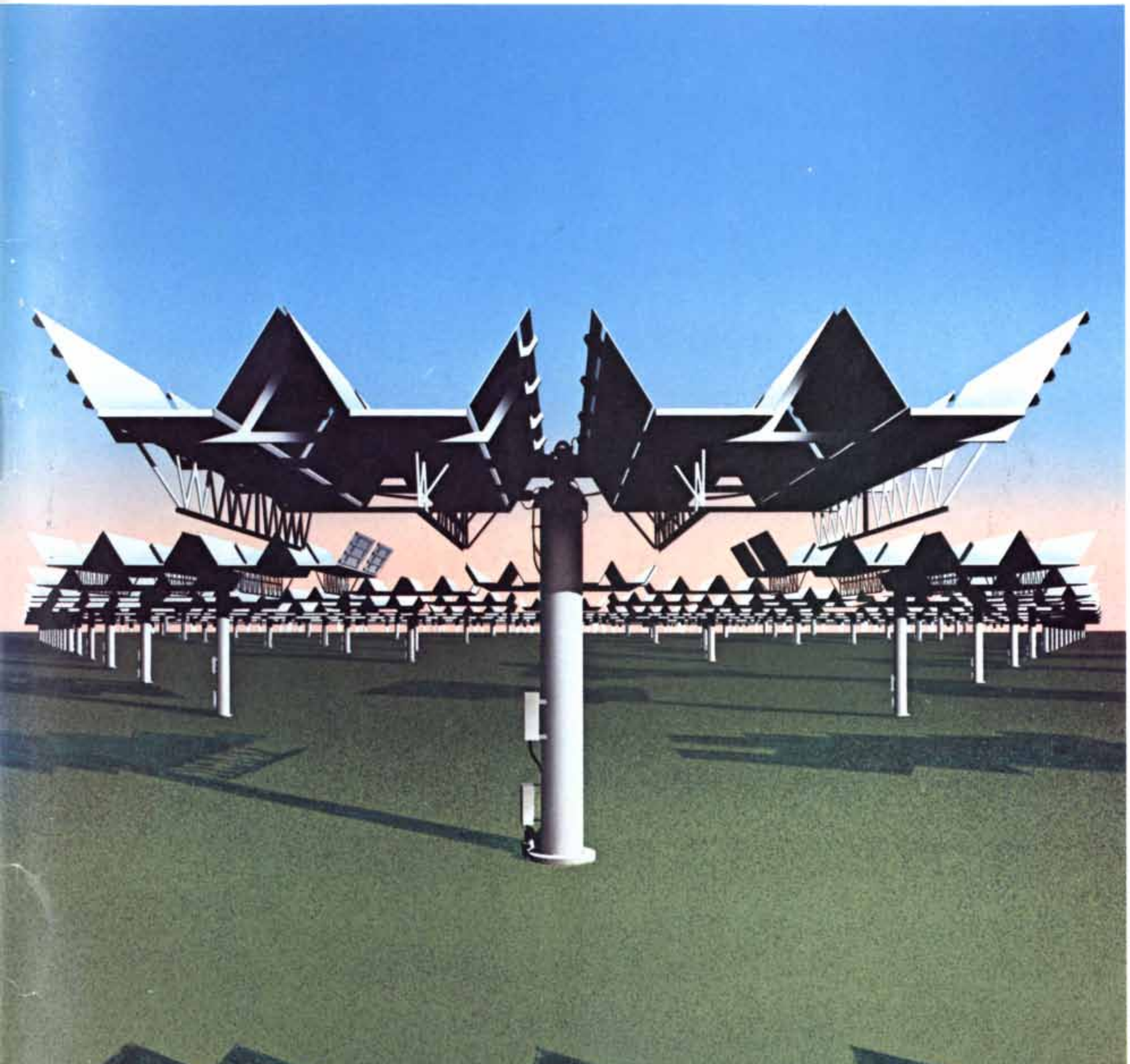
SCIENTIFIC AMERICAN

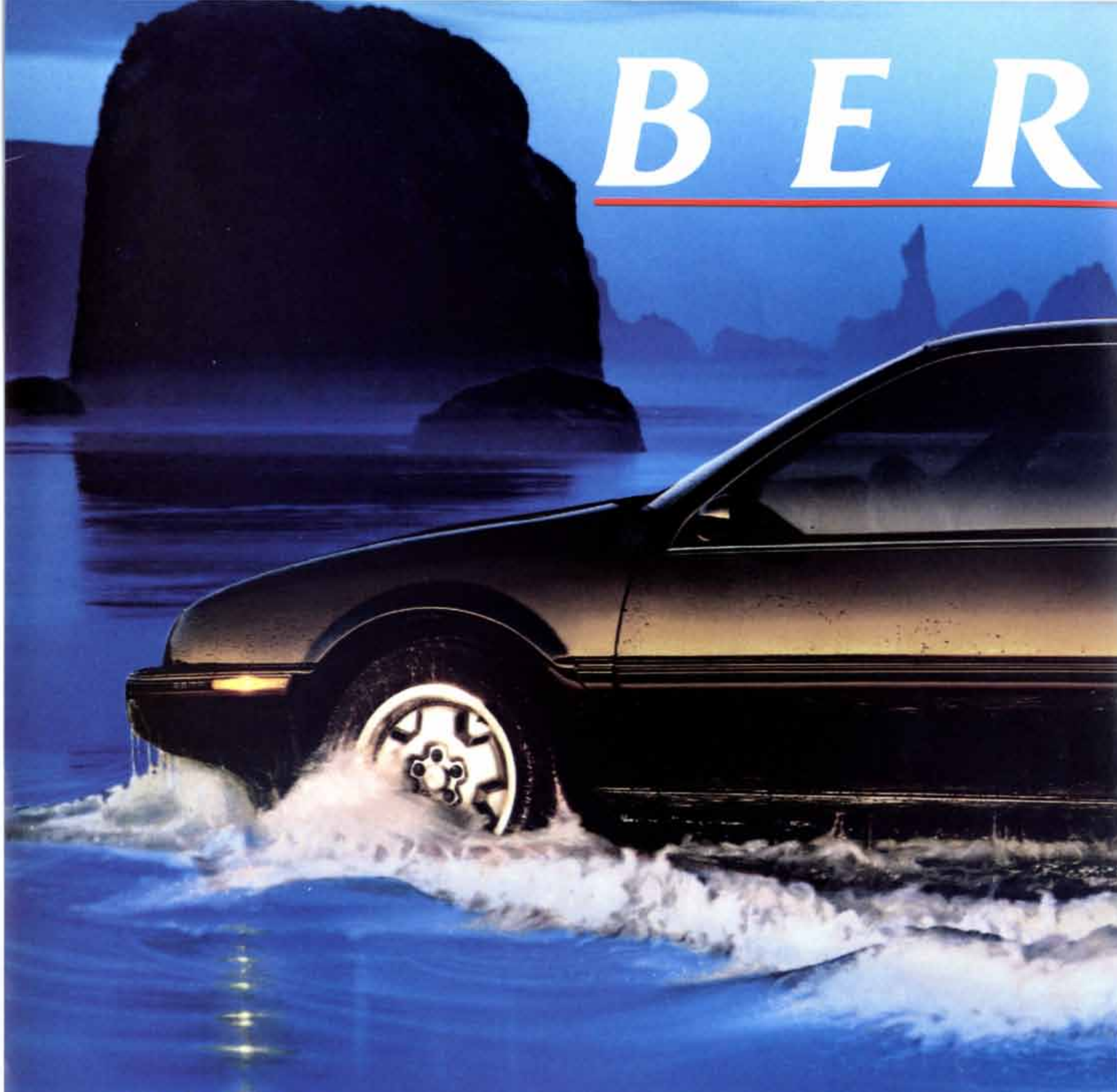


POWER FROM THE SUN

$2.50

April 1987


DISCOVER
A NEW SPECIES
FROM TODAY'S
CHEVROLET.


B E R


ETTA
BERETTA GT

Established 1845 **SCIENTIFIC AMERICAN** April 1987 Volume 256 Number 4

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1987 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

THE COVER

The cover shows part of the largest photovoltaic power plant now in operation, a 7.2-megawatt installation occupying 160 acres at Carissa Plains in central California (see "Photovoltaic Power," by Yoshihiro Hamakawa, page 86). The photovoltaic solar cells convert energy from the sun directly into electricity. Groups of solar cells wired together at the factory make up modules; each of the panels contains 16 of them. The plant was built in less than a year by ARCO Solar, Inc., and the Pacific Gas & Electric Company.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
14	Edward Bell
16	Music Publishing Services, New York
18	Edward Bell
20	Music Publishing Services, New York
30–39	Hank Iken
41	Scott Mittman and Maria T. Maglio
42	Carol Donner
43–44	Alan D. Iselin
45	Julie L. Schnapf and Denis A. Baylor
46–47	Alan D. Iselin
49	U.S. Geological Survey, Flagstaff, Ariz.
50	George Retseck
51	Jet Propulsion Laboratory, California Institute of Technology (*left*); George Retseck (*right*)
52	Jet Propulsion Laboratory
53	U.S. Geological Survey
54	Jet Propulsion Laboratory
55	U.S. Geological Survey
56	Jet Propulsion Laboratory
57	U.S. Geological Survey
58	George Retseck (*left*), Jet Propulsion Laboratory (*right*)
59	U.S. Geological Survey
60	George Retseck (*top*), Jet Propulsion Laboratory (*bottom*)
77	David Ho, Cedars-Sinai Medical Center
78–84	Thomas C. Moore
86	Sanyo Electric Co., Ltd.
88–92	Gabor Kiss
95	J. Wyatt Durham
96–98	Tom Prentiss
99	Mark A. S. McMenamin
100	Mark A. S. McMenamin (*top*), Geological Survey of Canada, Ottawa (*bottom*)
101	Tom Prentiss
104–104A 104A	Chinese Music Society of North America
104B	Andrew Christie
105	Chinese Music Society of North America
106–108	Andrew Christie
109	Andrew Christie (*top*), Chinese Music Society of North America (*bottom*)
110	Andrew Christie
112–113	Kathy Schick
114	Andrew Tomko (*top*), Patricia J. Wynne (*bottom*)
115	Ed Hanson
116	Nicholas Toth
118–119	Kathy Schick
120	Ed Hanson
121	Kathy Schick
123–126	Michael Goodman

LETTERS

To the Editors:

Behaviorism has not "traditionally treated instinct as irrelevant to learning," as James L. Gould and Peter Marler say in "Learning by Instinct" [SCIENTIFIC AMERICAN, January]. Seventy-nine years ago behaviorism's founder, John B. Watson, reported species differences in learning in "The Behavior of Noddy and Sooty Terns" and explained those differences as interactions between instinctive and learned behavior. He continued to study the relations between learned and unlearned behavior throughout his academic career. The leading behaviorist B. F. Skinner has had a continuing (but unrecognized) interest in innate behavior. His earliest papers, which appeared almost 60 years ago, were about innate factors in behavior; so is one of his most recent papers. His extended analysis of instinct and learning, "The Phylogeny and Ontogeny of Behavior," appeared in *Science* more than two decades ago.

The idea that all stimulus-response associations are equally open to conditioning was never tenable. Response preparedness has always been a concern in behaviorism. Watson studied burrowing and maze running because rats naturally do those things well. Behaviorists make use of pigeon pecks not because they are easy to teach but because they come "ready-made." John Garcia did not deal a "severe blow" to "the idea that any perceptible cue could be taught...as a conditioned stimulus"; he continued the behaviorist tradition of careful, detailed analysis of factors affecting learning.

Imprinting too is neither new to behaviorists nor foreign to their thinking. Watson described it in 1908, Neil Peterson's classic study on the relation between imprinting and operant conditioning appeared in 1960, and behaviorist journals continue to publish research on imprinting.

The contrast Gould and Marler paint between ethology and a mythical and unrealistic "early behaviorism" creates ill will and misunderstanding and does nothing for scientific communication. Their "new synthesis" can work only if both behaviorism and ethology are recognized as important specialties, neither viewing learning and instinct as "diametric opposites."

JAMES T. TODD

University of Kansas

To the Editors:

"Behaviorism" may mean different things to different people; our characterization, however, is well supported by textbooks and the writings of Watson and Skinner. For example, the most recent edition of the widely used text *Introduction to Psychology,* by Clifford T. Morgan and R. A. King, states that, in emphasizing learned behavior, behaviorism "denied the existence of instinct and inborn tendencies." Watson, in his enormously influential book *Behaviorism* (written long after his work on terns), summarizes his thoughts this way: "There are for us no instincts—we no longer need the term in psychology.... Think of each 'unlearned' act as becoming conditioned shortly after birth—even our respiration and circulation." As for Skinner, he never mentioned instinct in his most important book, *The Behavior of Organisms.* In *Science and Human Behavior,* however, Skinner does unburden himself on the subject, characterizing instinct as a "flagrant example of an explanatory fiction....an appeal to ignorance."

It is difficult to see Garcia as a disciple of the behaviorists rather than a revolutionary, particularly in view of the initial reception of his work. He could as well be regarded as the object rather than the subject of the "severe blow" to which Todd refers. One less resilient than Garcia might have been discouraged by the repeated rejections of his papers by respected journals. He describes how the criticisms were often "embellished with gratuitous personal insults." Another commentator has recorded that "one investigator, who had worked for years on delay of reinforcement, remarked publicly, 'Those findings are no more likely than birdshit in a cuckoo clock.'" Garcia entertainingly described his encounters with behavioristically minded editors in his Distinguished Scientific Contributions Award address to the American Psychological Association in 1980, which was titled "Tilting at the Paper Mills of Academe." It contains his account of the "fatal flaw that would ultimately dismantle Skinner's system"—hardly the sentiment of a disciple. We can only think that Todd must be mistaken when he describes Garcia as continuing the behaviorist tradition.

We agree that the principles we attributed to classic behaviorism were never tenable; nevertheless, our survey of pre-Garcia psychology texts reveals that those were the precepts generally taught and, as we suppose, held. Todd's insistence on burdening modern psychology with the epithet "behavioristic" seems ill-considered; we, like Morgan and King, put behaviorism in the past tense (it is not even in the index of some present-day texts, such as Henry Gleitman's *Psychology*) and welcome the new synthesis that is now developing between ethology and modern psychology.

JAMES L. GOULD

Princeton University
Princeton, N.J.

PETER MARLER

Rockefeller University
New York

To the Editors:

The very interesting article "Leonardo's Contributions to Theoretical Mechanics," by Vernard Foley and Werner Soedel [SCIENTIFIC AMERICAN, September, 1986], suggests that Leonardo was concerned about the nonlinearity of the bow. In fact, this is a great virtue of the bow's kinematics. Immediately after the bowstring is released much of the stored strain energy is converted into the kinetic energy of the bow tips. As the string straightens, however, the ratio of the velocity of the arrow to the velocity of the tips increases rapidly, so that eventually nearly all the kinetic energy is in the arrow. The energy is fed to the arrow by way of the string, which is actually at a greater tension than it was at the corresponding point of the drawing action, because it is *decelerating* the tips of the bow. On the other hand, at the instant of release the tension in the string drops below that in a string drawn and held at the corresponding position because the bow tips are accelerating.

Perhaps the most significant aspect of the article is its attempt to explain Leonardo's visual mode of thinking. The peculiar relation of design thinking to drawing is hard to capture in words, and the authors have done well in this difficult interpretative effort.

More studies of this kind would be most desirable, and they might help us to understand better the creative processes of the human mind. Rough drawings and notes often contain evidence about those processes that is generally totally obliterated in formal papers.

MICHAEL FRENCH

University of Lancaster
Lancaster, U.K.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

APRIL, 1937: "The vast empty spaces between the stars tempt the imagination. In overwhelmingly the greater part of interstellar space there is nothing that can be seen by direct observation. We know, however, that these open spaces are not altogether empty. Calcium and sodium had been known to be there, and it was recently discovered that titanium is another constituent of the interstellar gas. It is probable, indeed, that many, if not all, kinds of atoms are represented among the population of interstellar space."

"Molybdenum is a newcomer. It's so new that its name is strange to the lay world, yet it serves millions of consumers every day. Better automobiles, higher grade gasoline and lubricating oils and quality radio tubes are just a few of the essentials that molybdenum has made possible. It is recognized to be an exceptional alloying metal and an element having possibilities worthy of broad exploration."

"Shore erosion has become an acute problem along the water fronts of the Great Lakes, where the advance of the water upon the land has averaged, over a period of years, as much as four feet annually. When most of the solid jetties built out from the beaches were wrecked, there was doubt about what to do next. It was then that a new type of permeable jetty was developed. The results have been astonishingly satisfactory."

"When the data of sport are examined with the critical and skeptical eye which is continually focused upon the data of science, it is found that athletic measurements have not been guarded by the scrupulous precautions against error that are commonplace in the laboratory. Among the numerous errors afflicting measurements in the field sports there is none which is more systematically committed than that which pertains to the variation in the force of gravity. Any reasonably good broad jump is three-eighths of an inch broader in Texas than it would be in Massachusetts. A further geophysical phenomenon with a small but noticeable effect upon athletic performances is the rotation of the earth. To correct the performance of, say, a discus thrower for the effect of the earth's motion, it is necessary to take account not only of the latitude at which the throw took place but also of the direction in which the implement moved."

"It has taken about 110 years for the steam locomotive to develop its present effectiveness and capacity. With the advent of the electric locomotive and the beginnings of railway electrification 30 years ago it met its first competitor, and in the popular opinion of the day its doom was then sealed. Today less than 2 percent of the locomotives in service in the United States are electric and the limitations of railway electrification are well understood. Its more recent rivals have not impaired its position, and the steam locomotive is to-day, and is likely long to continue to be, the power mainstay of railroads the world over."

SCIENTIFIC AMERICAN

APRIL, 1887: "The mineral deposits of the great West have for many years attracted a good deal of attention. Besides the common minerals and compounds there are beds of pure white kaolin, mineral wax or ozokerite, borax, petroleum and bituminous coal. Sulphur is not recognized by text-books as existing to any extent in the United States, and yet, in the Territory of Utah, there is beyond doubt the richest and largest known sulphur deposit in the world. It would seem by the extent and purity of this deposit that at some remote period the sulphur must have poured out in a molten mass and flowed like lava into the valley beneath."

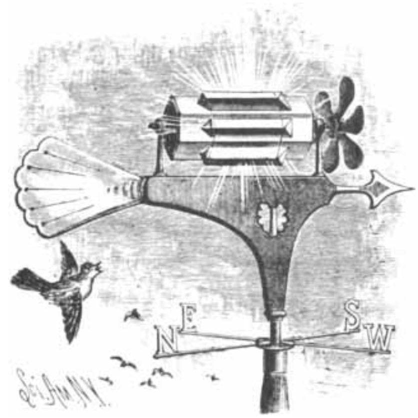
"M. Hermite, a French astronomer, has made some curious mathematical observations concerning the number of the stars. According to his computations, the total number of stars visible to the naked eye does not exceed 6,000. An opera glass will bring out 20,000; a small telescope will bring out at least 150,000, and the most powerful telescopes will show more than 100,000,000. Using data derived from the magnitude of stars and the law of total luminous intensity, the astounding result is reached that the sum of all the stars down to the twentieth and a half magnitude is some 66 millions of millions!"

"The electrical wires that now cover the roofs of almost every building in the large cities and cumber the streets have been the indirect cause of the destruction of much property by most seriously delaying and interfering with the firemen. The necessity for getting rid of this nuisance has caused inventive genius to devise means for accomplishing this object. A recent proposal is an underground conduit. It consists of rectangular tubular blocks, about four feet in length and a foot square, that can be joined. The conduit provides nine ducts, each of which will hold one hundred and fifty telegraph wires, or two hundred telephone wires, or seventy-five electric light wires."

"The supply of the natural oil of wintergreen or birch will soon cease to be of any commercial importance, since the artificial product (salicylate of methyl) is now being prepared of such good and uniform quality that it will undoubtedly replace the natural oil. Moreover, the artificial article can be produced at a cost below that at which the natural oil can be distilled profitably. Here is a chance for Congress to repress the improvement, as in the oleomargarine case."

"The Edison Company have closed a contract for electric lighting in Tokio, Japan. The central station will supply several thousand lights, a large number of which will be used in the Mikado's palace."

"A decided departure from the ordinary type of weather vane may be obtained by following the construction shown in the accompanying engraving. A hexagonal barrel, with sides of mirrors, is mounted on its axis, and a propeller or helix is connected with it at one end, whereby a rotary motion is imparted to the barrel by the action of the wind. Prisms are wired to the sides of the mirrors, which give to the rays a rainbow-like hue which is dazzling in the extreme."



A novel weather vane

M I T S U B I S H I

The sensation between here and there.

To satisfy those given to spontaneous fun often requires the most disciplined patience. For proof, consider the Mitsubishi Montero.

Montero's designed to give new freedom to your imagination. Patiently crafted with all the power of 2.6 liters, the independence of both 2- or 4-wheel drive and plenty of room for all your cargo. Even if all you're carrying is an impulsive imagination.

Montero's five-speed manual overdrive transmission and fully independent front torsion bar suspension help you navigate the shortcuts with confidence and ease. While its hefty all-terrain radials give you the ground clearance you need when you take the high road.

Of course in town, Montero still knows how to be a perfect gentleman. Its ergonomically shaped interior includes reclining front bucket seats and plush carpeting. And to make sure that cruising downtown is just as easy as cruising the backroads Montero comes equipped with power assist steering and power assist vented front disc brakes.

The Mitsubishi Montero. With prices starting as low as $9889,* it's not only a unique experience but an exceptional value. Patiently crafted to transport your imagination from here to there . . . even if there is nowhere.

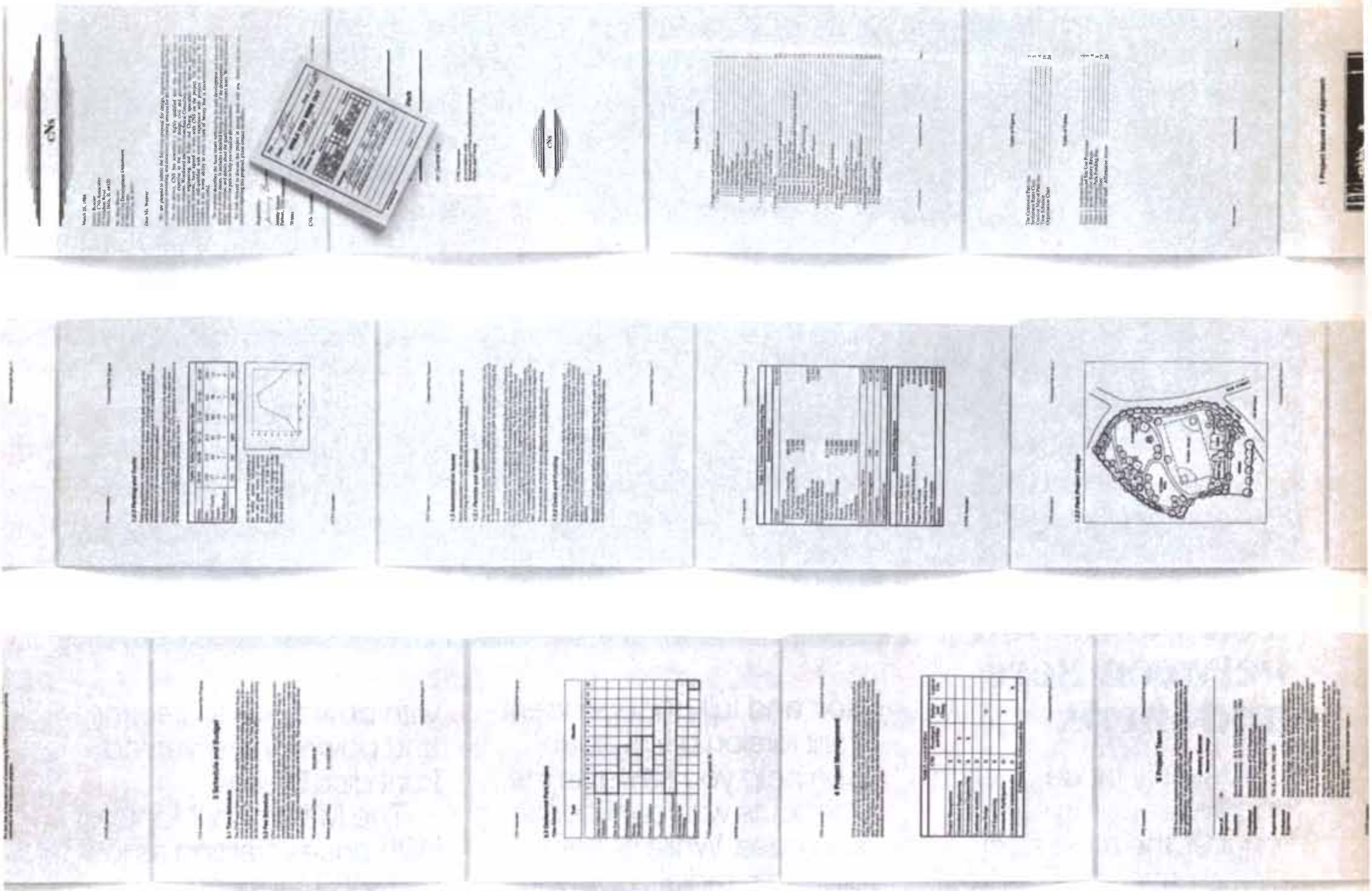
*Mfrs. Suggested Retail Price. Taxes, license, freight, options, other dealer charges extra. Call 1-800-447-4700 for your nearest dealer.

Mitsubishi. Perfecting the Experience.™

MITSUBISHI MOTORS

M O N T E R O

How to process docum



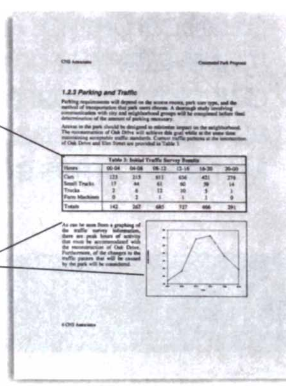
You may be content with your present word processor. And it's easy to understand why.

Even the clumsiest word processor is light years beyond the electric typewriter, the accepted standard only a decade ago. Plus, just the term word processing conjures up a rather modest expectation, ho hum, the ability to process words.

But, the mere processsing of words is hardly the challenge today. When you're creating a 20-page report, an 80-page proposal, a 200-page specification, or just a 2-page letter, you often have to work with graphics, equations, tables and more.

That's simply too much to ask of a product designed to process words. But, that's exactly what you can expect from Lotus Manuscript,™ a word processor that is truly a complete document creation system for technical professionals.

Very often documents require the mixing of text and graphics on the same page, graphics from Lotus® 1-2-3,® Symphony® and Freelance® Plus. With Manuscript you can import spreadsheets and charts, as well as diagrams and scanned images.



Spreadsheets, graphs and charts for 1-2-3® and Symphony® can easily be mixed with text on the same page.

