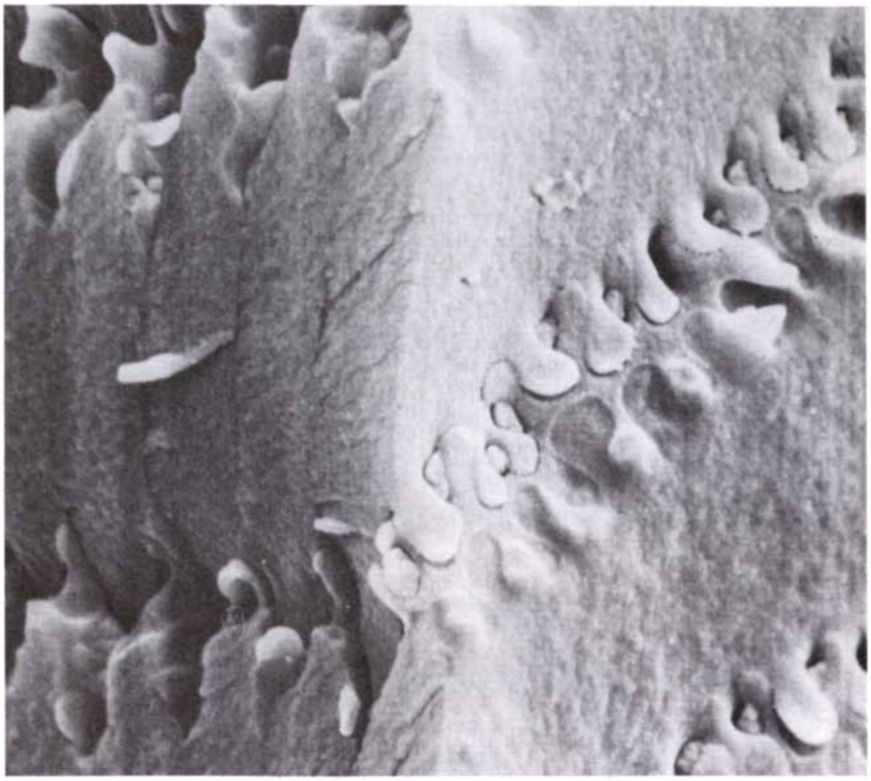
assumed that, because they are interlocking, the lens fibers cannot slide past one another: that the only way the lens can change shape is for the fibers to alter their curvature.

Our calculations indicated that all the forces acting on the surface of the accommodating lens body are approximately equal and are in a direction approximately perpendicular to the lens surface. This finding suggests that the capsule, which is subject to discretely applied stresses from the zonules and is the only material in direct contact with the lens body (the zonules do not penetrate the body), transforms these discrete stresses into a uniform compressive force against the entire surface of the lens. When the zonular stress on the capsule is released, the compressive force acting on the lens body is also released; the compressive force within the lens is reduced as well, and the lens undergoes elastic recovery. The finding that the force exerted by the capsule is perpendicular to the surface of the lens body is not entirely surprising. The zonules exert a force that has both a parallel (stretching) component and a perpendicular (compressive) one. The capsular fibers, however, resist stretching, and so only the perpendicular force is transmitted to the lens.

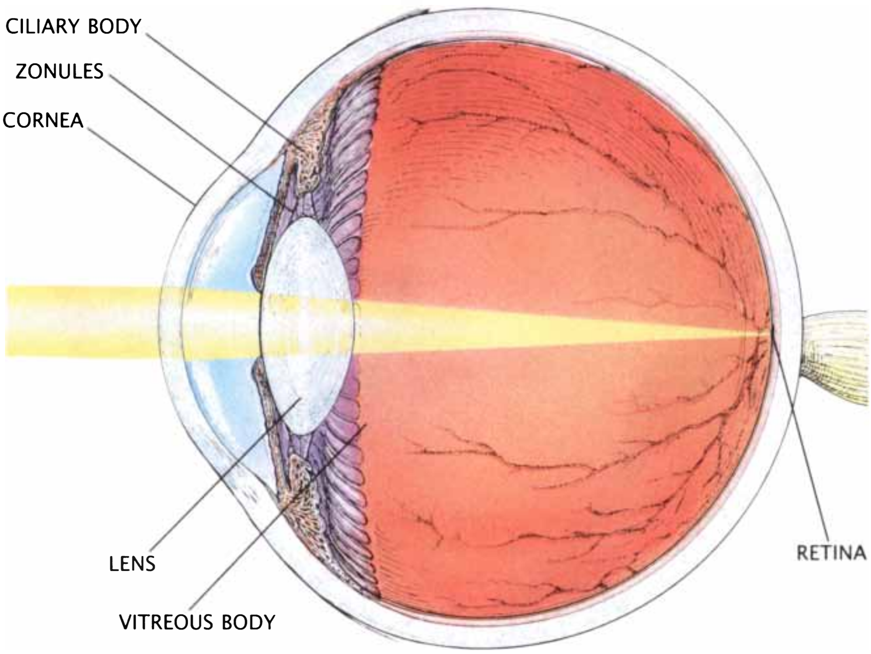
Another important implication is that the shape assumed by the lens at a given accommodation state is indeed different from the shape it would assume if it were merely a bag of fluid. Under uniform pressure on its surface a bag of fluid could not achieve the known curvatures of the accommodated lens. The observed changes in shape must therefore be influenced by the internal structure of the lens.

We were now ready to determine the forces exerted on the lens by other structures in the eye. Relaxation of the zonules can account for most but not all of the shape change seen in the lens during elastic recovery. This suggests that the vitreous humor may take part in the focusing process—by lending support to the posterior of the lens. Interestingly, at about the age when most people need reading glasses, the gel-like vitreous begins to liquefy, losing its ultrastructure. Exactly how such liquefaction would contribute to the loss of accommodative ability is not clear, but it may have some role to play.

The only way the zonules could produce the calculated change in forces during accommodation would be to lose tension themselves or to change the angle at which they apply stress to



LENS FIBERS, here enlarged about 8,000 times, interlock along their long edges, as is seen in this electron micrograph made by Richard G. Kessel of the University of Iowa. Each fiber is also linked to those above and below it, as is shown in the illustration on the opposite page. The interlocking of the fibers affects the distribution of forces within the lens and constrains the shape it adopts during accommodation.



EYE FOCUSES on an object by refracting, or bending, the light (*cone*) reflected from the object so that the light rays converge on the retina. Nerve cells in the retina transform the light into electrical signals that are transmitted to the brain for interpretation. Light is bent at the front and back surfaces of both the cornea and the lens, but only the lens can accommodate. Excessive refraction causes light to converge in front of the retina and thus impairs distance vision, whereas insufficient refraction causes light to converge behind the retina and impairs near vision. A progressive decrease in the refractive nature of the lens and in the ability of the lens to accommodate is thought to undermine near vision in middle age.

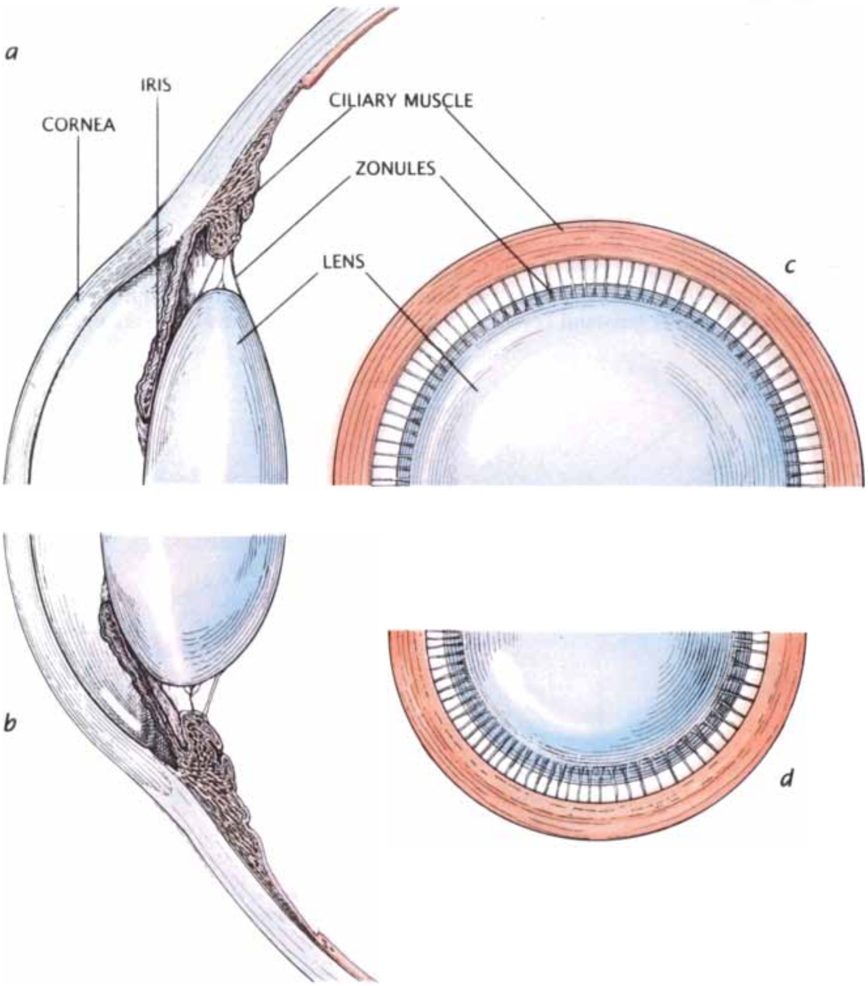
the lens in such a way that the compressive component of the stress is decreased. Our data suggest that they do both. In addition to becoming less taut, the zonules, which are attached to the capsule, move slightly, becoming more parallel to the surface of the capsule; the more parallel the zonules are, the less compression they cause in the lens. Experimental evidence from other investigators confirms that in at least one individual (who has a rare disorder of the iris that makes the normally obscured zonules visible through a microscope) the zonules hang quite loosely when the lens is maximally accommodated. Our calculations suggest that the forces acting on the lens never decline to zero, however; if they did, there would be no way to hold the lens in position.

The model based on the young eye enabled us to understand many of the details of accommodation, but it did not explain why the refractive power of the eye, and hence the ability to focus on nearby objects, declines with age. The eye's refractive power is measured in diopters, the reciprocal of the distance in meters between the eye and an object. For instance, an eye with a refractive power of 10 diopters can bend light sufficiently to focus on an object a tenth of a meter, or roughly four inches, away. In human beings who start off with normal vision the refractive power declines from about 14 diopters at the age of 10—when it is possible to focus on the tip of one's nose—to about nine diopters at 20, four diopters in the mid-thirties, one or two diopters in the mid-forties and close to zero by the age of 70. A zero-diopter eye cannot focus on anything nearer than infinity. The change from four diopters (roughly 10 inches) to two (roughly 20 inches) or less is the one people notice because it affects reading; most people hold books from 12 to 16 inches away from their eyes.

As a first step toward understanding the cause of this gradual age-related decline in near vision, we decided to gather as much information as possible about the way the lens changes shape with age and accommodative state. Having done this, we would attempt to determine both the effect of such changes on refractive power and the likely causes of the changes. We began by examining a series of cross-sectional photographs of the lens made by Nicholas Phelps Brown of the Institute of Ophthalmology in London in the early 1970's and by preparing 100 similar sets of photographs of our own. Brown, who gave us technical advice, provided four photographic sets from subjects 11, 19, 29 and 45 years old, who focused on objects at varying distances from the eye; our study included subjects ranging in age from 18 to 69 who had healthy eyes and normal distance vision. All the cross sections are vertical "slices" running from the front to the back of the lens and were made with the aid of a slit lamp, which shines a narrow ("slit") beam of light into the eye.

A quick look at the photographs confirmed the well-known fact that the size of the lens increases as a function of age. The unaccommodated lens of the infant has been found to be about 3.3 millimeters thick from front to back. As time passes, the cells that constitute the outer layer of the lens body (the epithelial cells lying just inside the capsule) grow and are transformed into the kind of ribbonlike fibers that constitute the bulk of the lens. As new epithelial cells are laid down over the older fibers, they undergo the same growth process as their predecessors, and so the lens thickens. By the time a person is 70 the unaccommodated lens can be as much as five millimeters thick.

The photographs also revealed a series of bands in the interior of the lens known as zones of discontinuity. The bands in the front of the lens exhibit roughly the same curvature as the front surface, and the bands in the



LENS INCREASES its focusing power by becoming more sharply curved. When the initially unaccommodated, or flattened, lens is viewed from the side (*a*), it looks fairly thin from front to back. It thickens as it accommodates (*b*); the front surface moves closer to the cornea but the back surface remains in place. The change in shape is effected primarily by the contraction of the ciliary muscle. A front view shows that the lens adopts the unaccommodated state (*c*) when the muscle expands so that its diameter is at a maximum. The expansion of the muscle pulls the zonules taut, and they pull on the lens and flatten it. When the muscle contracts (*d*), the zonules relax, and the lens rebounds into a rounded state, much as a foam-rubber ball rebounds after compression. The illustrations of accommodation are exaggerated for clarity.

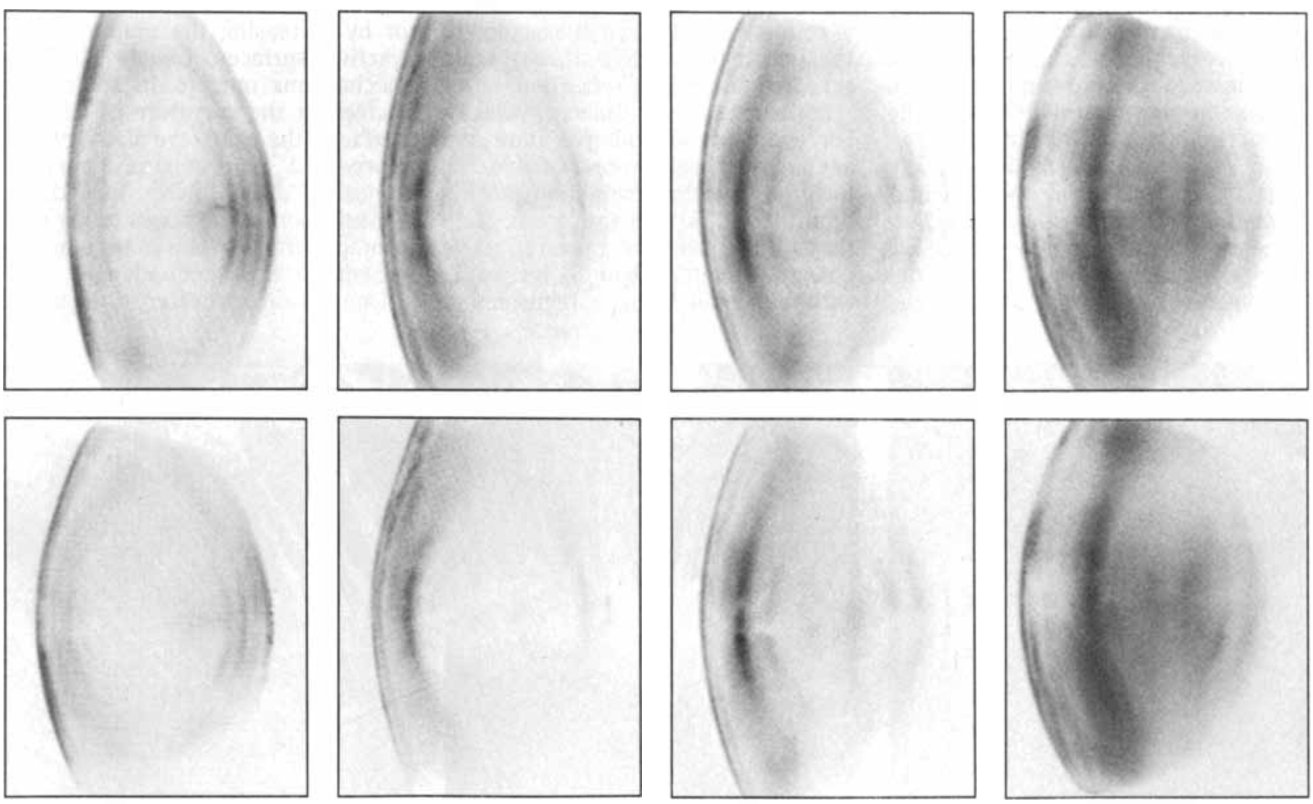
back are similar to the posterior surface, although the curves become progressively sharper as they approach the nucleus, or core, of the lens. In the young lens the bands are few and faint. As the lens ages, the number and clarity of the zones increase until sometime in or soon after the fifth decade, when the zones merge.

On the basis of the photographs and other sources of information, we created another model. This time we described the surface and interior curves of each lens in mathematical terms. (They were parabolas and hence could be re-created with simple equations.) Then we extrapolated from the cross sections to determine the shape of the entire lens in all accommodative states as a function of age. This enabled us to make extensive comparisons and to track the movement of selected points in the lenses during focusing.

The comparisons revealed that the unaccommodated lens adopts an increasingly curved shape as it grows in size throughout a life span. Moreover (as Brown had discovered with surprise in his subjects), in people younger than about 45 the older the lens was, the more sharply curved it was for any given accommodation state. For example, a 33-year-old lens curved more than a 19-year-old lens focusing at the same close distance. Brown called the phenomenon of increasing curvature with age the lens paradox, because one would expect more sharply curved lenses to produce a greater degree of refractive power than less curved ones.

When we traced the movements of selected points on the lens surface and in the interior as the eye accommodated, we found a parallel effect: equal amounts of movement during accommodation produced less change in focus in the older eyes than in the younger ones. Considered in another way, the data showed that in order to achieve, say, a one-diopter increase in focusing power, the points on an older lens had to move more than those on a younger lens.

A critical finding was that the total range of movement declines gradually with age. Indeed, in subjects older than about 45 the front of the lens was unable to change shape and so could not accommodate: it appeared to be locked in the unaccommodated state. This suggests either that the front zonules become progressively less able to relax when the ciliary muscle contracts in the attempt to accommodate, or that the zonules relax but become less able to exert much influence on the lens, or both. The zonules might be unable to relax if the front of the enlarged lens is so far from the ciliary muscle that the lens pulls the zonules taut. Moreover, just as the zonules meet the lens at a different angle when the lens is accommodated than when it is unaccommodated, so they meet the lens at a different angle when the lens is thick from aging than when it is younger and thinner. Ultimately the filaments may



SOME AGE-RELATED CHANGES are readily apparent in slit-lamp photographs of unaccommodated (*top*) and maximally accommodated (*bottom*) lenses from subjects (*left to right*) 19, 33, 45 and 69 years old. (The subjects' maximum focusing power is respectively 9, 4.5, 1 and .25 diopters; lower numbers indicate less power.) The photographs show vertical, front-to-back cross sections; the front of the lens is to the left. The changes include increased growth and curvature with time and, in the 45- and 69-year-old lenses, an almost complete failure to accommodate. Also apparent are dark bands called zones of discontinuity. As the lens ages, the bands multiply and become more prominent; when the lens is more than about 45 years old, the zones merge. The authors suggest that the increasing thickness of the lens and a rise in the fraction of insoluble protein in the zones contribute to an age-related decline in the refractive nature of the lens. For a time the increasing curvature of the lens may help to compensate for the decline; so might the extra refractive surfaces provided by the zones. Eventually, however, compensatory mechanisms fail—probably at about the time the lens ceases to accommodate.

come to exert a force that is tangential, or nearly so, to the surface of the lens. Once this stage is reached, relaxation of the zonules would have little effect on the shape of the lens, which in turn would exhibit little or no elastic recovery.

In other words, presbyopia—the name given to the age-related loss of near vision—appears to be a geometric disorder, stemming to a large extent from a change in the size and angular relations between the lens and the zonules.

We think geometry explains why the lenses of people older than 45 or so can no longer accommodate, but what explains Brown's paradox? Why is it that an older lens has to be more curved than a younger one to focus on the same object? One possibility is that the nature of the cytoplasm in the lens fibers changes in a way that decreases the lens's refractive index—a measure of a material's ability to refract light. If the refractive index declines with age, then the decline would solve the paradox: the increased curvature would not increase the focusing power of the lens; instead it would simply compensate somewhat for the decline in the refractive nature of the medium.

To explore this possibility, we undertook computerized "ray tracing" experiments that simulated the passage of light through every eye we had photographed. We did this by describing the factors that influence the light's trajectory, such as the curvature of each lens in accommodated (well-focused) states, the curvature of the cornea and the spacing between the cornea and the lens and between the front and the back of the lens. (Along with curvature, the distance between refractive surfaces influences the path of the light; closely spaced surfaces bend light more than widely spaced ones.) We then assigned to each part of the eye a refractive index—one that is generally accepted in the literature—and instructed the computer to bend the light at the boundaries between materials having different indexes of refraction. If the overall index of refraction of the lens remained constant throughout life, the simulations would have indicated that the light passing through each eye was bent in a way that focused an image on the retina.