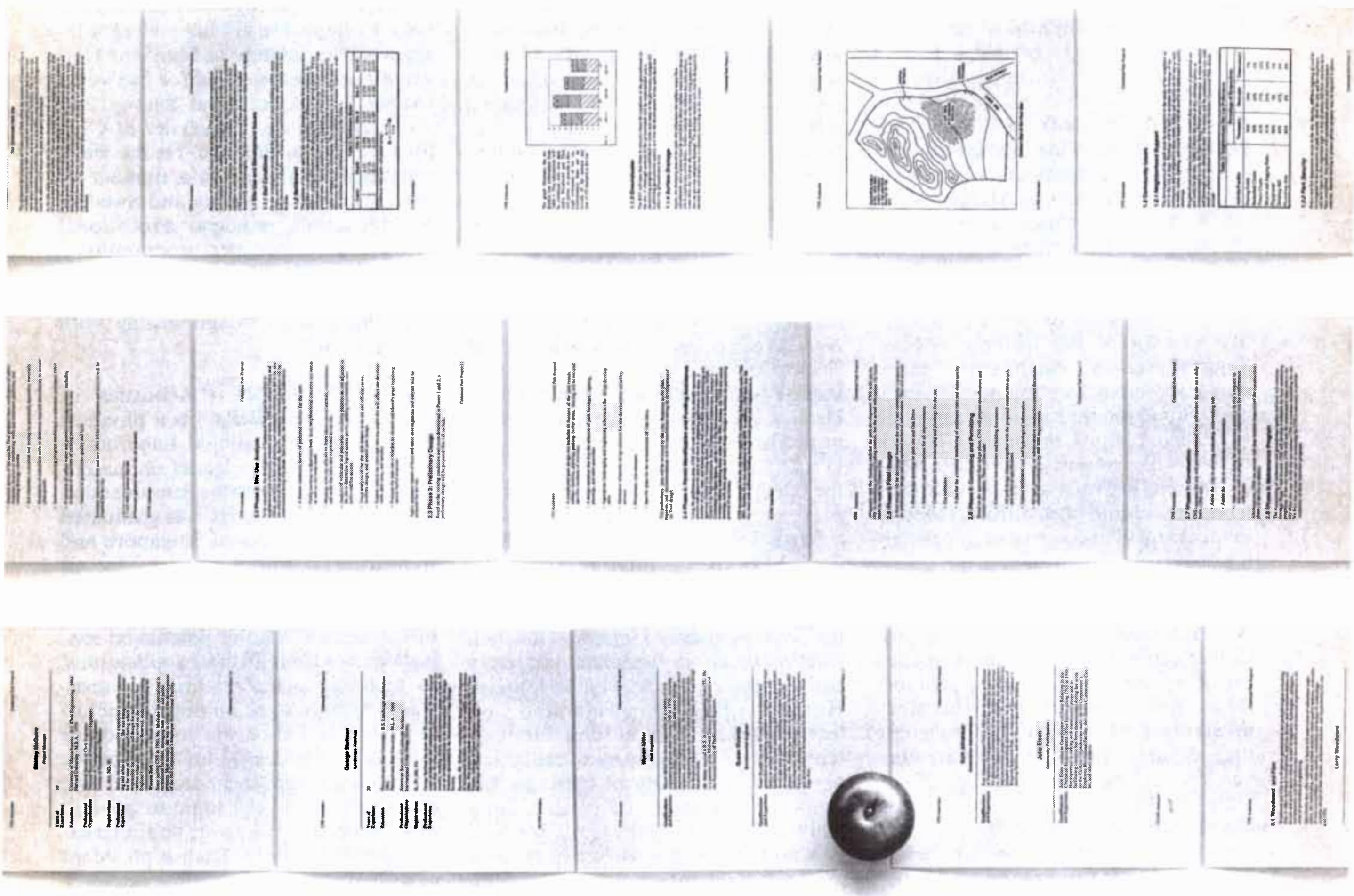
You can throw away your scissors and glue, cut and paste are a thing of the past.

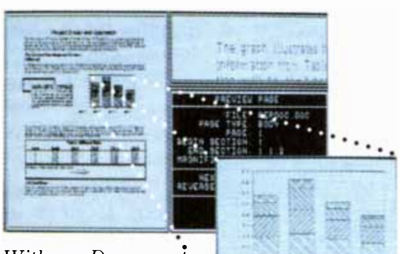
Another feature that goes beyond conventional word processing is Manuscript's integrated Outliner. When you collapse the document you're able to navigate the outline much easier than wading through the entire piece. You can move a page, or even an entire chapter with just a few keystrokes.

Because Manuscript understands the structure of your document, you can globally format an entire document or format by individual sections. Imagine changing all headlines from 14 point plain text to 16 point bold, by a simple menu selection.

Manuscript's Document Com-

›nts instead of words.





With our Document Preview feature you can see how any page will look before it's printed. Zoom capability lets you take a closer look at graphics and equations.

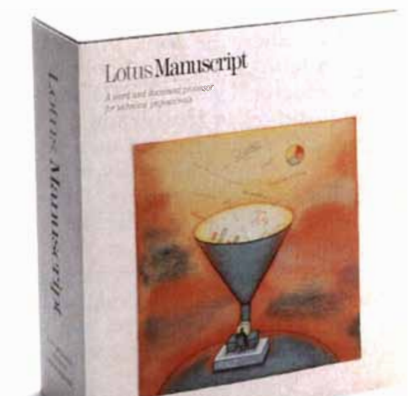
pare feature highlights changes between revisions for fool proof proof-reading.

Also, Manuscript automatically sizes and generates math equations as well as upper and lower case Greek symbols, diacritical marks and brackets.

With our powerful Print Formatter you have complete control over the look of your document, from position and size of graphics, to fonts and point sizes and more.

As for the quality of output, Manuscript takes full advantage of today's printing technology, from dot-matrix to laser, including PostScript® printers.

Manuscript is designed to work on most IBM® PCs and compatibles.* Its familiar 1-2-3 interface makes it easy to use. And our Manuscript evaluation kit makes it easy to try. For $10.00, you'll get a presentation disk, working software, and a tutorial manual. To get your evaluation kit, call 1-800-345-1043, ask for lot #BD-1450. Or, for more information, see your authorized Lotus Dealer, or write Lotus Development Corp., 90 Annex, Atlanta, GA 30390-0370.



Lotus Manuscript™

THE AUTHORS

THEODORE B. TAYLOR ("Third-Generation Nuclear Weapons") is the president and chairman of NOVA, Inc., a company that does research and development in the field of renewable energy sources. He earned a B.S. in physics at the California Institute of Technology in 1945 and a Ph.D. in theoretical physics from Cornell University in 1954; in the interim he spent time at the University of California at Berkeley and at the Los Alamos Scientific Laboratory, where he worked on the design of nuclear weapons. From 1956 to 1964 he was in the General Atomic Division of General Dynamics, and from 1964 to 1966 he was deputy director of the Defense Atomic Support Agency, which is now called the Defense Nuclear Agency. He went on to found the International Research and Technology Corporation. Taylor came to his present position in 1980, after spending four years as a visiting lecturer with the rank of full professor in mechanical and aerospace engineering at Princeton University.

JULIE L. SCHNAPF and DENIS A. BAYLOR ("How Photoreceptor Cells Respond to Light") have worked together since Schnapf was a postdoctoral student in Baylor's laboratory at Stanford University. Before going there Schnapf got her B.A. at the University of Rochester and her Ph.D. from Washington University; she also did postdoctoral work at the National Institutes of Health and the University of California at San Francisco. She has recently joined the faculty of U.C.S.F. as the Jules and Doris Stein Research to Prevent Blindness Professor. Baylor became interested in neurophysiology at the Yale University School of Medicine, where he took his M.D. after graduating from Knox College. He did postdoctoral work at Yale, the National Institutes of Health and the University of Cambridge and then joined the faculty of the University of Colorado School of Medicine in Denver. He moved to Stanford, where he is professor of neurobiology, in 1978. In his free time Baylor enjoys woodworking and jogging and is an avid football fan; Schnapf says that her two-year-old daughter leaves her no free time.

TORRENCE V. JOHNSON, ROBERT HAMILTON BROWN and LAURENCE A. SODERBLOM ("The Moons of Uranus") are members of the Voyager imaging team. Johnson and Brown are planetary scientists at the Jet Propulsion Laboratory of the California Institute of Technology; Soderblom is with the U.S. Geological Survey's Branch of Astrogeology. Johnson did his undergraduate work at Washington University in St. Louis and obtained his Ph.D. from Caltech. He is a senior research scientist at the J.P.L. and project scientist for the *Galileo* mission to Jupiter. Brown got his bachelor's and master's degrees in physics at Purdue University and his Ph.D. in astronomy from the University of Hawaii. He has made extensive telescopic observations of the Uranian moons. Soderblom received his undergraduate education at the New Mexico Institute of Mining and Technology and his Ph.D. in planetary science and geophysics from Caltech. He is a supervisory physical scientist in the Branch of Astrogeology of the U.S.G.S., and he is deeply involved in the upcoming *Mars Observer* mission.

MARTIN S. HIRSCH and JOAN C. KAPLAN ("Antiviral Therapy") are associate professors at the Harvard Medical School and also work at the Massachusetts General Hospital, where Hirsch is associate physician and Kaplan is associate bacteriologist. Hirsch's A.B. is from Hamilton College and his M.D. from Johns Hopkins University. He had postgraduate training at the University of Chicago, the Centers for Disease Control in Atlanta and the National Institute for Medical Research in London. Hirsch has been a member of the Infectious Disease Unit at Massachusetts General since 1969. Kaplan joined the faculty of the Harvard Medical School in 1971 after her postgraduate training there. Before that she had gone to McGill University, where she got her B.Sc., and Brandeis University, where she earned her Ph.D. in biochemistry. Kaplan's research has been in the areas of mammalian DNA repair, the activation of viruses and AIDS.

YOSHIHIRO HAMAKAWA ("Photovoltaic Power") is professor of electrical engineering science at Osaka University, where he has been since he received his Ph.D. there in 1964. He has also been a visiting professor at the University of Illinois at Urbana-Champaign, where he worked in the electrical engineering department and the materials research laboratory. Hamakawa's research centers on the physics of semiconductors, optoelectronics and solar photovoltaic conversion, with particular emphasis on the optical properties and band structure of solids, optoelectronic devices, solar cells and amorphous semiconductors.

MARK A. S. McMENAMIN ("The Emergence of Animals") is assistant professor of geology at Mount Holyoke College. He did his undergraduate work in geology at Stanford University and then worked for five years for the U.S. Geological Survey. His Ph.D. is from the University of California at Santa Barbara. He has done geological fieldwork in a number of places, including Alaska and Sweden. McMenamin's principal avocational interests are computer programming, biodynamic-French intensive gardening, cross-country skiing, bicycle touring, hiking and "roughhousing with my children."

SINYAN SHEN ("Acoustics of Ancient Chinese Bells") is a physicist at the Argonne National Laboratory, where he is research leader on materials, fuel from renewable resources and industrial acoustics. He was graduated from the University of Singapore and went to Ohio State University for his master's degree and Ph.D. He is the inventor of a damper for bowed string instruments and holds patents on sonic detectors for industrial applications. In addition to his research interests, which include work on novel structural materials and on the conversion of biomass, he is deeply involved in the diffusion of new technology, serving as an adviser on the topic to government agencies, industries and international organizations. Shen is president of the Chinese Music Society of North America.

NICHOLAS TOTH ("The First Technology") is assistant professor of anthropology at Indiana University and with his wife, Kathy Schick, now has plans to establish a Center for Research into the Anthropological Foundations of Technology there. He became interested in the study of human origins as a youth and went to Miami University and Western College in Ohio for undergraduate training in anthropology and archaeology, graduating in 1974. He did graduate work in prehistory at the University of Oxford, participated in the Flintknapping Fieldschool at Washington State University in 1978 and earned his doctorate (which was in anthropology and Old World prehistory) from the University of California at Berkeley in 1982. It was while Toth was at Oxford that he began to experiment with making and using stone tools.

Hitachi's advances in medical electronics extend to automatic blood chemistry analysis, a Picture Archiving and Communications System, diagnostic ultrasonics and Magnetic Resonance Imaging.

COMPUTER RECREATIONS

The sound of computing is music to the ears of some

by A. K. Dewdney

It was in 1965 that I first heard the sound of computing. An IBM 7090 at the University of Michigan Computing Center had been fitted with a simple electromagnetic pickup connected to a loudspeaker. As a computer program ran, a specific register in the machine changed its contents many thousands of times per second. The resulting pattern of miniature clicks was heard as an astonishing rush of alien sound that alternated among buzzing, screaming, burping, rumbling and whining. At times a grinding noise would change from bass to treble. This may have been the sound made by a double loop in the program; perhaps the inner loop executed ever faster, creating a tone of increasing pitch. It all sounded like a humpback whale.

The experience has suggested a variety of programs that explore aspects of melody, harmony and rhythm. Even though home computers do not normally come equipped with electromagnetic pickups poised over accumulator registers, the majority have small loudspeakers connected to a primitive tone generator. There are instructions in most popular computer languages that call forth a variety of sounds when they are executed. Such a simple facility can be exploited both to make programs audible and to make audible programs. In the former case a sound-generating instruction is added to a program originally intended for computational purposes. I shall discuss this topic briefly below. In the latter case a program is deliberately constructed to produce melodic, harmonic or rhythmic effects. Two of the programs described here resemble the former kind; a few sound-generating instructions are inserted in otherwise normal-looking code.

Melody, considered merely as a succession of notes, is easily generated by a program. In fact, the program can consist of a single loop. A decision process embedded in the loop decides what note to play next and how long it should be played. The process itself may be arbitrarily complicated. For all I know there are already programs of this type that generate convincing melodies in a given traditional style. If such programs exist, I should be glad to hear of them. In the meantime I shall take the tack of letting available algorithms shape the melody.

The humblest applicants for the job of melody maker are simple arithmetic algorithms. The numerical output of such algorithms is easily converted into notes under an enormous variety of possible encodings. The simplest of the encodings uses the linear congruential assignment, a process that is shorter than its own name:

$$x \leftarrow (a \cdot x + b) \bmod m$$

Here, when one specifies the parameters a, b and m in advance, an initial value of the variable x is converted into a succession of values by the continued iteration of the assignment. The expression "mod m" is an abbrevation for "modulo m," which means that the number computed inside the brackets should be treated like the hours of an m-hour clock. For instance, 10 modulo 8 equals 2. Thus if m is 8 and if a, b and the initial value of x are all integers, one obtains a sequence of numbers ranging between 0 and 7.

The resulting sequence of numbers is readily converted into a succession of notes by a simple table:

0	1	2	3	4	5	6	7
do	*re*	*mi*	*fa*	*sol*	*la*	*ti*	*do*

In the program I call SOLFEGGIO (from the practice of sightsinging) the quantities symbolized by the names of the notes are replaced by the frequencies, in cycles per second, of the C-major musical scale beginning at middle C:

0	1	2	3	4	5	6	7
262	294	330	349	392	440	494	523

The complete program can be summarized in the usual algorithmic form:

```
input a, b, x
for i = 1 to 100
  x ← (a • x + b) mod 8
  note ← notes (x)
  play note
```

In my version of SOLFEGGIO the numbers in the table called *notes* are employed directly by the instruction that generates a tone of the appropriate pitch. Most languages also allow the duration of the tone to be controlled. For the present it will suffice to set the duration to, say, half a second.

A world of maniacal melody now awaits those with an "ear" for arithmetic. Depending on what numbers are selected for the three parameters, one hears either boring staccato monodies, odd, repetitive melodies on a few notes or strangely wild music full of leaps and sudden runs. In the last category I have been caught short on occasion by not having made a note of the parameters responsible for a haunting little piece forever lost. The algorithm I have suggested produces melodies 100 notes long. Explorers of this new musical terrain may wish to shorten the length in order to investigate more possibilities.

SOLFEGGIO can and should be enhanced. Choose a larger value for the modulus m, large enough to embrace two or more diatonic octaves. There is no modal restriction as such; instead of diatonic scales one can choose 12-tone scales consisting of half-note steps. One can even choose the ultramodern form of linear-congruential music in which the numbers x produced by the algorithm specify frequencies more directly by the addition of a constant, say 100. If at one point x is 183, then 283 is the frequency of

Note	C	C♯	D	D♯	E	F	F♯	G	G♯	A	A♯	B	C
Frequency	261.6	277.2	293.7	311.1	329.6	349.2	370.0	392.0	415.3	440.0	466.2	493.9	523.3

Frequencies of notes above or below this octave are obtained by multiplying or dividing by 1.05946 and rounding off as appropriate. The number is $\sqrt[12]{2}$, the twelfth root of two.

Frequencies of the semitone scale from middle C to one octave above middle C

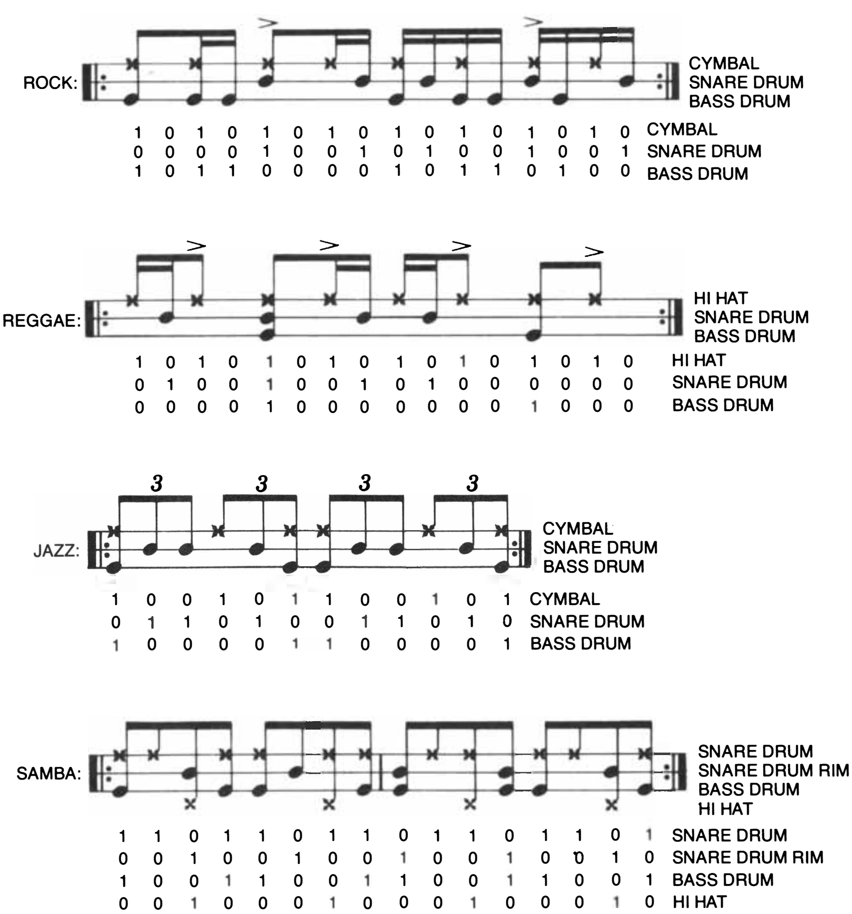
the note the program will play. I wonder what other possibilities the reader might invent. In the meantime here is an interesting question about boredom. For given values of *a*, *b* and *m*, how many notes will be played before the melody begins to repeat itself?

Harmony is now within the vocal competence of many home computers. Those equipped with two or more speakers will be able to play the entire repertoire generated by a program I call CANON. Indeed, a few enthusiasts will undoubtedly carry the project much further. Some readers may already have acquired a MIDI. The acronym stands for Musical Instrument Digital Interface, an electronic black box that converts signals generated by one's computer into commands for electronic instruments such as musical keyboards and individual synthesizer channels. (Readers interested in learning more about MIDI may write the International MIDI Association, 12439 Magnolia Boulevard, Suite 104, North Hollywood, Calif. 91607.)

CANON generates two-part harmony involving two almost identical melodic lines. Rounds such as "Row, Row, Row Your Boat" and "Frère Jacques" exemplify the type if not the species. CANON generates a canon in the academic tradition known as first-species imitation, the first stage in the serious study of counterpoint.

Such a harmony has two melodic lines that satisfy four criteria. First, all notes have the same duration. Second, one line begins after the other. Third, both lines are identical except that one line is transposed upward by some standard musical interval (unison, perfect fourth, perfect fifth or octave). Fourth, the two lines together must satisfy, note for note, certain rules of first-species imitation. All the rules are found in standard texts. Such rules generally include the allowed note-for-note harmonic intervals [*see illustration on page 18*] and establish the shape of participating melodic lines. In the interest of simplicity the latter has been omitted.

An example of a harmony in first-species imitation is shown in the upper illustration on page 20. The example was written by a human being, not by a machine. A glance makes it plain that after a certain point the composer tires of the strenuous demands imposed by all four criteria. Rule three is usually the first to go; it is enough if the second line imitates the first in spirit only.



Four sample rhythms for BEAT

A computer program, on the other hand, is undaunted by the constraints. CANON will grind on tirelessly until an entire piece is generated. The usual output of the program is tedious, however. To avoid this problem I have generated a great many short canons and have catalogued the more interesting ones. Some of these may then be strung together to make longer pieces [*see lower illustration on page 20*]. The experiment was done a decade ago with the assistance of Gregory Utas, one of my students at the University of Western Ontario.

CANON counts its way into a piece. Suppose, for example, it has been told to generate a miniature canon six notes long. For reasons that will soon be obvious, the tonal range spanned by each musical line is small, say six chromatic steps up and six steps down from the tonic, or beginning note. The set of 13 possible notes can be thought of as digits in a base-13 number system. The digit 0 represents the tone six steps below the tonic and the "digit" 13 represents the tone six steps above. CANON proceeds quite simply by counting its way through all six-digit base-13 numbers. Each one, after all, encodes a melody of a kind.

Every time CANON generates a new melodic line it makes a copy, transposes it upward by a fifth and translates it forward by a specified number of notes, say two. It then compares the pairs of notes brought into temporal juxtaposition by the operation. If none of the resulting pairs violates the rules of harmony, the line is accepted as canonic and is either printed out directly to the waiting composer or saved in a file for later printing.

In particular, CANON requires that the composer specify three parameters before running the program: *int*, the interval of imitation, *del*, the delay in starting time for the second melodic line, and *num*, the number of notes in each line. The notes of the line being generated are stored in an array called melody, or *mel* for short. Here is an algorithmic outline of CANON:

```
input int, del, num
mel(1) ← 7
for i = 2 to num
  mel(i) ← 1
found ← false
while found false
  increment mel
  for j = 1 to num − del
    compare mel(j) and
      mel(j + del) + int
  if harmonious then found ← true
output mel
option to continue
```

The algorithm starts by assigning to

MACSYMA solves mathematical problems symbolically.

Just as computers revolutionized numerical analysis, MACSYMA automates symbolic mathematics and yields enormous improvements in speed, accuracy and modeling power.

If you work with quantitative models in scientific or engineering disciplines, MACSYMA can increase your modeling power. MACSYMA combines symbolic and numeric computation. And enables you to accurately manipulate symbolic expressions in a fraction of the time required manually.

Wide range of capabilities

MACSYMA offers the widest range of capabilities for symbolic computations in applied mathematics of any commercially available program. For example:

Algebra: MACSYMA can manipulate large expressions, expand, simplify and factor expressions, handle matrices and arrays, and solve systems of equations.

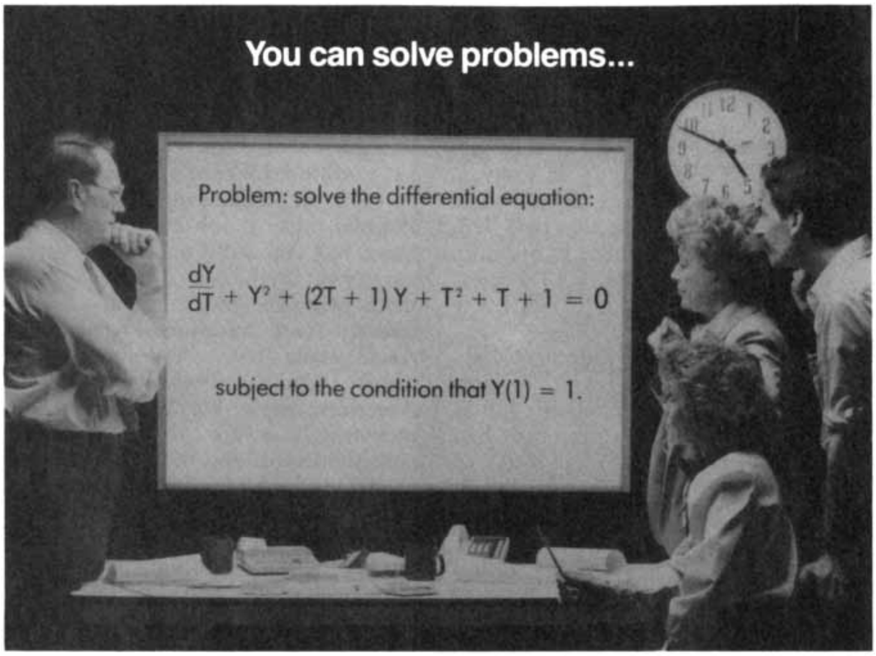
Calculus: MACSYMA can differentiate, perform definite and indefinite integration, take limits, expand functions in Taylor or Laurent series, solve differential equations, and compute Laplace transforms.

Numerical analysis: You can perform numerical analysis using MACSYMA® language, compute with arbitrary precision arithmetic, use a library of numerical analysis routines, and generate FORTRAN code.

Graphics and interfaces: MACSYMA can generate report-quality graphics in 2D or 3D, with perspective, hidden line removal, and captions. MACSYMA also interfaces with the mathematical text processors 'TeX' and 'troff'.

Broad base of applications

Throughout the world thousands of scientists, engineers and mathematicians are using MACSYMA in such diverse applications as aeronautical design, structural engineering, fluid mechanics, acoustics, CAD, electronic and VLSI circuit design, electromagnetic field problems, plasma physics, atomic scattering cross sections, control theory, maximum likelihood estimation, genetic studies, and more.



Symbolically...

```
(C1) DEPENDS(Y,T)$
(C2) DIFF (Y,T) + Y ^2 + (2*T + 1)*Y + T ^2 + T + 1;
(D2)     dY/dT + Y^2 + (2 T + 1) Y + T^2 + T + 1
(C3) SOLN:ODE(D2,Y,T);
(D3)     Y = - (%C T %E^T - T - 1) / (%C %E^T - 1)
(C4) SOLVE(SUBST([Y = 1, T = 1],D3),%C),NUMER;
(D4)     [%C = 0.5518192]
(C5) SPECIFIC_SOLN:SUBST(D4,SOLN);
(D5) Y = - (0.5518192 T %E^T - T - 1) / (0.5518192 %E^T - 1)
```

and Numerically.

```
(C6) FORTRAN(D5)$
        Y = -(0.5518192*T*EXP(T) - T - 1)
     1       /(0.5518192*EXP(T) - 1)
```

Available on many computer systems

Current systems include:

- Symbolics 3600™ Series
- VAX, VMS & UNIX
- MicroVAX II, VMS
- SUN-2 & SUN-3
- Masscomp MC5500

Other versions will be following soon.

For an information kit about all the ways MACSYMA can work for you, just call

1-800-MACSYMA.

In Mass., Alaska or Hawaii only, call (617) 577-7770.