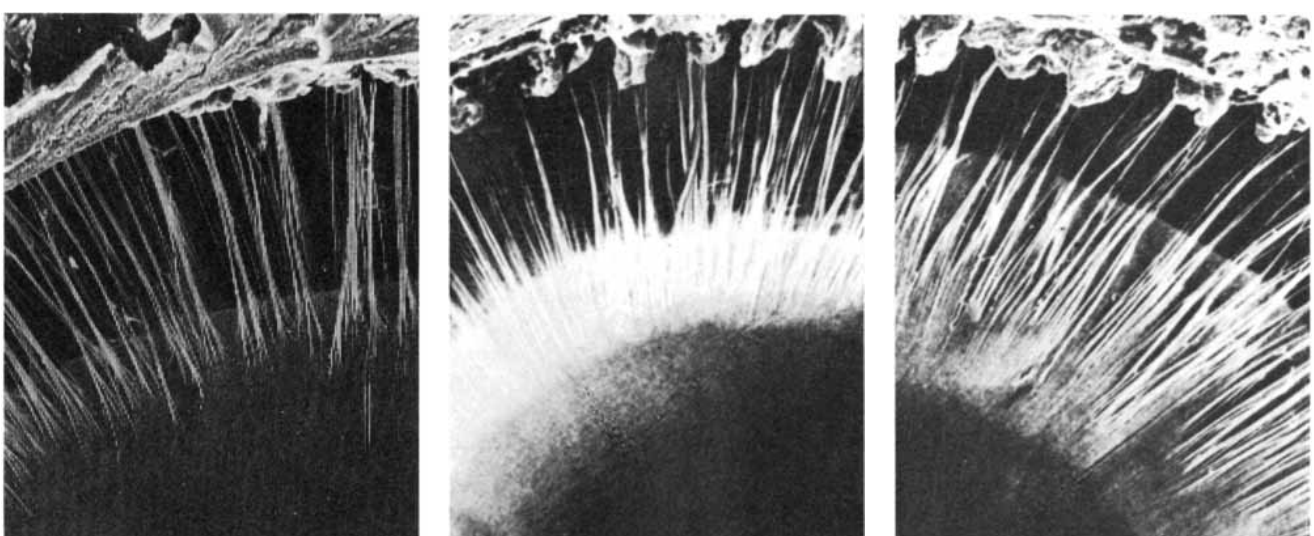
In the simplest simulation we had the light bend at the front and back of both the cornea and the lens. In another simulation we treated the boundaries between the nucleus and the outer "cortex" of the lens as additional refractive surfaces and assigned to the two regions different, but again accepted, indexes of refraction.

Both simulations failed abysmally. For every accommodative state and age at which data were available, the focal point of the simulation was behind the retina, as if the lenses of all the subjects had too little refractive power. Something had to be wrong with the model. Perhaps we needed to include each zone of discontinuity as yet another refractive surface.

We ran another set of simulations in which we kept the overall refractive index of the lens constant but allowed the indexes of the zones to vary. This time the simulation allowed the eye of a 45-year-old subject to focus, but the simulations for younger subjects remained underfocused. Moreover, the younger the subject was, the more the simulated eye was underfocused. The results suggest that the overall index of refraction for young eyes is actually somewhat higher than the values in the literature and that the younger the subject is, the more the index varies from the accepted value.

Looked at in another way, the results showed that the index of refraction of the lens decreases with age. This finding fits with the age-related changes in the lens discussed above. If the refractive index of the lens material decreases with age, the only way the reduction could be lessened would be by increasing the sharpness of curvature of the surfaces of the lens or the zones of discontinuity within the lens, or by increasing the number of refractive surfaces. Clearly all these mechanisms operate. In addition to increasing the curvature of the lens surfaces, the aging eye also develops more—and more sharply curved—zones of discontinuity. Indeed, the contribution of the zones to the eye's overall refractive power becomes increasingly important with age.

One remaining mystery is the nature



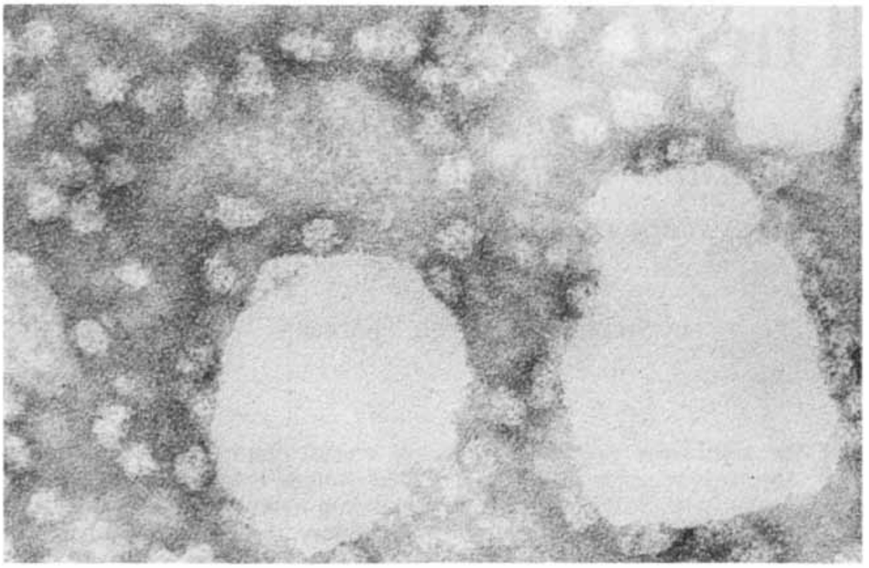
ATTACHMENT SITES of zonules on the front surface of the lens change as the lens ages and grows. In a 17-year-old (*left*) the zonules are close to the equator, but they move progressively onto the face of the lens in a 46- and an 85-year-old (*middle and right*). As the zonules shift, so must the angle formed by the lens surface and the filament. The authors propose that the lens ultimately loses its ability to accommodate in part because the zonules are almost parallel to the face of the lens. A release of tension on such zonules can have little effect on the lens. The photographs were made by Farnsworth, who was the first to propose that changes in the lens-zonule geometry could contribute to the failure of aged eyes to accommodate.

of the zones of discontinuity that are so apparent in slit-lamp images. Ultrastructural examination of the lens provides no evidence that they exist. Furthermore, if protein concentration is measured as a function of distance from the lens surface, one finds a small but smooth increase rather than the alternating increases and decreases that might be expected if the dark bands resulted from differences in protein concentration. Why, then, are these distinct regions clearly visible in photographs?

The answer, we concluded, had to lie in the nature of the photographic technique, which transmits the slit beam into the eye and produces an image from the light that is reflected into a camera located to the side of the original light source. (The film in the camera is tilted to compensate for the angular distortion.) We reasoned that in order for the zones of discontinuity to appear in the photographs, the material in these regions would have to interact with light differently from the material in adjacent regions. In particular, one would expect to find dark bands in the photographs if the protein in the corresponding regions scattered light to a great extent. Is such scattering possible?

At first glance the answer appears to be no. In the eye alpha-crystallin, the major protein component of the lens, is generally thought to assemble into aggregates that are smaller than the minimum size that would cause scattering. Yet the size of such particles is known to increase in response to increases in temperature, and it is thought to change also in response to such environmental factors as small alterations in *p*H level or calcium concentration. Many vision scientists have also shown that although the protein concentration in the eye remains constant over the course of time, the fraction of protein that is insoluble—and therefore particulate—increases.

Such findings raise the possibility that, with time, alpha-crystallin may form larger aggregates than has been suspected. If this is the case, the presence of a significant number of large, insoluble particles of alpha-crystallin would help to explain not only the appearance of the zones of discontinuity in our photographs but also a phenomenon known as glare, in which intense light whites out the entire visual field. The phenomenon is particularly common in people in their forties and older, who are most bothered by it on a sunny day or when they have to contend with the headlights of oncoming traffic while driving at night.



ALPHA-CRYSTALLIN, the major protein in the lens, is usually present as flat sheets (*larger shapes*), but these can break apart to form small, insoluble spheres (*smaller spots*). The spheres in turn can join to form rodlike bodies that can adhere to one another. If the aggregates grew very large—as it is thought they may—they would scatter a great deal of light. Such scattering could account for the presence of the zones of discontinuity in the photographs on page 97. The insoluble nature of the particles could also account for the apparent age-related decline in the refractivity of the lens. The spheres shown here are about 11 billionths of a meter in diameter.

Our data suggest that the degree of scattering is minimal in the first few decades of life but slowly increases.

The increasing presence of insoluble protein particles could also explain why the index of refraction of the lens declines with age. The index of refraction of a solution (such as the cytoplasm in lens fibers) depends on the nature and concentration of its solutes, or dissolved materials. Adding soluble protein to an aqueous medium increases the medium's refractive index. On the other hand, if a large fraction of the protein is converted into large, insoluble particles, the refractive index will decline appreciably. Such may be the case in the lens.

Both microscopic and macroscopic factors must be invoked to explain why the nearest point one can see with clarity moves progressively farther away as time passes. In our conception, the increasing amount of insoluble lens protein, the growing size of the lens (and hence the increased spacing between the front and the back of the lens) and the concomitant reduction in the index of refraction tend to interfere with near vision. Closeup viewing is also increasingly hampered by the declining ability of the lens to accommodate, probably because of gradual changes in the geometry of the lens-zonule-ciliary-muscle complex. These processes are somewhat counteracted—at least for a time—by the development of new refractive surfaces (the zones of discontinuity), by the overall sharpening of lens curvature and by further sharpening during accommodation. Eventually, at about the time the vitreous liquefies, such compensatory mechanisms fail and the lens loses its ability to accommodate.

Perhaps someday investigators will learn enough to reverse or prevent the natural, age-related decline of near vision. For now, however, the need for reading glasses is—like death and taxes—inevitable.

FURTHER READING

ANALYSIS OF HUMAN CRYSTALLINE LENS CURVATURE AS A FUNCTION OF ACCOMMODATIVE STATE AND AGE. J. F. Koretz, G. H. Handelman and N. P. Brown in *Vision Research*, Vol. 24, No. 10, pages 1141-1151; 1984.

ON THE HYDRAULIC SUSPENSION THEORY OF ACCOMMODATION. D. Jackson Coleman in *Transactions of the American Ophthalmological Society*, Vol. 84, pages 846-868; 1986.

MODELING AGE-RELATED ACCOMMODATIVE LOSS IN THE HUMAN EYE. Jane F. Koretz and George H. Handelman in *Mathematical Modelling*, Vol. 7, pages 1003-1014; 1986.

A POSSIBLE STRUCTURE FOR α-CRYSTALLIN. Robert C. Augusteyn and Jane F. Koretz in *FEBS Letters*, Vol. 222, No. 1, pages 1-5; September, 1987.

The Chaco Canyon Community

This ancient Anasazi site contains the ruins of nine multistory dwellings connected to the outside world by an elaborate system of roads. Why were they built and who lived in them?

by Stephen H. Lekson, Thomas C. Windes, John R. Stein and W. James Judge

The northwest corner of New Mexico is a barren desert, spectacularly beautiful but far from hospitable. Vegetation is scarce and rainfall averages about nine inches a year, most of it in the form of intense and often destructive downpours in late summer. In winter the temperature often falls as low as 20 degrees Fahrenheit below zero; in summer it may reach 100 degrees F. It is hard to imagine life in such a harsh and unforgiving environment. Yet along a 10-mile stretch of Chaco Canyon, in the heart of the San Juan Basin, there are remains of what was a thriving community built almost 1,000 years ago by the Anasazi, ancestors of today's Pueblos. The community once bustled with activity: irrigation systems were built, fields were leveled and a system of roads was laid out in and around Chaco. Now there are only the remains of several hundred stone-wall buildings. Among them—and still partially standing today—are the ruins of nine multistory dwellings known as the Great Houses of Chaco.

For more than a century archaeologists have puzzled over these ruins and the people who built them. The massive size and planning of the Great Houses was unprecedented and unequaled in the contemporary Anasazi world. Their elaborate design and size suggests a prosperous, almost urban community and a resident population in the thousands. How can the presence of a flourishing community in an isolated desert canyon be explained?

In 1972 the National Park Service, in cooperation with the University of New Mexico, assembled a multidisciplinary team to address the Chaco Phenomenon, a name that refers to the perplexing cultural history of this settlement. By bringing together specialists from different backgrounds, it was hoped, an accurate and coherent picture of Chacoan history might be developed. We wanted to know, for example, how the people of Chaco supported themselves in such a harsh environment. We hoped to learn how the population changed over time and what interaction the Chacoans had with other prehistoric groups. Most of all, we wondered what purpose the Great Houses and their roads served in the lives of the Chacoans.

Constructed in periodic bursts of activity from about A.D. 900 to 1115, the Great Houses are clearly the centerpiece of Chacoan culture. There are nine of them: Peñasco Blanco, Pueblo Alto, Kin Kletso, Hungo Pavi, Pueblo del Arroyo, Pueblo Bonito, Chetro Ketl, Una Vida and Wijiji.

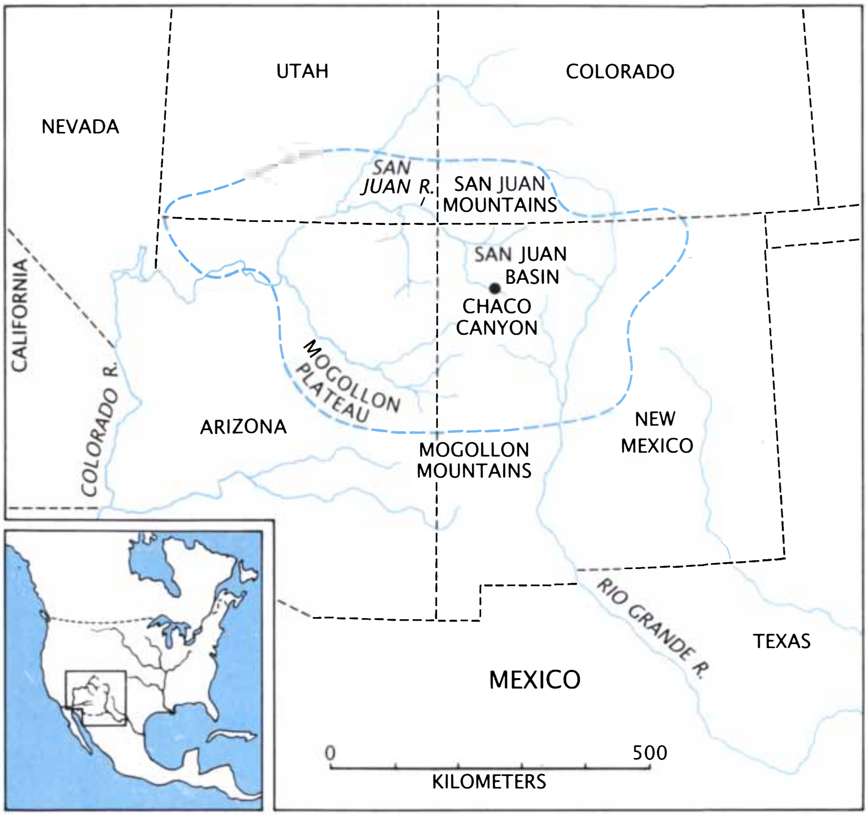
Of the nine Great Houses the most famous and most thoroughly studied is Pueblo Bonito. Situated at the base of a 100-foot-high mesa on the north side of the canyon, it has a distinctive D shape and occupies almost three acres of ground. Although little remains today of its upper floors, several sections that are four or five stories tall are still standing. When it was finally completed in the 12th century, it contained more than 650 rooms and was the largest of the Great Houses.

Pueblo Bonito and the other Great Houses in the canyon are constructed from blocks of sandstone, carefully coursed (horizontally aligned) in beautiful patterns. These masonry patterns, which are both intricate and period-specific, are a well-known feature of Chacoan architecture. The sandstone blocks fit together tightly, forming walls that are as much as a meter thick at ground level and decrease slightly with each successive story. The walls are far stronger than those at other Anasazi settlements. Some 45,000 kilograms of sandstone cut by hand from the surrounding cliffs, for example, were needed to build a single small room in one of the Great Houses; as many as 50 million pieces of sandstone may have been cut to build Chetro Ketl alone.

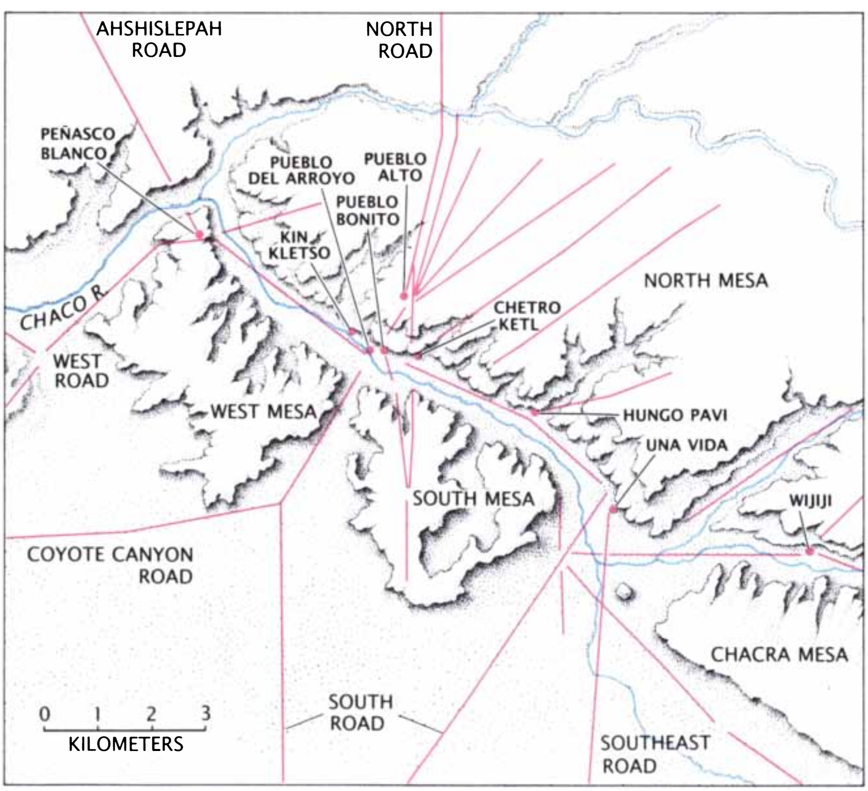
Overengineering, typified by the masonry walls, is a prevalent feature at Chaco Canyon. It presumably reflects the cultural importance of the Great Houses; it may also have meant that the buildings needed little upkeep and so could be maintained without difficulty by a small or intermittent population. The formal geometry of the

STEPHEN H. LEKSON, THOMAS C. WINDES, JOHN R. STEIN and W. JAMES JUDGE worked together on the archaeology of Chaco Canyon for more than 10 years. Lekson has a B.A. degree from Case Western Reserve University and got his M.A. in anthropology at Eastern New Mexico University in 1978. He is currently a visiting scholar at the Arizona State Museum in Tucson and expects to receive his doctorate from the University of New Mexico in 1988. Windes is a staff archaeologist with the National Park Service in Santa Fe. He has a B.A. from the University of North Carolina and an M.A. from the University of New Mexico. He is currently investigating the interrelation between Chacoan and Mesa Verdean cultures. Stein got a B.A. from the University of New Mexico in 1974 and is now an archaeologist for the Navajo Tribal Utility Authority in New Mexico. He is interested in the Chacoan road system and the extracanyon Great Houses. Judge is director of the Fort Burgwin Research Center at Southern Methodist University. He got his Ph.D. from the University of New Mexico in 1970 and was director of research at Chaco Canyon from 1977 to 1985.

PUEBLO BONITO, shown in this aerial photograph, is at the base of Chaco Canyon's north wall. One of nine Great Houses at Chaco, it was divided into more than 650 rooms, many of which may have served a purpose other than habitation. Great Houses are noted for their numerous circular ceremonial chambers called kivas. Two Great Kivas, which are characteristic features of Chaco Canyon, are visible here.



CHACO CANYON is at the center of the San Juan Basin in the northwest corner of New Mexico, where Utah, Colorado, Arizona and New Mexico come together. The region of the Southwest that was once occupied by the Anasazi is indicated by a broken line.



GREAT HOUSES of Chaco Canyon are connected to one another and to outlying communities by a system of roads that extends across the San Juan Basin. Where the roads encountered cliffs, scaffolding, wood ramps and stairs were built.

floor plans is another prominent feature of Chacoan architecture. Rooms and kivas (circular chambers believed to have had a ceremonial function) are arranged in gridlike patterns; even the doorways and air vents are constructed at regular intervals along each wall and from room to room and story to story. Such linear precision and complexity could not have been accomplished without advance planning and the active involvement of architects.

The sense of massiveness that pervades Chaco Canyon stems not only from the masonry itself but also from the amount of labor one imagines was required for the project. We estimate that in order to construct the floors and roofs of these multistory houses 215,000 trees, some of them 25 centimeters (10 inches) in diameter, were cut from forests as many as 80 kilometers away; how they were carried across the desert for such great distances remains a mystery.