Or please write to us at

Computer Aided Mathematics Group
Dept. M-SA2
Symbolics, Inc.
Eleven Cambridge Center
Cambridge, MA 02142

MACSYMA

The most comprehensive software for mathematical computing.

symbolics™

Your next step in computing.®

mel(1), the first note of the canon, the value 7. This is the tonic note, and it will not change. The remaining entries in *mel* all start at the value 0. A while loop tests a Boolean, or logic, variable called *found*, which is initially set false. *Found* is made true if a valid canonic melodic line is discovered by the instructions in the body of the loop. First, the array *mel* is incremented. This can be done by scanning the array from right to left. In the process of scanning, the counting procedure looks for an array entry that is less than 13. On finding such an entry, it adds 1 to it and changes all the entries to the right (if there are any) to 0. This is precisely what happens in ordinary counting, where 9 replaces 13. For example, $3572 + 1 = 3573$, $3579 + 1 = 3580$ and $3599 + 1 = 3600$.

The next job of the loop is to compare each note $mel(j)$ with the note $mel(j + del) + int$. In other words, the program adds *int* to the note that is *del* notes after $mel(j)$ and looks up the difference of the two note values in the table of rules for first-species imitation. If the difference is considered harmonious in all cases from 1 to $num - del$, the Boolean variable *found* is made true. (The notes beyond $num - del$ are played alone with no accompanying harmony. Thus there are no note differences to look up in the table.) Once such a melodic line is found and printed out, the program asks the composer, "Do you want to continue?" If the answer is yes, the program branches back to the $found \leftarrow$ false instruction and the count picks up where it left off.

One can, of course, play the melodic lines discovered by CANON through the tiny loudspeaker of one's home computer. Readers of a musical bent, however, will develop the knack of humming the line or of transcribing it, along with its canonic companion, onto sheet music. The canon can then be tested at another keyboard in all its harmonic glory.

Rhythm is a more sophisticated musical form than some readers may realize; traditional Western music has never been very elaborate rhythmically. Popular musical culture, on the other hand, has embraced an extraordinary variety of rhythmical forms. Most of them originate either directly or indirectly from traditional African or Asian music. This includes most rock music, jazz, Caribbean and Latin-American music. Westerners are also increasingly aware of the complex contribution of the tablas (a pair of drums played by the fingers) to Indian musical forms such as the raga.

The program I call BEAT enables one to specify simple rhythms as sequences of 0's and 1's. These are translated into sounds by the simple expedient of running through the sequence repeatedly. Each time through, the presence of a 1 triggers a brief tone pulse. A 0 triggers nothing.

Actually BEAT is simple enough to describe without further ado. Structurally it is rather similar to SOLFEGGIO (the program that plays linear-congruential music). A single array called *pulse* holds the rhythm as specified by the programmer at the start of the run:

```
input pulse, num, dur
for i = 1 to 25
  for j = 1 to num
  k ← 1
  while k ≤ dur
    k ← k + 1
  if pulse(j) = 1
    then sound
```

The variables called *num* and *dur* refer respectively to the size of the input array and the duration between sounds. The outer loop specifies that the basic rhythmic interval determined by *pulse* will be played 25 times. This number can easily be altered by readers who stumble onto rhythms they would like to hear for a longer period of time. The next inner loop controls the array index; the algorithm will consider each entry in turn in order to decide whether or not to play a tone. How long to wait between sonic events? That much is determined by a special waiting loop that simply counts up to the specified duration, *dur*. Then, if $pulse(j)$ is 1, the program BEAT will play a tone, a buzz or whatever may be available. If *dur* is small, the rhythm will be fast. If *dur* is large, the rhythm will be slow.

Like its predecessors, BEAT can also be enhanced. Something like a percussion ensemble is possible if different tones (preferably low ones representing various kinds of drums, bells or cymbals) are employed. Although the sounds may be far from realistic, the rhythms will be the real McCoy. One can complicate the array called *pulse* by using integers such as 0, 1, 2 and 3 to represent silence, a high drum playing alone, a low drum playing alone and both drums playing together. In each case the entry of *pulse* being examined in the inner loop must be decoded by a set of if statements that control the playing of no sound, one sound or both sounds. In the last case two notes are sounded consecutively as a substitute for simultaneity. If it is found that playing two consecutive sounds throws the timing off, then the waiting loop must be moved inside the if statements so that there is one copy for each of the four possibilities. In the case of no sound and two sounds, the limit *dur* must be replaced by new limits, one longer and one shorter than *dur*. In both cases the difference would be the time value assigned to an individual note.

A more sophisticated version of BEAT will certainly be able to play some of the interesting rhythms displayed on page 16. The various scores

a

LOWER NOTE \ UPPER NOTE	C	C#	D	D#	E	F	F#	G	G#	A	A#	B	C
C		X	X			X	X				X	X	
C#			X	X			X	X				X	X
D				X	X			X	X				X
D#					X	X			X	X			
E						X	X			X	X		
F							X	X			X	X	
F#								X	X			X	X
G									X	X			X
G#										X	X		
A											X	X	
A#												X	X
B													X
C													

b

NAME	NUMBER OF HALF-STEPS
1	0
1+/2°	1
2	2
3–	3
3	4
P4	5
4+/5°	6
P5	7
6–	8
6	9
7–	10
7	11
8	12

Notes sounded together are consonant, according to table *a*, if the combinations do not fall on an X. The names and sizes (in half-steps) of the resulting harmonic intervals are found in table *b*. The intervals of unison (1), minor and major third (3– and 3), perfect fourth (P4), perfect fifth (P5), minor and major sixth (6– and 6) and octave (8) are considered consonant in first-species imitation. The corresponding numbers of semitone (half) steps are 0, 3, 4, 5, 7, 8, 9, and 12.

Table of consonant harmonic intervals

The Tandy® 3000 HL makes 286 technology affordable.

We beat the competition . . .

If you're using MS-DOS® based computers and want to expand with more power for less, take a good look at the Tandy 3000 HL. It beats out the IBM® PC/XT-286 in price, performance and choice of options.

on performance . . .

Operating at 8 MHz (vs. 6 MHz for IBM's 286), the 3000 HL's advanced, 16-bit microprocessor delivers up to seven times the speed of a standard PC's microprocessor. This means you can run the business software you've grown accustomed to faster than ever.

on flexibility . . .

And you can choose your options. Expand with more memory, hard disk drives, communications and more. And you can connect the 3000 HL with other MS-DOS based computers with the ViaNet™ Local Area Network.

and on price!

The floppy-based Tandy 3000 HL starts at only $1699. The IBM PC/XT-286 costs $3995*. *And a comparably equipped Tandy 3000 HL with a 20-megabyte hard disk and other options still costs less than IBM's 286.*

Compare the Tandy 3000 HL. It delivers today's advanced computer technology . . . for less.

Send me a Tandy 3000 HL brochure.

Mail To: Radio Shack, Dept. 87-A-1187
300 One Tandy Center, Fort Worth, TX 76102

Name ______________________

Company ______________________

Address ______________________

City ______________ State ________

ZIP ____________ Phone ____________

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

TANDY COMPUTERS: In Business . . . for Business™

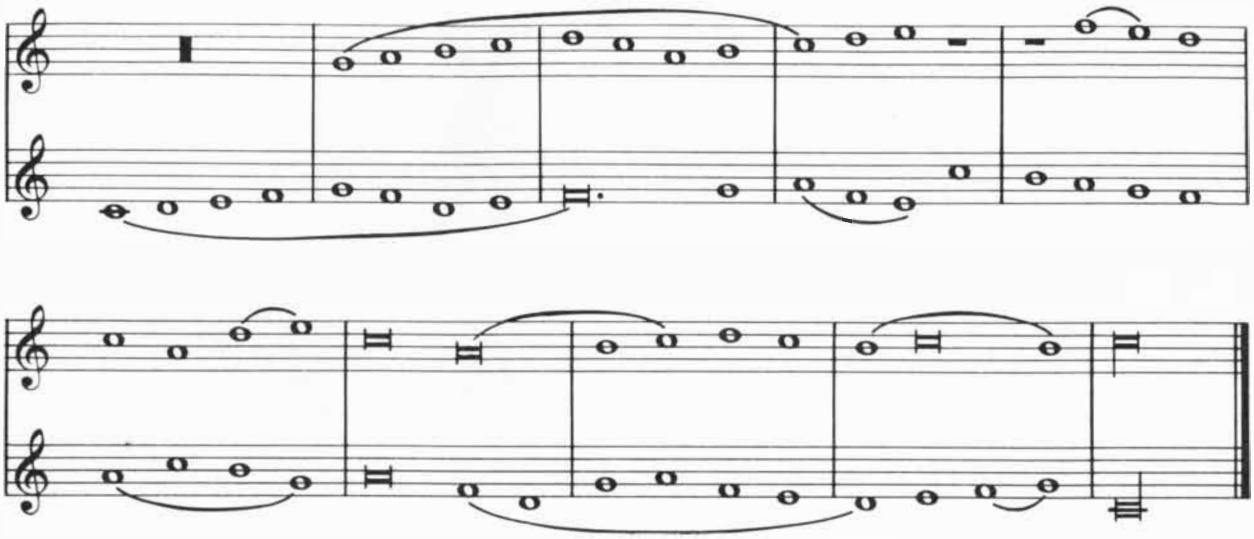
Radio Shack® COMPUTER CENTERS

A DIVISION OF TANDY CORPORATION

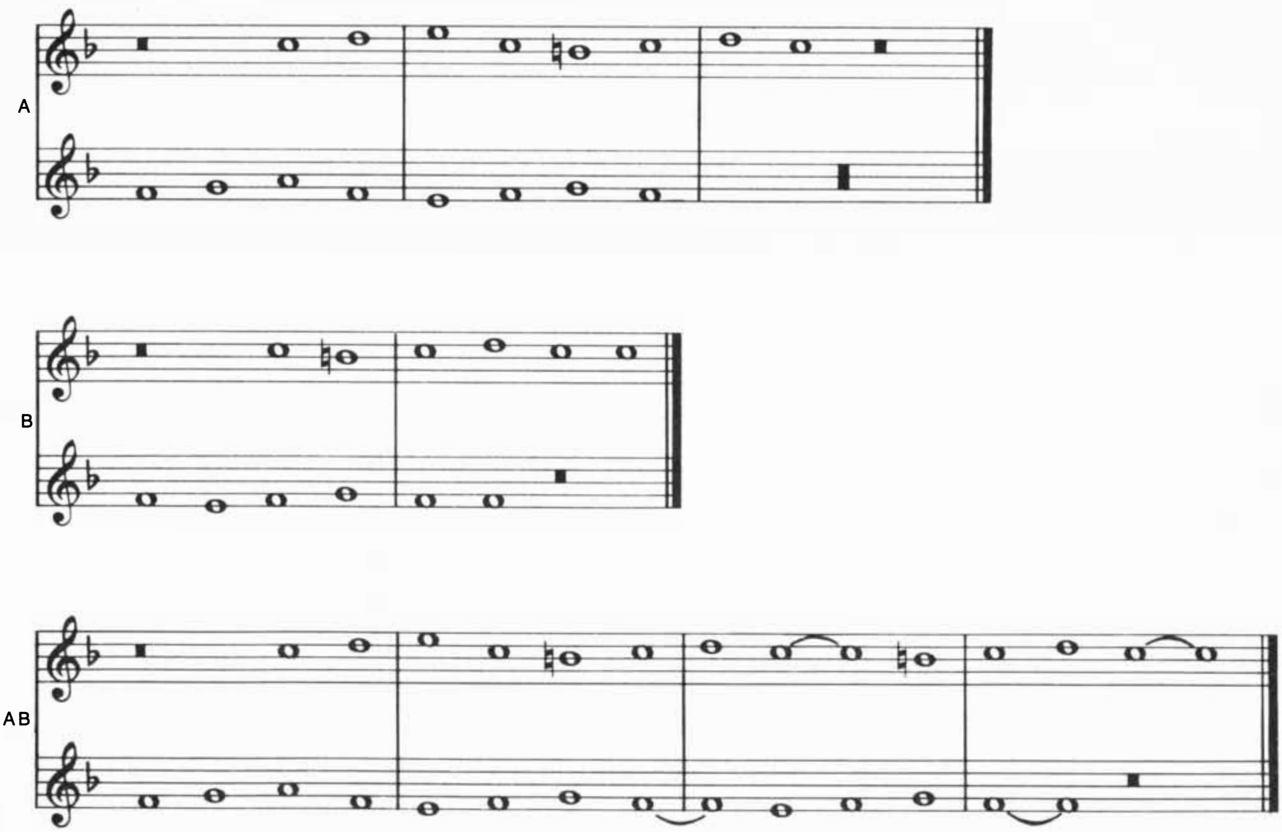
*Based on IBM price list as of Sept. 2, 1986. Tandy 3000 HL price applies at Radio Shack Computer Centers and participating stores and dealers. Monitor sold separately. MS-DOS/Reg. TM Microsoft Corp. IBM/Reg. TM International Business Machines Corp. ViaNet/TM ViaNetix, Inc.

shown there have been interpreted as sequences of discrete events strung along a common time base. They are easily translated into the contents of *pulse* under encodings arranged by the aspiring composer.

In addition to the preceding forays into melody, harmony and rhythm, I continue to be fascinated by the notion of endowing every computer program with two roles. In one role the program computes what it computes. In the other role the program is fitted with a tone or two next to its inner loops, outer loops and conditional (if) statements. The program has a song to sing for each problem it is given. Those who listen regularly to their favorite program (be it recreational or commercial) will develop an ear for its performance. I have no doubt that some bugs



A textbook example of two-part first-species imitation



Two computer-generated canons are spliced to make a longer one in new versions of a program may be detected by ear. Those readers with tales to tell in this area are hereby invited to put a bug in my ear.

In last month's column on Valentino Braitenberg's vehicles and their electronic brains, I posed a puzzle involving a golden sphere dropped from a UFO; once it is activated, it rolls along a smooth floor with a buzzing sound. How does it move if it has no exterior appendages such as wheels or legs? The sphere also avoids obstacles by rolling around them when it encounters them. How does it do so?

A number of years ago a toy called "Moon Walker" appeared for a few months in North American stores. It consisted of two hollow, transparent plastic hemispheres that snap together to enclose a small, battery-driven car. To operate the toy, the car's electric motor was switched on before the car was placed in one of the hemispheres. The resulting 10-centimeter sphere rolled along the floor because the car inside continuously tried to climb the inner wall of the sphere in the direction in which the vehicle happened to be pointing. The sphere rolled in the direction of the consequent imbalance.

When the sphere encountered an obstacle such as a chair leg, the car would climb still higher in the sphere. At such a point a built-in slippage on the right rear wheel would take effect. The car would skid continuously to the right, working its way around the inside surface until it headed in the direction in which the sphere encountered no resistance. In this way the sphere dodged obstructions. The puzzlement of onlookers was complete only when the sphere was spray-painted to obscure the little car. From that point on there was no limit to what people might guess inhabited the mysterious ball from outer space.

Since the description in January's column of the first international Core War tournament, membership in the International Core Wars Society has quadrupled, according to its director, Mark Clarkson. Expressions of interest continue to be welcomed by Clarkson at 8619 Wassall Street, Wichita, Kan. 67210. At the same time William R. Buckley, editor of *The Core War Newsletter,* has noticed a corresponding jump in subscriptions. Buckley explains that the publication will act as the principle organ of communication for the society members as standards are improved, new battle techniques are developed and new levels of play emerge. A subscription to the newsletter can be obtained by writing to Buckley at 5712 Kern Drive, Huntington Beach, Calif. 92649.

BOOKS

The origin of shape, the shaping of clay, the great dying and the great migrations

by Philip Morrison

Symmetry through the Eyes of a Chemist, by István Hargittai and Magdolna Hargittai. VCH Publishers, Inc. ($95). **Symmetry: Unifying Human Understanding**, edited by István Hargittai. Pergamon Press ($115).

A witty cartoonist shows us the atelier of the Creator, who is judging with pleasure the wall where young angelic designers have posted their competing projects for the decorated hexagons that will become snow crystals: "Impossible to choose; produce them all."

The cosmopolitan eye caught by the Parisian cartoon, good-humored and clearly delighted by diversity, informs this entire book. The work offers a broad new perspective that looks outward over the rich and often demanding monographic literature. Its authors are two structural chemists, research workers at the Hungarian Academy of Sciences in Budapest, who have seen everything relevant. They show most of it to us in photograph and drawing and tell us about it lightly and well. As an instance, their figures present clear although spare prototypes of the 17 plane groups that describe all possible repeating wallpaper patterns; each array of black triangles is accompanied by a strictly corresponding example documented from Hungarian needlework. All 17 groups are found in that needlework, as well as all seven classes of the border decorations that are their one-dimensional analogue.

The book has three roughly equal parts. The first part sets out one after another the symmetries whose combinations impose such interesting limits on form. The opening sections are studded with examples, both fresh and familiar, from the Van Dyck portrait of Charles I painted in fullface and in both profiles, through a photograph taken by her chemist-parents of little Eszter Hargittai along with her reflection, to the intricate Pueblo pottery that has only rotational symmetry.

Inversion, polarity and chirality (handedness, as in a DNA molecule) are treated at some length, now with molecules in mind. Two closely related organic crystals are shown, the pointy one anchored by a center of symmetry, the other smugly lacking it: a gas-solid reaction corrodes one face of a narrow slot cut in a crystal distinctly faster than it eats at the other. Drawings and tabulated properties of the five regular convex and the 13 semiregular polyhedrons are also presented; they serve both as part of the classical lore and as examples for discussion of symmetries and their notation. Kepler's fit to the planetary orbits is examined along with the goodness of the fit. The lists of references are long and useful. They offer the reader the simple canons of the subject—for example, Hermann Weyl's masterly popular lectures, Martin Gardner's writings and Edwin Abbott's *Flatland*—as well as up-to-date textbooks and technical papers.

The shapes of molecules are the next broad topic. A disarming and unforgettable start is made in the account of the standard projectional representation of isomers that differ only in the relative orientation of their parts. Two drawings after Degas are invoked: one dancer with arms flung apart at the end of an arabesque, who shows the staggered conformation; the eclipsed conformation is embodied in another dancer, who bends to adjust her shoes.

Cubanes and pentaprismane have been prepared, hydrocarbon molecules shaped as they are named. Adamantane is a hydrocarbon analogous to diamond, whose carbon framework resembles a set of four cubes, one inside the other. It is a real substance; even multiply joined adamantanes have been synthesized. (The book includes an index of formulas that is a page and a half long. Most of the entries are organic molecules, but there are plenty of boron and fluorine compounds as well.)

The first part of the book ends by treating electron-cloud shapes in molecules according to a simple and popular method known somewhat Germanically as the valence-shell electron-pair repulsion model. This pragmatic and qualitative scheme is rather faithful to the results of genuine ab initio quantum calculations for a wide class of molecules, those dominated by a few central atoms. An alternate name for the model suggests one reason: a recent proposal sought to recognize explicitly the role of the exclusion principle by calling the scheme Pauli mechanics.

The closing third of the book extends the ideas of combined symmetry from operations about one point to those that build infinite lines, planes or volumes. We have already mentioned the border and plane symmetry groups. Next there is a useful introduction to classical crystallography, sweetened by stacks of cannonballs and the old cleavage figures of the pioneers. There is also a dazzling long quotation (illustrated with his own drawing) from the writer Karel Čapek describing a visit to the minerals at the British Museum. "Even in human life there is a hidden force towards crystallization. Egypt crystallizes in pyramids and obelisks, Greece in columns; the middle ages in vials; London in grinny cubes..."

Easily understandable drawings of crystal forms and the corresponding projection diagrams illustrate the 32 crystal point groups. The restrictions imposed by combining rotations, mirror reflections and translations are nicely described. It is easy to grasp why by these rules nothing mineral is fivefold. The text is sufficiently contemporary to present a brief comment on the recent discovery of a quasicrystal or two with a penchant for fivefold symmetry. Our authors are far from dismayed, for the tendency of their final chapter is to view classical crystallography, with its rules derived from the mathematics of infinity, as an altogether too restrictive framework. The moderns could be counted on to depart from that hog-tied posture of perfection as materials of new kinds were made and examined.

The helical molecules of life are discussed (but rather too briefly), and there is something about the viruses as finite versions of symmetrical packed structures. The full 230 crystal space groups are presented; wisely they are not tabulated in detail.

The last sections begin to add notions of energy to the deep accounts of form. The Soviet physicist A. I. Kitaigorodskii has for decades treated molecular organic crystals as examples of packing. Using sphere close-packing as an introduction, he examined the ways molecules of arbitrary shape could be packed into lattices under the various space groups. It is not always possible to fit molecules close together without overlap and without allowing some convex surfaces to face other convex surfaces. Clearly such packing schemes cannot have high molecular density. Molecules with some added symmetry of their own (for example a plane of symmetry) admit closer neighbors even in unfavorable space groups. From such arguments Kitaigorodskii shows that in only six of the 230 space groups could planes of molecules be formed that have six neighbors in contact. Just those few groups describe the bulk of the known organic crystals. The classical groups are geometrically equal; energetically they differ, even grossly.

In the middle third of the book the treatment suddenly becomes abstract. Most illustrations are diagrams and graphs; virtually gone are the scenes from the real world. It is a helpful introduction to the algebra of group representations, complete with matrix multiplication, reducibility and group-character tables. This expertise is not carried to the point of much calculation, yet most readers will hardly make their way smoothly past a projection operator without having paper and pencil at hand.

It is the dynamics of molecules that flows from this excursion away from visualization into algebra. A fine detailed look at the normal vibrations of the water molecule provides a halfway house. The text then pursues up-to-date ideas about the use of symmetry to understand single-step reactions between molecules. The scheme is to take some account of the tendency for merger in collisions that match the symmetries of the approximate wave functions. These constraints lead to general rules in simple collisions. Symmetry-related rules of reaction have been growing in utility for two decades, and the Nobel lists recite the names of their formulators: Fukui, Hoffman, Woodward. This part of the book is too technical to win many general readers.

The Hargattais wrote in English, mostly during a two-year stay at the University of Connecticut at Storrs. The book had a modest progenitor in Hungarian published by the academy as part of a paperback series in popular science. It has since had grander and bulkier issue as well: the ambitious symposium on symmetry, orchestrated by István Hargittai, to which more than 70 mathematicians, artists and a variety of scientists from many lands contributed.

I cannot profess to have attended in full to the 1,000-page volume with so many articles of differing level, topic and length, from particle physics to country dance; yet a very few pieces quickly caught this rather daunted eye. Musical and literary symmetries are here. Two papers generalize the Kekulé benzene ring, a carbon-atom hexagon in which alternate links are doubled. The puzzle of how many ways there are to join benzene rings has already led to a voluminous literature and some remarkable formulas. The phenomenon of moiré is described with interesting works of moiré art; the simpler examples plainly invite more mathematical analysis.

The big regular jointed basalt columns found in Fingal's Cave and in similar showy formations here and there are hard to explain. Their strange pseudocrystallinity is often accounted for by a theory of the 1880's, which proposes a process of growth so unstable that it would require perfection. Now there is a new understanding, if not yet a full answer. At their base the columns are four-sided, displaying the same kind of irregular T joints as those seen in surface-tension cracks in mud and in some pottery glazes, not the Y joints of the hexagonal mesh that become so conspicuous well above the base. Here the ideal stress-relief pattern is being approached through the third dimension.

CERAMIC MASTERPIECES: ART, STRUCTURE, AND TECHNOLOGY, by W. David Kingery and Pamela B. Vandiver. The Free Press ($49.95).

Ceramic ware has been intimately part of living since before 6400 B.C. when pottery making became a mature village craft. Bowl, plate or vase is sized and shaped for the hand and rings sonorously to the ear. Matte or glossy, colorful or earthy black, it attracts the eye. Such qualities anchor a ceramic securely to the human scale, whether it is a mere commonplace of hasty mealtime or a precious object of art. To fashion any ceramic whatever demands more than hand and eye: earth, water, air, motion, fire.

The duality means that a piece of handmade pottery takes its nature, more than most of the material objects we encounter, from several scales of size. What we see and feel at arm's length is important: there dwell its provenance and intention in style, decoration and graceful or utilitarian form. Decisive features are there half-hidden on the submillimeter scale, at the margin of vision and the finger's touch. "A magnified view is essential.... If the only result of this book were to get a good 10X loupe and a penlight into the hands of each reader, that would signify a great success."

The authors pursue that end, among many others, by setting us a fascinating example. Two dozen color plates present a good look at (although no chance to heft, ring or feel) 10 ceramic museum pieces. The oldest of these is a blue lotus chalice about seven inches high, made in Egypt during an expansive and creative dynasty early in the New Kingdom. There follow Chinese, Middle Eastern and European examples. The latest is a Wedgwood chocolate pitcher of blue jasperware made in about 1790.

To each of the 10 the authors devote a chapter that carefully documents how the particular qualities of the piece were achieved by the potter. The result is an informed appreciation, an understanding participation in what the artist did long ago out of tradition, experience and intuition.

The senior author is Kyocera Professor of Ceramics at the Massachusetts Institute of Technology; Dr. Vandiver is a research scientist in the history and technology of ceramics and glass at the Smithsonian Institution. They have examined the objects they display with the full arsenal of modern materials science, beyond the penlight, to electron micrograph and X ray. The spirit of the admirer is shiningly clear. Their explanations and the detailed supporting evidence form a rich, nontechnical introduction to the nature of matter and of light and vision. Around this core the authors have gathered much of the history of ceramics; they assess and rationalize the worldwide development of that technology by inferring its intentions from its subtle achievements.

All 10 of the masterpieces are high-fired white-bodied ceramics, in which clear colors and white ground underlie a glaze. This cross-cultured unity reflects a clear analogue to evolutionary convergence. The goal of Egyptian faience can be traced by its forms over the millenniums; it was the emulation of the prized lapis lazuli and turquoise in shapes and sizes expensive or unavailable from the lapidary. The celadon jar from Sung China was frankly made as a "ceramic recreation" of semiprecious jade in form and color, seeking the translucent, all but tangible luster characterizing that tough fibrous mineral.

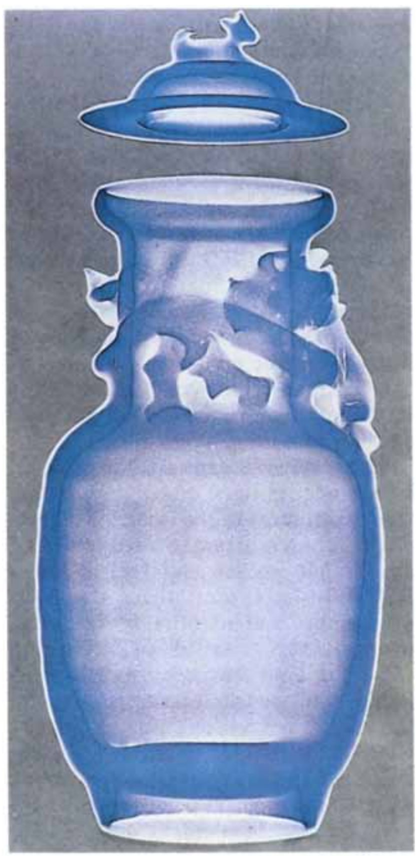
The faience was the earliest "contrived" nonclay ceramic body, made from crushed and ground white quartz pebbles with a small admixture of desert salts. The material was first made in about 4500 B.C., and blue donkey beads are still produced from it every day at Qom in Iran. The Chinese white porcelains arose more out of opportunity: natural white "china stone" is found there, although to be sure it was chosen for use in white-bodied ware.