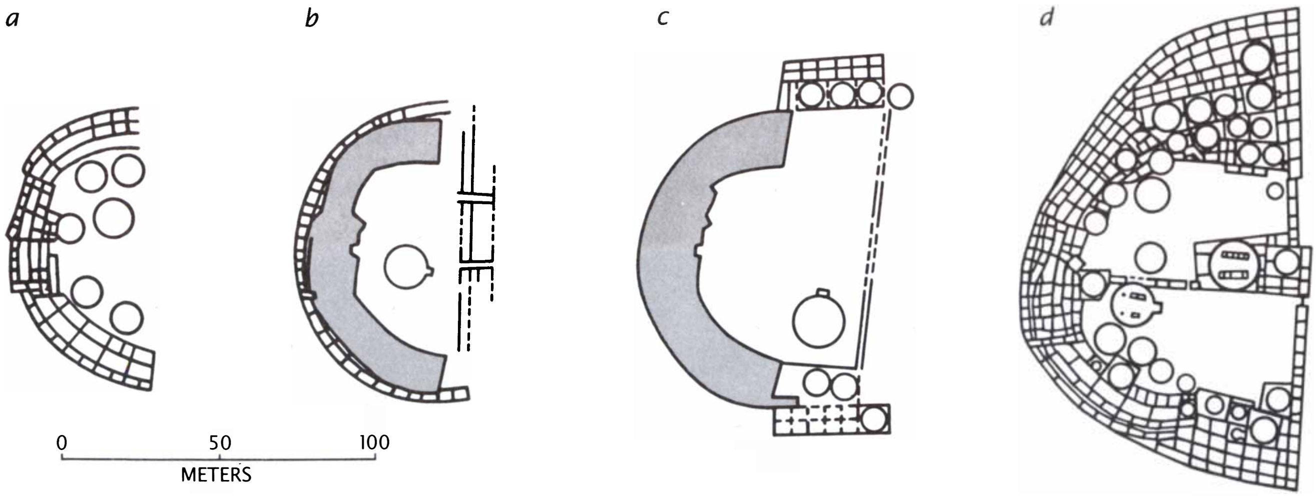
Clearly considerably more labor was needed to build the Great Houses than was needed for the smaller houses in the canyon. The Great Houses were much more labor-intensive even per unit area than other buildings of their day. Many of the actual construction techniques seen at Chaco were commonplace at smaller Anasazi settlements in the Southwest, but the large scale and grand design of the Great Houses is unique. Not only are the buildings taller, the walls thicker and the stone patterns more intricate but also the individual rooms are larger than rooms constructed elsewhere during the same period.

How many laborers were needed for construction of the Great Houses? We know (from wall abutments and masonry styles) that major sections of the houses were constructed as single projects. Did the Great Houses require a small army of laborers, as some have suggested? Based on calculations of the amount of stone, clay, water and wood beams needed to build a single room, we estimated that over a 10-year period the equivalent of about 30 men working between two and four months a year could have successfully completed a major construction project at Chaco. Although the work would have been arduous (carrying sandstone blocks from nearby cliffs and logs from a distance of 40 to 80 kilometers could not have been easy), we believe it did not require a resident population in the thousands or a large influx of outside laborers.

The Great Houses were clearly an integral part of Chacoan culture, and in order to understand their importance we set out to excavate one that had not yet been studied. We selected Pueblo Alto in part because it was unexcavated but even more because it appeared to be a terminus for several roads from the north, including the Great North Road. By studying Pueblo Alto we hoped to gain greater insight into the regional road system and its relation to the Great Houses of Chaco. Between 1976 and 1979 we excavated 10 percent of the site in order to investigate the different types of rooms and the activities associated with them. We were reluctant to excavate more than 10 percent, not only because the effort would have been considerable (Pueblo Alto covers more than two acres) but also because we wanted to disturb the site as little as possible.

CHETRO KETL appears almost camouflaged to an onlooker standing near the rim of the canyon facing west. The Great Kiva at Chetro Ketl, which is a little more than 18 meters (60 feet) in diameter, is clearly visible in the foreground.

SMALLER KIVAS, like those at other Anasazi sites, are abundant at Chaco. At Pueblo Bonito 37 have been excavated. Bonito was built opposite a gap in the canyon's south wall (*horizon at center right*), through which the South Road passes.



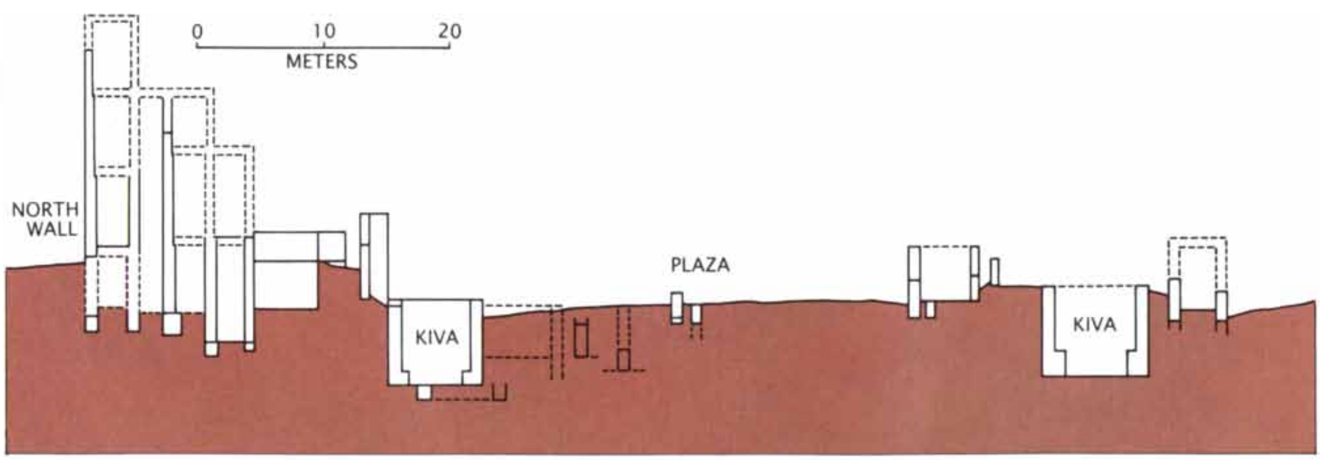
GREAT HOUSES were built in stages; the additions at each stage at Pueblo Bonito are shown. Older construction is in color. Pueblo Bonito was built in the 10th century as a semicircular suite of rooms and kivas (*a*); in the 1040's road-related suites were added along the north wall and a Great Kiva was built (*b*). In the 1050's and 1060's outer wings were added, razed and remodeled (*c*). It was completed by 1140, when more rooms, another Great Kiva and several smaller kivas were added (*d*).

Situated on top of the north rim of Chaco Canyon about a kilometer from Pueblo Bonito, Alto commands a vast 360-degree panoramic view of the San Juan Basin, a position that is unique among the Great Houses. It also differs from the other Great Houses in the canyon in having only one story. Unhindered by the confusion created by the collapse of multiple floors, we were able to obtain an unparalleled view of architectural organization and suite planning at the site.

By utilizing a variety of dating techniques we determined that most of Pueblo Alto was built between 1020 and 1060, although minor construction and remodeling continued into the early 1100's. Three different configurations of connecting rooms, called suites, were identified: household suites, big-room suites and road-related suites. Together they accounted for at least three-fourths of Pueblo Alto's floor area. To our surprise household suites, which consist of a living room connected to a storage room, represented a relatively small percentage of the total number of rooms at Pueblo Alto. We were able to determine that only five of the approximately 85 rooms originally constructed at Alto were for habitation and all of them were in the west wing of the building.

In two of the household suites we found stone- and adobe-lined hearths, storage pits, mealing bins for grinding corn and special niches for food. The presence of pollen and burned seeds from food plants indicated that food had been processed and eaten in those rooms. The floors, which had been replastered many times, were badly worn—presumably as the result of heavy domestic use. Indeed, the level of activity in those few rooms greatly surpassed that of other rooms we excavated at Pueblo Alto.

Once we had identified the key features of household suites in Pueblo Alto we were able to look for evidence of habitation in the rooms of other Great Houses. A similar pattern emerged: we found only five household suites at Pueblo Bonito, five at Pueblo del Arroyo and no more than 11 at Una Vida. All had apparently been constructed between 920 and



CROSS SECTION of Pueblo Bonito shows its layout and topography. Near the canyon's north wall Pueblo Bonito is thought to have been five stories tall; the number of stories decreased closer to the center. Some kivas shown here were excavated into the central plaza; others were built aboveground and the area around them was subsequently filled in with dirt.

1095 (a period that coincided with construction at Pueblo Alto). We estimate that in each of these Great Houses the household suites may have been home to no more than 100 people, a far cry from the 5,000 or more projected by earlier archaeologists.

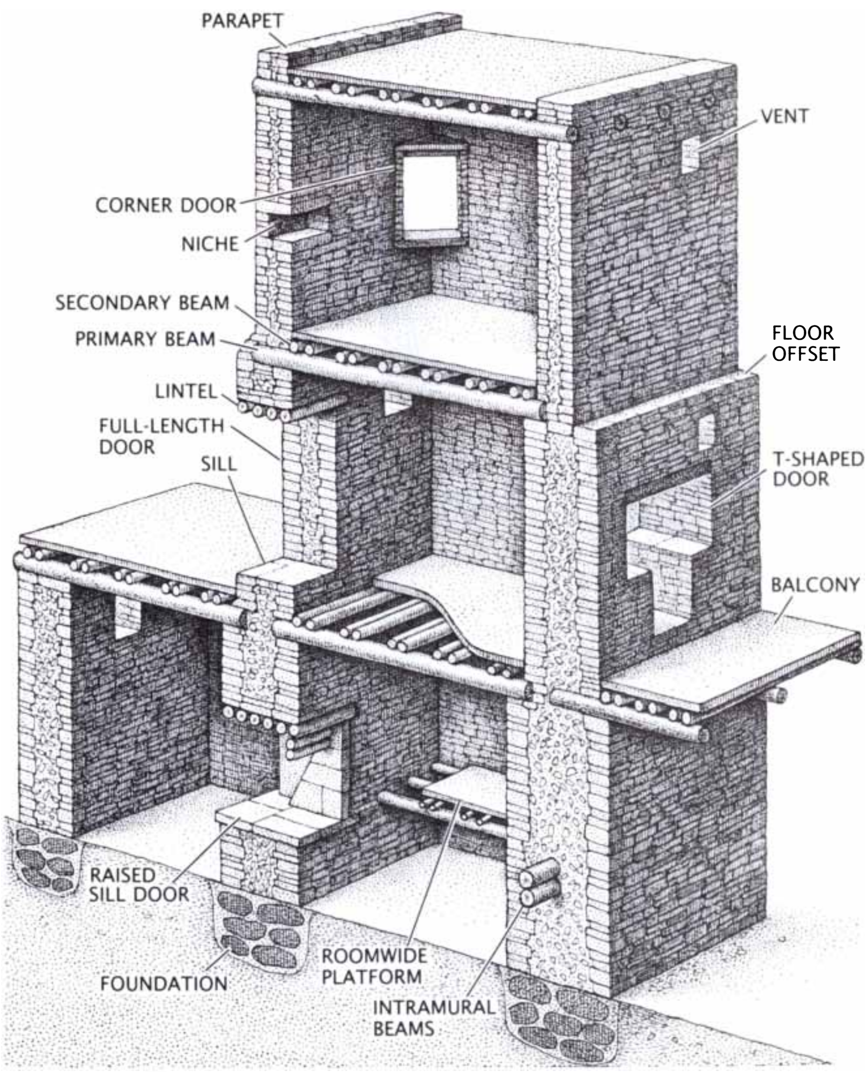
The big-room suites at Pueblo Alto and at other Great Houses are also suggestive of a limited resident population. Big-room suites consist of a storage room connected to a second room that may have been inhabited on a limited basis, and they are associated with a small kiva. As their name implies, the rooms of a big-room suite are larger than those of household suites; the overall suite may be as much as 100 square meters in area, or almost twice the size of a household suite. But within a big-room suite the amount of space allocated for habitation is only about 15 percent of the total area, whereas in household suites it is 50 percent or more. Moreover, there are no doors connecting one big-room suite with another—evidence that they were defined in some way by political or social boundaries—and no indication that cooking or eating took place in them. Five big-room suites were built when Pueblo Alto was first constructed in the early 1000's and three more were added in subsequent years.

Road-related suites are by far the most intriguing of the three types of rooms we found at Pueblo Alto. They are small, interconnected storage rooms that were built in rows along the exterior of the building and opened directly onto adjacent roads. An unusual aspect of these suites is that they are completely inaccessible from inside the house. If the Great Houses were primarily for habitation, one would expect storage rooms to be accessible to their inhabitants. The existence of doors that open only to the exterior of the building implies instead that the suites are road-related storage rooms of some kind and that the residents of the Great Houses had limited control over them. The fact that road-related suites were constructed at Pueblo Alto, Pueblo Bonito and Chetro Ketl, and probably at Penasco Blanco, at approximately the same time (the 1040's) strongly suggests that they were built for a common purpose.

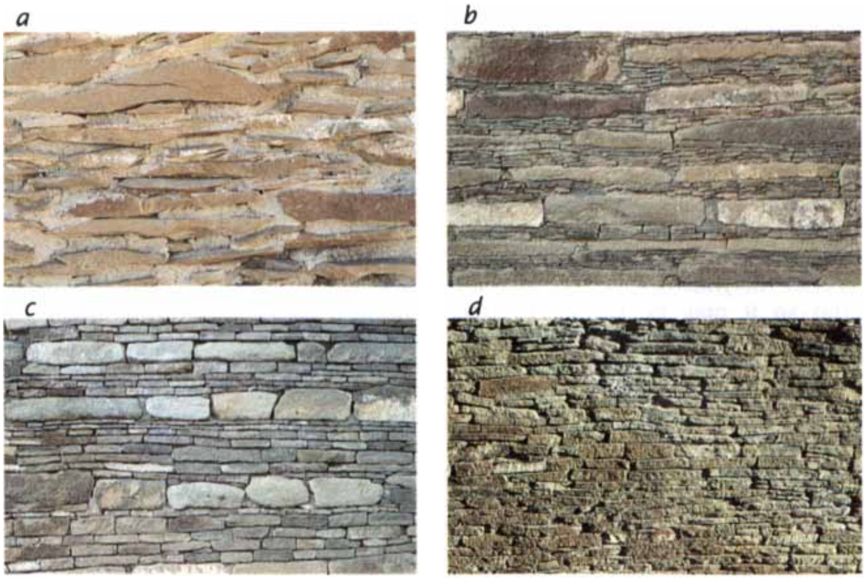
Other aspects of Pueblo Alto architecture are of interest. Like all the larger Great Houses, it has a central plaza: a large open area within the building's walls. Unlike the plazas at smaller dwellings in the canyon, which were of soil compacted by foot, the plaza at Pueblo Alto was periodically resurfaced with thick coats of clay. We assume that the plaza served a special purpose in the lives of the Chacoans, perhaps as an important gathering place at certain times of the year, but we know little else about it. A second plaza was also identified at Pueblo Alto, but it is outside the building and marks the juncture of several roads, and so it may be related to trade rather than to ceremony.

Our systematic excavation of Pueblo Alto resulted in the discovery of an extraordinary number of artifacts. More than 204,000 pieces (including potsherds, flaked stones and food remains) were recovered, 70,000 of them from a nearby trash mound that occupied 2,400 cubic meters of space and stood four meters tall. How could the small permanent population in the household suites have produced such a disproportionate number of artifacts? We estimated that over a 60-year period more than 150,000 pottery vessels alone were discarded in this one trash mound. If our estimate of the number of permanent residents at Pueblo Alto is correct, that amounts to 2,500 vessels per year or 25 pots per person per year! How might such a quantity of refuse be explained?

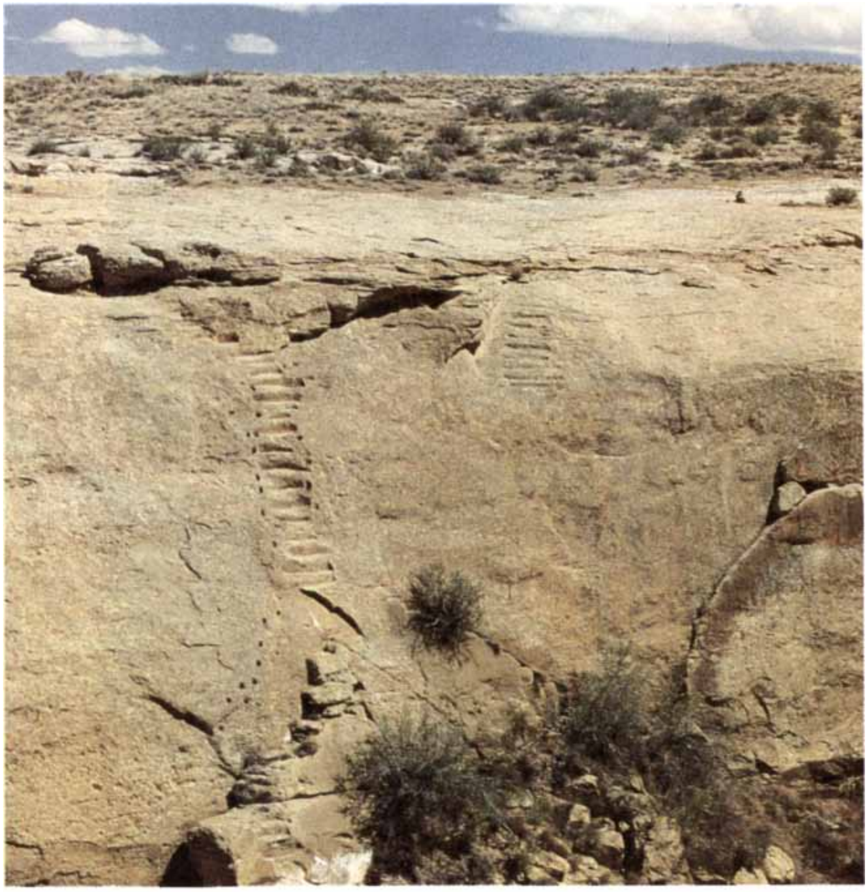
The trash mound provides interesting clues to the activities that may have taken place at Chaco. The trash appears to be layered in the mounds as if it was deposited intermittently



CROSS SECTION of several rooms in a typical Chacoan Great House is shown in this schematic drawing. The strength of the structure is suggested by the width of the walls. Floors were constructed from primary and secondary pine beams (laid at right angles to one another) and then covered with a mud plaster. Rooms were well vented to provide adequate air circulation. A variety of doors, including small corner doors, T-shaped doors and full-length rectangular doors, can be found in each Great House.



MASONRY of the Great Houses is highly stylized and falls into four basic types. The oldest (*a*) is fairly crude, consisting of irregular chunks of sandstone held together by large quantities of mud (replaced here with concrete); the second consists of blocks of sandstone chinked or reinforced with small, flat chips of rock (*b*). The third and fourth types are most recent, having been built in the late 1000's. The third is characterized by rows of large sandstone blocks alternating with rows of smaller pieces (*c*); the fourth is made from flat, highly regular pieces of sandstone (*d*).



STAIRS were sometimes cut into the cliffs surrounding Chaco to facilitate movement into and out of the canyon. This staircase was built about a kilometer north of Chetro Ketl and runs some 75 feet vertically up the cliff face. To the left of the stairs handholds, also carved into the rock, are visible. To the right of the staircase at its top one can see the beginning of a second set of stairs, which was never completed.

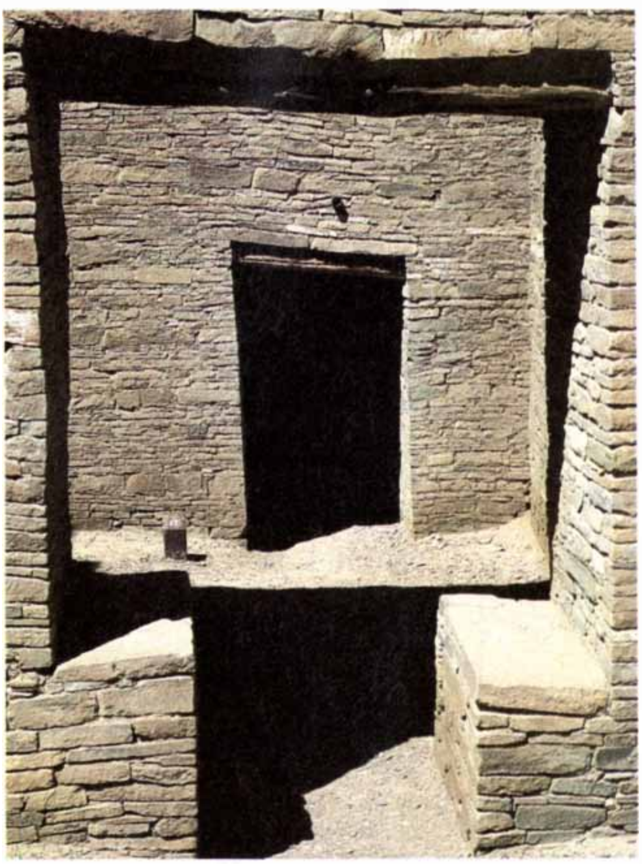
rather than daily. Not all the Great Houses have trash mounds, implying that they were not essential to or indicative of everyday living. Houses that do have mounds show a similar discrepancy between the number of artifacts and the size of the resident population. We therefore think the mounds reflect human activity unrelated to daily occupation. Could the mounds have been created by seasonal gatherings of large numbers of people, who may have converged on Chaco for ceremonial reasons? Once created in that way, they may have become important features of the landscape, serving as visible reminders of past ceremonial events.