There were some green glazed wares in China 2,000 years before the celadon jar. In either case, however, as cause or consequence of white materials, the imitation of the precious was a major impetus.

Follow one chapter: that celadon jar. It was the fruit of some 1,000 years of craft-based empirical technology, practiced on quite a large scale, witnessed by wide foreign trade and by the archaeology of the old kilns in Zhejiang. Those "dragon kilns" are long sloping hillside tunnels fueled at many fire holes. The fire moved uphill during the firing process. The air entering at the base is preheated by already hot ware to produce high temperature efficiently, 10,000 pieces to a firing.

The key to the celebrated jadelike luster is newly found. There are tiny bubbles in the glaze just within reach of eye and lens. They add brilliance to many glazes, here too, but that does not single out the unmatched Sung celadon, so much more like jade than any earlier or later wares. The electron microscope shows that this glaze is crowded with clumps of fine submicron crystals, just large enough to scatter light. The X-ray microprobe shows variations in calcium content at the scale of a tenth of a millimeter or so. The two crystal phases present can form out of glass of the right chemical composition that is viscous enough and is held at temperature for a considerable time. Given that the mean composition was about right, the needles formed in clusters because the glaze materials were not all finely powdered and were not thoroughly mixed; the compositional variations, probably around large particles of lime, led to the crystal clumping that optically reproduces the jadelike silkiness that delights the eye.

"European porcelain glazes of the same overall composition are clear, transparent, and bright, just like the overfired Longquan rejects"—plenty of which remain in the waste dumps. Much earlier green glazes were too limy, and they melted to a gloss; later the grind was too fine and the temperature too high, resulting again in bland transparency. The eye can judge invisible structure by such summed effects; the Sung craftsmen would not rest content with simpler efforts. What they wanted was the true look of jade, and they learned to manufacture reproducibly from their raw minerals a ceramic analogy to the microstructure of jade that modern technology finds difficult to reproduce. It required control of many steps to produce in the glaze a mixture of fibrous and granular inclusions on a fine scale. Repeated student efforts at M.I.T. to reproduce the Sung glaze have not yet succeeded.



Chinese 12th-century celadon jar and a Xerox radiograph giving clues to how it was made

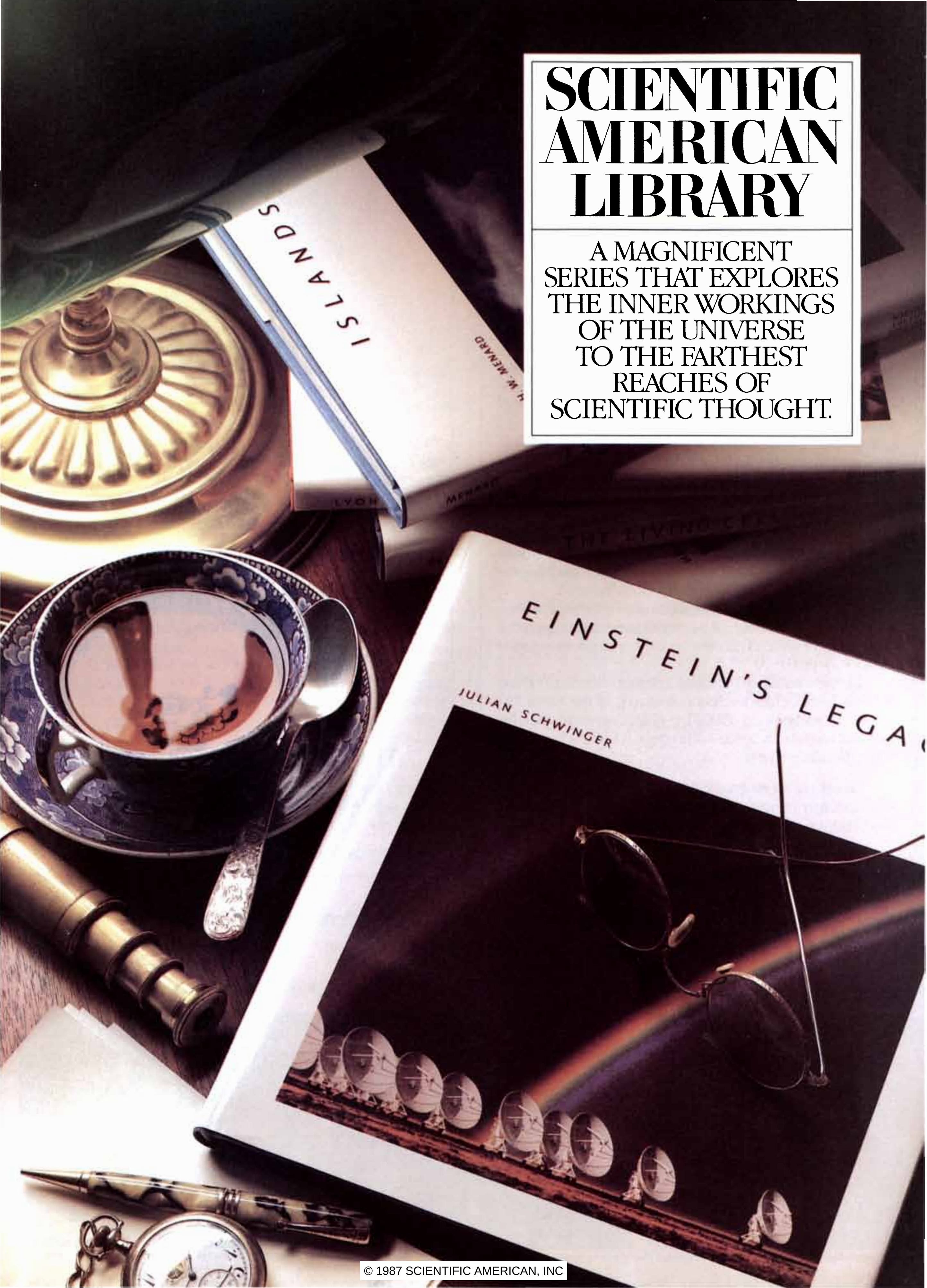
Permeable earthenware is the oldest of pottery. The plasticity of clay, its fine platy particles lubricated by layers of water, is the beginning of it all. Clays vary greatly, and with them the product. As the water evaporates from the worked and finished clay pot, the ware becomes weak and permeable. Soluble impurities tend to concentrate at the points of contact between the particles, where the water film remains longest.

When fire replaces water, liquid appears exactly there anew. It is a different fluid, a hot fluxed mineral, which sinters the particles together at red heat. Yet at orange heat the ware becomes overfired and slumps. A good product therefore requires a fairly narrow range of temperature. If air is present, the ubiquitous iron impurities will form their rust red oxide at red heat; air lacking, black magnetite will form. Control and alternation of these two conditions marked many ancient styles of pottery. The famous Attic black-on-red decorated pottery was made by combining such control with painted designs in a fine-particle nonporous slip that remained black, while the porous unpainted surfaces of the ware turned red as air diffused inward once the kiln was opened.

There is much more that might be reported, such as a systematic introductory look at the mineral compounds used in ceramic and the phase changes in certain mixes induced by temperature. This is indispensable matter for the potter; for an admirer of ceramic it makes possible understanding of the varied types of ware. Three chapters provide a fine summary of ceramic science and technology.

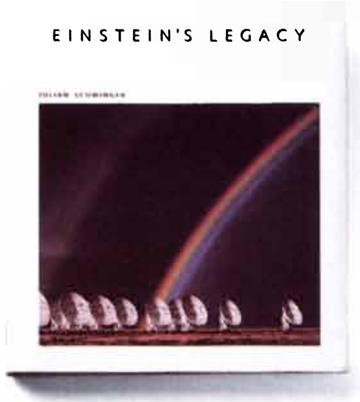
The last chapter is about seeing and studying ceramic objects. It evokes the pleasure of holding any fine piece (enough to motivate a collection of one's own, even if only a miniature one). From that point the story moves to visual examination and then to lens and penlight. The next step is a binocular microscope and an ultraviolet illuminator lamp in the study. Finally comes the modern laboratory, whose powers are great for dating, authenticating, replicating and in the end understanding.

There is a rather endearing discussion of the aesthetics of sampling a unique work of art. It covers nonintrusive tests as well as tests that require


SCIENTIFIC AMERICAN LIBRARY
A MAGNIFICENT SERIES THAT EXPLORES THE INNER WORKINGS OF THE UNIVERSE TO THE FARTHEST REACHES OF SCIENTIFIC THOUGHT.
ISLANDS
H. W. MENARD
JULIAN SCHWINGER

EINSTEIN'S LEGACY

Julian Schwinger



Nobel laureate Julian Schwinger tells the story of one of the twentieth century's greatest achievements—the theory of relativity—which is largely the work of one man: Albert Einstein.

The groundwork was laid in the seventeenth century by Isaac Newton, who unified motion on Earth and in the heavens in a framework of absolute space and time; then, in the nineteenth century by James Clerk Maxwell, who unified electricity, magnetism, and light. But it was left to the sixteen-year-old Einstein to glimpse, for the first time, that the theories of these two giants were incompatible.

Schwinger makes a lively narrative of Einstein's quest for the reconciliation of this conflict, a quest that led to the unification of matter and energy, and of space and time in his special and general theories of relativity.

The special theory has had its awesome confirmation in mankind's command of the nuclear force. Schwinger shows how the general theory, in turn, has stood up for 70 years to experiments drawn from the theory by Einstein, himself, and by his successors, employing ever more ingenious instrumentation and carrying the proofs ever farther beyond the decimal point.

In the last chapter, Schwinger describes space-age experiments made possible by the technologies that incorporate the theory itself. Their outcomes may not only secure further confirmation of Einstein's legacy, but face it with difficulties that lead on towards a still more comprehensive theory.

Julian Schwinger was awarded the Einstein Prize in 1951, the National Medal of Science in 1964, and the Nobel Prize for physics in 1965.

He is currently University Professor of the University of California, Los Angeles. He received his Ph.D. from Columbia University and has been on the faculty at Purdue University and Harvard University. Through the years, he has done theoretical work in various areas of both classical and quantum physics.

"I am happy to report, that his book makes for delightful and instructive reading. Technical demands on the reader do not exceed the most elementary algebra.

"It is particularly gratifying that the subject's intricacies, including glimpses of non-Euclidean and Riemannian geometries, and the theory's early successes, . . . are so well conveyed here in simple language. Altogether, this well printed and pleasingly illustrated book is an ideal gift for the curious non-expert." —NATURE

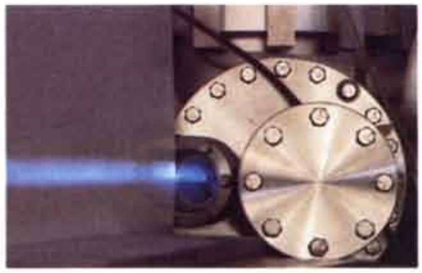
250 PAGES, 200 ILLUSTRATIONS

COVER PHOTO: The rainbow arches over the Very Large Array radio telescope of Soccoro, New Mexico. For nearly three generations, Einstein's legacy has withstood experimental test by instrumentation not yet invented when Einstein did his thinking. © Doug Johnson

TO GUARANTEE SWIFT DELIVERY OF YOUR FIRST VOLUME OF THE SCIENTIFIC AMERICAN LIBRARY, WE WILL SHIP *EINSTEIN'S LEGACY* VIA UPS.

Radio telescopes
Stephen Northup/Black Star © 1980

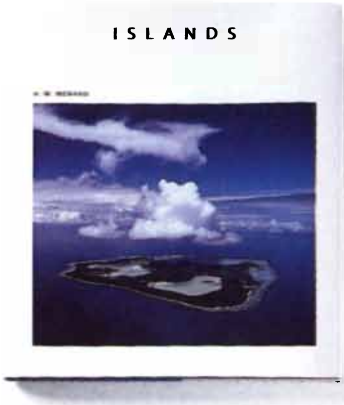
Einstein and Chaplin in Hollywood, 1931
The Granger Collection



Synchrotron light produced at Brookhaven National Laboratory

ISLANDS

H.W. Menard



In a lifetime of scientific adventure on the broad reaches of the Pacific Ocean, H.W. Menard discovered in the Earth's most romantic islands new insights into the forces deep in the Earth's interior.

His central contributions to plate tectonics, the unifying theory of geology, explain the layout of the island arcs, the atolls and the seamounts on the map of our globe. He also tells the story of the evolution of life, of human settlement and of the exploration, in the latest centuries, of these islands.

Sharing in Menard's adventure, the reader will also share his understanding of the Earth as a dynamic system. Photographs and cross section diagrams illuminate the text and make vivid connection between the beauty of nature at the surface and the processes driving the system in its depths.

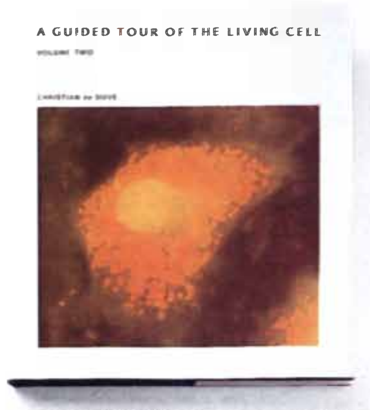
H.W. Menard was a founding member of the Scripps Institution of Oceanography, where he was professor of Geology. In the course of his distinguished career, he served as Technical Advisor in the President's Office of Science and Technology and as Director of the U.S. Geological Survey.

"In a very real sense, this book is his scientific legacy."—ROGER REVELLE, PROFESSOR OF SCIENCE AND PUBLIC POLICY, UNIVERSITY OF CALIFORNIA, SAN DIEGO

230 PAGES, 152 MAPS, PHOTOGRAPHS AND ILLUSTRATIONS

THE LIVING CELL

Christian de Duve



In A GUIDED TOUR OF THE LIVING CELL, Christian de Duve invites us to accompany him on an engrossing expedition through the trillions of cells that make up the human body.

The journey is divided into three itineraries. In the first, we examine a cell's outer and inner membranes and their two-way commerce with the exterior. In the second, we call at all the major organelles—the cell's own organs—and observe them manufacturing their own special products. Our third itinerary takes us into the nucleus itself, where we see DNA steering the reproductive biochemical machinery of the cell.

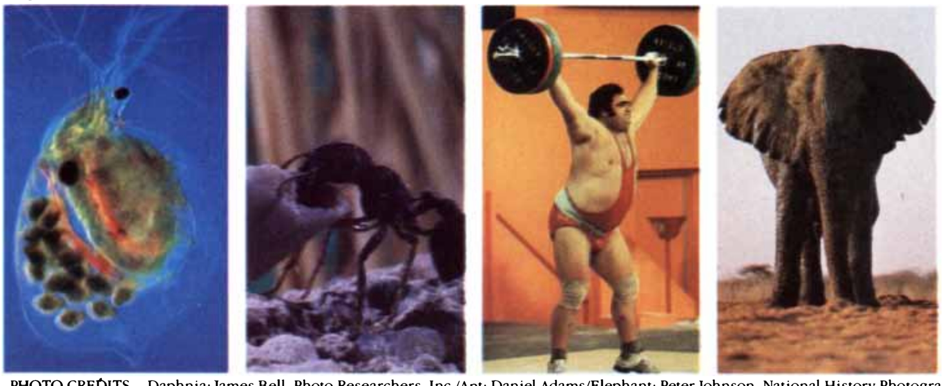
Christian de Duve is Andrew W. Mellon Professor at The Rockefeller University in New York. He was a joint recipient of the Nobel Prize for medicine in 1974.

"It is hard to imagine a better introduction to the world of the cell and all it means to everyone."
—JOHN BARKHAM REVIEWS

2 VOLUMES, 422 PAGES, 425 ILLUSTRATIONS

"The difference [between this and other book series] comes in the quality of
With the handsome volumes in front of us, what response can

The largest forms of life reproduce slowly, more ponderously, and live much longer than smaller organisms. On Size and Life *examines why.*



A molecular view of the plasma membrane. Electr
through an axopod of the heliozoon Echinosphae

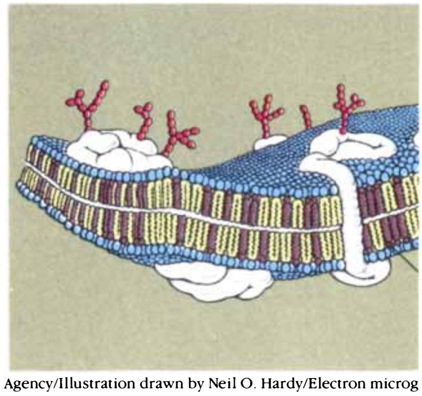
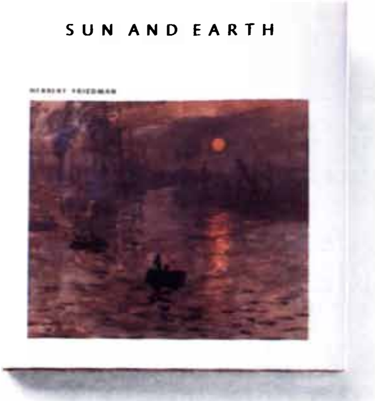


PHOTO CREDITS—Daphnia: James Bell, Photo Researchers, Inc./Ant: Daniel Adams/Elephant: Peter Johnson, National History Photographic Agency/Illustration drawn by Neil O. Hardy/Electron microg
Mutual Research Corporation

SUN AND EARTH

Herbert Friedman



As we can tell with our unaided senses, the distant sun bathes the Earth in radiant energy. From instruments on the ground and in orbit, we have now learned that the Earth revolves around in the Sun's extended atmosphere—the two bodies are in contact. An out-rushing solar wind engages our outer atmosphere and sets into action the processes that reach all the way to the planet's surface.

In SUN AND EARTH, the space-scientist Herbert Friedman combines an historical narrative with interpretive accounts that retain the awe and wonder with which scientists and poets have regarded the sun's life-bestowing characteristics.

SUN AND EARTH also provides a comprehensive picture of how the sun shapes our environment and explains the impact of solar variability on the contemporary problems of survival.

Herbert Friedman is chief scientist emeritus of the E.O. Hulburt Center for Space Research of the Naval Research Laboratory. In 1969, Dr. Friedman received the National Medal of Science—the nation's highest honor for scientific achievement.

"It covers more ground in an authoritative fashion than any contemporary book I've seen."
—WILLIAM C. LIVINGSTON, ASTRONOMER

251 PAGES, 195 ILLUSTRATIONS

OTHER VOLUMES IN THE LIBRARY:

SUBATOMIC PARTICLES
Steven Weinberg

MATHEMATICS AND OPTIMAL FORM
Stefan Hildebrandt and Anthony J. Tromba

POWERS OF TEN
Philip and Phylis Morrison and the Office of Charles and Ray Eames

HUMAN DIVERSITY
Richard Lewontin

PERCEPTION
Irvin Rock

DRUGS AND THE BRAIN
Solomon H. Snyder, M.D.

THE SECOND LAW
P.W. Atkins

THE SOLAR SYSTEM
Roman Smoluchowski

THE SCIENCE OF MUSICAL SOUND
John R. Pierce

FOSSILS
George Gaylord Simpson

CONSTRUCTING THE UNIVERSE
David Layzer

FIRE
John W. Lyons

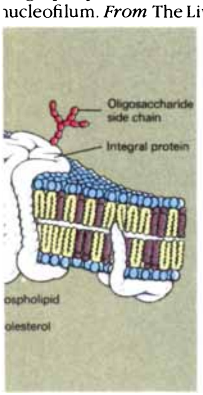
ON SIZE AND LIFE
Thomas A. McMahon and John Tyler Bonner

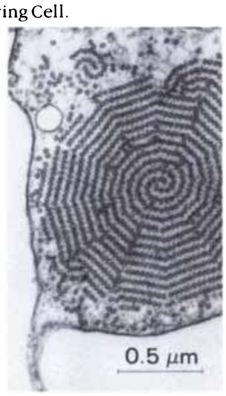
THE TIMING OF BIOLOGICAL CLOCKS
Arthur T. Winfree (in preparation)

EXTINCTION
Steven M. Stanley (in preparation)

authors, who combine eminence in their fields with proven gifts of exposition.
e be but immediate surrender to this firepower?" —Nigel Calder, NATURE

crograph of a section ucleofilum. *From* The Living Cell.





The "crowning" of a forest fire. Laser-beam spectroscopy. The copper converter. An experimental warehouse fire. From Fire.

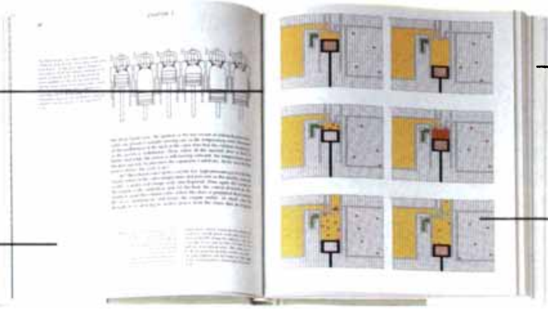
ourtesy of Dr. L.E. Roth and Dr. Y. Shigenaka/Fire in trees: L. Burr, Gama-Liaison/Copper Smelter: John Moss, Photo Researchers, Inc./Laser: Courtesy of Sandia Labs/Plastics Fire Test: Courtesy of Factory

Now in its third year, SCIENTIFIC AMERICAN LIBRARY has won recognition as an exciting new genre in the literature of science. Created out of the collaboration of scientists and editors that distinguishes SCIENTIFIC AMERICAN itself, each volume in the LIBRARY:

☐ Covers a topic of major interest with immediate relevance to ongoing research in its own and related fields.
☐ Carries the authority of an author who has made a significant personal contribution to its subject.
☐ Proves there is no topic in any field of science that cannot be explained to readers who want to understand.

Impeccably printed on high quality paper. Heavy cloth covers and binding to insure years of trouble-free reference.

Ample margins for visual relief and notes.



Large, inviting format ($8\frac{1}{2}'' \times 9\frac{1}{4}''$) opens flat for convenient reading.

Handsomely illustrated with many full-color charts, diagrams, maps and photographs.

JOIN OUR LIBRARY. THERE IS NO RISK INVOLVED.

A Two-Month Cycle.
As a library member you will receive a new volume approximately every two months. (Your first volume, EINSTEIN'S LEGACY, will be shipped UPS for fastest delivery.)

15-Day Free Examination.
Examine each new volume at your leisure. If you decide within 15 days not to own a particular volume, simply return it, and your account will be promptly credited.

No Prepayment Necessary.
Pay only for the volumes you decide to keep.

No Minimum Purchase.
You may choose to keep as many or as few volumes as you like. There is no yearly minimum.

Low Members' Price.
Members may purchase any volume at the exclusive low price of $24.95—well below the bookstore price.

Your Guarantee.
You may cancel your membership at any time without notice, for any reason, without penalty.

MEMBERSHIP RESERVATION

Free with your paid order:
TELESCOPING POCKET TRAVEL ALARM.

Its sleek black case slides open to reveal a 7-function alarm and calendar clock with a dependable quartz movement. Comes with mini-support leg for your viewing convenience.

Please speed my premier volume—EINSTEIN'S LEGACY—on its way via UPS and enroll me as a member of the Scientific American Library. Thereafter, send me a volume approximately every two months for 15 days' free examination.

I need send no payment now. For each volume I keep, I will be billed $24.95 plus postage and handling. This price is guaranteed for at least one full year. I understand that I may cancel my membership at any time, for any reason, and without penalty.

Yes, I'm looking forward to receiving my LCD Travel alarm after you receive payment for my first volume.

Signature ______________________

Name ______________________

Address ______________________

City ______________ State ________ Zip ________

487PAP

SCIENTIFIC AMERICAN LIBRARY

Mail this form to: SCIENTIFIC AMERICAN LIBRARY, P.O. BOX 646 HOLMES, PA 19043.
Or, to join right away call toll-free 1-800-345-8112 (1-800-662-2444 in Pennsylvania).

mere crumbs, which are usually easy to find but not ordinarily representative. The largest sample needed in today's laboratories is one taken by a core drill three millimeters in diameter, down to the same depth. "It seems to us, however, that it would be as irrational to say that a macrosample should be available from every object as to say that no sample should be allowed on any object."

This book is essential for anyone who is attentive to ceramics of any period but has not been trained as a ceramicist. It is a delight to inspect, read or ponder. A reader might wish the authors had included among their examples just one earthenware masterpiece, perhaps even an alumina-based spark plug. The tabulated data and the references are intriguing and up-to-date. (For the next well-deserved printings the proofreader should keep a sharper eye out for the many small typographical errors.)

THE GREAT DYING, by Kenneth J. Hsü. Harcourt Brace Jovanovich, Publishers ($17.95).

"It was a long drive from Houston, but we never had a dull moment." Hsü was a fresh geology Ph.D. at the Shell Research Laboratory on his way to a three-day field trip with old Doc Stenzel, a paleontologist as Prussian in bearing and as dogmatic as he was illustrious. Doc loved argument and chose the topic. "'You know nothing about fossils, young man,' he roared, 'nor has your venerated teacher learned much paleontology.... Yes,'" Doc continued, "'many groups appeared suddenly, not because of the incompleteness of the geologic record, but because they *did* appear suddenly.'" In the mid-1950's the discussion was only a taxing exercise for the ambitious young sedimentologist, who was quite unpersuaded by the tirade. By now we all know Doc was right. We usually hear of it in the reverse: many famous groups—all the dinosaurs, all those horn-curled ammonites—*did* just as suddenly disappear at the same level in the record.

At that boundary between two worlds of life, as Lyell and Darwin knew, both old species and newly evolved ones, large and small, on sea and land, amounting to three-fourths of the existing life forms, vanished and were replaced by another biota. The classical explanation, a rationalization, is just what Ken Hsü was taught. There must be a big gap in the record; enough time passed to allow sweeping change, but the pages in the Book of Sediments are missing.

Among the tourists who come to Assisi to admire the famed Giotto scenes from the life of St. Francis, few detour to provincial Gubbio nearby for a couple of altarpieces. But Gubbio draws the geologists. In the cliffs near the town that critical boundary—the record of the great dying—is "so sharp you can mark it with a razor." It is written in the microfossils, the one-celled marine creatures called foraminifera, typical of the Cretaceous seas at depth and on the surface. A single thin layer of clay a centimeter thick threads an entire cliff of limestone. Below it a mixture of big, vigorous forams abounds, up to the very last millimeter. Then comes the clay, free of limy fossils. Above that for a few meters it is small, plantlike plankton that form the calcareous rock. The newer forams, of totally different types, appear in number only a few meters higher in the limestone.

In about 1970 the layers around the boundary were searched with new intensity and a new ultrafine sieve as well. The sudden change was documented, better than ever before. Yet the new sieve recovered new forams immediately above the clay, the shells so small that they had been overlooked earlier. Within a few meters the little animals had evolved to include many diverse and larger forms. Now that sequence shouted against a gap in the record; it seemed to be a rapid but quite complete sequence in foraminiferal history after the Cretaceous. The rock showed that the old Cretaceous forms had gone away in less time than it took to deposit a layer of fine clay; that might be 10,000 years in a normal example. The events could be fixed by radiometric-dating techniques at 65 million years ago, plus or minus a couple of million.