The roads that radiate out from the Great Houses provide further—and indeed the most compelling—evidence that Chaco was not an isolated community but was physically connected to distant Anasazi communities. Although the roads were first noticed in the early 1900's, when the Navajo who live in the region today reported finding traces of them in remote areas of the San Juan Basin, it was not until remote-sensing techniques became available in the 1970's that the roads were studied on a large scale.

In 1981 the Bureau of Land Management began a systematic study of the Great North, Ahshislepah, South and Coyote Canyon roads, major roads known to cross the vast area of public lands surrounding Chaco Canyon. At the time it was widely believed the roads, which are noticeably linear (the Great North Road, for example, follows a bearing close to true north for almost 50 kilometers), were little more than simple trails made visible by prolonged wear. As the study progressed, however, it became apparent that the roads had been carefully engineered: not only were the roadbeds excavated below the surface particularly in areas of uneven topography, but also they were flanked by linear mounds of earth called berms. The width of the roads varies in some places, becoming greater in the vicinity of the Great Houses, but for large distances it is a consistent nine meters (almost 30 feet) across. Wide stairways were cut into the cliff faces of Chaco Canyon, and wood ramps and scaffolds were set at various points along the cliffs to facilitate movement into and out of the canyon.

We are just beginning to comprehend the extent of the roads beyond the San Juan Basin. Ongoing studies by Jon R. Roney of the Bureau of Land Management, Michael P. Marshall of the Solstice Project, Andrew P. Fowler

GREAT KIVA at Chetro Ketl is one of the largest at Chaco. It is a massive structure with a wall about three feet thick. On the floor four masonry-lined holes held enormous wood roof supports; at the center was a square firebox. At regular intervals around the circumference there are recessed niches, each about a foot square, whose function was probably ceremonial.

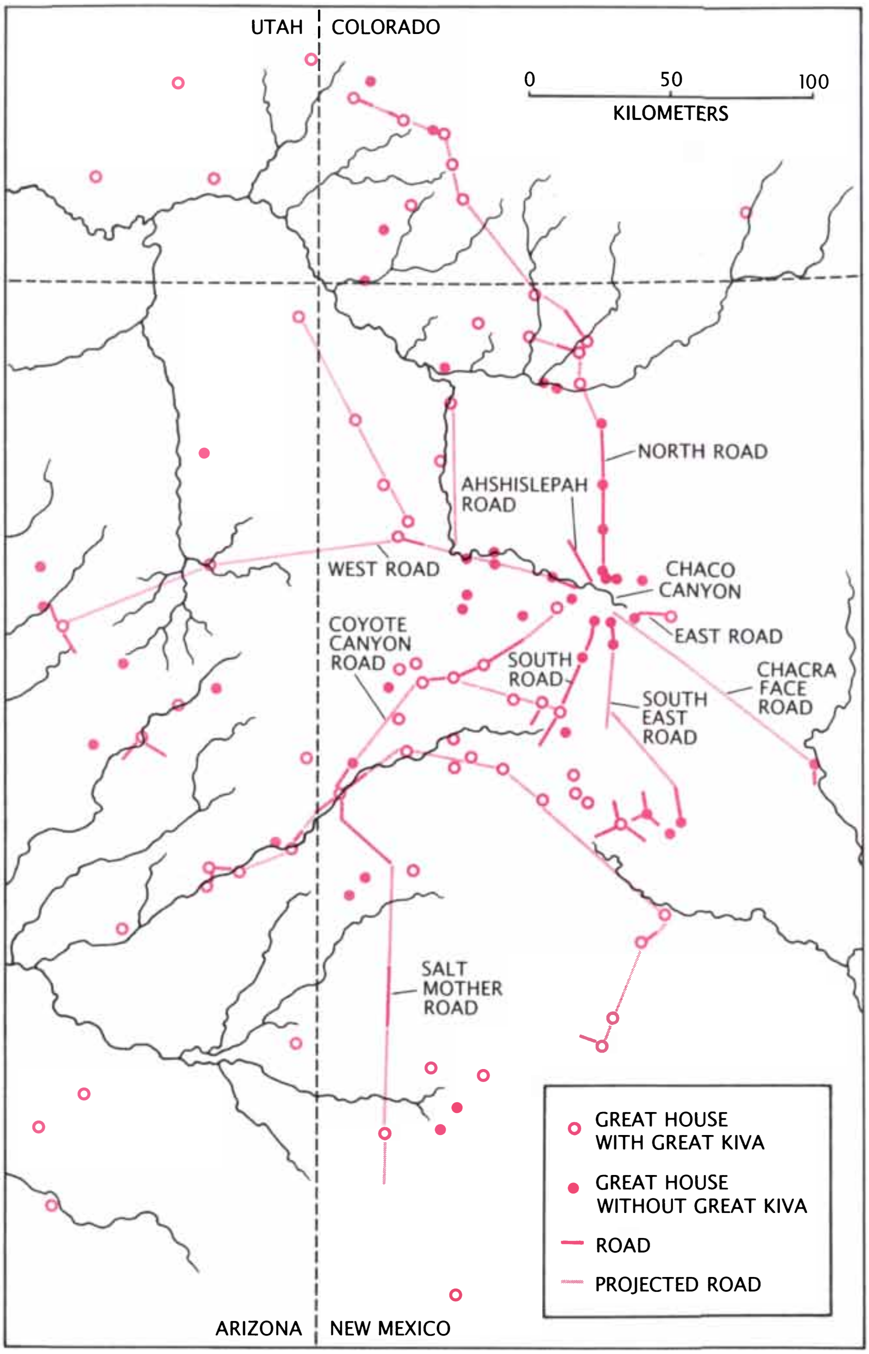


DOORS are quite distinctive in the Great Houses, where several types were built. T-shaped doors (*left*) were constructed in various dimensions. This one, from Pueblo Bonito, is unusually wide; its function, like that of most T-shaped doors, is not known. Perhaps the T shape made it easier to carry a large burden through the doors easily. Corner doors, such as one from Pueblo del Arroyo (*right*), were less common; they may have facilitated movement through the maze of storage rooms.

of the Zuni Archaeological Program and us suggest that the roads probably extend as far north as the San Juan Range in the Rocky Mountains and as far south as the Mogollon Mountains. Although we are less certain of their paths east and west, we believe they may have extended as far east as the turquoise mines near Santa Fe and as far west as the Little Colorado River valley, perhaps even as far as the San Francisco peaks near Flagstaff.

What purpose might these roads have served? We believe they identify Chaco Canyon as both the ceremonial center and the hub of a vast regional network. The probability that Chaco was a ceremonial site is indeed likely. Great Kivas, conspicuous elements of Chacoan architecture, are found within the walls of some of the larger Great Houses. At least 18 have been found at Chaco; two of them have been excavated at Pueblo Bonito. These masonry-walled subterranean structures are twice as large as the other kivas at Chaco: they average from 15 to more than 20 meters in diameter and are as much as four meters deep. At the bottom of each there are masonry-lined holes that once held massive roof supports; one of the timber supports found at Chetro Ketl is known to have measured 68 centimeters (26.5 inches) in diameter, a hefty load to carry for a distance of 40 or more kilometers. Recessed niches are spaced at set intervals in the walls of the Great Kivas. Their function is not known, but they may have had ceremonial importance.

Evidence that Chaco Canyon was the center of an extensive regional system can be derived from the types of artifacts found there. Exotic items, including turquoise from the Santa Fe region 160 kilometers to the east, ornamental shells from the Pacific coast and copper bells and macaw feathers from Mexico, are more abundant at Chaco Canyon than at other contemporary Anasazi sites. More remarkable is the number of utilitarian goods that were imported from outside the canyon. As much as a third of the chipped-stone debris found at Pueblo Alto came from distant quarries, such as Washington Pass in the Chuska Mountains 80 kilometers away. Up to half of the many thousands of pottery cooking vessels are made from a special clay that has also been identified as coming from the Chuska area.



CHACOAN REGION, defined by the presence of Great Houses, Great Kivas and roads, covers northeastern New Mexico and parts of Colorado, Utah and Arizona and is much more extensive than previously recognized. Although the road system is incompletely mapped, the authors are currently expanding their research in this area.

Almost a century ago structures similar to the Great Houses of Chaco Canyon were recognized some distance from the canyon. The site called Aztec Ruin, for example, which is similar to Chetro Ketl in size and layout, was discovered 85 kilometers north of Chaco and excavated in the 1910's and 1920's. Over the years other Chaco-like sites have been reported. By the 1970's the number of extracanyon Great Houses had grown to more than 20, most of them severely eroded by the forces of nature and reduced to rubble mounds. After a decade of fieldwork the list of possible extracanyon houses has grown to more than 150 and covers an area of at least 150,000 square kilometers and perhaps as much as 300,000 square kilometers.

We think this entire area was part of the Chacoan regional system during

PUEBLO ALTO is situated above Chaco Canyon on the north mesa, where it commands a panoramic view of the surrounding San Juan Basin. Only a vague outline is visible from the air, in part because much of Pueblo Alto has been eroded and in part because much of the site has been filled in to prevent further decay. The remains of four prehistoric roads leading north from Pueblo Alto can be seen faintly (in contrast to modern roads, which are readily visible). The road system was carefully planned: the roads are surprisingly straight and for much of their length they are a constant nine meters wide.

the 11th and 12th centuries. We think so not only because the extracanyon houses are architecturally similar to the Great Houses of Chaco but also because many of them are connected to Chaco by the elaborate system of roads. The Chacoan system, as we now define it, is from eight to 10 times larger than we had previously recognized, and the scale of our studies has become irreversibly much larger.

Before the discovery of Great Houses and roads beyond the San Juan Basin, we thought the evolution of the Chacoan regional system could best be explained on the basis of local ecological conditions. We knew that during the 11th and 12th centuries the San Juan Basin was a bleak and largely uninhabited desert and that the road system extended for many kilometers through terrain uninhabited by the Anasazi. Eventually the roads reached the rim of the basin, where greater rainfall and more productive soils encouraged densely populated settlements. We knew from tree-ring data that in any given year the amount of rainfall might have varied considerably around the basin and that a good agricultural year at one spot (producing a food surplus) might have been a bad year (leading to famine) at another. We theorized that Chaco Canyon was a central storage site for the entire San Juan Basin. Surplus food from around the basin would be brought to the Great Houses, where it would be stored until needed by communities affected by scant rainfall. In such a context the big-room suites, the road-related suites and the layered trash mounds make perfect sense.

Now, given the dramatic increase in the scale of the Chacoan system, we must reevaluate our hypothesis. Although we are certain that Chaco Canyon was very different from the isolated proto-urban community it is often portrayed as being, we are less certain of how it fits into the world beyond the San Juan Basin. We know that it was the heart of a vast regional system and a community of unprecedented complexity; we also have reason to believe it was much more than a communal center for the San Juan Basin. Just how large was the Chaco system? No one can say for sure, but it is clear that the picture now emerging signals the beginning of a new generation of Chacoan studies.

FURTHER READING

THE ARCHITECTURE OF PUEBLO BONITO. Neil M. Judd. *Smithsonian Miscellaneous Collections,* Vol. 147, No. 1, Publication 4524; 1964.

GREAT PUEBLO ARCHITECTURE OF CHACO CANYON, NEW MEXICO. Stephen H. Lekson. *Publications in Archaeology 18B, Chaco Canyon Studies.* National Park Service, Division of Cultural Research, P.O. Box 26176, Albuquerque, N.M., 1984.

PEOPLE OF CHACO: A CANYON AND ITS CULTURE. Kendrick Frazier. W. W. Norton and Company, Inc., 1986.

ARCHAEOLOGICAL SURVEYS OF CHACO CANYON. Alden C. Hayes, David M. Brugge and W. James Judge. University of New Mexico Press, 1987.

SCIENCE AND BUSINESS

Groupware

Beyond number crunching: computers aid management

Managing groups, satellite struggles, chemical freeze, comical physics

Some 50 miles from General Motors' Detroit headquarters, six managers hunch over an oval table in which personal computers are embedded. Their attention is fixed on a large common computer screen on the wall, their conversation punctuated by the clicking of computer keys as they take turns typing in one another's comments. At the end of the meeting each will leave with a copy of the document they are creating: an updated 60-page management manual. Thanks to the electronic meeting room, "we are getting superior results in a somewhat shorter time," says David H. Hill, a GM executive.

The episode marks another step in the emergence of "collaboration technology," or "groupware," as a driving force in information processing. The idea is simple. Personal computers already give individuals command of more information and boost their efficiency; now they should help groups.

Roles for collaboration technology basically fall into four categories, says John Seely Brown, a vice-president at the Xerox Palo Alto Research Center (PARC) in California. The technology can facilitate joint projects such as coauthored papers; it can also intelligently disseminate relevant information such as design changes to team members building a product. By linking people who might not otherwise communicate with one another, it could serve as an incubator for new ideas. Finally, by capturing group debates and discussions when a product is in gestation, collaboration technology could create a design history that the company can use when servicing the product, evaluating design changes or conceiving the next product.

To those ends, workers at universities, companies and research institutes are building software tools that run on personal computers linked by local or wide-area networks. Some people are testing theories of group interaction, others are trying to ascertain whether groupware will improve their organization's communication and still others are building products for the market.

Researchers at Xerox PARC, for example, put together the first electronic meeting room, called Colab, some four years ago to experiment with software that might change the texture of meetings. One program, "Cognoter," is intended to spur brainstorming. Individuals contribute phrases or thought fragments to a central screen, and afterward the group jointly identifies related ideas and organizes them into a logical form. Recently Xerox decided to put into practice some of the lessons it has learned: the company is building two similar rooms for its design engineers to use routinely.

Designers at the Center for Machine Intelligence (run by Electronic Data Systems, a GM subsidiary) intended to plunge theory into the real world when they built the electronic meeting room used by the GM managers. Investigators monitor the interactions of managers working on actual projects. Marilyn M. Mantei, a senior researcher, says she has observed that meeting participants typically perceive issues in greater detail than they would if they just talked. They also encounter fewer surprises after the meeting because most problems have been resolved, she adds.

By building on electronic mail, workers have also devised groupware that knits together individuals working in different places, even if they work at different times. Thomas W. Malone, a professor at the Sloan School of Management at the Massachusetts Institute of Technology, led the development of an experimental system that enables an individual to build what is almost the computer equivalent of an efficient secretary. Called the Information Lens, the system sorts electronic mail into files established by its user (such as "Priority" or "Meeting appointments"); it also extracts from the network public messages that match an individual's interests. For example, Lens can rapidly sift through a manufacturer's internal design-change notices and ship them to the relevant groups in the company. Since each user writes the rules for his or her Lens-based application, it falls to the individual to ensure the system does not filter information either excessively or incompletely.

Closer to coming on the market is a software tool kit developed by Lotus Development in Cambridge, Mass. More than a dozen potential customers are testing the software, code-named Notes, which Lotus hopes to introduce officially later this year. Like Lens, Notes is a tool for building an information-management system. It is geared for team members who, regardless of their location, work on the same data. Notes records changes in a file, ensures that appropriate individuals learn about those changes and creates an audit trail.

The only groupware applications currently on the market are sold by small companies. One of the best-known is "The Coordinator," marketed by Action Technologies in Emeryville, Calif., and codeveloped by Terry A. Winograd of Stanford University. The Coordinator strongly encourages those sending electronic messages to request some action from the recipients. Typically message recipients must commit themselves to finishing a task by a particular date, decline the task or propose an alternative one. The Coordinator then organizes a user's message files based on his or her commitments. The Coordinator has come under fire, however, for "demanding" that recipients pledge specific actions. Winograd argues that, particularly in an office, clarity is often critical. "There are circumstances where it is good to be vague, but I intuitively feel that vagueness is tied to face-to-face conversations," he adds.

Workers developing collaboration technology stress that it can improve the efficiency of face-to-face meetings without supplanting them. But groupware alone will not invariably make groups more effective, cautions Lucy A. Suchman, an anthropologist at Xerox PARC. "People have been working together since day one," she says, and in many cases very effectively. Both those who are building the technology and those who adopt it must understand enough about how people work

together to know when computer-supported collaboration will help.

"No one is going to say 'We need groupware,'" adds Esther Dyson, an industry analyst. "Instead they'll say 'We need a tool that will help us do a particular task,'" such as running a project, defining a marketing plan or updating policy changes. "There's no one archetypical groupware application," Dyson says. Irene Greif, advanced-products manager at Lotus Development, expects that all software will eventually have to be tailored to fit groups as well as individuals.

Investigators say that real gains will come when groups use collaboration technology to reshape the way they work together. Brown predicts that companies employing collaboration technology to "capture and reflect on" how they make decisions will find they have a "learning-curve accelerator" that will hone their competitive edge. —*Elizabeth Corcoran*

Messages from on High

Satellite positioning and data relay enter the marketplace

Two companies are battling over the potentially lucrative market for satellite-based systems that can track and communicate with trains and trucks.

Last year the Geostar Corporation of Washington, D.C., began offering a system that sends short messages from a miniature keyboard and portable transmitter anywhere in the world to Geostar's headquarters by way of satellite. The messages, along with a signal indicating the transmitter's location, are then passed over telephone lines to the subscriber—for example, a shipper of produce who wants to know whether a truckload of fruit is on schedule. Commercially available software can display the location of the vehicle on a map on the subscriber's computer terminal.

Until recently messages took an hour or two to be relayed, because the satellite, in low earth orbit, had to store them until it passed within transmitting range of Geostar's headquarters. From a transponder on a newly launched GTE geostationary satellite positioned over North America, Geostar messages can now be relayed within a few minutes from mobile terminals almost anywhere on the continent. A similar satellite, to be launched in September, will expand coverage of Mexico and the Caribbean.

At present Geostar tracks the location of its mobile units in North America through the U.S. Coast Guard's LORAN-C system. Low-frequency transmissions from LORAN-C stations enable a special receiver built into a mobile unit to determine its position to within a mile. The position is then automatically encoded in the signal sent to the satellite. Several trucking companies have already tested the system for keeping track of their vehicles. Geostar predicts that when pocket-size transmitters are available, its system will be able to track everything from trucks and trains to backpackers and joyriding teen-agers.

Geostar's existing transmitters can be relatively small and low-powered because small amounts of information are transmitted, encoded in a single "burst." The more difficult technological challenge of maintaining satellite voice links, which require a greater rate of information flow, is thereby avoided. (A consortium that intends to offer satellite voice links in the 1990's, the American Mobile Satellite Consortium, was established in May.)

In 1990 Geostar plans to launch its own geostationary satellite. Then the company will be able to offer positioning based on the timing of satellite signals, which is expected to pin down location to within 50 meters. Geostar envisions specialized applications for this more accurate service, such as telling which track a train is on and positioning oil and gas drilling platforms. The new satellite will also make two-way communication possible between the mobile terminals and Geostar headquarters. Until then Geostar will send messages to its mobile stations by way of linkups with existing satellite-based paging systems, which work only in some metropolitan areas. Geostar's T. Stephen Cheston says the decision to offer this interim one-way communication service was based on "launch schedules and business considerations": two-way transponders consume much more power than one-way versions, and so leasing satellite space for them is more difficult.

Geostar's strategic decision has created an opening for its major competitor, Omninet, a Los Angeles company. Apparently planning to steal a march on Geostar, Omninet is ready to offer both LORAN-C positioning and two-way data communication. The company says the Department of Energy has successfully tested the Omninet system for tracking hazardous cargoes such as radioactive wastes. By May, Omninet had received only an experimental license from the Federal Communications Commission but was expecting a permanent one within months. Unlike Geostar, Omninet indicates it will stick with a satellite-aided version of LORAN-C for positioning instead of adopting the potentially more accurate system of positioning by timing satellite signals. It remains to be seen whether the market is big enough for both approaches. —*Tim Beardsley*

Keeping Cool

Industries struggle to prepare for the upcoming CFC crunch

The hue and cry over the environmental damage caused by chlorofluorocarbons (CFC's) has left refrigerator and air-conditioner manufacturers facing more questions than answers in a world that proposes to use fewer CFC's.

According to the Environmental Protection Agency, roughly 28 percent of the ozone depletion attributed to CFC's is caused by the CFC-12 coolants in commercial and residential refrigerators and mobile air conditioners; another sizable part is due to the CFC-11 used in foam insulation for refrigerators. Production of these CFC compounds and others must be cut in half by 1999 to meet the Montreal Protocol, ratified by the U.S. in March (see "Science and the Citizen," May).