Dinosaur bones, unlike marine microorganisms, are not found in every pinch of sample. We cannot date their final appearance closely enough to match the discrimination of the microscopes. The magnetic reversals calibrated through sea-floor spreading matched to the magnetic epoch captured in the limestones define a period of only a few thousand years during which the clay layer was deposited. The layer that holds the last of the big saurians in Alberta now appears to match the Gubbio magnetic epoch to better than half a million years. So do the fossil pollens of an extinct assemblage of land plants; a layer rich in iridium is found at the right time in a New Mexico borehole, just at the level where the telltale pollens were extinguished. "The time of transition is laid out as a stripe on the seafloor, and as a layer in both oceanic and continental sediments. It has width, it has depth. [It] is real." There was no chasm in the record at Gubbio; there is no chapter missing, not even a page.

In the spring of 1980 four papers independently put forward the E.T. hypothesis. One was our author's, who attributed the oceanic extinctions to chemical pollution of the seas after a comet's impact and postulated a temperature rise to catch the dwellers on land. Another was the wonderful result of the Alvarez team, which had found the trace of iridium in the clay layer at Gubbio, a trace that might be carried in only by a comet or an asteroid. "Simultaneous publication...is neither coincidence nor the result of communication. When the data point to the obvious, all those quick on the draw will come up with the same answer at the same time."

By now the topic has grown to notoriety. The scenario favored here—the evidence and argument are never slighted—is complex. Darkness at noon after the impact: dust if by land, water vapor if by sea. Surface plankton perish; their output of oxygen no longer keeps down the acidity of the carbonated sea. The shocked air holds the oxides of nitrogens and down comes the acid rain, enough to inhibit "all but the toughest of plankton," in particular those that secrete calcareous shells. Eventually the lime of the ancestors neutralizes the sea. New plankton forms spread from rare spots of safety to inherit the empty resource, perhaps after tens of thousands of years; they evolve. Dinosaurs, ammonites, corals, brachiopods, land plants—how can we know? Each of these goes extinct by causes that may be individual and unique, species by species. Here is a lottery of survival. The ultimate cause is a universal catastrophe; how that feathers out into the myriad niches of life, particularly niches outside the more nearly mixed ocean, may provide many tests and no single answer.

This is "not an anticlimax" just because we lack the death certificate of the dinosaurs; we have learned of a profound event whose lesson is written large in our mammalian legacy and even on human life today, wetted by its own acid rain and shadowed by a hint of nuclear darkness. (The claims of periodicity, the star Nemesis and the rest, have no appeal for Professor Hsü; he offers odds at 100 to one against them.)

Hsü writes with unusual clarity and intimacy. He spells out the full logical content of his arguments with more attention to how he came by his premises than is common in popular ex-

position. He does not leave out his own responses—he recollects even the failings—as conversations, meetings, ideas and opportunities come and go. We learn that he is no patient fisherman for clues but a man eager to dive in. One endearing recollection goes back to his first postdoctoral year in Zurich (where he is now professor at the Swiss Federal Institute of Technology). A lecturer, faithful to the memory of his own professor, a giant of geochemistry, handed out a long table of element abundances. Then he chose to read it word for word, a droning tedium of number after number. "I did not even worry about courtesy; I walked out of the room before he finished proactinium.... If I had gone to his second lecture, I was almost certain to have learned [about the noble metals that are rich in the meteorites]." Had Hsü done so he might have accepted the proposal someone put to him for trace analysis of his ocean cores at the strange boundaries—a decade before it was done for the Gubbio clay. "Only in later years did I realize...that new data lead sooner or later to new levels of understanding."

The book offers more, a first chapter and a last one. They are attractive, strong, compassionate and well argued. This brilliant geologist has divided a lifetime among three cultures, in China, America and Switzerland. He was certainly an achiever in his American phase, "what my daughter calls a *Funfziger*." Now, however, he leans toward the Tao.

The aim of his more philosophical years is to recast the metaphor of evolution, particularly the social Darwinism whose issue was in part the Third Reich. First of all, Darwin asserts the common descent of all life. That unity is triumphant; science cannot doubt it. But the uniformitarianism of the geologists is wrecked by modern work, spectacularly the boundary between the eras of reptilian and of mammalian radiation. Agreed: gradual change is not all. Catastrophes can be decisive.

Hsü goes further, though. In a cogent but brief discussion he finds no validity in natural selection as a driving force for evolution. Adaptation is real, but not natural selection. That is deeper water. The Spencerian talk of fitness is unpleasant enough, to be sure, and it retains little meaning. But the modern view is much cooler, and differential reproduction does seem to describe much of what we see: convergence, parasitism, mimicry—even if chancier isolation drift is also needed. "'The Tao that can be talked about is not the true Tao.'" We might believe that, but for science we must talk to one another, although we do not aspire to the true Tao but rather to a partial Way nearby.

"'The race is not to the swift, nor the battle to the strong...but time and chance happeneth to them all.'" The Preacher is eloquent, but did that shrewd observer maintain that the battle is never to the strong, the race never to the swift? A scientist cannot think so. Evolution is a historical science; we shall know it only in part, yet more and more. A physicist-reader is wary of principles, even welcome ones, that assert rules entirely without quantity.

MIGRATIONS IN PREHISTORY: INFERRING POPULATION MOVEMENT FROM CULTURAL REMAINS, by Irving Rouse. Yale University Press ($20).

The disconformity was striking. At several shore sites in Puerto Rico the lowest level in the old deposit of refuse was full of crab shells. They were usually found in association with distinctive potsherds painted white on a red ground. Above that was a distinct layer; there the debris was hard clamshells, and the pots were modeled and incised. A new people had come to the place, along the island chain from the Orinoco plains, long-known home of the distinctive white-on-red potters.

It was an easy inference, a commonplace of the archaeologist, its counterpart on a thousand museum walls and in the pages of as many textbooks. The inference here was altogether too easy: the style had in fact evolved gradually through red-on-white to undecorated, and then on to the incised, again gradually. The collectors who first indicated the sites had a selective eye for attractive pottery, although the shift from crabs to shellfish turned out to be pretty local as well. That change was a movement of neighboring and interacting villages on one shore, not of entire peoples across a sea.

How do we know? This lucid and tightly argued book sets out the means for a serious test of the hypothesis that migration causes a given culture change. The author has 50 years of reflective work to draw on; he puzzled out the change from crabs to clams 35 years ago as part of his own fieldwork. His study is concise and to the point; its unusual readability derives from four varied and fascinating case studies, all of them, he argues, examples of a degree of success in the tested reconstruction of the movements of peoples before history.

Single grand hypotheses have been the course of explanation in archaeology for a long time, particularly large schemes, "ruling theories" of migration. They are not without the taint of cultural bias. Rapid, long-distance migration of peoples is the historical source of gross cultural change in Africa, the Americas, India, Oceania, Siberia, all within a few centuries. The Westerner knows all about colonialization and falls into such a model. Eastern scholars instead prefer as a model the long local development of central, stable cultures in China and India; change is the product of acculturation on the fringes. The scholars of fringing lands, say Japan, tended to seek synthesis between foreign additions and local endowment.

Migration theories today seem overblown, somehow insensitive to local identities and strengths; they are out of fashion. Yet they deserve attention. We seek to know as much about human prehistory as we can, whether it fits our taste or not. Can past migrations be shown to be probable and significant? The examples suggest that success can come only with ample evidence. The test is rigorous only if a hypothesis is treated not as a self-evident truth, to which many pieces of data can be assimilated, but as one postulate in open competition with its peers.

Three kinds of data dominate the reconstruction of prehistory. Language manifestly offers a wealth of information about the unwritten past. Physical anthropology, from skulls to genes, is a classical contributor. Archaeological data, the most varied, include artifacts and the social structures inferred from them. These disciplines offer rather independent tests of migration theories. Irving Rouse constructs for each of his hypotheses a graphical page for two or three test disciplines, time plotted from bottom to top, spatial regions across the page and recognized cultural units entered in the matrix. That space-time plot is the summation of the archaeologist; the question is, of course, its credibility, defined as consistency under the several tests.

The most recent migration in prehistory is the courageous mastery of the Pacific Ocean by the virtuosos of the sailing canoe. Our present chronicle begins in New Ireland in Melanesia, where we have a preceramic assemblage from about 2000 B.C. Before that, we cannot say. The whole of Melanesia seems to have been occupied at about that time. The archaeological evidence of dated settlement moves only generally eastward across the Pacific, proceeding through much change and a good deal of to-and-fro along the very "route of migration worked out by the linguists." Although their studies of word and grammar do not provide secure comparison dates, all the current languages of the region do appear to have developed locally since about A.D. 900. The last major movement was the occupation of New Zealand.

For that case there was a famous alternative hypothesis. The linguists held for a long time that the languages of the islanders must have come from the west. The only issue might be whether the route would have been first northward or southward from the Indonesian region. It was Thor Heyerdahl on his unkempt raft, the *Kon Tiki,* who ended complacency by forcing on scholarly and public attention the new notion that migration had come first from Peru, later from North America. Heyerdahl did not deny the basic linguistic evidence; his voyagers had to have come around the subarctic rim of the Pacific long before the oceanic islands were settled.

In the north the linguists found no trace of the speech they sought; instead the evidence showed a clear path across Melanesia. Recent blood-antigen studies confirm the linguistic distances well. The dates of first settlement dug out by the spade concur. The islands that would have been the first ones occupied on Heyerdahl's model turn out to be the first by all three tests. The islets and the atolls were desert, virgin, until the canoes came; there were no human beings in place to blend their genes, their speech or their ways with those of the immigrants. This case is therefore the best we have, for there is no noise.

The Eskimos and their cousins, the Aleuts, are related physically and by language, albeit a long way back, to the Chukchis of Siberia. The old view that their ancestors were Magdalenian hunters of ice-edge Europe who went west all the way to Alaska has no support. The story is by no means simple. The Eskimos were not the first to settle the north. Their roots are in the Diuktoid peoples, who inhabited the lands on both sides of the Bering Strait at the end of the Ice Age. The strait was dry land, Asian then, cut off by the ice from Canada. Millennium by millennium a number of cultures began to occupy the entire region as the ice dwindled. Human beings reached Greenland in about 2000 B.C. from the Canadian islands. "For us, the Bering Strait is a boundary, to be traversed in order to reach the places...where developments are taking place. For the Eskimos and their predecessors, the strait was a center from which traits spread."

The modern Inuit carry mainly the Thulean cultural inheritance. That ingenious people, their tools so varied and deftly made as to "recall those of modern dentists," settled across the entire New World Arctic rim. They adapted as they went, yet they are recognizable by the portable artifacts through which we can trace the Inuit migration. For 1,500 years and more they moved from east to west; they "did not pass through the Bering Strait; they started from it."

Migration and adaptation are here commingled. Professor Rouse puts it that the anthropological approach called structuralism attends to patterns and styles in kits of tools, whereas functionalism addresses the ecological adaptation to everyday life. He sees them at least in part as complements, not alternative methods, the one tracing the relics of the past, the other ordering the present.

The case of Japan is treated next. The best evidence suggests that the Japanese people did not, as all the schoolbooks state, enter the islands from outside in about 300 B.C. Instead they spread to the northern islands after they had developed the Japanese language and prototypical culture on Kyushu at an earlier date. In this example there is a discrepancy between linguistic and archaeological results, not yet understood. The conclusion must be held provisional.

The fourth case is that of the people who greeted Columbus, the Taino (also called Arawak) of the Caribbean. We owe to them the words hurricane and cannibal, as well as the crops tobacco, maize and pineapple. In exchange they got mainly smallpox and measles, diseases that decimated the Classic Taino. This Caribbean example is more complex still. Whence did they come? It is conceded that their origins are somewhere on the plains of the Orinoco. But the tale is tangled, with much varied local development and evidence of interactions with the mainland. The chronological chart is complicated enough to require two pages: work in progress.

Not many books that intend to discuss method are as full of good sense as this one. It deals with explicit cases, tough cases at that. A reader is particularly caught by a point mentioned only lightly: population movements should not be studied without knowing a good deal about how people move. A landlubber is struck by the fact that in the West Indies the two shores of a sea passage between islands are home to the same cultural series more often than the two ends of one island are. The mixture of detailed information and clear argument makes this critical essay, professionally aimed, attractive to a general reader.

Established 1845 SCIENTIFIC AMERICAN April 1987 Volume 256 Number 4

Third-Generation Nuclear Weapons

Unlike deployed nuclear weapons, which unleash their explosive energy indiscriminately, future nuclear weapons may selectively produce certain types of energy and concentrate them on targets

by Theodore B. Taylor

During the early 1950's American weapon laboratories were exceptionally productive. They not only achieved dramatic improvements in the performance of fission bombs, which represent the first generation of nuclear weapons, but also succeeded in establishing a second generation of nuclear weapons by harnessing the explosive power of fusion in the form of the hydrogen bomb and its various derivatives. By the end of the 1950's the warheads in the U.S. nuclear armament bore little resemblance to the bombs that had ushered in the nuclear age over Hiroshima and Nagasaki.

Today a third generation of nuclear weapons is technologically feasible. By altering the shape of the nuclear explosive and manipulating other design features, weapons could be built that generate and direct beams of radiation or streams of metallic pellets or droplets at such targets as missile-launch facilities on the ground, missiles in the air and satellites in space. These weapons would be as removed from current nuclear weapons in terms of military effectiveness as a rifle is technologically distant from gunpowder.

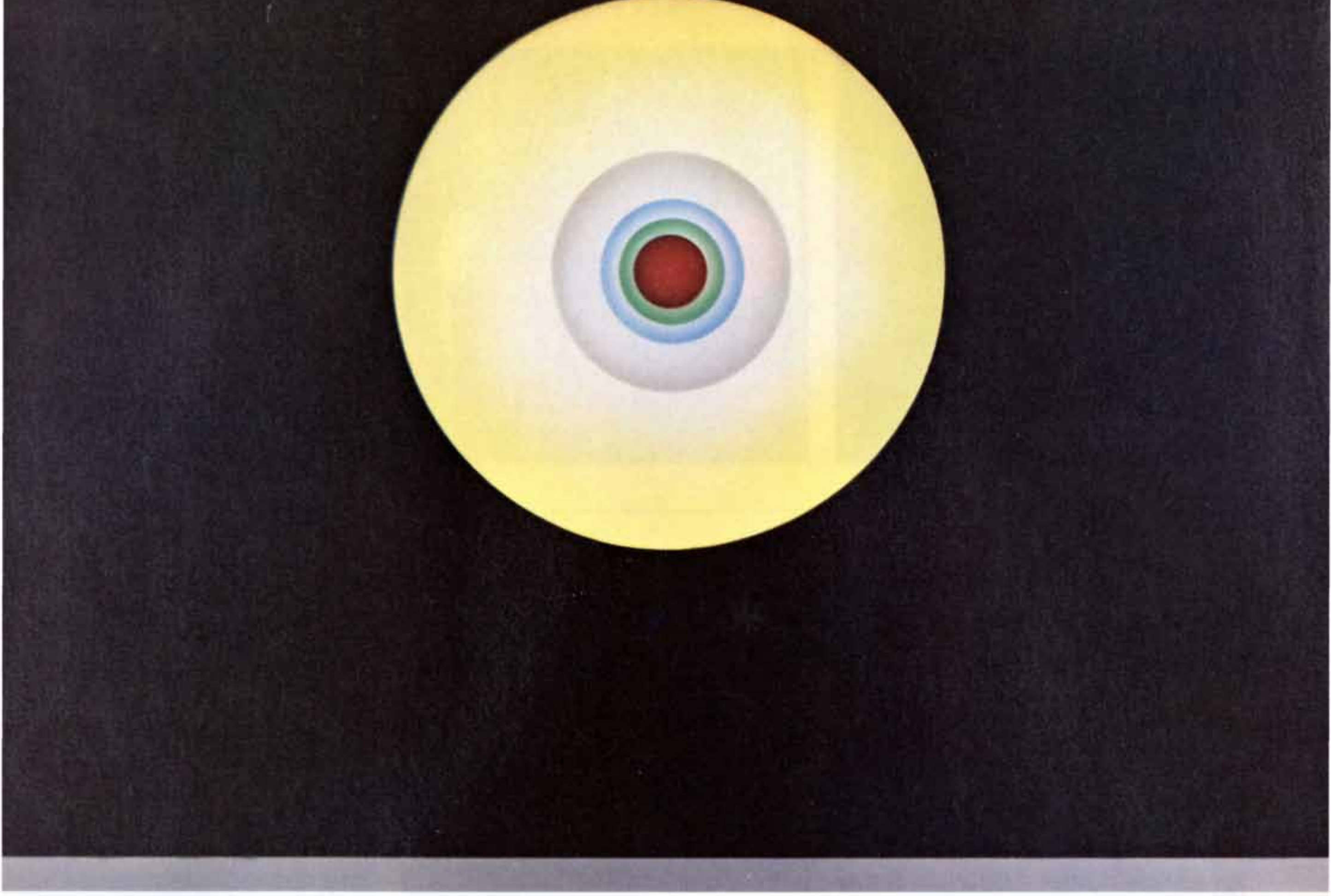
The surge of technical creativity that produced the first two generations of nuclear weapons can be explained largely by the fact that the national laboratories had massive funding, a mandate to pursue new weapon possibilities and unqualified Government support. Yet speaking as one who worked at that time on the design of nuclear weapons, perhaps the most stimulating factor of all was simply the intense exhilaration that every scientist or engineer experiences when he or she has the freedom to explore completely new technical concepts and then to bring them into reality.

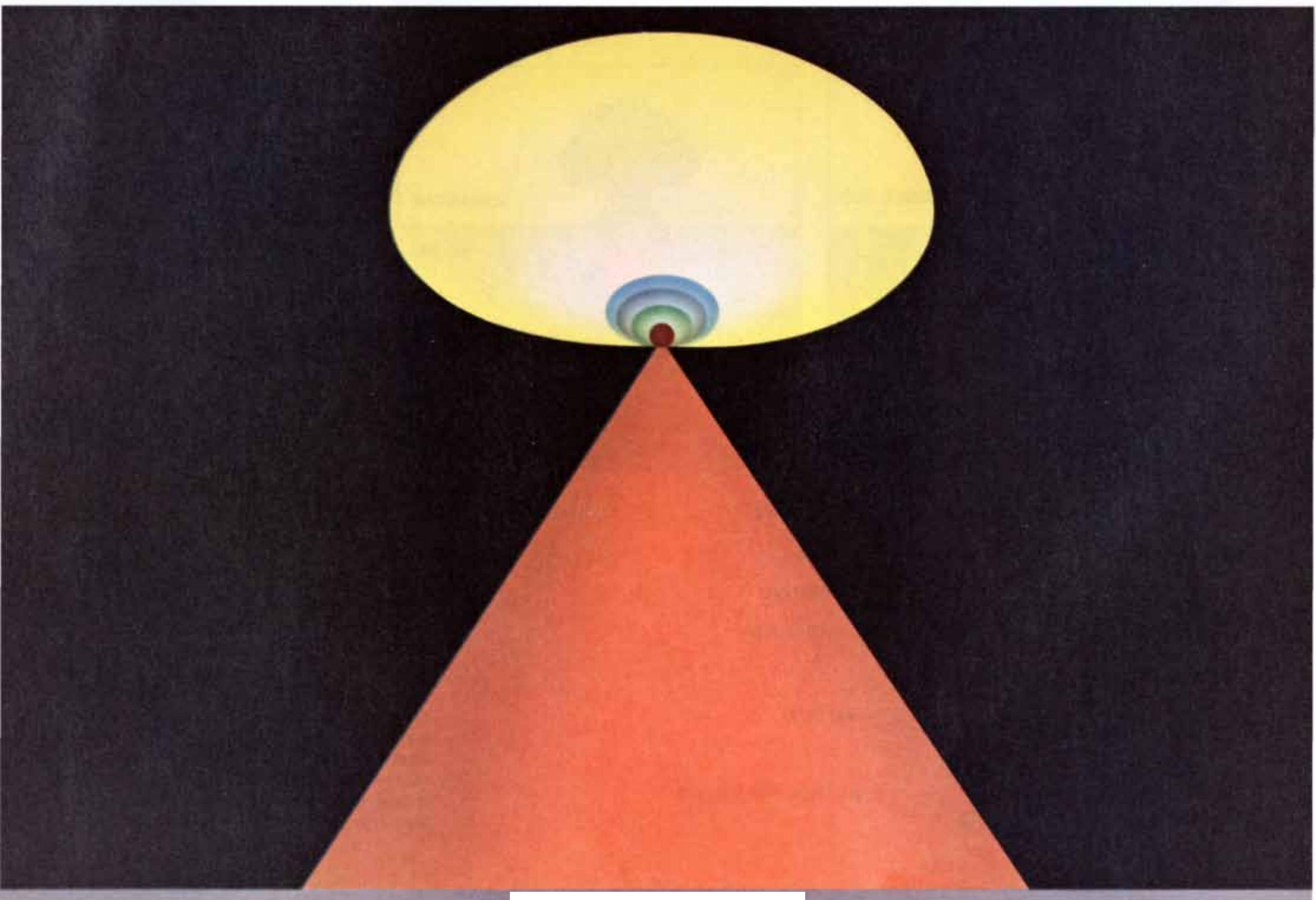
The Strategic Defense Initiative, under which a vigorous military research and development program is currently being carried out, could well generate conditions at the U.S. weapon laboratories similar to those in the 1950's. The daunting technical challenge implied in President Reagan's call to search for a way to defend the nation against ballistic missiles is likely to spur modern-day weaponeers to consider radically new types of nuclear weapons—quite apart from concurrent advances in delivery and command-and-control systems.

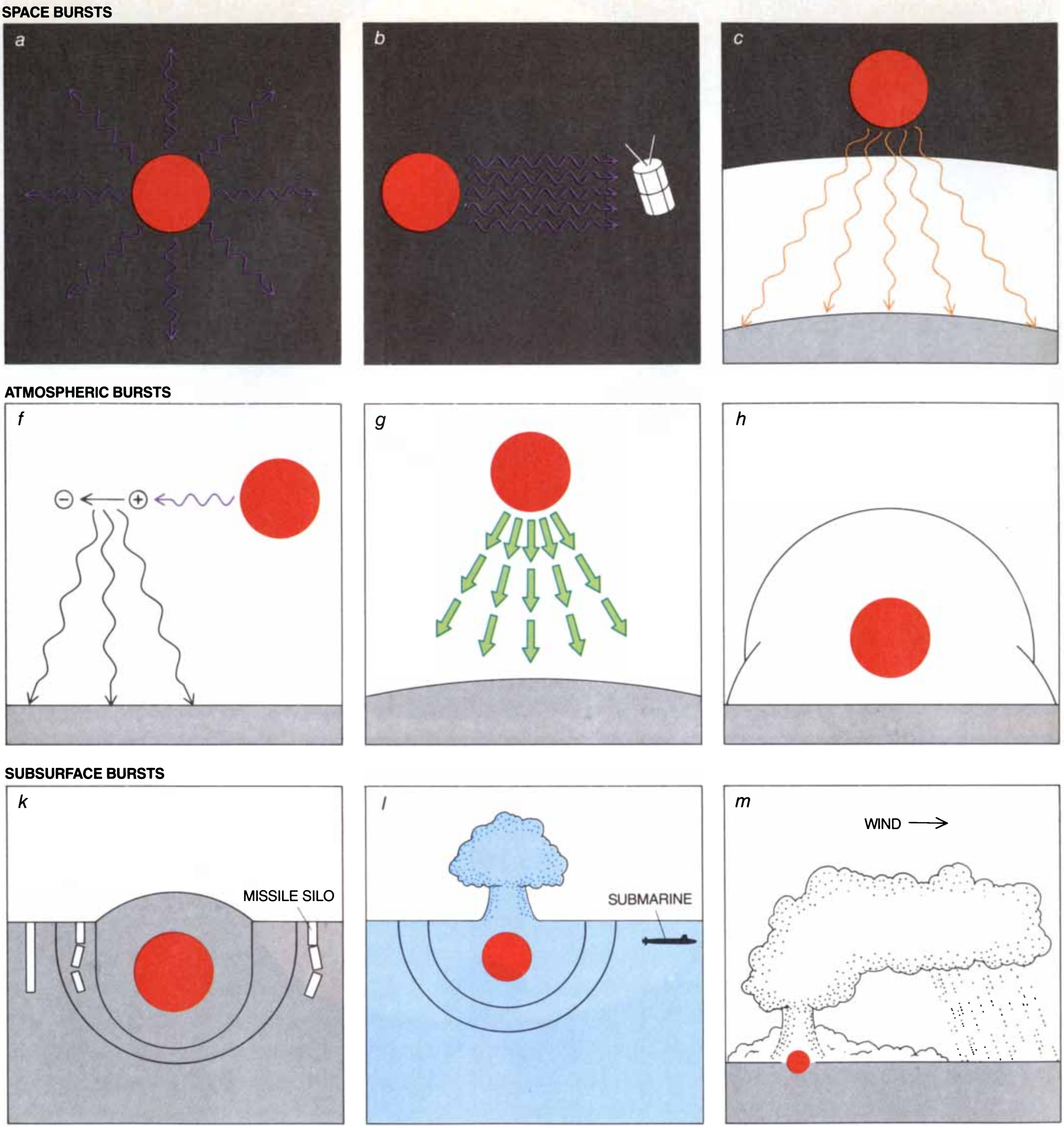
It would be logical for a weapon designer to build on the legacy of the first- and second-generation nuclear weapons, all of which transform mass into an abundance of energy that is then uniformly dissipated in a roughly spherical pattern. Such a new generation of nuclear weapons might selectively enhance or suppress certain types of energy from the vast energy source provided by a nuclear explosion. Moreover, the lethal effects of a selected energy carrier (such as electromagnetic radiation, subatomic particles or expelled material) might be increased by distorting its normal pattern of emission into a highly asymmetrical one—in essence concentrating the energy in a certain direction.

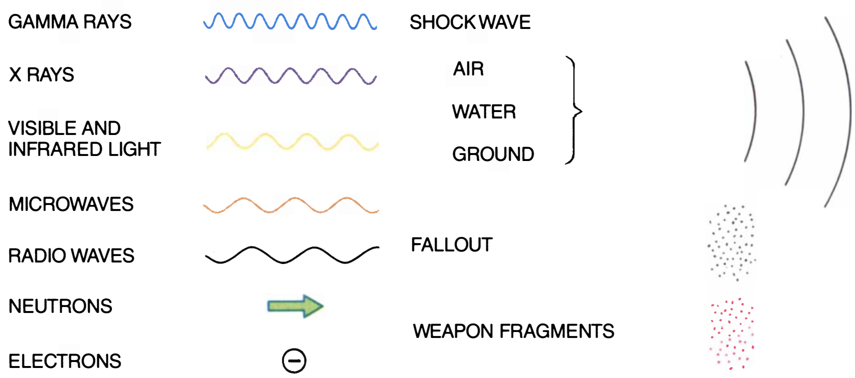
Indeed, nuclear weapons that deliver 1,000 or more times the energy per unit area on a target than does a conventional nuclear weapon are entirely plausible. Special components or materials attached to the exterior of a nuclear device could convert the energy released by its detonation into a different form; configuring the nuclear explosive and its casing in certain ways could channel most of the energy in certain directions. Alternatively, the energy released from a nuclear explosion could be converted and directed by exploiting the effect such an explosion has on natural surroundings. Regardless of their original intent, if such weapons are built, they will undoubtedly be modified for application in a

PATTERN of energy emission distinguishes current nuclear warheads from those likely to be developed in the near future. Current warheads (*top*) release their explosive energy in many forms, each of which is radiated uniformly outward. Hence the region in which military equipment would be destroyed or incapacitated for each of the major energy types (*color key above*) can be roughly represented as spheres. In contrast, warheads of future nuclear weapons could be equipped with devices that suppress, convert and direct energy, enabling a significant fraction of the explosive energy to be transformed into microwaves that are then concentrated on targets (*bottom*).