CFC producers say that most alternatives will not reach the market until the mid-1990's. Yet CFC users expect to face shortages of the compounds next year when the EPA launches its plan for meeting the Montreal Protocol. The EPA will not finalize its plan until August. So far it has proposed to put production limits on specific CFC compounds based on how much each compound apparently depletes the ozone and then let supply and demand run its course. "We expect that the increased prices will encourage users to find alternatives and that the least valuable uses will fall out first," says Stephen O. Anderson, an EPA economist.

From the vantage of Ray Bohman, chief engineer for refrigeration products at Amana Refrigeration in Amana, Iowa, the situation "has all the earmarks of a calamity." Last year Congress announced that appliances must meet higher energy-efficiency levels by 1990. According to the Association of Home Appliance Manufacturers, some 75 percent of the 1986-model refrigerators and freezers do not meet the higher requirements. Consequently manufacturers plan to use more foam insulation in each refrigerator and freezer, upping their use of CFC's as much as 20 percent, says Frederick H. Hallett, a vice-president at White Consolidated Industries. "We can't meet the energy standards on the books with the proposed substitutes unless there's another technological breakthrough," he adds.

Howard Geller, associate director for the American Council for an Energy-Efficient Economy, says that under the EPA proposal "whoever needs CFC's will get them." Moreover, since retail and residential refrigerators together used only about 7,000 metric tons of CFC-12 in coolants in 1985 (about 2 percent of the regulated CFC's), "it's not going to be a problem if they use a little more."

For its part, the foam industry is scrambling to find ways to reduce the percentage of CFC's in foams. That goal is formidable, Anderson notes: CFC's give foam desirable characteristics including chemical stability, density and rigidity, as well as good insulation properties.

Mobile air-conditioner manufacturers meanwhile are looking for ways to conserve or recycle CFC's, even as they assess proposed alternatives. Robert McFadden, a senior analyst with the Motor Vehicle Manufacturers Association of the U.S., views recycling pragmatically. "Up until this point there was no economic reason for doing this," he says, "but now, with the price going up, the economies will come about."

McFadden remains cautious about CFC substitutes. Du Pont and 13 other chemical manufacturers have formed a consortium to begin more than five years of toxicity testing of two promising alternatives, FC-134a and HCFC-123. But since the automobile industry "doesn't want to get locked into a substitute that will bomb out somewhere along the line," McFadden expects it will be at least three years before car manufacturers are sufficiently confident in the proposed substitutes to start redesigning mobile air conditioners to rely on the new compounds. "We don't want an interim solution," he concludes. —*E.C.*

Laughing Matters

The Japanese race ahead in superconducting lore

It was a dark and empty laboratory. Enter an industrial spy who stumbles across documents that describe new superconducting materials. Little does this hapless fellow realize he is about to embark on an intercontinental chase after an ever higher-temperature superconductor that will leave him battered and bruised but technically up-to-date.

Does it sound farfetched? Not according to a Japanese comic book on superconductivity. Within the past few years several publishers in Japan have begun to capitalize on the popularity of comic books and to churn out light works on weighty subjects. And so, next to translated versions of *What They Don't Teach You at Harvard Business School,* the New York Kinokuniya Book Store offers comic books on superconductivity.

The 220-page hardback published by Goma Shobo wastes little time on introductions (our spy gets trounced by thugs on page 9). Still, the didactic purpose of the story is clear from the first chapter, titled "What Is Superconductivity?" to the last, on "How Will the Economy Be Changed by Superconductors?"

The spy, a former physicist who lost funding for his research when it failed to produce results, soon explains the rudiments of superconducting behavior to a college friend turned businessman. Seeing the gleam of a profit, the businessman hires the spy to crash laboratories in pursuit of the latest information.

The brunt of the story's technical content is presented in a university lecture overheard by the spy. Readers learn the right-hand rule for finding the direction of a magnetic field near a current, along with the importance of yttrium-barium-copper oxides. Several pages show magnets floating above superconductor wafers, demonstrating the Meissner effect. For those with a greater scientific appetite there are a dozen or so footnotes. The comic book also does what most scientists studiously avoid: it discusses in grand terms the presumed payoffs of the new superconductors.

The story does not shy from the realpolitik of science either. At one point the spy explains to his friend that unless Japan develops high-temperature superconductors, the U.S. could put Japanese researchers "out of the race" by banning exports of liquid helium, essential for earlier superconductors.

Later the spy proclaims, "This is not the time to conceal superconducting materials or lock them up in secret rooms."

"You are too naive," his friend retorts. He adds that American politicians were already advocating secrecy, aimed at locking out the Japanese.

Eventually both the roving spy and the profit-hungry businessman get their just deserts: the businessman finds he cannot patent a trivial superconducting application and the spy falls into the arms of a pretty girl.

At the Kinokuniya Book Store the comic has met with mixed reactions. Store managers say patrons want to see the superconductivity comics but few are buying them. What are they taking home instead? Installments of the four-part comic series on Japan's economy. —*E.C.*

It looks mysterious to the women too, but be assured that the man in this Goma Shobo comic book has no intention of building a bomb. Making a high-temperature superconductor in his hotel suite will keep him busy until dawn.

Once involved in a project many of us tend to lose sight of the overall context in which that project began. NASA has called the Space Station: "The Next Logical Step." What most of us forget is what it is a logical step towards.

With Pioneering the Space Frontier, the President's National Commission on Space reminded us of the overall context in which the station fits. With the National Space Policy directives of 1988, it has become clear that our overall goals are "to expand human presence and activity beyond Earth orbit into the solar system."

One organization stands alone in its lasting commitment to space settlement: the National Space Society. Our members reflect a growing public concern for vigorous and comprehensive space activities leading to permanent human habitation of the Solar System. If you share this concern, we urge you to join us and become visible in support of this objective.

Whether it is taking the next logical step or preparing the way for the footsteps of future generations, a lot depends upon **your** involvement. Join us today!

Test firing of an experimental cryogen-fueled rocket as part of NASA's Project Pathfinder to develop new propulsion for Lunar and Mars landers. (Courtesy: NASA)

☐ I am enclosing $30 regular annual dues

☐ I am enclosing $18 youth/sr citizen (under 22/over 64 my age is ______

Please charge my credit card:

☐ MasterCard ☐ VISA ☐ Cards ☐

NAME ______

ADDRESS ______

CITY ______ STATE ______ ZIP ______

CREDIT CARD NUMBER ______ EXP. DATE ______

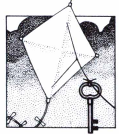
SIGNATURE ______

MAIL TO

NATIONAL SPACE SOCIETY ★ DEPT SA7 ★ 922 PENNSYLVANIA AVE SE ★ WASHINGTON, DC 20003 ★ 202/543-1900

THE AMATEUR SCIENTIST

Shadows cast on the bottom of a pool are not like other shadows. Why?



by Jearl Walker

Most shadows are easily understood, but those that are cast on the bottom of a pool of water sometimes have strange features. For example, when a leaf floats on water, its shadow is normal if the water is one or two centimeters deep, but in somewhat deeper water the shadow has a peculiar bright border. Equally unusual are the fleeting shadows left on the bottom of a shallow pool when an object is drawn through the water quickly and then removed: nothing opaque is left in the water, and yet dark disks with bright borders play on the bottom. A pencil can also produce an odd shadow. Stick the pencil partly into the water and then tilt it in various directions. For many orientations the shadow consists of two sausagelike shapes separated by a band of light.

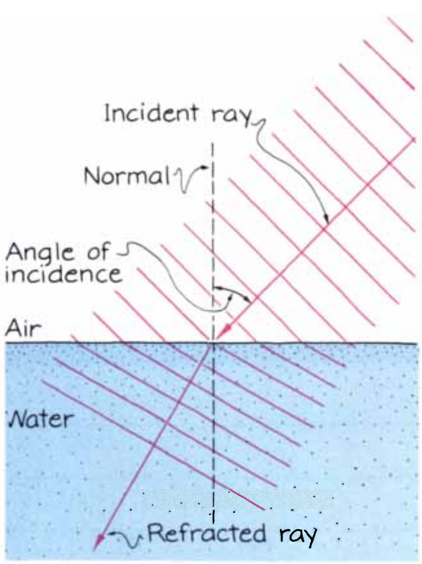
The puzzling aspects of these shadows are due to the fact that light is refracted at the air-water boundary. That is, its direction of travel changes because its speed changes. In a vacuum light travels at 3×10^8 meters per second, the ultimate speed in the universe. In air it travels slightly slower because it occasionally interacts with air molecules along its path. In water, where it is slowed by interactions with much more densely packed molecules, it travels at only three-fourths its speed in a vacuum.

Textbooks commonly treat light as being a wave with a series of straight wavefronts; the direction of travel is represented by a ray perpendicular to the wavefronts. When the light is refracted through a boundary, the initial ray is called the incident ray and the final ray is called the refracted ray. The orientation of the incident ray is measured with respect to a line, called the normal, that is perpendicular to the boundary. If the boundary is curved, the normal is perpendicular to a tangent to the boundary at the point where the ray crosses. The angle between the normal and the incident ray is called the angle of incidence.

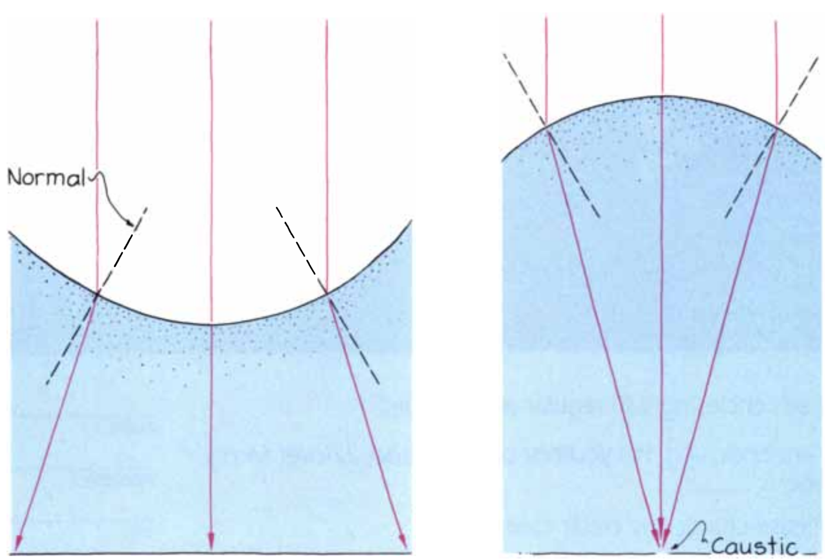
If light travels from the air into a flat pool of water with the incident ray aligned with the normal, each wavefront slows uniformly as it crosses the boundary, and so the light's direction of travel does not change. For any other orientation of the incident ray, each wavefront crosses the boundary gradually. The part that first enters the water slows down before the rest of the wavefront does. The nonuniform reduction in speed produces a kink that moves along the wavefront and alters the wavefront's direction of travel: the light is refracted. The larger the angle of incidence is, the more the light is refracted.

When the water surface is flat, a bundle of parallel rays is refracted equally and the rays illuminate the bottom of the pool uniformly. When the water surface is curved, however, the refraction is not equal. The rays have different angles of incidence, because the normals are oriented differently where different rays cross into the water. The nonuniform refraction gives rise to nonuniform illumination on the bottom of the pool.

A concave surface diverges light rays, thereby decreasing the illumination in the region of the bottom where the rays would otherwise have gone. A convex surface converges light rays; the region where the rays cross, and where the light is therefore brighter, is called a caustic. Depending on the shape of the water surface, the caustic



Refraction



The divergence (left) *and convergence* (right) *of refracted rays* can be a point, a line or a three-dimensional "surface." If the bottom of the pool happens to intercept the caustic, you see a bright point or line on the bottom. If the bottom intercepts the rays either before or after they cross, it is illuminated somewhat, but not as brightly as it would be at the caustic.

Small waves provide curved surfaces that throw fleeting, complex patterns of light on the bottom of the pool, but the motion is too rapid to follow. A better way to study refraction through a curved water surface is to float a small object whose edges will draw the water surface into a concave or convex shape. Where an edge is slightly higher than the normal water level, water is pulled up to make a concave surface; where the edge is lower, water is pulled down to make a convex surface.

In 1983 Michael V. Berry and J. V. Hajnal of the University of Bristol described how the curved water surfaces surrounding a floating object influence the object's shadow. To repeat their experiments, partially fill a white container with water. Then gently lower a razor blade (not the kind that has a heavy reinforced edge) onto the water. If you are careful to keep the blade flat and not to break the surface, the blade will float. The water under the blade is depressed below the normal water level by the weight of the blade, and so the water surface along the edge of the blade is convex.

Illuminate the blade and water with an overhead lamp at least a meter away and examine the shadow cast by the blade on the bottom of the container. If the depth of the water is less than three centimeters or so, the shadow is normal: a dark rendition of the blade. If the water is deeper, the shadow has a bright border. Berry and Hajnal describe a way to change quickly from one type of shadow to the other. Add water to a depth of five centimeters and then put a sheet of white paper on the bottom of the container; the blade's shadow on the paper has a bright border. Raise the paper until it is one or two centimeters below the blade; the shadow becomes normal. I prefer to change the depth of the shadow by adding or removing water in the container. Pouring water into the container usually does not disturb the blade; I remove water by sucking it out through a drinking straw.

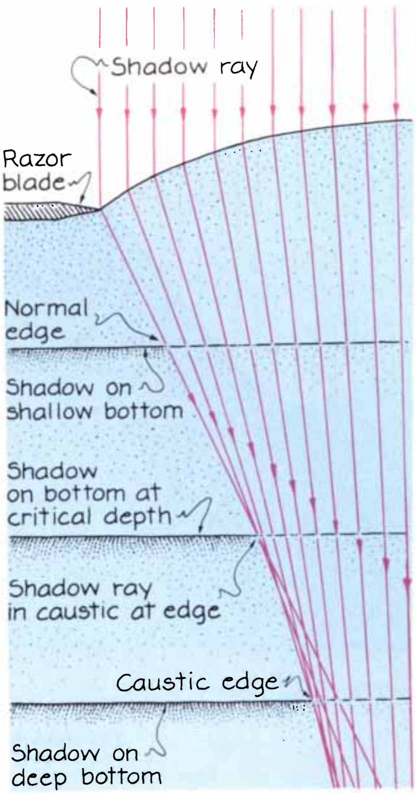
Along the curved water surface next to the blade's edge, rays are refracted so that they converge, but their point of focus depends on where they traverse the surface [*see top illustration at right*]. The surface curvature is greatest near the blade, and so rays entering the water there converge earlier than rays entering the water farther from the edge. Rays traversing the nearby flat surface do not converge at all.

The ray that enters the water right at the edge of the blade is known as the shadow ray; if there were no water, that ray would always mark the border of the shadow. Consider another ray that is slightly farther from the razor's edge than the shadow ray is. The curved water surface focuses that ray and the shadow ray to form a caustic at a certain depth below the normal water level. Call that depth the critical depth. (In the illustration the bottom of the container is shown at three depths, the middle one of which is the critical depth.)

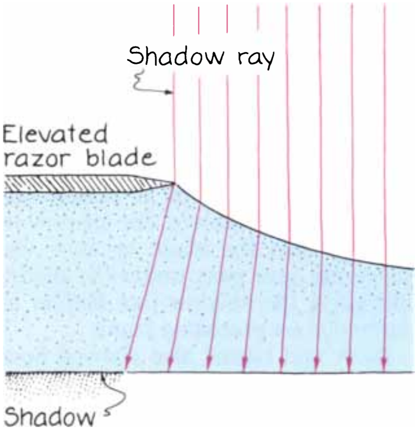
Suppose you begin with about two centimeters of water in the container. The depth is less than the critical value (which in my experiments is about three centimeters), and so the bottom of the container intercepts the shadow ray and its neighboring ray before they cross. The blade's shadow is normal and the shadow ray marks its border. Slowly add water to the container while you monitor the shadow. As the water reaches the critical depth, the border of the shadow develops a caustic because it is at the focus of the shadow ray and its neighbor.

Add more water to the container. A caustic remains at the shadow's border, but it is not formed by the shadow ray and its neighbor. The caustic is now due to two closely spaced rays that pass through the water surface somewhat farther from the blade's edge. Since the surface curvature is less there than it is at the edge, the rays focus at a greater depth than the shadow ray and its neighbor. The shadow ray no longer skims the border of the shadow but falls in the illuminated region on the bottom of the container somewhat away from the shadow. The width of the shadow is no longer set by the shadow ray, as in a normal shadow, but by the caustic that encroaches on the shadow. If you continue to add water, the location of the pair of rays responsible for the caustic at the border of the shadow gradually shifts away from the blade's edge. The distance to the bottom of the container is increasing; the caustic is at the convergence of rays that pass through the water surface where it is less curved.

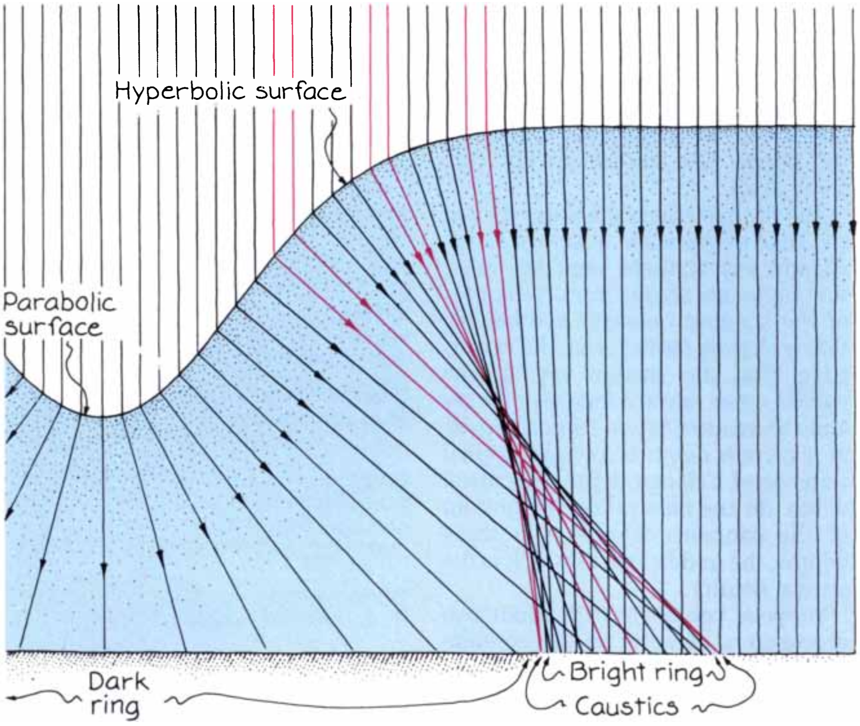
The critical-depth value depends in theory on the weight of the blade. A heavier blade should depress the water more, increasing the curvature of the water surface next to the blade. The shadow ray and its neighbor should cross earlier—at a shallower critical depth. In an experiment designed to check the theory I drained the water to a depth of 2.5 centimeters and floated a single blade as a control. The blade produced a normal shadow. Then I stacked three blades and lowered them onto the water. Although the intact stack floated for only a few



The shadow of a floating razor blade



The shadow of an elevated blade



The shadow cast by a vortex

seconds, I was able to study its evanescent shadow. The border of the shadow was bright. The weight of the stack depressed the water level more than the single blade did, decreasing the critical depth so that a caustic illuminated the border of the shadow. The caustic was wider than the caustics I had seen with a single blade. Once water seeped into the stack, the lower blades sank, leaving the top blade floating. The shadow immediately became normal.

Berry and Hajnal pointed out that a variety of floating objects, such as leaves and insects, can create shadows with bright edges if the water is deep enough and if the objects depress the water surface. When floating objects elevate the water surface so that the surface is concave, the shadows have normal borders because the light rays diverge rather than forming a caustic. The shadows are also different in size. A floating razor blade produces a shadow that is larger than itself. If you lift the blade so that it makes the water surface concave, the shadow is smaller than the blade because the diverging rays pass under the blade, shrinking the shadow [*see bottom illustration on preceding page*].

Berry and Hajnal also showed how shadows with caustic borders can be produced by the curved sides of a vortex in water. They created the vortex by spinning a magnet placed under the container, which in turn rotated a bar magnet in the water, on the bottom of the container. When light from an overhead source passed through the vortex, refraction formed a wide bright ring that surrounded a dark interior. The inside and outside edges of the ring were caustics.

A few months ago M. H. Sterling, Michael A. Gorman, P. J. Widmann, S. C. Coffman and Robert M. Kiehn of the University of Houston and John A. Strozier, Jr., of the State University Empire State College described how similar vortex shadows can be made without elaborate equipment. You can repeat their observations the next time you take a bath. After letting the water settle, briskly move an object horizontally through the top layer and then remove it. For a few seconds you will see dark circles, each with a bright border, playing on the bottom of the tub. The border is actually a narrow version of the bright ring seen by Berry and Hajnal. Try a variety of objects to find the one that best generates vortexes. Sterling and his co-workers used a circular paddle. As it moved, water circulated from the front of the paddle around the edge to fill in the space left at the back of the paddle. The visible swirls remaining on the water surface were connected by a vortex tube below the surface.

The group's investigation was initiated by an observation Kiehn made at an outdoor swimming pool. When he came out of the water, he left two dark disks on the bottom of the pool, each surrounded by a narrow bright ring. The shadows lasted for as long as 10 minutes. Kiehn reasoned that the shadows were caused by the refraction of sunlight through vortexes he had left in the water. But exactly what kind of vortex shape was responsible?

Members of the group considered two types of shape, one concave (parabolic) and the other convex (hyperbolic). If the vortex were entirely parabolic, the light rays would diverge; the region below the vortex might be dim enough to appear dark, but the divergence could not create a bright ring. If the vortex were entirely hyperbolic, the rays would focus to form a bright ring, but the water surfaces at the center of the vortex would meet at a sharp angle—a situation that is physically unlikely. The group found that the best model for the vortex is a blend of the two shapes: the vortex has a parabolic core that is surrounded by a hyperbolic surface.

The core creates the dark disk of the shadow. The bright ring that limits the size of the disk is due to closely spaced rays that pass through the hyperbolic surface and focus on the bottom. The rays that pass through the hyperbolic surface closer to the core converge too early to contribute to the ring and end up spread over the illuminated region outside the ring. The rays that pass through the surface farther from the core do not have a chance to converge and also end up outside the ring.

Which rays are responsible for the caustic depends on the depth of the water. If the vortex glides over water of decreasing depth, the responsibility for the caustic shifts to the rays closer to the core; the dark disk shrinks to some minimum size and then disappears if the water becomes even shallower. If the vortex moves over deepening water, the disk widens. The ring may grow wide enough so that you can distinguish its caustic edges.

I found I could make a similar shadow on the bottom of a container of water by blowing on the water through a straw. The resulting dimple on the surface usually casts a circular shadow with a bright ring, but when I blow almost horizontally, the shadow develops intriguing distortions.

The divergence of light rays by a concave water surface is responsible for the odd appearance of the shadow of a pencil inserted into water. The effect was reported in 1967 by Cyrus

Adler of New York, who discovered it while idly playing with a pencil during a bath. He noticed a curious divided shadow of the pencil, cast on the bottom of the tub by light from the single lamp in the room. He called it "the shadow-sausage effect" and correctly explained the illuminated gap between the sausages as being caused by the concave water surface along the sides of the pencil.

Intrigued by Adler's observation, I toyed with a pencil in a container of water. To follow my play, fill a container with water to a depth of about six centimeters, place it directly below a light and insert an unsharpened pencil vertically into the water. When you first insert the pencil, you may see a large shadow with a bright edge, because the pencil initially depresses the water surface. Either wait until the water climbs the shaft or pull the pencil up slightly so that the water surface around the shaft is concave. Consider the rays that pass through the water surface on one side of the pencil. They diverge in various directions and some are blocked by the submerged segment of the pencil. The ray that just skirts the end of the pencil is the shadow ray. The rays on the opposite side of the pencil undergo similar divergence and also have a shadow ray.

When you insert the pencil several centimeters into the water, the two shadow rays do not cross before they reach the bottom of the container; you will see a shadow of the pencil that is narrower than the pencil. Gradually lift the pencil. The shadow narrows and then disappears, replaced by a bright spot. The shadow rays now cross before they are intercepted by the bottom of the container. The bright spot is caused by light rays from opposite sides of the pencil that overlap when they reach the bottom of the container. Lift the pencil barely above the normal flat level of the water. If the liquid bridge clings to the end of the pencil, the bright spot is maintained. If you lift the pencil too far, the bridge fails and a normal shadow of the pencil immediately appears.

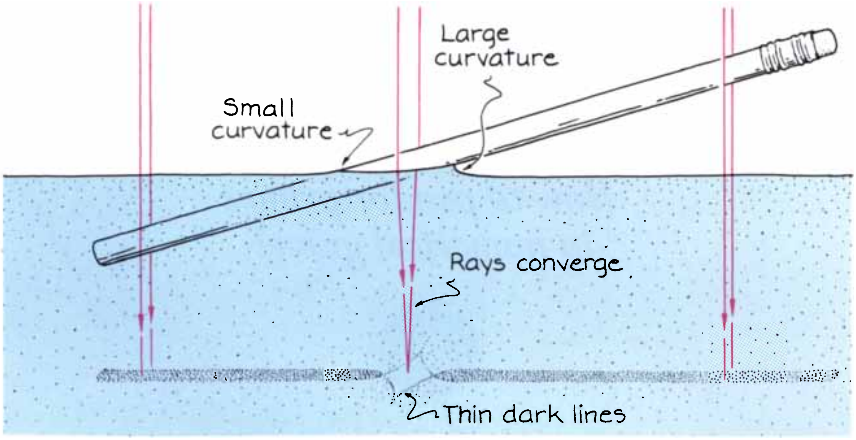
Next insert the pencil at an angle of 45 degrees and again either wait for the water to climb the shaft or pull the pencil slightly upward. The water surface is again concave near the pencil, but now the extent of the curvature varies around the sides of the shaft [*see illustration above*]. It is greatest where the pencil forms a 45-degree angle with the water, smallest at the obtuse angle on the opposite side of the pencil and intermediate at intermediate points around the pencil.

Both the dry and the submerged parts of the pencil cast normal shadows, but the short segment of the pencil surrounded by a curved water surface does not. Consider the shadow rays that pass through the two regions of intermediate curvature on opposite sides of the pencil. They converge toward the area on the bottom that lies between the shadows cast by the dry and the submerged parts of the pencil. If the water is shallow enough, the shadow rays are intercepted by the bottom before they cross; you will see a narrow shadow connecting the two wider shadows. If the water is deeper, however, the shadow rays cross and eliminate the connecting shadow. You see an illuminated gap between the two shadows, as is shown in the illustration.

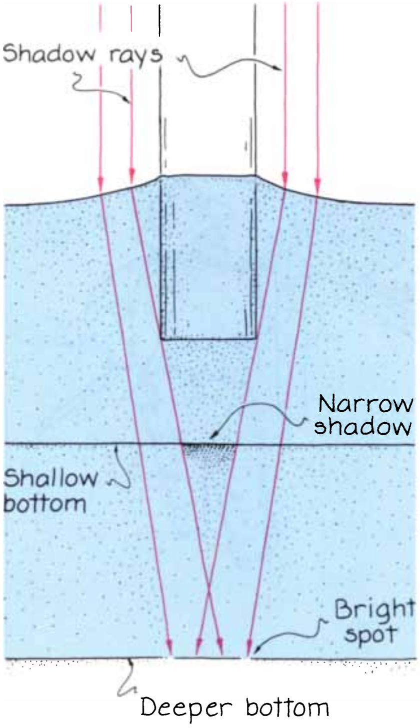
Adler noted that the gap had a complex structure of bright and dark regions; those near the shadows were brighter than the interior of the gap, which was grayish. I saw two dark lines radiating from the shadows' tips—as if "anticaustics" were being produced.

I poured a thick layer of corn oil over the water in my container and inserted a pencil through both liquids at a 45-degree angle. Now the pencil's shadow had three parts separated by illuminated gaps. One gap was from the refraction at the air-water interface, the other from the refraction at the water-oil interface.

You might investigate how other objects cast shadows on the bottom in shallow water. For example, a floating hair creates a string of different shadows. Some are dark and others have bright borders. Can you tell from the shadow which segments of a hair are fully submerged and which lie above the normal water level?



The "shadow-sausage effect"



The shadow of a vertical pencil

FURTHER READING

SHADOW-SAUSAGE EFFECT. Cyrus Adler in *American Journal of Physics*, Vol. 35, No. 8, pages 774–776; August, 1967.

THE SHADOWS OF FLOATING OBJECTS AND DISSIPATING VORTICES. M. V. Berry and J. V. Hajnal in *Optica Acta*, Vol. 30, No. 1, pages 23–40; January, 1983.

WHY ARE THESE DISKS DARK? THE OPTICS OF RANKINE VORTICES. M. H. Sterling, M. Gorman, P. J. Widmann, S. C. Coffman, J. Strozier and R. M. Kiehn in *Physics of Fluids*, Vol. 30, No. 11, pages 3624–3626; November, 1987.

COMPUTER RECREATIONS

How to pan for primes in numerical gravel

by A. K. Dewdney

No recreation embodies the lure of pure number better than the search for primes. Like nuggets of gold, they hide in the gray gravel of ordinary numbers. A prime is elemental: it cannot be divided evenly by any numbers other than 1 and itself. Primes are precious because they are rare. Common enough among the small numbers near the source of the great Continuum River, they thin out rapidly in the downstream banks.

One can pan for primes, even build a sluice box to mine these nuggets, but no one knows where they all are without looking. There is no formula for primes. There are patterns of sorts, a primitive kind of geology by which we can guess the deposits. Just as amateurs flocked to California and the Yukon to pan distant streams for the elusive yellow, so ordinary readers can set out for Number Country with little more than this primer tucked into a spare pocket.

Few mathematical ideas are as accessible to the average person as the concept of a prime number. It takes about a minute to explain primes to the man or woman in the street. Buy them a coffee and with a little encouragement they will write the primes on a paper napkin: 2, 3, 5, 7, 11, 13, 17 and so on. The number 1 is not normally considered to be prime.

Can one tell a prime just by looking at it? If there are many numbers in the pan, does a prime flash yellow to the eye? Some people think so. Numbers that end in 1 are often precious, such as 11, 31, 41 and 51. But one must beware of such fool's gold as 21 and 81, for example. Eventually the numbers that end in 1 fool us with increasing frequency, so that it is possible to wonder, as a few ancient Greeks did, whether the primes eventually thin out to nothing. There comes an end finally, or does there? Euclid has passed down to us the first proof that there is no end to the prime numbers.

The proof is so simple that one can imagine Euclid drawing forth the demonstration, Socratic fashion, from a slave. I prefer the conversation between the tyro and the old-timer on the banks of the Continuum River:

TYRO: Hey, mister! How far downstream do the primes go?

YUKE: Why, boy, all the way to the Sea of Infinity.

TYRO: I don't believe you. Here we are at the millions and I haven't seen color all day.

YUKE: You tenderfeet gotta be told everything. Look, suppose you came to the last prime. No more after that, right?

TYRO: Uh, right.

YUKE: Call it *n*. You take and form the product of all the primes there are right on up to *n*. O.K.? That's $2 \times 3 \times 5 \times \ldots \times n$. Now add 1 to the product and call the number you finally get *p*.

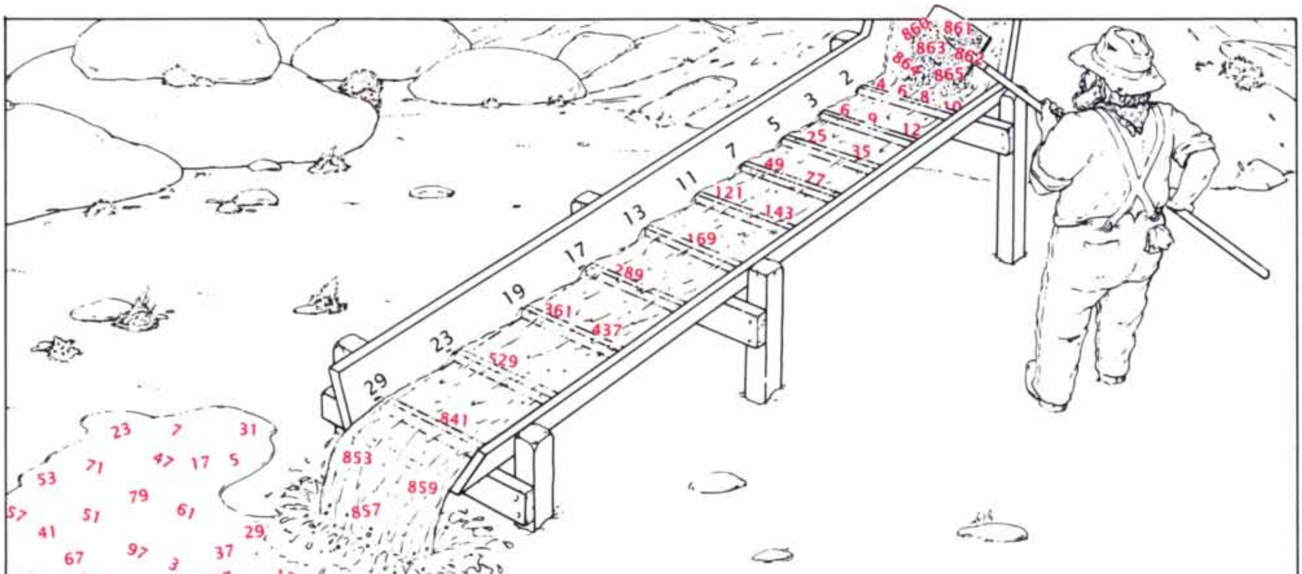
TYRO: Don't tell me that *p* is prime!

YUKE: Sure is. Prime as all get-out. Look. You can't divide it by 2 because there's 1 left over. You can't divide it by 3 because there's 1 left over. There's always 1 left over, right up to *n*. There's just no getting around it.

TYRO: Gosh, I guess that's what keeps you going.

YUKE: Yup. Well, don't just stand there yammering. Help me with this sluice box here.

Even if there is no largest prime, there is certainly a largest known prime. The distinction confuses some readers and even a few journalists. The fault lies with those back-page



A sluice box mines primes

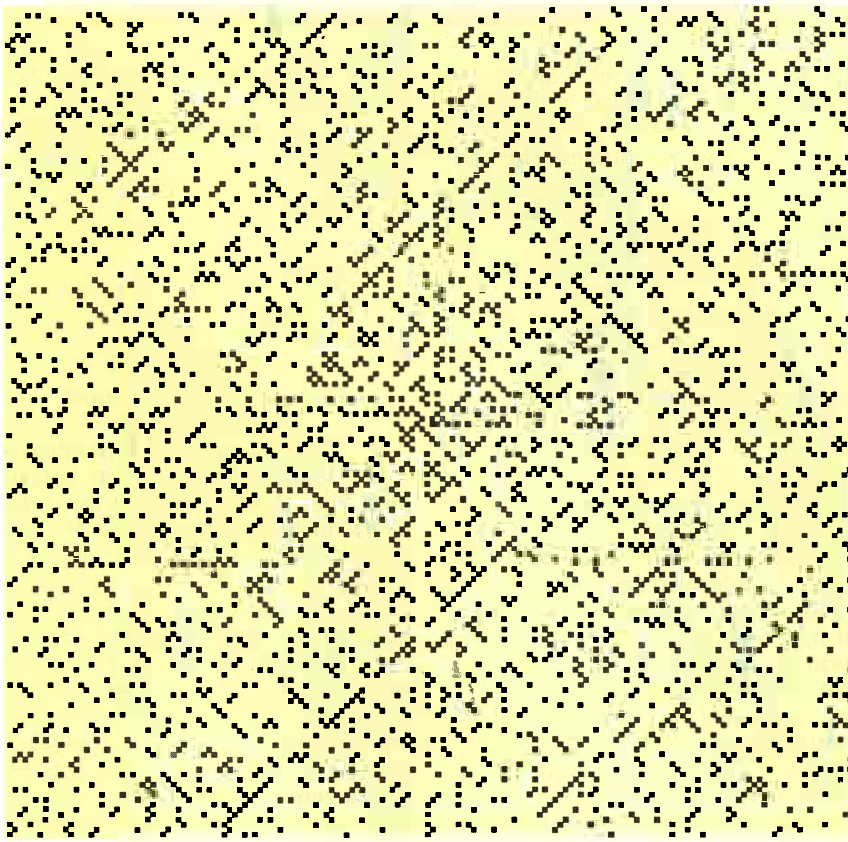
headlines: LARGEST PRIME FOUND. The confusion sometimes continues into the story. We learn that a new supercomputer has just shown that the 7,067-digit number $5 \times 2^{23,473} + 1$ is prime. It has no divisors except 1 and, of course, itself. The story might well omit (or the reader might miss) the fact that this is merely the largest known prime; soon a new, larger prime may well be found.

I hesitate to mention the largest prime number currently known. It may no longer be current by the time these words appear in print. As of this writing, the largest known prime is a 65,050-digit number found by David Slowinski of Cray Research, Inc., in 1985: $2^{216,091} - 1$. Prime numbers that have the form $2^m - 1$ are called Mersenne primes after the preeminent French mathematical amateur Marin Mersenne. All known primes greater than 1,000 are collected by another amateur, Samuel D. Yates of Delray Beach, Fla. The Yates collection is definitive. He has generously offered it to readers who send the cost of copying and postage ($3) to 157 Capri-D, Kings Point, Delray Beach, Fla. 33445.

How quickly do prime numbers thin out along the banks of the continuum? Within the first 10 numbers, four, or 40 percent, are prime. Within the first 100, the percentage drops to 25 percent. The percentage continues to drop more or less progressively. In general, the number of primes less than or equal to a number n is approximately $n/\log n$. (In this context the approximation is asymptotic. In other words, if the number of primes less than or equal to n is represented by the symbol $p(n)$, the ratio of $p(n)$ to $n/\log n$ approaches 1 as n gets larger. To quote old Yuke: "Downstream the primes thin out by the factor of a natural log.")

A few trials give some feeling for the rule. How many primes are there, according to the formula, between 1 and 100? Between 1 and 1,000? In the first case the formula yields an approximate value of 22. In the second case the formula predicts something like 145 primes.

Not surprisingly, the phenomenon of thinning-out produces more and longer stretches of numbers where there are no primes at all. To find a stretch of a million consecutive nonprimes, for example, one need only travel downstream, as Martin Gardner once did, to the number 1,000,001!. The exclamation mark does not indicate admiration but elaboration: it stands for $1 \times 2 \times 3 \times \ldots \times 1{,}000{,}001$. Teen-age terminology applies, namely humongous. But with ease we detect the prime-free stretch. If n runs from 2 to 1,000,001 in the simple formula $1{,}000{,}001! + n$, each of the resulting numbers happens to be composite. After all, in each case n divides both 1,000,001! as well as itself. Thus it divides the sum.



Stanislaw Ulam's spiral reveals a number of prime lodes

I stated above that there is no formula for prime numbers, no combination of algebraic operations on n that will produce, with a fixed number of crank turns, the nth prime. Many people have fallen prey to vain imaginings brought about by initial success. A well-known joke among mathematics undergraduates illustrates the point. It involves three ways of showing that all odd numbers are prime:

Mathematician: "Three is a prime, 5 is a prime, 7 is a prime... The result follows by induction."

Physicist: "Three is a prime, 5 is a prime, 7 is a prime, 9—experimental error, 11 is a prime..."

Engineer: "Three is a prime, 5 is a prime, 7 is a prime, 9 is a prime..."

Engineers may have the last laugh, since mathematicians depend increasingly on computers to probe for large primes.

Would it be enough to produce a formula that itself produces only primes? Pierre de Fermat, the famous 17th-century French mathematician, thought he had such a formula when he wrote $2^{2^n} + 1$. Plug in any value of n, he believed, and a prime number would emerge. Fermat's bubble was burst after his death when the Swiss mathematician Leonhard Euler found factors for the fifth Fermat "prime": $4{,}294{,}967{,}297 = 641 \times 6{,}700{,}417$.

As old Yuke might remark, "There's more than one way to skin a cat." Sometimes a visual pattern suggests a formula. Such a pattern was doodled one day by Stanislaw Ulam, the Polish-American mathematician. Attending a boring lecture, he began absentmindedly to draw a grid of horizontal and vertical lines. He numbered one of the resulting squares 1 and proceeded to number succeeding squares in a spiral around the first one:

5	4	3
6	1	2
7		

When the spiral of numbers had wrapped around itself several times, Ulam began to circle the primes with no particular purpose in mind. He sat up rather quickly, however, when he noticed an odd pattern developing. Straight lines had begun to appear out of nowhere! Ulam was immediately aware, of course, that such lines hinted at formulas for primes. The computer plot on the preceding page duplicates Ulam's pencil-and-paper experiment by replacing nonprimes with small white squares and primes with black ones.

The prominent diagonal lines correspond to prime lodes. How could one express this geology symbolically? Near the center of the diagram one such deposit proceeds down and to the left. It consists of the number sequence 7, 23, 47, 79,... The formula for this sequence happens to be quadratic: $4x^2+4x-1$.

Those with some memory of high school algebra can develop the formula for virtually any line in the diagram. It may be that the formula is rich in primes well beyond the limits of the plot. Euler (rhymes with "spoiler"—and he ruined a number of careers by anticipating so many mathematical results) had stumbled on a similar formula in the 18th century: x^2+x+41. The formula does not show up on Ulam's spiral unless one uses a different central number. A spiral that starts at 41 reveals a vein that contains 40 consecutive prime numbers before it peters out!

Perhaps it is only city slickers who mine primes by formulaic methods. Those who work the banks of the Continuum River prefer pans or, better yet, sluice boxes. In these devices, also known as number sieves, numbers are shoveled in at one end; only primes emerge from the other end. Wood ribs catch the composite numbers by a divisibility test [*see illustration on page 120*]. Sluice boxes work perfectly well inside computers, naturally.