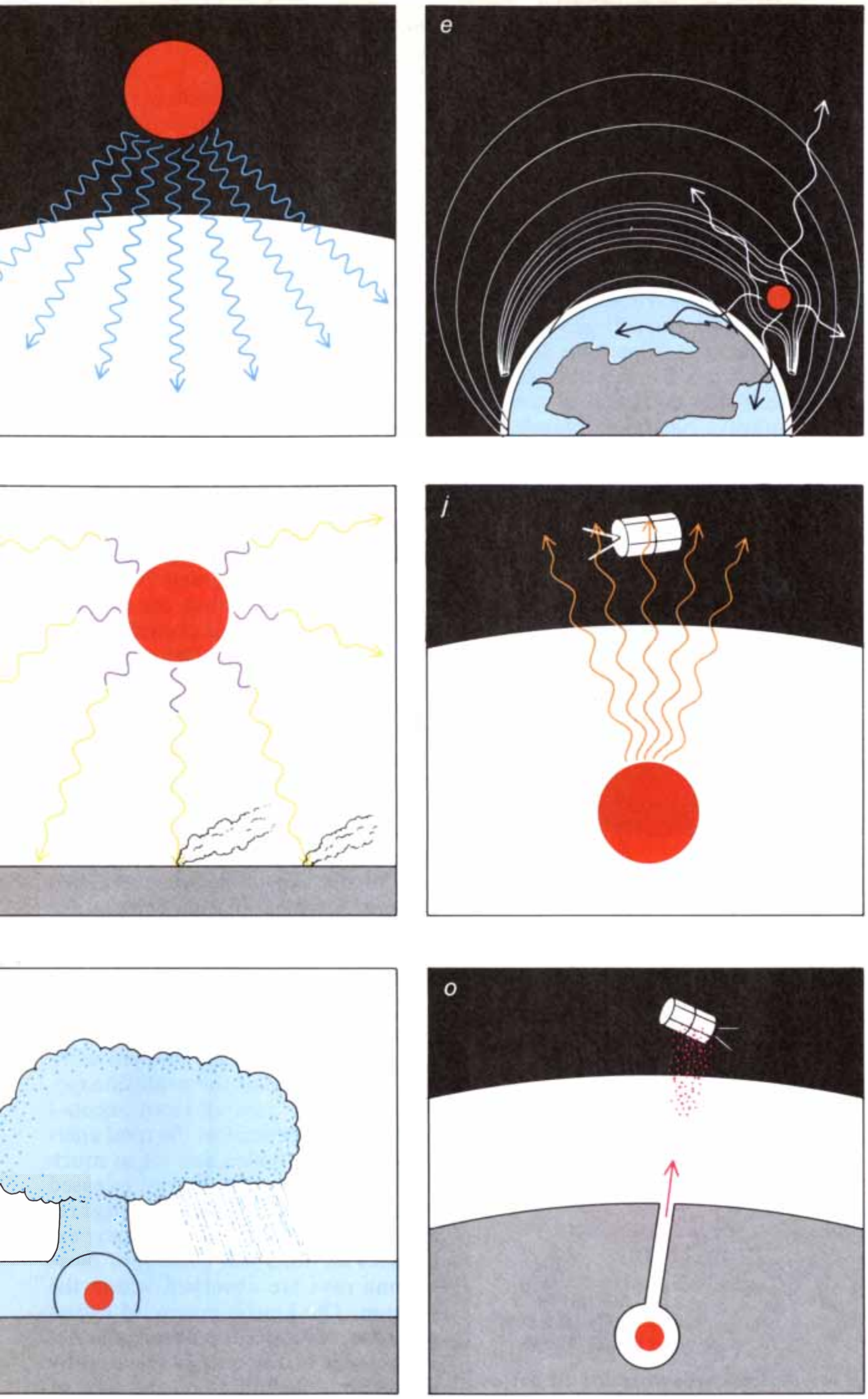








ARRAY OF EFFECTS listed in the key at the left could be militarily exploited by the next generation of nuclear weapons, which would suppress certain effects, heighten others and perhaps channel them in certain directions as well. In space (*top row*) nuclear weapons could radiate incoherent X rays in all directions (*a*) or coherent X rays in a particular direction (*b*). Microwaves can readily penetrate the atmosphere and could therefore reach the surface of the earth from space, particularly if they were concentrated (*c*). Gamma rays also travel a certain distance through the



air and could be directed to targets in the upper atmosphere (*d*). The ionized weapon debris produced by a nuclear explosion above the atmosphere but within the earth's magnetic field could produce a powerful pulse of long-wavelength electromagnetic radiation as it distorts the field (*e*). A similar effect can be achieved in the atmosphere (*middle row*): X rays can knock electrons loose from air molecules to create a sudden current surge through the air, which results in the emission of the radio-wave pulse (*f*). The more familiar neutron-emission (*g*), air-blast (*h*) and incendiary (*i*) effects of nuclear weapons could also be enhanced. Targets in space could be engaged by microwaves beamed upward (*j*). The energy of subsurface bursts (*bottom row*) could interact strongly with the surrounding medium to produce enhanced ground (*k*) or water (*l*) shock waves. The amount and distribution of radioactive fallout from nuclear weapons could be controlled, depending on the materials chosen to encase the weapon as well as on whether the weapon is detonated underground (*m*) or underwater (*n*). Finally, the blast of a subterranean explosion could conceivably propel projectiles through a "cannon barrel" and into space (*o*).

wide variety of strategic and tactical missions—offensive as well as defensive—in all kinds of environments.

Like previous generations of nuclear weapons, members of the new generation would derive their enormous explosive energy from fission (the splitting of a nucleus by a neutron into two nuclei of comparable size) or a combination of fission and fusion (the joining of two light nuclei to form a heavier nucleus). Fission explosions are easier to produce and essentially amount to bringing together, in the space of about a microsecond (a millionth of a second), enough fissile material (such as uranium 235 or plutonium 239) in a sufficiently small volume so that a huge number of fission-inducing neutrons can be quickly generated in the material. The high-speed assembly of the fissile material is generally achieved by precisely detonating chemical-explosive charges in such a way as to propel subunits of the material together to form a single compressed mass.

Initiating a fusion explosion is a much more complex affair, because extremely high temperatures (on the order of hundreds of millions of degrees Kelvin) are required. In fact, the only practical mechanism by which to generate such temperatures in a transportable device is a fission explosive. A pure-fusion explosive—without a fission trigger—reportedly still eludes weapon designers.

Fusion reactions not only release substantially more energy per unit weight than fission reactions but also produce more high-energy neutrons. The additional neutrons can in fact "boost" the yield of a fission weapon if they are allowed to interact with uranium or plutonium in the weapon's core. Hence placing small quantities of thermonuclear fuel such as tritium or deuterium (both are isotopes of hydrogen) in a fission weapon increases the overall yield-to-weight ratio of the weapon, since the added weight needed for boosting is insignificant.

Unlike boosted weapons, in which the energy released by fusion does not significantly contribute to the overall weapon yield, so-called thermonuclear weapons derive a substantial part of their explosive energy from fusion reactions. The relative amounts of energy attributable to fusion and fission depend on the design of the weapon. If a considerable amount of lithium deuteride (which, when it is irradiated with neutrons, produces tritium) is compressed and heated by the energy released from a small fission-explosive trigger, the fraction of the total yield due to fusion in relation to the fraction due to fission can become very large. Such weapons are sometimes called "clean" thermonuclear weapons, because they release relatively few radioactive fission products.

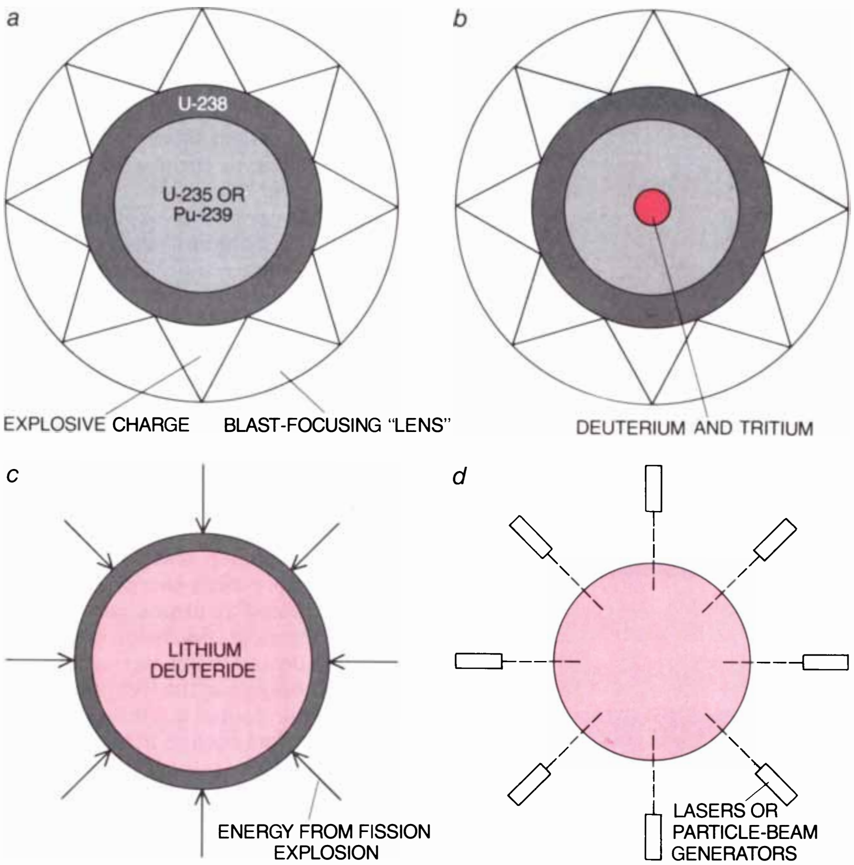
At the other extreme are weapons in which the thermonuclear fuel is enclosed in a substantial quantity of ordinary uranium (uranium 238). The high-energy neutrons produced by fusion in the thermonuclear fuel can induce fission in the surrounding uranium, multiplying the total fission yield considerably.

The yield-to-weight ratios of pure fission warheads have ranged from a low of about .0005 kiloton per kilogram to a high of about .1 kiloton per kilogram. (One kiloton is equivalent to the detonation of about 1,000 tons of TNT.) The overall yield-to-weight ratio of strategic thermonuclear warheads has been as high as about six kilotons per kilogram. Although the maximum theoretical ratios are 17 and 50 kilotons per kilogram respectively for fission and fusion reactions, the maximum yield-to-weight ratio for U.S. weapons has probably come close to the practical limit owing to various unavoidable inefficiencies in nuclear-weapon design (primarily arising from the fact that it is impossible to keep the weapon from disintegrating before complete fission or fusion of the nuclear explosive has taken place). Yet even the lowest yield-to-weight ratio of a pure fission weapon is orders of magnitude higher than the ratio of chemical explosives.

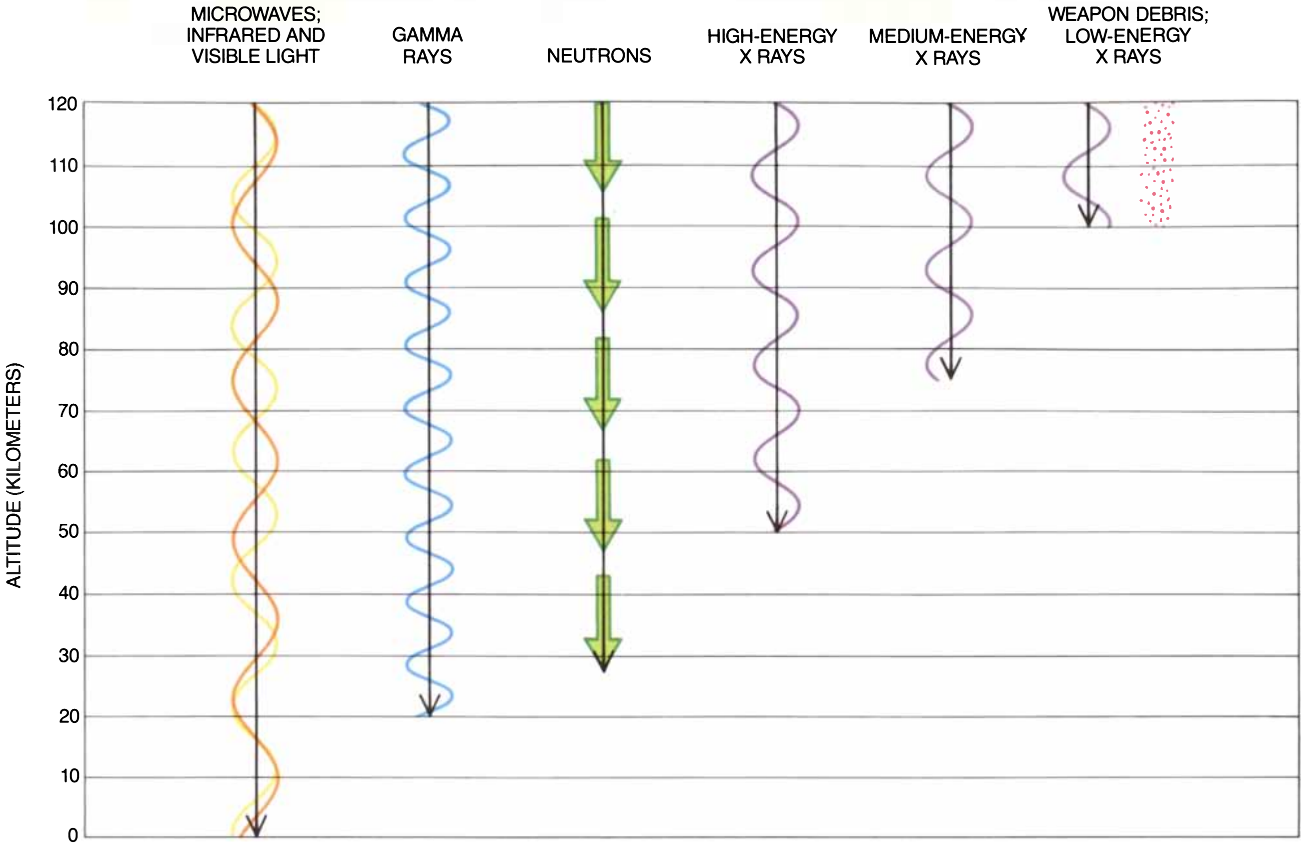
Indeed, the discharge of energy from a detonated nuclear weapon is so massive and violent that it immediately vaporizes and ionizes the weapon itself, converting it into plasma: an extremely hot gas of positively charged ions and negatively charged electrons. In addition substantial quantities of gamma rays and neutrons are emitted as by-products of the fission and fusion reactions. The kinetic energy of the weapon-debris plasma as well as the nuclear emanations constitute what could be called the primary effects of a nuclear explosion; they arise in any nuclear burst, regardless of the environment in which it takes place.

Plasma at the temperatures prevailing just after a nuclear explosion radiates X rays. Indeed, about 70 percent of the energy emitted in the first few microseconds after an explosion consists of this radiation. The exact fraction of the total explosive energy released in the form of primary X rays tends to increase with the yield-to-weight ratio, since the ratio determines the overall temperature of the weapon-debris plasma. The greater the amount of energy dissipated in the form of X rays, the less the kinetic energy of the expanding weapon-debris plasma. A typical plasma velocity for a thermonuclear weapon with a high yield-to-weight ratio would be about 1,000 kilometers per second, representing some 10 percent of the total explosive energy.

Gamma rays that are emitted within a second or so of the explosion (so-called prompt gamma rays) account for about 3.5 percent of the total energy released by fission and for as much as 20 percent of the energy released from some cycles of thermonuclear reactions. In current types of nuclear explosives all but a few percent of these gamma rays are absorbed within the weapon. The kinetic energy of excess neutrons accounts for about another 1.8 percent of the energy released by fission and, depending on the type of thermonuclear fuel, between 40 and 80 percent of the energy released by fusion. High-energy neutrons, however, tend to be slowed down by inelastic scattering or collision with light elements in the materials of implosion systems. The average energy of the neutrons that actually escape capture in the weapon materials and are released into the environment is therefore typically much lower. This effect is particularly pronounced in thermonuclear weapons, since the fuel consists of light elements. Indeed, in such weapons the energy of the neutrons is



FOUR TYPES OF NUCLEAR EXPLOSIVES are depicted schematically; all but one rely on fission (the splitting of a nucleus by a neutron into two lighter nuclei). A weapon relying solely on fission for its explosive energy (*a*) consists of a core of fissile material (uranium 235 or plutonium 239) surrounded by chemical-explosive charges and inert structures that focus the charges' blast energy inward, causing the core to implode and thereby initiate a runaway fission reaction. The yield of fission explosives can be "boosted" (*b*) by placing deuterium and tritium (isotopes of hydrogen) in them. The temperatures produced on detonation of a fission explosive cause the hydrogen isotopes to undergo fusion (the joining of nuclei), releasing substantial quantities of neutrons, which induce more fission reactions. In boosted weapons the fusion reaction does not contribute significantly to the total yield of the weapon. Fusion reactions can account for most of a nuclear weapon's yield, however, if a substantial amount of such a thermonuclear fuel as lithium deuteride is exposed to the energy released by fission (*c*). An outer shell of normal uranium (uranium 238) serves to hold the warhead together just a fraction of a microsecond longer before it blows apart, enabling the nuclear reactions to produce more energy. Also, when it is irradiated with neutrons produced by fusion, the U-238 itself undergoes fission. A pure-fusion weapon (*d*), which dispenses with a fission trigger by applying laser, electron or ion beams to implode thermonuclear fuel, reportedly eludes weapon designers.



ATMOSPHERIC PENETRATION of the energy emitted by a nuclear burst in space depends on the energy type. Radiation in the microwave, infrared and visible ranges of the electromagnetic spectrum could reach the ground with relatively little attenuation.

deliberately deposited within the thermonuclear fuel, since neutrons play a vital role in maintaining the elevated temperatures needed to achieve high reaction rates.

Most nuclear-weapon development for the past 40 years has not had the aim of significantly enhancing or suppressing particular forms of energy other than by adjusting the relative amounts of fission and fusion taking place in the warhead. One exception is the so-called neutron bomb [see "Enhanced-Radiation Weapons," by Fred M. Kaplan; SCIENTIFIC AMERICAN, May, 1978]. A neutron bomb is a low-yield thermonuclear explosive specifically designed for an increased output of high-energy neutrons per kiloton of total yield. It is intended to be a nuclear antipersonnel weapon that produces minimal concomitant blast damage and radioactive fallout.

Yet just as a nuclear weapon can be designed to enhance its output of primary neutrons at the expense of blast and radioactive fallout, virtually any other primary energy released by a nuclear explosive could similarly be enhanced by placing appropriate materials in suitable geometries close to the explosive. Significant control over the amount and energy of X-radiation, for example, could be achieved by changing the average molecular weight of the materials in the weapon, the weapon's exterior surface area and the way the energy generated in its core is distributed over the expanding front of weapon debris after detonation.

Changes in the design of thermonuclear weapons could also substantially increase the energy accounted for by prompt gamma rays. One possibility is to encase the weapon with an isotope that, when it is bombarded with neutrons, emits gamma rays. In this way excess fission or fusion neutrons escaping from the weapon's core could induce the emission of gamma rays, nearly half of which would leave the expanding explosion debris. (The other half would radiate inward and be absorbed by the debris material.)

The quantities of radioactive fission products (the main component of fallout) among the weapon debris could similarly be controlled over very wide ranges, particularly for thermonuclear weapons with yields greater than a few hundred kilotons. Furthermore, by blanketing the weapon with isotopes that, when they are irradiated with neutrons, produce radioactive nuclei having selected half-lives and decay modes, the lethality of the radioactive fallout could be increased.

The effects of a nuclear explosion could also be made directional in the same way high-explosive devices such as conventional shaped charges can produce armor-penetrating jets of molten metal or directional shrapnel. By considering how explosive charges of nonspherical shape release their energy some insight can be gained on how this could be done [*see illustration on next page*].

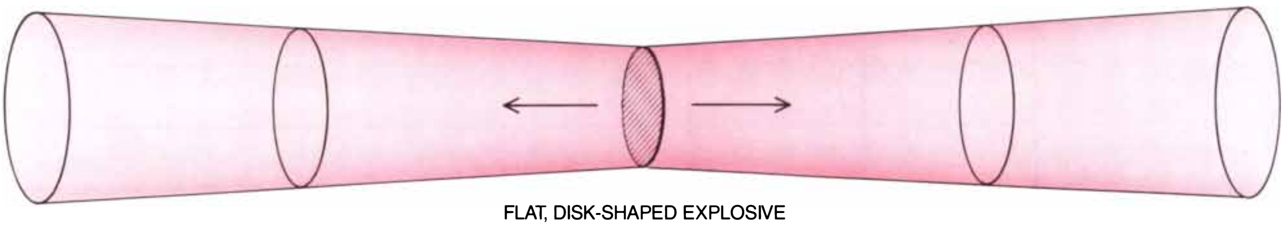
Detonating a disk of high explosive all at once, for example, causes the explosion products to be flung out in a characteristic double-cone pattern. The reason is that the velocity of the explosion products in a direction perpendicular to the disk's two surfaces will be higher than their radial velocity. The apex angle of the cones will be determined by the ratio of the thickness of the disk to its diameter. The average total kinetic-energy flux (energy per unit area per unit time) of the explosion products crossing a plane perpendicular to the axis of the double cone could therefore be considerably greater than it would be if the same mass of high explosive expels its products spherically. If the average velocity of the explosion products in the direction of the cone's axis is 40 times their average radial velocity (corresponding to a cone angle of about three degrees), the enhancement factor would be about 3,000.

Another example is the detonation of a long, thin cylinder of high explosive. In this case the highest explosion-product velocities would be perpendicular to the axis of the cylinder. Hence the explosion products would tend to preserve a cylindrical pattern; the energy-flux enhancement factor in this example tends to be smaller than the factor in the preceding one.

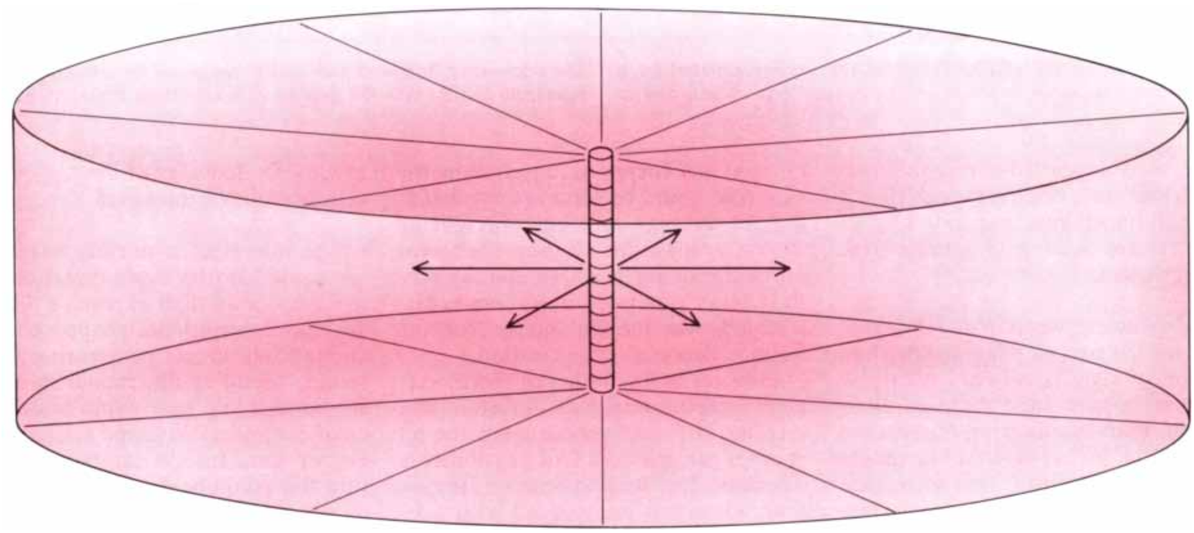
A final example is a charge of high explosive that is tamped, or restricted, by dense material in all directions except forward. In such a case the explosion products would be projected primarily forward. The additional weight entailed by the inert mass around the explosive is more than balanced by the concentration of the energy through the opening in the tamper. That is why a rifle bullet can produce much greater damage to a target than the detonation of a mass of high explosive having the same weight as the rifle.

Of course, nuclear reactions release many more forms of energy at much higher intensities than chemical high explosives, including gamma rays, X rays, neutrons and a wide variety of radioactive nuclei. It is clear that even nuclear explosives of very low yield offer many more opportunities than chemical explosives to produce such directional effects.

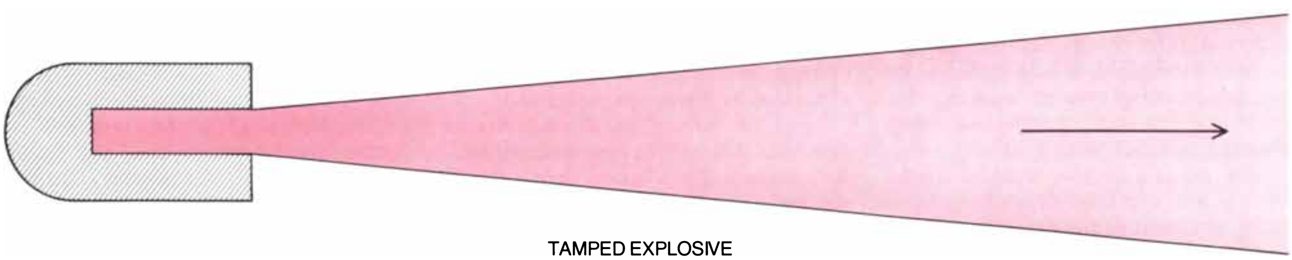
Most of a nuclear explosion's lethal effects are actually secondary effects resulting from the interaction of the kinetic energy of the weapon-debris plasma and the initial radiation (namely X-radiation) with the medium in which the detonation takes place. Hence many nuclear-explosion phenomena of military interest are determined by properties of the medium



FLAT, DISK-SHAPED EXPLOSIVE



LONG, THIN CYLINDRICAL EXPLOSIVE



TAMPED EXPLOSIVE

SHAPED CHEMICAL CHARGES can eject their explosion products (primarily blast and weapon debris) in markedly nonspherical patterns. A flat disk of chemical explosive, for example, emits its products in a characteristic double cone. Setting off a long, thin cylinder of explosive produces a cylindrical pattern of emission. Finally, by tamping, or restricting, the effects of the explosion with inert, dense material in all but one direction, the explosive products can be concentrated in that direction. Nuclear explosives could presumably apply such directional effects to control the pattern in which their explosive products are emitted.

such as its pressure, density and composition. It is the variations in these properties that account for the widely divergent responses associated with nuclear bursts in space, in the atmosphere, on the surface of the earth and below the earth's surface. By choosing the appropriate primary effects to be enhanced or suppressed, depending on the prevailing environmental conditions, the secondary effects of the weapon can be more efficiently transmitted to targets.

Because space is essentially empty, there is no medium with which to interact, and the primary products of a nuclear explosion (X rays, weapon-debris plasma and nuclear radiation) continue to travel in the same directions in which they were released until they hit something or are deflected by the earth's magnetic or gravitational field (depending on whether they have respectively electric charge or mass). That is why initial asymmetries in the distribution of mass in an explosive set off in space tend to be preserved out to great distances in the pattern of the energy radiated.

If a nuclear explosive is detonated above the atmosphere but within the earth's magnetic field, the plasma expanding in directions more or less perpendicular to the magnetic field lines will distort the field. When this happens, a large fraction of the kinetic energy in the weapon debris is converted into electromagnetic energy, resulting in the emission of a sudden burst of radiation with a broad range of wavelengths—from a few meters to hundreds of kilometers or more. Such an electromagnetic pulse (EMP) can represent a substantial fraction of the total energy of the explosion and can propagate with little attenuation through the atmosphere to the earth's surface.

Nuclear explosions in space or in the high-altitude regions of the atmosphere can produce another type of EMP. In this case gamma or high-energy X rays striking the upper part of the atmosphere cause electrons to be ejected from air molecules. Such a sudden cascade of electrons is equivalent to a huge surge of electric current. Since the current would not be spherically symmetrical (it would flow predominantly in the direction of higher air density, namely downward) and would vary with time, it would generate transient magnetic fields that in turn would produce electromagnetic radiation in the form of an EMP.

As a result of the approximately exponential increase in the density of the atmosphere with decreasing altitude, much of the energy radiated downward by a nuclear explosion above the atmosphere is deposited in the atmosphere's upper reaches. Deposition of this energy can sometimes produce severe secondary effects that then propagate to the surface of the earth. X rays and weapon debris at sufficiently high fluences (total energy per unit area) can, for example, heat the atmosphere to such high temperatures that it radiates visible light and infrared radiation. Gamma rays, neutrons and X rays released by the weapon, as well as the decay products of radionuclides, can directly or indirectly generate electric currents in the layer of the atmosphere where they deposit their energy. These currents can then generate other EMP's whose wavelengths and instantaneous power levels extend over a very wide range. Heating of the atmosphere can also initiate complex chemical reactions that affect its transmission and reflection of radio waves.

In the lower atmosphere, underground or underwater the primary X-radiation leaving an exploding nuclear weapon is absorbed by the atoms and molecules of the surrounding medium within a few meters of the point of detonation. Consequently the medium is quickly heated, forming a fireball, which in turn reemits electromagnetic radiation of lower frequencies. Most of this radiation is in the visible and infrared regions of the spectrum and can travel considerable distances through the air.

The radiative energy also combines with the kinetic energy of the outwardly expanding plasma to produce a pressure impulse of tremendous force on the surrounding medium. Such an impulse forms a shock, or blast, wave that propagates through the medium. The denser the medium, the greater the amount of energy transformed into the shock wave. Hence for explosions in water or earth a larger percentage of the explosion's energy is converted into a shock wave than is the case for explosions in air.

Surface, subsurface or very-low-altitude explosions can also fling huge quantities of dust, crater debris, manmade structures or water into the air that can directly or indirectly cause considerable destruction. Moreover, much of this material is likely to be rendered radioactive, thereby severely contaminating extensive areas through fallout.

Forms of energy that are not normally released as primary or secondary effects can also be generated from the vast energy supply provided by a nuclear burst. Furthermore, such energy can be channeled into small emission angles. The key question about such weapons (which cannot be answered in detail here because the subject is classified) is how to convert a substantial fraction of the energy of a nuclear explosion into a particular energy that can be emitted with high directional enhancement. Suffice it to say that electromagnetic energy with wavelengths typical of gamma rays, X rays, visible light and microwaves can be focused by the equivalent of lasers: devices that cause the atoms or molecules of a material to radiate in phase. Longer-wavelength radiation can be emitted directionally if such weapons are equipped with the equivalent of antennas. The problem in either case is how to channel the torrential flow of energy from a nuclear explosion into an energy-conversion and -direction device in the few microseconds before the entire weapon assembly disintegrates. Another option, which may simplify the problem somewhat, is to set off nuclear devices in a reusable containment structure from which the explosive energy could then be tapped. Such structures, designed to withstand explosions with yields of up to perhaps one kiloton, have in fact been under study for several decades. The Lawrence Livermore National Laboratory has recently considered a proposal to construct such a chamber in which a variety of nuclear effects could be studied.

For ground-based weapons intended to attack targets in space the weight of the needed equipment is not critical; for space-based weapons it is, however. It is therefore to be expected that the technical approaches for developing ground-based directed-energy nuclear weapons will be different from those required for similar weapons in space. Some advantages that ground-based weapons have over weapons placed in space include avoidance of treaties banning nuclear weapons in space, accessibility to large and heavy conversion equipment (with associated higher directivity and greater efficiency of conversion of the explosion energy into the form radiated), much lower cost and possible reusability of the equipment.

Conversion of the explosion energy into more tractable electrical-energy pulses can be accomplished by magnetohydrodynamic generators: devices that convert a plasma's kinetic energy directly into electricity. (Such devices have been proposed for converting fusion energy in a power reactor into electricity.) The pulses of electrical energy could then drive devices for conversion of the electricity into electromagnetic radiation (with or without an attendant self-destruction of the device) that could be tightly focused toward targets in space. In most cases the low efficiency of such energy conversion can be more than compensated for by a high degree of focusing in the direction of a target.

An extremer possibility is the use of a relatively small nuclear explosion deep underground to accelerate very large projectiles through the equivalent of a cannon barrel. These so-called hypervelocity projectiles would reach velocities close to earth-escape velocity (about 10 kilometers per second). Appropriately shaped, compact projectiles can thus penetrate the atmosphere in a way that is somewhat analogous to penetration of the atmosphere by large meteorites. Such proposals were studied as long ago as the late 1950's as a method for placing massive loads of materials in space at relatively low cost.

The kinetic energy of, say, 10 tons of material moving at 10 kilometers per second is the equivalent of about 100 tons of TNT. This suggests that reasonably efficient use of a nuclear explosion with a yield in the vicinity of one kiloton could provide more than enough propulsive energy. If the "cannon barrel" were a few hundred meters long, the average acceleration of the projectile would be on the order of 10,000 times the acceleration of the earth's gravity, which is not beyond the strain-bearing capacity of a compact, high-density projectile. Subsequent fragmentation of such a projectile into solid chunks or liquid droplets could make it a highly effective weapon for destroying satellites or ballistic-missile warheads in space.

Another possibility is to design nuclear weapons so that the act of detonation itself directly accelerates material on the weapon that immediately fragments into small pellets or droplets moving at velocities substantially greater than 10 kilometers per second. Such weapons could readily focus the hypervelocity fragments into a conical volume, but they would have to have a mechanism to control the acceleration process in order to avoid vaporizing the fragments. In addition they would probably be limited to attacking targets in space or in the upper atmosphere, since at low altitudes the ranges of such fragments are much less than the distances at which the detonation's air blast causes severe damage.

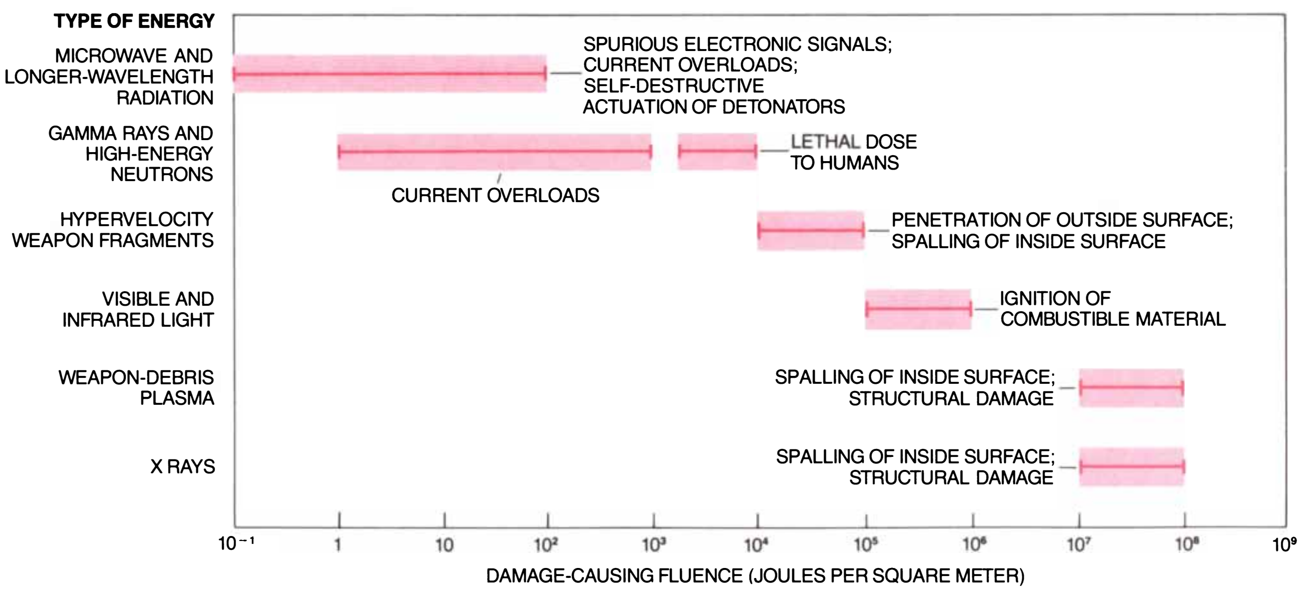
The damage an object is likely to suffer when it is exposed to the gamut of energy types emanating from a nuclear explosion can be roughly calculated by estimating the type of energy likely to reach the object, the way in which damage could be done and in many cases the rate of deposition of the energy. This aspect of the effects of nuclear explosions is extremely complex and often not well understood.

Ranges of total energy fluence that can cause temporary malfunction or permanent damage in military or civilian targets vary over nine orders of magnitude [*see illustration below*]. The effects of the longer-wavelength radiation (such as that produced by an EMP) at the low end of the energy-fluence scale are the subtlest and the most difficult to assess and are therefore the most uncertain.

A fluence of .1 joule per square meter is one million times greater than an easily detectable one-second radio signal emitted by a 10-kilowatt spherically symmetrical radio transmitter 100 kilometers away. Yet commercial and military communications and radar transmissions producing smaller fluences have been known to cause accidental firings of high-explosive detonators and malfunctions in computers and other electronic and electrical equipment. These effects would be similar to those produced by the EMP from nuclear explosions. Indeed, the effects of electromagnetic radiation on military ordnance have prompted efforts to protect against it. Some measures include enclosure in conducting shields and avoidance of components that can be sensitive to even small pulses of current induced by electromagnetic radiation that has leaked in. Yet these measures have not always been entirely successful.

Some components of electronic systems, such as transistors, can be very sensitive to small currents and other effects resulting from gamma-ray and neutron bombardment. These effects can be minimized by shielding or by avoidance of highly sensitive components. Yet the general lack of protective measures in nonmilitary space systems makes them particularly vulnerable to such nuclear radiation.

Gamma rays, neutrons, high-energy X rays or radionuclides impinging on targets in space can also cause the target to become charged to a potential that is on the order of the maximum energy of ejected charged particles. It is possible that the electric field



DESTRUCTIVE EFFECTS of different types of energy are listed in this chart as well as the fluence (total energy per unit area) necessary to achieve such effects on military equipment. Since relatively small fluences of microwave or longer-wavelength radiation are sufficient to cause damage, such kinds of radiation may be the energy types emphasized in third-generation nuclear weapons.

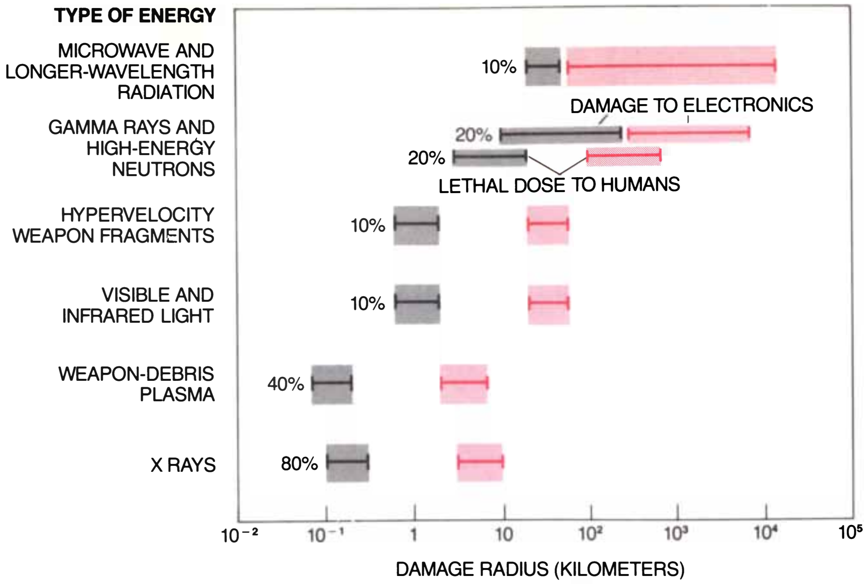
strength near the surface could reach values on the order of one million volts per meter, sufficient to induce malfunctions or permanent damage in some types of internal electrical systems that are not well shielded.

Unlike neutrons or gamma rays, hypervelocity fragments would pit the surface of a target. Exceedingly rapid ejection of the material during the pit formation drives a strong shock wave into the target. Because of their high velocities, which are up to about 100 times faster than a high-speed rifle bullet, hypervelocity fragments weighing much less than one gram can do considerable damage when they are aimed at targets in space.

Visible light or infrared radiation released as a secondary effect from the heating of the atmosphere primarily causes damage by igniting combustible materials on the surface of targets. Even if the target surface is not combustible, nonuniform heating of the surface can nonetheless cause damage from the resulting thermal stresses.

Incident high-energy X-radiation or weapon-debris plasma damages a target in space principally by the rapid blowoff of vaporized material from the target's surface. If X rays are the agent, the resulting shock can be transmitted through the outer layers of the object, causing the inside surfaces to shatter, presuming the time necessary to deposit the incident energy is short compared with the time required for the shock to reach the inner surface. Such a process is called spalling. For incident weapon-debris plasma, however, spalling does not generally occur. The reason is that it takes too long for the weapon-debris plasma to deposit its kinetic energy. In any case, the overall momentum transferred inward from the surface blowoff can result in incapacitating damage even if there is no interior spalling.

To help make these estimates more accessible, one can consider the range within which a particular energy carrier can produce destructive effects [*see illustration on this page*]. Potentially huge damage ranges (or, equivalently, large fluences at a given distance) can be readily achieved by emitting energy within a narrow angle. Microwaves that have wavelengths between three centimeters and one meter are particularly suited for such directional enhancement because the atmosphere is essentially transparent over this range, making it possible to use the radiation for ground-to-space, space-to-ground and space-to-space applications. Also, the ranges of the microwave-energy fluence needed to cause damage to many types of military and civilian targets are the lowest of all forms of electromagnetic radiation.



MAXIMUM DISTANCE from the detonation of a nuclear weapon at which damage can be done to military targets in space depends on the type of energy causing the damage and how much of the total explosive energy it represents. Two cases are considered: a one-kiloton weapon (*black*) and a one-megaton weapon (*color*). (A kiloton is the energy equivalent of the detonation of 1,000 tons of TNT; a megaton is 1,000 kilotons.) The bars indicate the range of damage-radius estimates for plausible third-generation weapons, whose energies have been enhanced but not directed. The percentage of the total explosive energy funneled into each particular energy type is indicated next to each pair of bars. Much greater damage radii could be achieved if the weapons focus their energy.

The military potential of directed microwave beams is therefore awesome. Suppose, for example, it should become possible to convert 5 percent of the energy released by a one-kiloton explosion into three-centimeter radiation that is emitted by a 50-meter-diameter antenna or an equivalent microwave laser. The explosion of such a device in a 30,000-kilometer geosynchronous orbit would deposit about 800 joules per square meter over an area of 250 square kilometers on the earth's surface (larger than the area of Washington, D.C.). This estimated energy fluence is greater than the level known to cause severe damage to many types of electrical equipment—computers, antennas, relays and power lines. Of course, at much shorter distances the energy fluence would be much larger, about five million joules per square meter at a distance of 400 kilometers.

The development and deployment of such a microwave weapon would greatly complicate both offensive and defensive military tactics and strategy. It could, for example, cause temporary malfunctions or permanent damage in the complex electronic and electrical equipment that is typically found in military systems for surveillance, tracking, communications, navigation and other command-and-control functions. Because the atmosphere is virtually transparent to microwaves, either the beam-generating device or the intended target could be based in space, in the atmosphere or on the earth's surface. In any event, the deployment of such weapons is likely to undermine confidence in the wartime reliability of strategic and tactical forces, including those forces that constitute the ultimate deterrent to nuclear war.

How likely is it that these third-generation nuclear weapons will actually be developed and deployed? The answer depends largely on the character and extent of support provided by both the U.S.S.R. and the U.S. to their respective national weapon laboratories. Since developments in the military realm of one country invariably elicit emulative responses from the other, the likelihood strongly depends on what is perceived to be the pace of the adversary's research and development in this area.

One key indicator of the extent of a country's effort is the frequency of nuclear testing. If the U.S. continues and the U.S.S.R. resumes underground nuclear testing even at levels substantially lower than the 150-kiloton limit stipulated in the Threshold Test Ban Treaty, it will probably be just a matter of time before these new types of offensive and defensive nuclear weapons are developed.

How Photoreceptor Cells Respond to Light

New information about how light energy is changed into neural signals shows how an individual photoreceptor cell of the eye registers the absorption of a single photon, or quantum of light

by Julie L. Schnapf and Denis A. Baylor

Vision begins with the conversion of packets of electromagnetic energy called photons, or quanta, into neural signals the brain can analyze. The translation is accomplished by the photoreceptor cells of the eye. They lie in a mosaic at the back surface of the retina, the plate of neurons lining the inside of the eyeball. The cornea and lens of the eye form an image of the outside world on the layer of photoreceptors. Each cell absorbs the light at one point of the image and generates an electrical signal that encodes how much light has been absorbed. The signals are transmitted through an elaborate array of synapses, or neural junctions, in the retina and brain. At these junctions signals from the population of photoreceptors are pooled and compared. The process enables the visual system to obtain information about form, movement and color in the outside world.

Given the key role of the photoreceptors in vision, it is surprising that for a long time not much was known about how they operate. The situation has changed dramatically over the past quarter century or so. Improved methods for making electrical recordings from individual photoreceptors have provided detailed information about the mechanism by which light energy is transduced into neural signals. The new techniques have made it possible to observe directly the signal triggered by the absorption of a single photon. Such measurements have also led to simple explanations for several features of overall visual performance, such as why we perceive dim stimuli more slowly than bright ones, why we sometimes see light in complete darkness and why certain mixtures of different wavelengths evoke the same color sensation as light of a single wavelength does.

In the eyes of most vertebrates there are two types of photoreceptors: rod cells and cone cells. Rods mediate vision in dim light but are so sensitive that they become overloaded and incapable of signaling in ordinary daylight. Daylight vision is mediated by cones, which operate successfully at high light levels. Cone vision is richer in spatial and temporal detail and makes it possible to sense colors.

Rods and cones bear specialized organelles for transducing and transmitting signals. At one end of the cell (farthest from the lens) is the so-called outer segment, which absorbs light and generates electrical signals. At the other end of the cell is the synaptic ending, which relays the signals to other neurons (bipolar and horizontal cells) in the retina by secreting a chemical transmitter. Between the outer segment and the synaptic ending lies a region called the inner segment.

The outer segment of a rod is cylindrical, whereas the outer segment of a cone usually tapers—hence the names rod and cone. Both kinds of outer segment contain a large expanse of photosensitive membrane studded with light-absorbing pigment molecules. Rods contain the reddish pigment rhodopsin. In the human retina there are three kinds of cone, each of which contains a pigment that absorbs strongly in the short-, middle- or long-wavelength region of the visible spectrum. The differences in the absorption bands of the three cone pigments provide the basis for color vision. In starlight, when vision is mediated by rods, all objects appear colorless.

In the rods the photosensitive membrane consists of an orderly pile of disks inside a separate surface membrane, resembling a stack of coins inside a test tube. In the cones, on the other hand, the photosensitive membrane consists of one large, elaborately folded sheet that also serves as the surface membrane. The membrane topology of the rods indicates that a diffusible substance, an "internal transmitter," relays information from the disks, where light is absorbed, to the surface membrane, where the electrical signal is generated. Evidence from many laboratories now indicates that the transmitter is a nucleotide, cyclic guanosine monophosphate (cGMP), which also takes part in transduction in cones.

How does the absorption of light by a rod or a cone generate an electrical signal? The answer requires an understanding of how the photoreceptor behaves in darkness. One might naively think the cell would be dormant in the absence of light; in reality, however, the cell is abuzz with activity. The membrane of a photoreceptor, like the membrane of other cells, separates solutions that have different concentrations of ions (atoms with a net electric charge). The solutions both outside and inside a photoreceptor contain positively charged sodium ions and positively charged potassium ions.

ROD CELLS AND CONE CELLS in the retina of the tiger salamander are enlarged 2,000 diameters in a scanning electron micrograph made by Scott Mittman and Maria T. Maglio of the University of California at San Francisco. The cylindrical cells are the rods, the smaller conical cells the cones. The photoreceptor cells of the human retina are roughly four times as thin.

Outside the cell the concentration of sodium ions is high and of potassium ions low; inside the cell the concentration of potassium ions is high and of sodium ions low. The concentration differences are maintained by the action of a "pump" that uses metabolic energy to extrude sodium and draw in potassium.

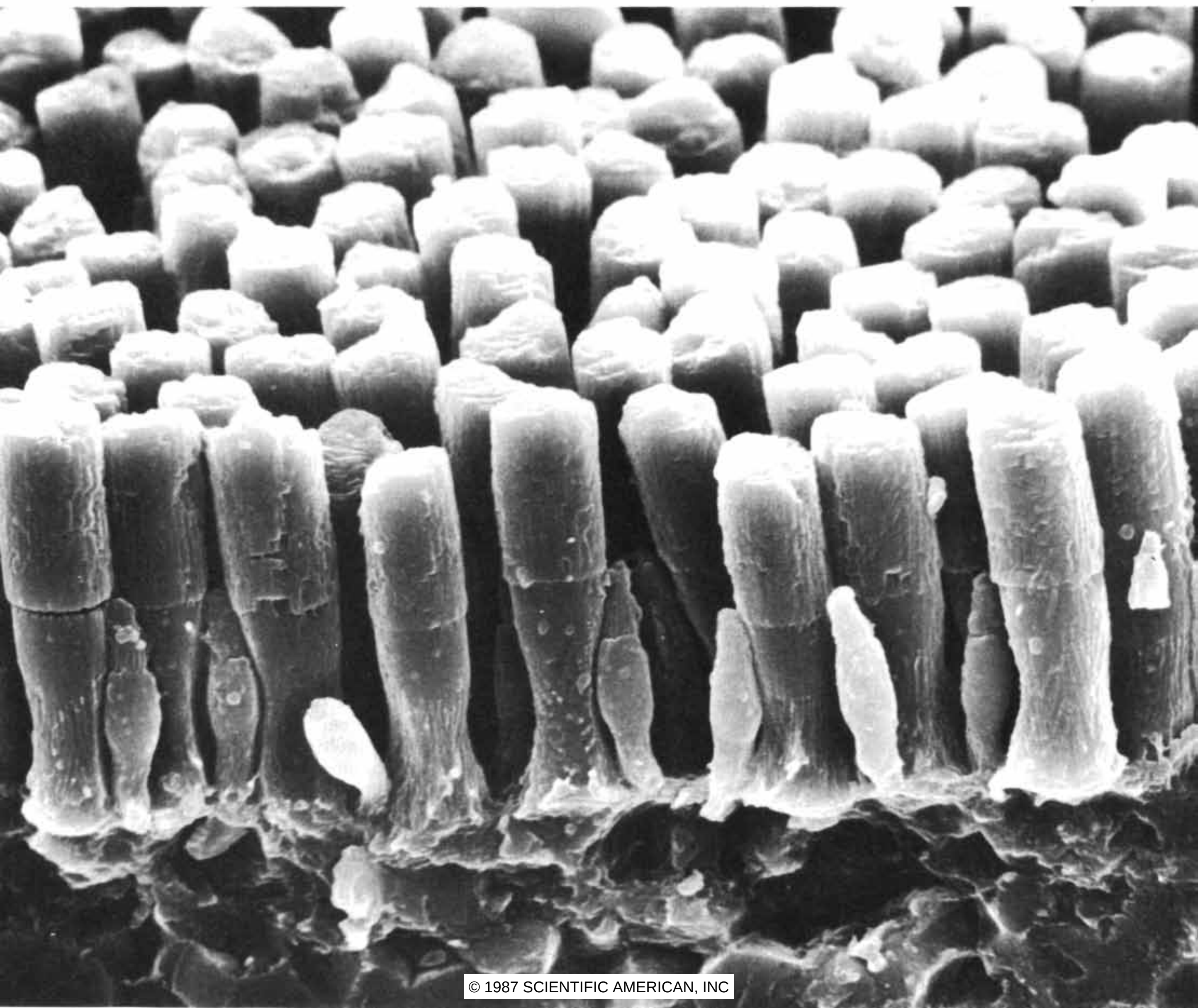
In the resting state the membranes of most neurons allow potassium ions to cross them more freely than other ions. Because potassium ions are more concentrated inside the cell, they tend to diffuse across the membrane to the outside. As the diffusion proceeds, charge is moved from the inner to the outer surface of the membrane. The transfer of charge causes the internal potential to become negative with respect to the outside, typically by as much as .1 volt. In a photoreceptor the permeability to the potassium ions is highest at the inner segment and synaptic ending.