The simplest sluice box separates primes by dividing by 2, 3, 4 and so on. If one inputs the number n at one end, the sluice box tests whether n is divisible by 2, by 3, by 4 and continues until one of the tests succeeds or the count reaches n. In the first case the number is not prime. In the second case it is. An algorithm for this model of sluice box provides the simplest framework for home-computer programs. It is called SLUICE1:

```
input n
f ← 1
for k ← 2 to n − 1
    test ← rem(n/k)
    if test = 0 then f ← 0
if f = 1 then output "prime"
```

The program accepts a number n that is typed in (input) by the human user. Then the program sets the variable f (which acts as a flag) to 1; if f still has the value 1 when the program reaches its last line, the number n must be prime. A single *if* statement is executed repeatedly inside a loop. The index k runs from 2 to $n-1$. For each such value SLUICE1 performs the division n/k, takes the remainder (rem) of the division n/k and stores the result under the name *test*. Usually *test* will be nonzero at the end of the loop. In other words, the number k does not divide n evenly. But if the division ever results in a zero remainder, SLUICE1 will immediately set the flag f to the value 0, holding it there until the loop has been completed. If the second *if* statement has a positive outcome, the program will print "prime." If f has been set to 0 somewhere along the line, only a grim silence will follow.

Although it is easy to understand, the foregoing program is too slow, particularly if it is adapted to produce a sequence of primes. The adaptation would merely involve replacing the first input statement by a loop statement, such as "for $n \leftarrow 3$ to 1,000." The final statement would be modified to print not "prime" but the value of n that made it all the way through the sluice box without being divided evenly. One by one all the prime numbers from 3 to 997 will come tumbling out, but very slowly!

Things move much more swiftly after SLUICE1 has been subjected to some tinkering. First, there is no point in testing whether the number n is prime by the division n/k if k happens to be larger than the square root of n; at least one of n's factors does not exceed its square root. The famous "fundamental theorem of arithmetic" also tells us that every whole number is the product of a unique set of prime numbers. It is not composite unless it can be divided evenly by a prime less than itself. Putting the two facts together results in a much shorter loop that uses only prime values for the index k and only those primes that are less than the square root of n.

The new algorithm, called SLUICE2, differs enough from its simpleminded counterpart to require a relisting:

```
r ← 1
p(1) ← 2
for n ← 3 to 1,000
    k ← 1
    f ← 1
while f = 1 and p(k) ≤ sqrt(n)
    test ← rem(n/p(k))
    if test = 0 then f ← 0
    k ← k + 1
if f = 1 then r ← r + 1
    p(r) ← n
```

Because SLUICE2 needs a list of primes in order to function properly, it stores these as it generates them in an array called p. The variable r keeps track of the index for the last entry of p. That way SLUICE2 always knows where to put the next prime it generates. The first line of the algorithm sets the index to 1. The next line specifies that the first member of the prime array will be 2. Then comes the loop command discussed above. It controls the testing of all numbers from 3 to 1,000. The variable k keeps track of which array element is currently being tested against n. Inside the main loop is a common kind of loop that uses the word "while"; as long as the flag is 1 and the current prime to be tested does not exceed the square root of n, the inner loop keeps chugging through successive values of k. Outside this loop f will equal either 1 or 0. In the first case a prime has been found. SLUICE2 adds the prime to its list. In the second case the main loop will simply go on to the next value of n.

Readers programming this kind of sluice box have two choices for structuring the program to print out all the primes found. SLUICE2 may print the array p all at once when the main

67	1	43
13	37	61
31	73	7

3	61	19	37
43	31	5	41
7	11	73	29
67	17	23	13

Henry Ernest Dudeney's prime square **(left)** ***and that of Allan W. Johnson, Jr.*** **(right)**

3	1	3	9	9	1
9	8	3	9	2	9
1	6	4	3	1	2
5	1	7	4	7	1
7	1	5	9	7	1
9	3	7	3	3	9

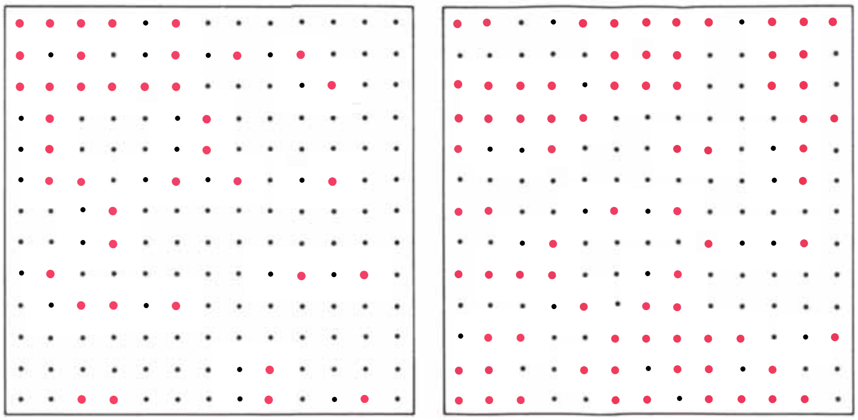
Gordon Lee's 6-by-6 prime grid

algorithm is complete. The experience is akin to opening a poke full of nuggets. It is more adventurous, some would say, to place a print command right after the line $p(r) \leftarrow n$. Then the user watches individual nuggets appear as soon as they are found.

I have been unadventurous in suggesting an iteration limit of a mere 1,000 numbers to test. There is no reason the limit should not be increased to 100,000 or even one million. Or is there? It all depends on how large an array one's system allows. The size of the array hinges on the number of primes one expects to generate. Here the prime-ratio formula comes in handy. The formula suggests that approximately 72,382 of the numbers less than 1,000,000 are prime. A computer that has only 64K of memory will not make the grade.

Prime numbers have figured in countless mathematical recreations. To continue the theme of primes in square arrays, two diversions come to mind. The first originates with Henry Ernest Dudeney, the distinguished English puzzle creator. Magic squares will be familiar to many readers as square arrays of numbers whose entries have the same sum along each row, each column and the two main diagonals. Are there magic squares consisting only of primes? The answer is yes. The 3-by-3 magic square shown on the opposite page sums to 111 (which happens to be prime) along all rows, columns and the diagonals. The 3-by-3 square is accompanied by a 4-by-4 companion. Squares of orders higher than 4 have been found. Readers who discover such squares for themselves are invited to send them in. The best example will appear in a future column; larger squares are superior to smaller ones and, for the same size, smaller sums are better than larger ones.

Another British purveyor of puzzles has issued a challenge to readers. Gordon Lee, who writes a column called "Winners and Losers" in a computer publication titled *Dragon User*, has constructed a 6-by-6 square of digits that conceals a great many prime numbers, 170 to be exact [*see illustration at right on opposite page*]. To find a prime number on Lee's grid scan along any row, column or diagonal in any direction. It may happen that a sequence of digits thus encountered is a prime number, one of the 170 counted by Lee. No more than 616 numbers (prime or otherwise) can be found in a 6-by-6 grid of digits. Repeated primes are only counted once. Lee counts 1 as a prime.



David H. Fax's blown matrix **(left)** ***and Patrick E. Kane's*** **(right)**

Can readers come up with a 6-by-6 square of digits that contains more than 170 primes? Those who write and run SLUICE1 or SLUICE2 will have a slight edge on the task. Lee suggests a strategy of seeding the square with the digits 1, 3, 7 and 9, since a prime must end with one of them. On the other hand, a square composed of only those digits would be relatively poor in primes. A judicious scattering of even numbers, including 0, might improve one's chances of meeting Lee's challenge. I shall publish the best solution sent to me (as long as it contains more than 170 primes).

In March this department dealt with a home-computer laboratory in which simulated gas molecules bounced within a closed vessel to mimic the effects of pressure. Molecules also diffused digitally from one side of a container to the other. Finally, make-believe atoms of the dangerously unstable substance I called gridium were gathered into a critical mass. A number of readers developed their own pressure vessels and diffusion chambers, but most found gridium too hard to resist. From all over North America and later from other parts of the world, reports of "computer explosions" crossed my desk.

By specifying the track of an escaping neutron with a linear equation, I inadvertently implied that two neutrons travel in opposite directions from the splitting atom. Most readers chose one of the two directions. In either case, it was not easy to get all the atoms in a plane n-by-n grid to blow up.

By plotting the results of numerous experiments, Robert Castle of Webster, Tex., found that 90 percent of the atoms usually fissioned in a 16-by-16 grid. By the time n reached a value of 32, 99 percent of the atoms were atomized. Robert M. Martin, a professor of philosophy at Dalhousie University in Nova Scotia, found a 90 percent criticality when n was equal to 19, and 99 percent of the gridium disappeared when n was equal to 39. Most other nuclear experimenters found values somewhere between these. Of those people sticking to the double-neutron burn in the original recipe, Richard W. Smith of Ann Arbor, Mich., was typical in finding lower critical masses. He reports a 98 percent burn when the grid measured 15 by 15.

Few readers developed a system as sophisticated as the program called SHAKEY. Developed by Robert B. Merkin of Northampton, Mass., SHAKEY not only can handle very large grids but also incorporates a great many time-saving techniques to speed up the essential boom. SHAKEY's home reactor allows different grid spacings to be tried and emits clicks at every decay like a Geiger counter. Merkin says he will be glad to share SHAKEY with readers who write to him at 55 Milton Street, Northampton, Mass. 01060.

Dramatic samples of blown matrixes were sent by David H. Fax of Pittsburgh, Pa., and Patrick E. Kane of Champaign, Ill. Their productions are shown above.

FURTHER READING

THE QUEEN OF MATHEMATICS. Eric Temple Bell in *The World of Mathematics*, edited by James R. Newman. Simon and Schuster, Inc., 1956.

AMUSEMENTS IN MATHEMATICS. Henry Ernest Dudeney. Dover Publications, Inc., 1970.

THE SEARCH FOR PRIME NUMBERS. Carl Pomerance in *Scientific American*, Vol. 247, No. 6, pages 135–147; December, 1982.

ONE MILLION PRIMES THROUGH THE SIEVE. T. A. Peng in *Byte*, Vol. 10, No. 11, pages 243–244; Fall, 1985.

BOOKS

Terrae and maria, the spiky virus, discovering Luie, old wives' botany

by Philip Morrison

The Geologic History of the Moon, by Don E. Wilhelms, with sections by John F. McCauley and Newell J. Frank. U.S. Geological Survey Professional Paper 1348, U.S. Government Printing Office. Books and Open-File Reports Section, U.S. Geological Survey, Federal Center, Box 25425, Denver, Colo. 80225 ($33).

Earth history is a romantic epic. It resembles a bravura opera, offering entangled, diverse themes and bizarre orchestral color. At one time it is dominated by the drift of great plates, at another by ice and water. Everywhere are seen the signs of life in its astonishing variety. Contemporary change is still lively, and it evokes the past.

In contrast the geology of the moon seems an early work of chamber music, austere, contrapuntal, closed. That score is here, a very readable if technical account of the exotic stratigraphy that has developed since the formation of the entire small planet. The giant paperback is crammed with evidence: the photogeological images of the airless and waterless moonscape, marked by complex rock forms; the telling geochemical and geophysical analyses, some made in remote overflight, some made of samples returned from those distant rocks. The surface of the moon is here geologically periodized and carefully mapped. The information comes from almost 40 spacecraft, Soviet and American, sent there between 1959 and 1976; nine of them were Apollo missions.

Two forces have shaped the lunar crust. Almost every highlight is traceable to the impact of projectiles from orbit. Their shocks and rock splash have carved and littered the rugged surface at every scale, from the universal impalpable dust to a strike that patterned half of a hemisphere all at once. The darker features, seas of lava, are the work of the moon's interior. Visible to every human eye, they have been recognized as smoother surfaces, ever since Galileo first examined them telescopically from some garden in Padua. The two kinds of terrain are still called by the old names. The uplands, or terrae, that cover five-sixths of the whole, are the bright, rough parts; the seas, called maria, are the darker ones.

A dozen chapters tell the tale. The first six outline the methods and use case histories of particular formations to make the arguments clear. Many hypotheses have their origins in eyeball geology, the only kind of investigation possible before the era of sampling and detailed visual coverage. Some pioneer earthbound judgments were prescient, particularly the work of Ralph Baldwin in the 1950's and 1960's; some were not. The latter half of the book presents and supports a systematic view of the lunar surface as it formed. All the features we see are grouped into five periods, called systems. They are successive, superposed sets of layers. The last big spread of the book shows them mapped in color from earliest to latest.

The moon is scarred with big craters shaped by energies simply too large to have been stored within the moon rocks themselves. An impact crater comes from "the meeting of an irresistible force with an immovable object." There are tens of thousands of sizable craters, small and large. They have rims, walls, peaks and rays. The ejecta make new craters, limited in scale by the escape velocity from the light moon. The presence of those secondaries allows the ordering of crater events in time. By now it is clear that counts of primary craters similar in size on comparable surfaces give reliable relative ages. Craters serve more or less the same role in the study of the history of the moon that fossils serve in geology; a big, much eroded crater is almost as significant as a trilobite. Famous anomalies fade. The apparent strangeness in some crater alignments dissolves once the craters of a given age are mapped separately.

Exploration made it clear that the smooth, flat maria were in fact made of darker material, quite rich in iron and titanium, like basalt on the earth; all nine sampling missions brought back such samples. The maria are congealed flows of lava. Mare Imbrium shows clear lobate patterns like those found on the earth, patterns of congealing flow in shallow relief that mark wide surfaces. The thickness of the flows is harder to be sure about, but the flooding of preexisting craters and penetration by newer ones give hints; both gravity and radar data suggest layers up to a couple of kilometers thick. The old impacts and their consequent crustal defects gave rise in time to extrusion of molten magma that flooded forth after partial melting at depth. Many details diagnostic of lava emplacement, such as scabland forms, sills and dykes and collapsed lava tubes, can be associated with curious features seen on the maria.

All the maria lie within basins: shallow old craters of great extent, now detectable mainly through the outlying rings of highland terrain. This is the most novel feature of the present synthesis. The sight of double rings outside Mare Orientale was the first clue to such structures. They probably arise from waves in shocked, fluidized postimpact surfaces, only minutes after some ancient asteroidal collision. A display of a few dozen lunar basins arranged by increasing size from 300 kilometers up to eight times that diameter (one at the South Pole) prepares the reader for a table of less definite ones, marked by rim-form features and other near-circular ridges and rings that today can be made out only in part. The greatest of the obscure basins occupies the north and left quarter of the face of the moon we see from our Northern Hemisphere. Within it are Mare Imbrium and much more; the basin preexisted the big lava flows, and its formation weakened and thinned the crust locally to make them more likely.

There is a strange global asymmetry on the moon. About 30 percent of the visible face is under dark maria but only 2 percent of the far side is. This is no longer an enigma: it is simply that the biggest impact we know took place by chance right where we can see it. Then within that huge basin the basalt flowed time and again. One of the most impressive photographs, albeit a small one, is a telescopic view of the full moon. Ewen A. Whitaker of the University of Arizona has marked the image so that even a tyro can just about make out the many-ringed basin of Procellarum, the name given this oldest and largest of known impacts. Mountain building from any internal processes seems nearly absent on the moon. All the ridges and highlands turn out to ring some half-ruined basin of the deep past.

In our time the moon is as tranquil as it looks. *Apollo 17* landed near, and sampled, a landslide area. Its age was determined from cosmic-ray exposure of rocks brought to the surface by the slide, which may have been triggered by the impact of projectiles flung up at the formation of bright crater Tycho, not far away. A few such projectiles can be seen near the site. Don Wilhelms assigns this event to his last period of lunar history. The period began with formation of the crater Copernicus, .8 billion years ago. The date comes from radioargon analysis of samples brought back by the crew of *Apollo 12*. The Tycho explosion started that slide .1 billion years ago, and cosmic-ray exposures show that three small craters, sampled by *Apollo 14, 16* and *17,* are all less than 50 million years old. The adopted ages for the named lunar periods are listed; all absolute ages are based on samples. Meteorites still strike; the lunar seismometers picked up the signal from a large one that hit the far side in 1972. Only 100 or so bright, conspicuous big craters and a portion of the basalt flows in Mare Imbrium and westward of it have formed during the past three billion years.

The book ends as it ought to, with a geologist's field proposal: a list of 18 places where future samples of the regolith should be sought and returned by robot probes. Top priority goes to the old basin of Mare Nectaris, whose relative age is well known; an absolute date would determine the ancient cratering rate before the maria were formed. Speed the day!

INFLUENZA, by Edwin D. Kilbourne. Plenum Publishing Corporation ($45).

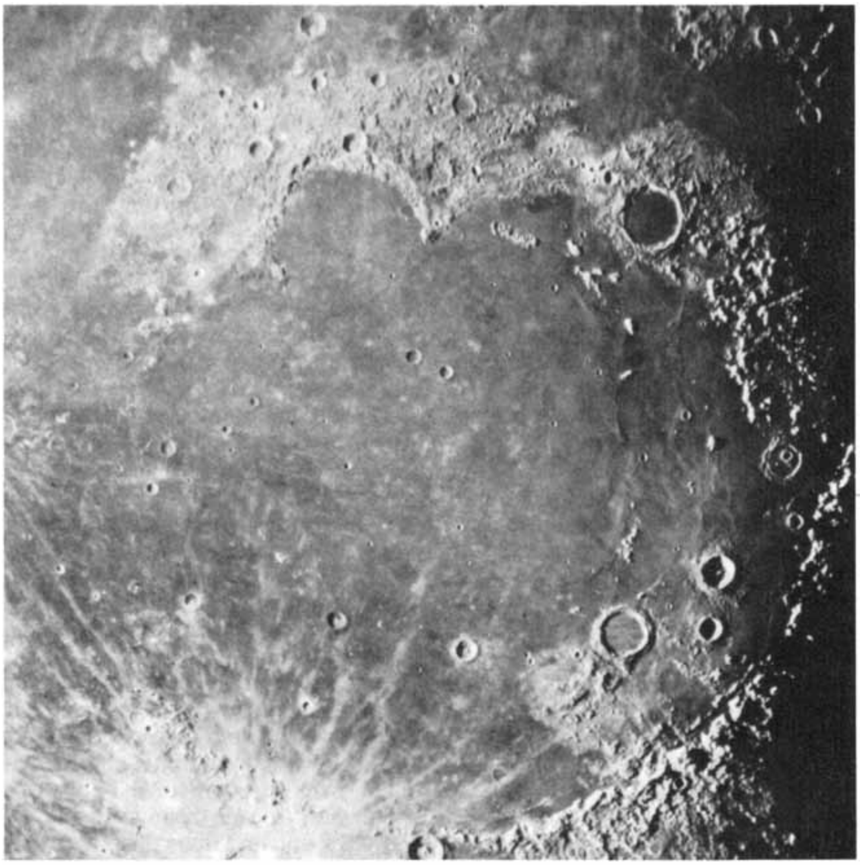
The electron micrographs tell a dramatic story. Influenza is a skin-deep but profoundly incapacitating malady; the cells that are attacked and swiftly killed are essential ones. They are the short-lived surface cells that line the respiratory passages. They bear a dense fringe of beating cilia that sweep out secreted mucus, steadily cleansing the moist and intricate airways of nose, throat and lungs. Infected, the busy cell is doomed; it buds and bursts with the virus, and is sloughed off. The smooth vital cells below are then vulnerable. But in a week or so the missing fringed cells are replaced and the disruptive illness is gone. Influenza is a brief disease from which "recovery is rapid and uneventful." (Runny nose and sore throat are directly understandable, and fever is a debilitating although therapeutic reaction of the body, but those characteristic aching muscles remain a puzzle, since the virus is not usually found far from the respiratory tract.)

Why worry about flu? Indeed, except for the professionals and those of us who are actually infected, we do not seem to worry much. Even the elderly, who along with young children are most at risk, are usually unreached by the cheap, effective and constantly evolving vaccines now in use. In most years, after all, fewer than one patient in 10,000 will die of influenza. The virus passes easily, however, from person to person, so that when it is going around, a great many people will fall ill. Correspondingly sharp spikes of mortality rise every few years well above the tremorous background of deaths from respiratory infection in American cities. In 1918 the November spike became a terrible four-month tower. Worldwide 20 million died, half a million of them in the U.S.

It appears that most fatalities came from primary viral pneumonia; the virulent strain did not limit its growth to the upper respiratory passages but heavily infected the bronchial passages of one patient in 100 or more; there were and are other complications, such as bacterial pneumonia, particularly in that war-weakened world. Similar "undisputed pandemics," if an order of magnitude less virulent, were endured in 1946–47, 1957 and 1968. There was a false alarm in 1976. In 1968 about 50 million Americans fell ill, and death came to about one in 2,000 flu sufferers; 1957 was rather worse, 1946 rather better.

Edwin Kilbourne has spent a virological lifetime in battle with the flu; he notes that his immune cells are marked with the imprint of the 1918 strain, and his brain cells with all the rest. His book is an effort to make overall sense of the baffling disease in this time of powerful molecular biology. He covers in 11 chapters the history, the virus, the human disease and the paths to influenza control. No disease has been better studied (although by now the human immunovirus might enter a counterclaim) and few are "less well understood."

There is plenty of credible, if not



The vast Imbrium basin, rimmed by highlands

conclusive, clinical prehistory of flu epidemics beginning with the horse colds and cattle murrains that for four centuries went along with acute human "catarrhal fevers." The modern period begins with serological archaeology. Since the 1930's blood tests have been detecting the circulating antibodies people formed to flu as far back as the outbreak of 1889. These inferences are given strength by the later recurrence of the antigenic patterns under the eyes of the virologists. Virus subtypes of the major virus known as influenza A dominate attention; the B type is a very similar, less important but by no means negligible virus; C is now seen as a distinct and simpler parasite, a parainfluenza.

In 1918 the A virus was of the antigenic pattern now called H1N1; the subtype came again in 1946, when it was first seen in Australia. Each time the old form to which many were immune was supplanted and all the new susceptibles were attacked until the subtype was burned out. Then the next variant came, and the next. In 1976 clinical and serological tests showed that the 1918 subtype seemed to be back in a form of swine flu first seen in New Jersey. But the agent failed to "poke its nose into the tent" and displace its predecessor. Forty million people were given a quickly made vaccine, within two months. The campaign was contentious: there were some bad side effects, but no epidemic disease appeared. Better that, the author says, than the other way around. "At this writing" viruses of the H1N1 subtype continue to circulate right along with the Hong Kong subtype of 1968.

The text at hand, although technical for the nonspecialist reader, is rather exciting. You can see an electron micrograph of a neatly twisted skein of molecular yarn, the RNA of influenza A, coupled tightly to nucleoprotein. That virus holds eight such single-strand pieces of RNA, packed along with the specific enzymes. Once a host cell has been entered and taken over, the enzymes will read out the message stored in the RNA to make the proteins of the next generation. The envelope of the middle-size virus, which can take near-spherical or even filamentary form, is a complex lipid bilayer that the virus derives by budding out through the wall of the host cell. That membrane is lined by a special dense viral protein and armed on the outside like a sea mine with 400 or 500 external spikes, all together half the mass of the virus. They are its entry weapon into host cells and confer antigenic specificity on the virus; they are the target of most vaccines, and of the immune system.

There are two kinds of spikes, one glycoprotein making each spike. One form, 135 angstrom units long, has a triple globular end, held by a set of protein helical rods. It is suited to bind certain cell-surface receptors of the host. It comes in a dozen antigenic varieties. Most of them are limited to the flu viruses that infect various warm-blooded animals, mice to mallards. A few, called H1, H2, H3, are common human forms. The second spike type is a mushroom-shaped implant about 100 angstroms long, bearing an enzyme able to attack the mucoproteins of the host cell surface to enforce entry. Its forms are labeled N1, N2, N3.... All this nanoarchitecture is well known; no virus protein has been better characterized. The RNA text too has been deciphered letter by letter, and its replication studied.

The battle is joined. This virus has the genetic capability to reassort its eight RNA genes. It may take up host information as well. It can also mutate at any single point. Sometimes those point mutations, which steadily accumulate over time, are insignificant typos; sometimes even a single amino acid change can make a known difference. Thus the fight against flu is a deeply evolutionary one. Virus strategy is to hide in some unusual host, in some town or farmyard, and vary this way and that, until it is able to avoid the common antibodies already out there. Then it can spread. Human immunity is solid once the system is informed, but that first strike on entry is our Achilles' heel. Nearly always the host survives, new virus is shed to seek out newly susceptible hosts, and both organisms live on, the host immune as compensation for its brief discomfort.

We still do not understand the polygenic origins of influenza virulence; even the quantitative yield of infective virus multiplication could make all the difference between a disagreeable sore throat and a fatal pneumonia. Human fatalities seem after all to be only an exceptional and unintended by-product of virus adaptation. The swift flu virus seems a simpleminded enemy in contrast to HIV. That slow and sneaky retrovirus disguises itself to deceive the immune system in half a dozen ways, even attacking the cellular corps of guards. Maybe we can adapt first to each of its protean new varieties, and end them on entry by an ever adapting arsenal of vaccines. This is an arms race we must undertake; unless we do, we need to fear someday another pandemic, the accidental but fateful airborne influenza, every person at risk.

DISCOVERING ALVAREZ: SELECTED WORKS OF LUIS W. ALVAREZ, WITH COMMENTARY BY HIS STUDENTS AND COLLEAGUES, edited by W. Peter Trower. The University of Chicago Press ($37.50).

Some 20 papers in experimental physics first published from 1932 to 1980 are reprinted in this kaleidoscopic collection. To each is adjoined a personal commentary by a participant in the research presented. The papers show the extraordinary power the ideas and methods of physics have gained since the golden year of 1932. The rise of statistical physics, the successes of the quantum atom, the understanding of the nucleus are evident. So disciplined was the ballet of theory and experiment, of insight with instrumentation, that we are not able to resolve cause from consequence. More than that, the physicists have freely exported their new powers of synthesis and invention. The technology of war gained only the most urgent of those bestowals. Other activities have also been transformed, including navigation and cartography, geology, archaeology, biochemistry, even the systematic study of everyday affairs such as accident and crime.

That sounds like a plausible basis for an ambitious anthology. But in fact this book is a selection of the work of just one original, audacious and penetrating man, accompanied by the personal and readable recollections of dozens of his friends and partners. Luis Walter Alvarez was a Chicago undergraduate in physics. He earned his Ph.D. under Arthur Compton for cosmic-ray studies (including a collaboration with the balloon-borne Piccards as they ventured up into the stratosphere). He came to Berkeley at about the same time that the 60-inch cyclotron went into operation. He used it well. Indeed, he has outlasted and outshone it for 50 years. There is a style in every big laboratory, a style that in our time first arose at Berkeley. Although the style is strongly modulated by a particularly vivid personality, it can be heard throughout this volume.

A few samples will show what you can learn about one physicist, his ideas, his community, his failures and his triumphs. Of the prewar reports, the use of the cyclotron as mass spectrometer is the most delightful. The idea was to make an easily detected high-energy beam out of the possible but then unknown rare isotopes of hydrogen and helium that had an atomic mass of 3. One page carries two brief publications, over the signatures of Alvarez and his colleague Robert Cornog, whose good-humored memories of those days accompany the reprints. The helium 3 showed up as a steady if weak beam from both gas-well and atmospheric helium gas. It was clear then that the hydrogen isotope, the fateful tritium, would be unstable. Cornog reports that among those of the Berkeley staff who helped him with his 1940 thesis, five won a Nobel prize. "The easiest way to win a Nobel prize at Berkeley," Cornog concludes, "was to have helped me with my thesis work!" Hans Bethe comments on Luie's admiration for Bethe's famous reviews, in which the issue of the stability of the isotopes of mass 3 was nicely treated: "My articles mainly stimulated him to prove me wrong! Whenever I had stated that a certain experiment could not be done, he went ahead and did it."

Almost 30 years later the first beam of radioactive tritium was produced at Berkeley (in an 88-inch cyclotron) by another young colleague of Luie's, Richard Muller. He writes of it here as part of a fruitless but ingenious search for free quarks. The accelerator technique is now standard, applied worldwide for radioisotope dating of small samples and slow-decaying isotopes. "Many apparently wild jumps in research are really the revival of old ideas...given new life by developments in theory or technology."

Luie is a magician of these wild cards, not random but artfully played for big stakes and small. Once he organized the X-raying of the pyramid of Chephren, son of Cheops, to seek its treasure chamber. Directional spark chambers below the great structure mapped cosmic-ray mesons arriving from the sky. The masonry above proved solid, although the coarse-grained limestone turned out to be much less dense than the handbook values—"surely the hard way to measure density," Luis writes.

Here are a few of his inventions, such as variable-focus eyeglasses and image stabilizers for binoculars. Here is his calibration of the uncertain speed of the home movies taken of the stricken president's car as it sped away after the shooting in Dallas, based on understanding that the energy spent in applause rises with the cube of the frequency of the visible clapping. Here is a continuing story, informal testimony and the first formal paper, of what came out of Luie's wonderful idea for checking the deposition time of the layer of clay found the world around at the Cretaceous-Tertiary boundary. Neutron-activation analysis showed parts per billion of iridium, brought in by an errant comet. Nearly a decade later such intervention from the skies remains a major evolutionary issue, a celestial joker in the game of adaptation.

A few lines must be spent to outline Alvarez' experiences in World War II. A nuclear physicist early caught up in radar work, he devised the scheme that is now a keystone of air traffic control. Aircraft carry a coded transponder that automatically signs each plane's own blip on a distant screen; on arrival they can be safely talked down to fog-hidden runways by the quiet guidance of operators using a precision radar system Alvarez swiftly adapted to the task. A civilian at war, he flew in escort along with the nuclear bombs over the Japanese cities. The book reproduces the handwritten note that parachuted into Nagasaki along with Alvarez' instrument microphones; it is an amateur's appeal to Japanese physicists for quick reasonable surrender.

With the help of the physicist at the Virginia Polytechnic Institute and State University who was the editor and spirit of this collection, Luis has published a recent autobiography as well. It is good reading. This exhibition of the stuff itself is even more apt. Richard Garwin, himself no tyro at daring analysis, describes Luie's exemplary process of inquiry: "The physicist *always* asks, 'Do I know this?' 'How do I know this?' and 'Is this still true?'" Like Luis Alvarez himself, such curiosity is never idle.

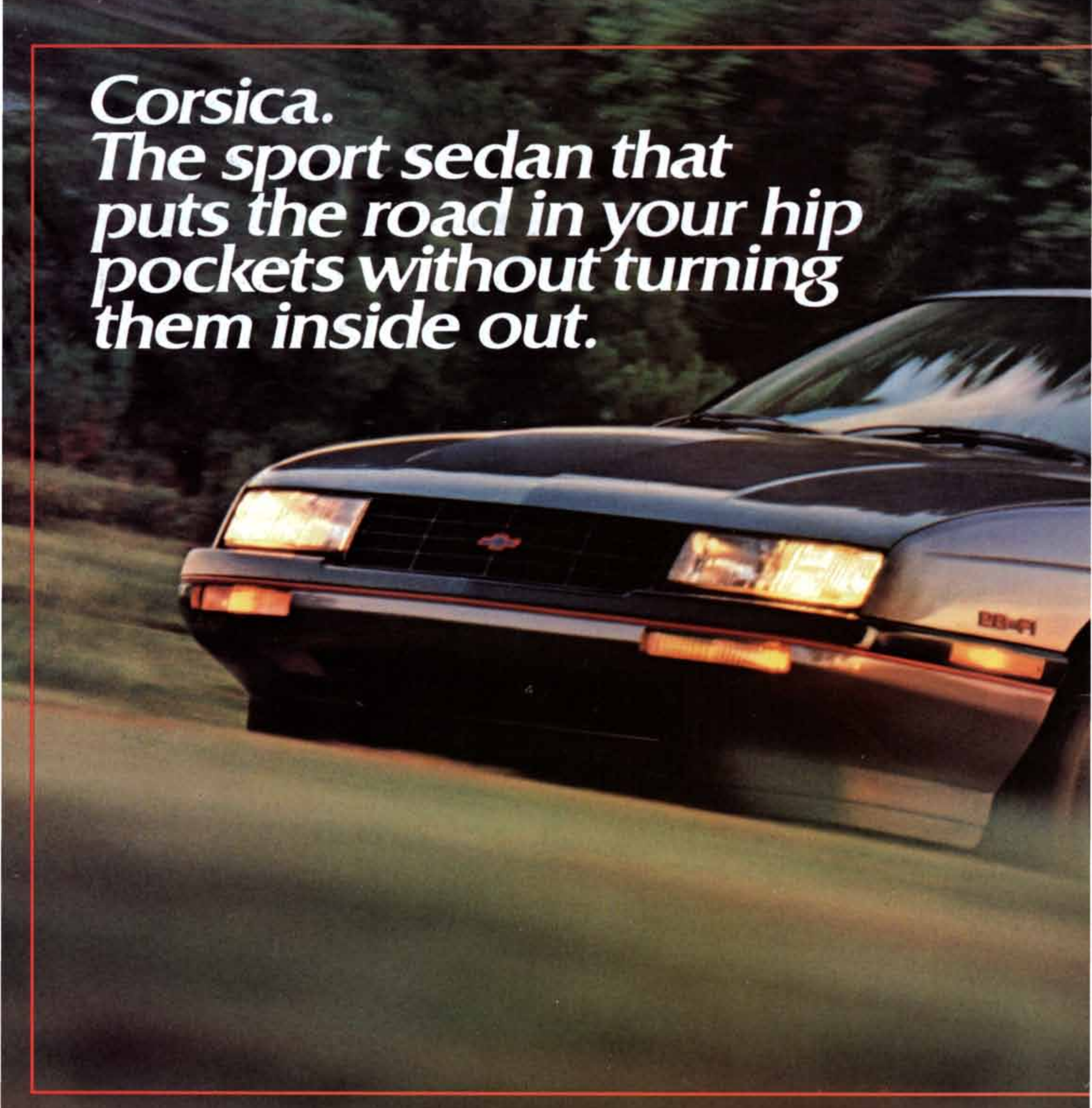
HERBALS: THEIR ORIGIN AND EVOLUTION, by Agnes Arber. Third Edition. Cambridge University Press (paperbound, $24.95).

The Renaissance has two aspects: learning long dormant in Europe was revived, and the freshness of rebirth then quickly carried learning far beyond its earlier limits. Exemplifying the trend is this famous work of scholarship first published in 1912 and then revised in 1938. The "gracious and...erudite" author is a paleobotanist and plant morphologist of high distinction (the second woman scientist admitted to the Royal Society of London over a mere 280 years). She turns her unfailing taste, lucid style and critical insight to "a chapter" in the history of her science. The book is a welcome reprint in a new series of science classics; with 130 old figures and 20-odd plates in black and white, it is a pleasure for eye and mind.

A set of figures makes rebirth plain. The first reproduction is that of a page in a Byzantine manuscript, now in Vienna. The plant, a robust thistle, is recognizable to any gardener. The artist has seen the living plant and drawn it well. The classical text illustrated, a treatise on medicinal plants, was written by Dioscorides, a physician of Asia Minor in the "time of Nero and Vespasian." The illustrations carried on the tradition of an even earlier doctor, one Krateuas (Cratevas), who on the authority of Pliny had "produced...a herbal containing colored pictures of plants." Dioscorides' guide remained famous, in use even up to our times. In 1934 the director of the Kew botanical conservatory worked with an Official Botanist, a monk at Athos, who carried a manuscript copy of Dioscorides to help name the plants he gathered.

That is prologue; Agnes Arber's volume is a study of early printed herbals, botany made publicly visible. She presents in evidence a series of five drawings of the water lily. First, out of a ninth-century manuscript, we see the plant all symbol, three symmetrical circular blossoms on straight stems. Next is a 1481 printed version, a woodcut much the same as the drawing. Third is another woodcut done in 1499; the triune form is still dominant. But this later artist has looked at the flowers; they are no longer neat circles. Instead the plant is somewhat lunate; its stem is realistically scaled. Next we see in Otto Brunfels' herbal of 1530 the white water lily drawn lightly and compellingly from life. Finally, an equally clear likeness appears in a woodcut by Mathias de l'Obel of Antwerp. Made in 1581, the woodcut emphasizes texture and presents a darker, coarser line. The world outside has come into plain view, although always as seen through the eye of an artist. By 1623 the momentum of botanists had rolled on: a Basel collector-artist prepared a chart that identified 6,000 species.

The classical tradition of medicinal plants thus became the takeoff point for what we know today about the plants of the fields, and more. It is agreeable to recall that intelligent collection in the service of healing was for a long time the province of old countrywomen; mother passed on the lore to daughter. Brunfels, author of one of the first herbals to enter print (the gifted artist was Hans Weiditz), alluded in his book to those "highly expert old women" who could act as source.



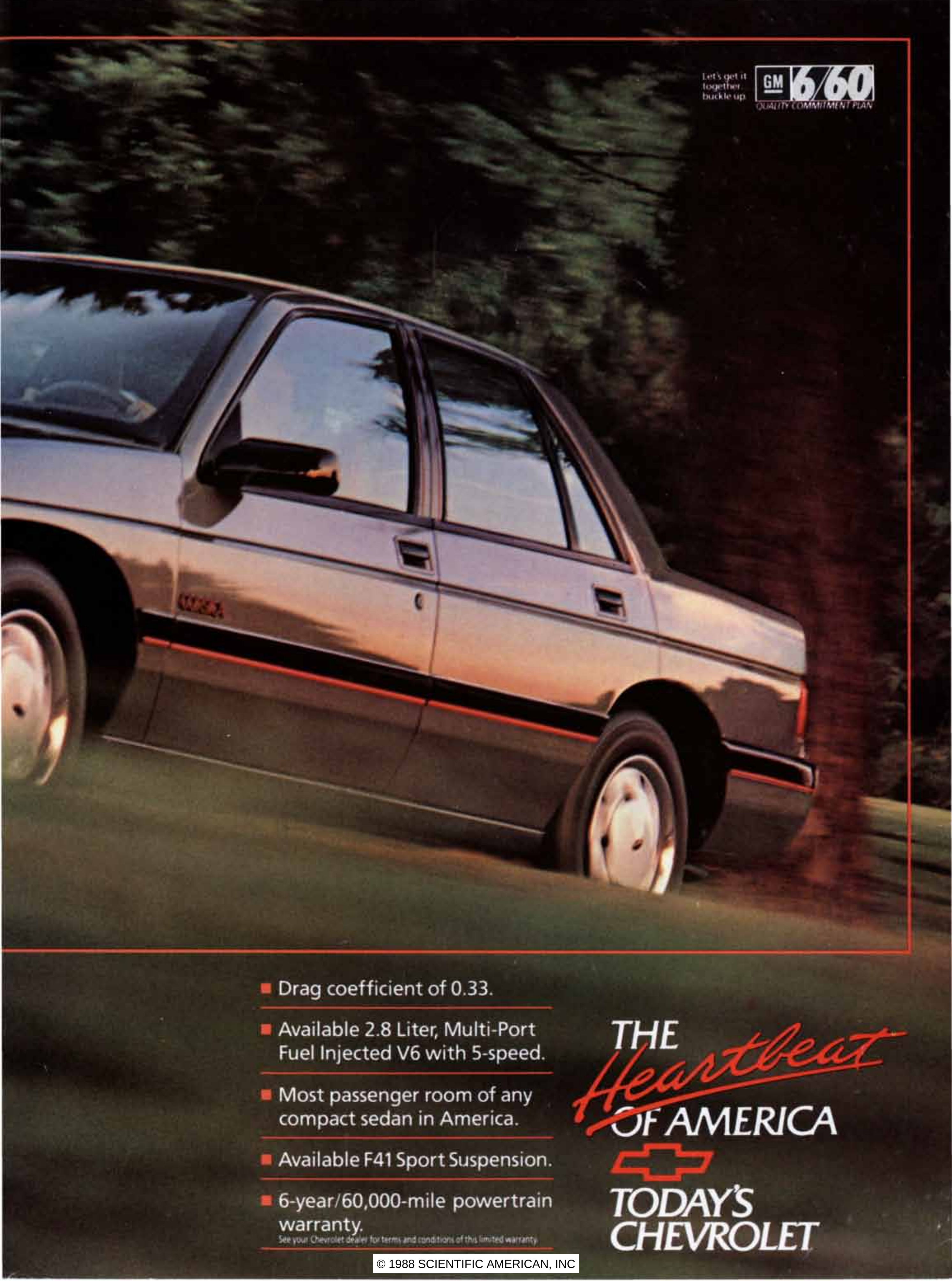
Corsica. The sport sedan that puts the road in your hip pockets without turning them inside out.

Unlike some, we do not measure a sport sedan's performance capabilities by the number of figures to the right of the dollar sign.

Hence, new Corsica. It has a drag coefficient of 0.33. Sleek enough to beat the Porsche 928 S in the wind tunnel.

But since you probably don't often drive in a wind tunnel, Corsica also offers an optional 2.8 Liter, aluminum-head, 130-horsepower Multi-Port Fuel Injected V6. Just in case you're interested in power. (Obviously, we are.) It has a Getrag-licensed manual 5-speed that's as quick as any hairpin you're likely to encounter. And it has rear trailing arm suspension to leave tight curves asking what just went by.

Which leaves only the price. And while Corsica takes great pride in matching or exceeding certain other sport sedans in many areas, that of price isn't one of them.


Let's get it together. buckle up.
GM 6/60
QUALITY COMMITMENT PLAN
Drag coefficient of 0.33.
Available 2.8 Liter, Multi-Port Fuel Injected V6 with 5-speed.
Most passenger room of any compact sedan in America.
Available F41 Sport Suspension.
6-year/60,000-mile powertrain warranty.
See your Chevrolet dealer for terms and conditions of this limited warranty.
THE Heartbeat OF AMERICA
TODAY'S CHEVROLET