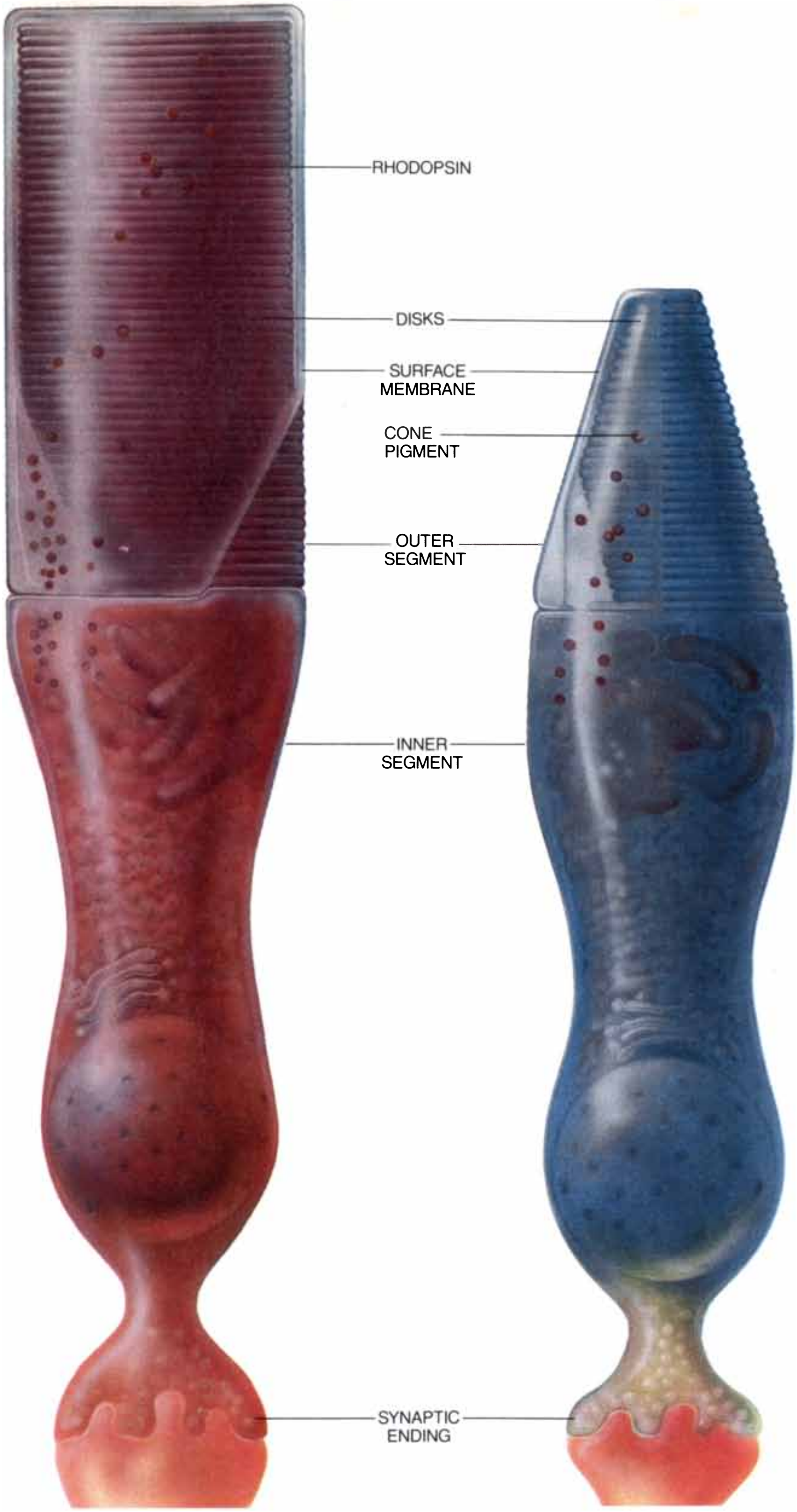
In darkness a photoreceptor also has an appreciable permeability to sodium ions. The sodium ions flow from the more concentrated external solution into the outer segment, carrying an inwardly directed electric current. The inward current is balanced by the outward current of potassium ions from the rest of the cell. The loop of current is called the dark current.

When a rod or a cone absorbs light, the influx of sodium is blocked. This reduces the dark current and allows the negative polarization of the cell interior to increase. The negative swing of the transmembrane voltage is called a hyperpolarization and the reduction in dark current is known as a photocurrent. The dark current and photocurrent were first described in about 1970 by William A. Hagins, Richard D. Penn and Shuko Yoshikami, then at the National Institute of Arthritis and Metabolic Diseases. They measured the currents around a large population of rods.

The light-evoked hyperpolarization is generated at the outer segment but spreads to the synaptic ending, where it is communicated to other retinal cells. The hyperpolarization can be recorded with a microelectrode placed inside one of the relatively large rods or cones found in the retina of certain fishes, amphibians and reptiles. Experiments of this type were pioneered in the mid-1960's by Tsuneo Tomita and his colleagues at Keio University in Japan. These recordings show that the transmembrane voltage is about −40 millivolts (mV) in the dark. A flash of light causes a hyperpolarization that increases with the strength of the flash. After a very bright flash the response reaches a limiting size of about 30 mV, at which the membrane voltage is −70 mV.

In 1975 Robert Fettiplace and one of us (Baylor) at the Stanford Univer-





RODS AND CONES differ in both form and function but have certain similarities. The upper part of the cells, which is called the outer segment, contains light-absorbing pigment molecules; the lower part, the inner segment, contains mitochondria and the nucleus. The synaptic ending links the photoreceptors to other retinal cells. Rods mediate vision in dim light; cones mediate vision in daylight. In the human retina there are three types of cone. Each of them incorporates a pigment that absorbs strongly in the blue, green or red region of the visible spectrum, providing the foundation for color vision.

sity School of Medicine showed that the hyperpolarization is indeed necessary and sufficient for controlling the flow of information across synapses to other visual neurons. In order to simulate and prevent the light-evoked hyperpolarization, we employed an intracellular electrode to pass an electric current into a single photoreceptor. Simultaneously we monitored the responses of another cell, called a ganglion cell, farther along in the chain of retinal neurons. We successfully reproduced the response of the ganglion cell to a small spot of light applied to the photoreceptor by artificially hyperpolarizing the photoreceptor in darkness. Moreover, the ganglion cell failed to respond when we blocked the light-evoked hyperpolarization by injecting a depolarizing current.

How does the absorption of light block the influx of sodium ions at the outer segment? In the dark both rods and cones have a high concentration of cyclic guanosine monophosphate. This substance binds to pores in the surface membrane and opens them, allowing sodium ions to enter. In the light the concentration of cGMP drops, cGMP leaves the binding sites and the pores close. The permeability of the membrane to sodium atoms is thereby decreased and the membrane hyperpolarizes.

The chain of molecular events leading to the reduction of cGMP consists of three steps. George Wald and his colleagues at Harvard University showed some years ago that the pigments in both rods and cones contain a light-absorbing component called 11-*cis* retinal, coupled to a protein that "tunes" the absorption to a particular region of the visible spectrum; the proteins of the rod pigment rhodopsin and the three cone pigments are different. When the retinal in rhodopsin absorbs a photon of light, it changes configuration, causing the protein part of the molecule to become enzymatically active. As Lubert Stryer and his fellow workers at Stanford subsequently showed, the active form of rhodopsin catalytically activates many molecules of a protein Stryer and his co-workers named transducin. The activated transducin molecules in turn activate an enzyme that cleaves cGMP. The system behaves like a chemical photomultiplier. Absorption of a single photon by rhodopsin causes the rapid breakdown of hundreds of molecules of cGMP and blocks the entry of a million sodium ions.

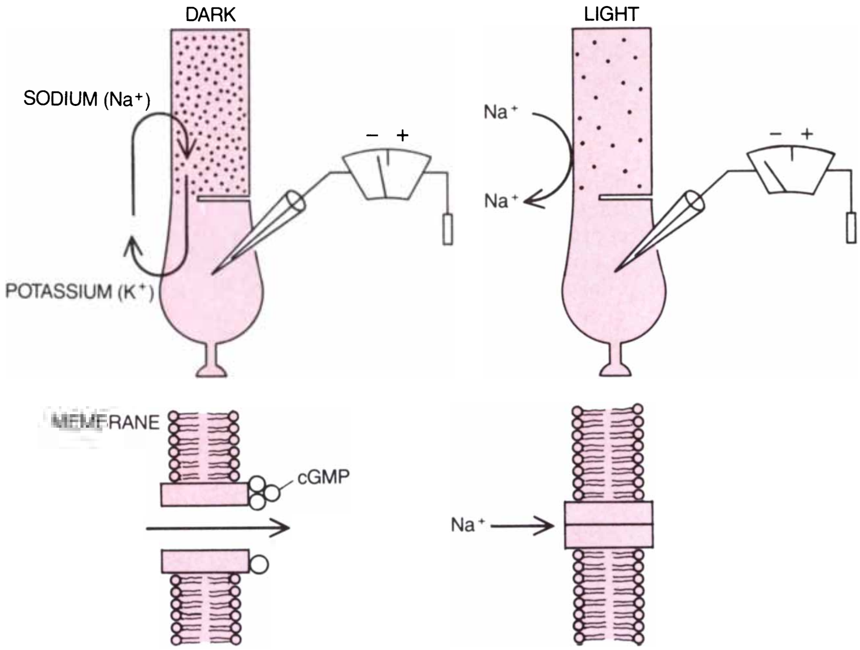
Recently strong evidence in support of the idea that cGMP does indeed control the transport of sodium ions through the surface membrane was obtained by Evgeniy Fesenko and his colleagues at the Academy of Sciences of the U.S.S.R. They touched a patch pipette—a glass capillary with a tip about a micrometer (one millionth of a meter) in diameter—to the surface membrane of the outer segment of a rod from a frog retina. By applying gentle suction to the pipette and rapidly withdrawing it, they excised the patch of membrane that adhered to the tip. They found that when they exposed the patch to cGMP, it became permeable to sodium.

The molecular mechanism of the cGMP-regulated sodium movement was not clear until Anita Zimmerman in our laboratory and Lawrence Haynes and King-Wai Yau of the University of Texas Medical Branch at Galveston succeeded in showing that the current passing through a single permeation site can exceed a million sodium ions per second. Such a flux surpasses by two orders of magnitude the transport rates of membrane carrier molecules that must undergo a configuration change to translocate ions one or a few at a time. The large flux shows instead that ions cross the membrane by diffusing through water-filled pores. The opening of an individual pore appears to be triggered by the cooperative binding of three or more molecules of cGMP. In other words, each pore behaves like an efficient molecular switch designed to detect infinitesimal changes in the concentration of cGMP.

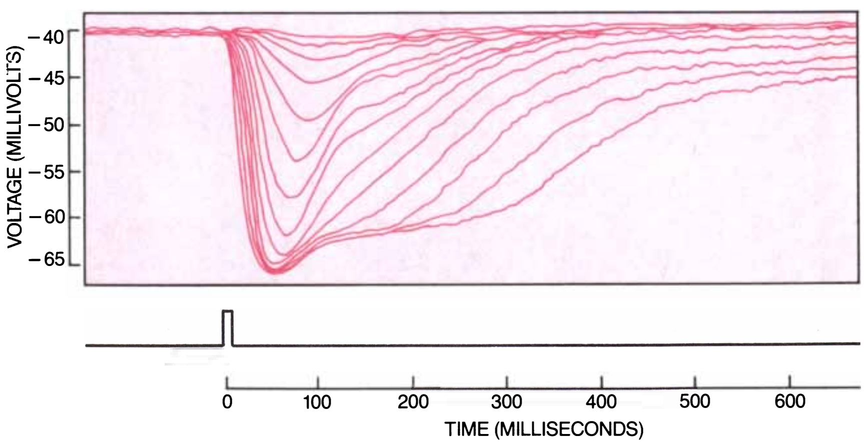
Under appropriate conditions a rod in the human retina signals the absorption of a single photon, which activates only one of the 100 million rhodopsin molecules in the rod. This remarkable performance was first demonstrated in psychophysical experiments done in the early 1940's by Selig Hecht, Simon Shlaer and Maurice H. Pirenne of Columbia University. They directed dim flashes of light into one eye of a subject who was sitting in complete darkness. By varying the strength of the flash they found that the subject usually perceived a flash when only seven photons were absorbed. Because a population of 500 rods absorbed the photons in a random spatial pattern, there was virtually no chance that any rod had absorbed more than one photon. The investigators therefore concluded that a rod must produce a detectable signal when it absorbs a single photon.

What are the amplitude and form of the electrical signal that is triggered by the absorption of a photon? How likely is it to occur after an absorption occurs? Do similar signals occur in darkness? In the early 1970's workers in several laboratories attempted to record the quantal voltage response of rods. The initial efforts failed. The reason is that rods "pool" their signals: specialized connections called gap junctions link neighboring rods in the retina and allow electric currents to flow freely among their interiors. As a consequence the hyperpolarizing response to a single photon is distributed to 10 or more rods, making it too small to detect.

In order to overcome the problem created by pooling, Yau, Trevor D. Lamb and one of us (Baylor) decided to use a different indicator of the rod



ELECTRICAL RESPONSE TO LIGHT of a rod or cone results from a reduction in the outer-segment surface membrane's permeability to sodium ions. In darkness the sodium ions, which carry a positive charge, flow into the cell and steadily reduce the negative charge density on the inside of the cell membrane (*top left*). An outward flow of potassium ions through the inner segment and the synaptic ending completes a continuous loop of "dark current." The high sodium permeability is maintained by the action of the nucleotide cyclic guanosine monophosphate (cGMP), whose concentration in darkness is high (*stippling*). In darkness several cGMP molecules bind to a pore and cause it to open (*bottom left*). In light the concentration of cGMP drops, the nucleotide leaves the binding sites and the pore closes (*bottom right*). The influx of sodium ions is thereby blocked and the internal voltage of the cell hyperpolarizes: it becomes more negative (*top right*).



HYPERPOLARIZING VOLTAGE RESPONSES from a red cone in a turtle retina were recorded with an intracellular electrode. The traces are superposed responses to brief flashes of increasing strength. The voltage difference across the membrane is plotted as a function of the elapsed time after the flash, which is shown in the lower trace. The strengths of the flashes were increased by factors of two; the weakest flash activated about 50 molecules of the light-absorbing pigment in the cone. Bright flashes caused the response amplitude to saturate, the membrane potential reaching about −65 millivolts.

response. We measured a rod's photocurrent rather than its voltage. Our choice proved to be a good one, because the photocurrent is effectively independent of membrane voltage and is therefore not influenced by coupling among rods.

To identify the response to a single photon, we observe a rod's response to very dim light. At first we worked with rods from toad retinas; more recently the two of us, in collaboration with Brian J. Nunn of Stanford, have worked with rods and cones from retinas of the macaque monkey (*Macaca fascicularis*). To make the measurements we draw an individual outer segment into a close-fitting glass capillary tube. We then record the rod's photocurrent with a sensitive amplifier connected to the capillary.

We repeatedly shine on the rod a flash so weak that it activates on the average one molecule of rhodopsin. The resulting photocurrent varies, assuming values near zero, one, two and three picoamperes (trillionths of an ampere). Such variation is expected, because the emission of photons from the source fluctuates randomly: sometimes the flash fails to activate a rhodopsin molecule and at other times it activates one, two or three molecules. Statistical analysis and calibration of the flash strength show that the one-picoampere response is triggered by the activation of a single rhodopsin molecule. The size and shape of the response are remarkably constant, suggesting that the gain of the enzyme cascade is subject to an elegant control. Measurements also show that there is a good chance—about 50 percent—that a photon will trigger a response when it is absorbed.

Photon counting by rods is impressive but not quite perfect. Even in complete darkness rods give an occasional signal identical with that triggered by the absorption of a photon. In a rod from a monkey retina, for example, signals appear randomly at a rate of one every two and a half minutes on the average. The signals seem to arise because thermal energy, or heat, can activate a rhodopsin molecule just as light can. This process sets the ultimate limit on a rod's ability to reliably encode very dim light. Fortunately, however, thermal activation proceeds quite slowly: from the frequency of the signals and the number of rhodopsin molecules in the rod we find that the half-life of the process is 420 years at body temperature. (Only because a rod is packed with rhodopsin molecules can the error signals be studied experimentally.) Nevertheless, the signals are perceived by the visual system and give rise to sensations of very dim light in complete darkness. Psychophysicists have quantified such "dark light" and find an activation rate similar to what we have measured.

The response of a cone to a single photon cannot be measured because it is too small. Background fluctuations overwhelm it. The quantal response can nonetheless be estimated from the response of a cone to flashes that activate many of its pigment molecules. We estimate that in a cone one absorbed photon produces a photocurrent approximately 10 femtoamperes (10×10^{-15} ampere) in size. This is about 100 times smaller than a rod's quantal response. The characteristic difference in the response sizes helps to explain why human cone-mediated daylight vision is less sensitive than rod-mediated night vision.

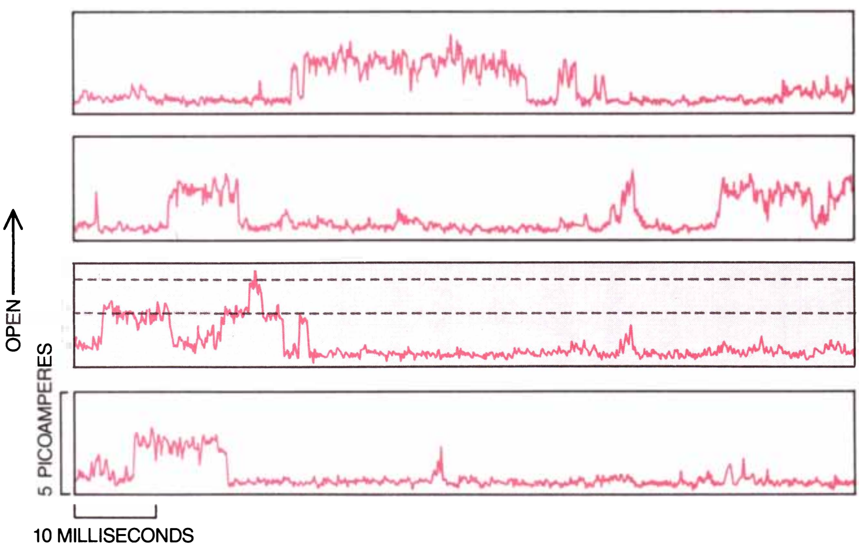
On the other hand, a cone's quantal response is roughly four times as fast as that of a rod. A primate rod, for instance, takes 300 milliseconds to finish signaling the absorption of a photon. In that time a pitched baseball travels most of the way to home plate. Because of their greater response speed, cones are better at encoding rapidly changing visual stimuli.

Visual transduction appears, then, to involve a tradeoff between sensitivity and temporal resolution. The small, fast quantal responses of cones enable the visual system to detect rapid changes in intensity or rapid movement of objects when the level of illumination is high and the rods are saturated. The slower and larger rod signals, on the other hand, are optimal for counting photons when the level of illumination is low.

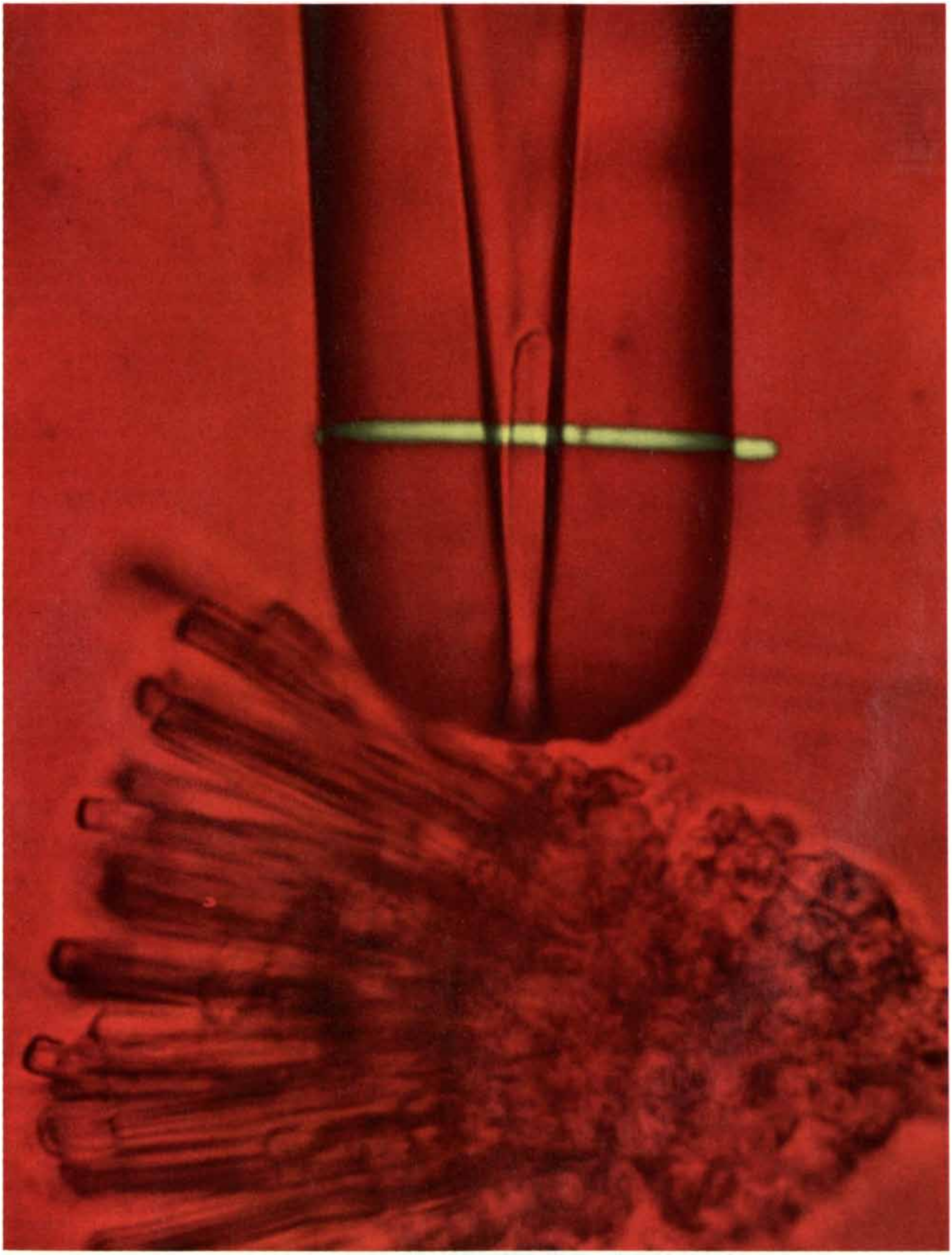
A striking increase in visual sensitivity occurs at low levels of illumination because of a switch from cone vision to rod vision. On entering a dimly lighted room, for example, we are initially blind because of the insensitivity of the cone system. Slowly the rod system becomes more sensitive, and as it assumes the primary role objects become visible. Even in pure rod vision, however, visual sensitivity rises as the level of background light falls. Does this change in sensitivity take place within the rods or in other neurons that process the rod signals?

The effect of background light on the sensitivity of a primate rod can be determined by recording responses to dim flashes from a rod that has been adapted to darkness. The peak amplitude of the response is divided by the flash strength to yield a measure called the flash sensitivity. Steady background lights are then turned on and the sensitivity is again determined. The sensitivity drops as the intensity of the background lights increases. The desensitization is accounted for by a simple saturation mechanism. As the background intensity rises, more of the sodium channels in the surface membrane close, making fewer channels available to be closed by the flash.

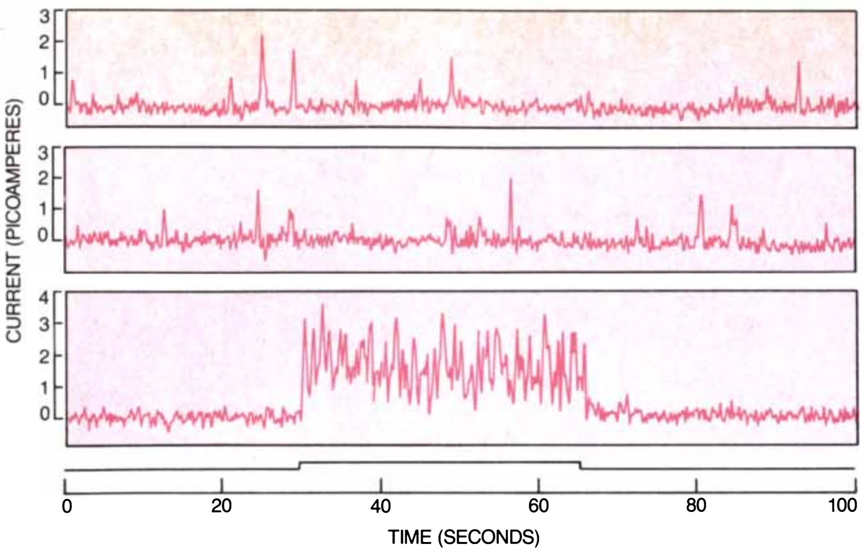
In psychophysical experiments the sensitivity of the human rod system has been measured by determining the strength of a flash that is barely de-



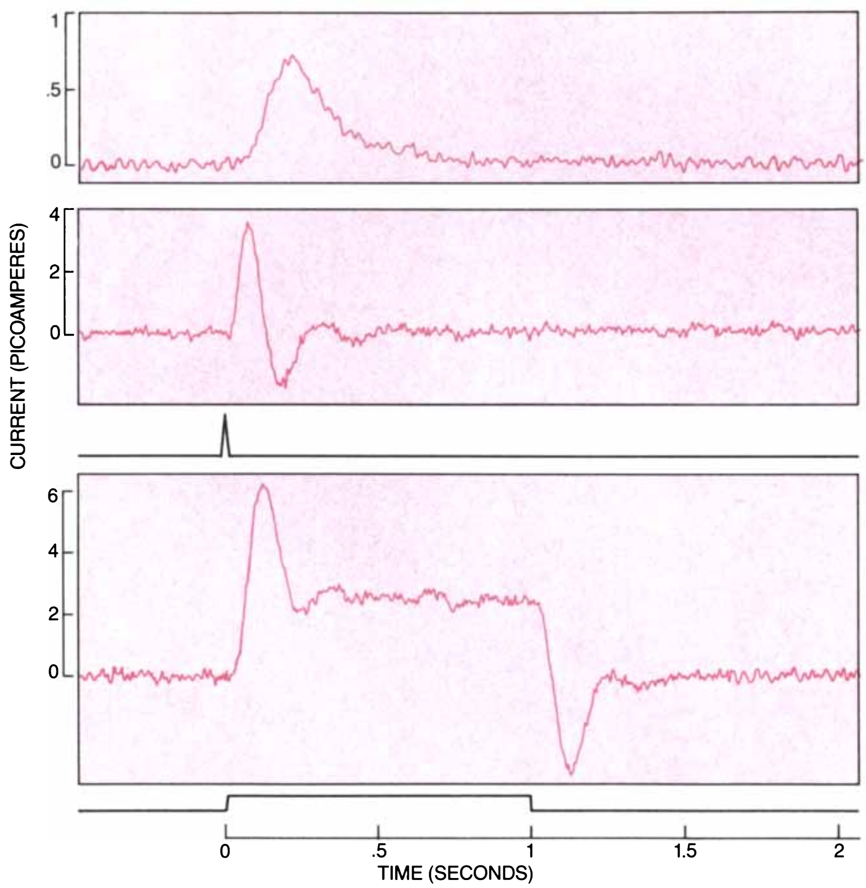
MEASURABLE ELECTRIC CURRENT results when a single pore in the outer segment of a rod opens. Here a patch of the membrane of a salamander rod was exposed to a solution containing cGMP while the membrane voltage was held at +75 millivolts. Upward deflections in the traces correspond to the opening of a single pore. In the third trace from the top two pores open simultaneously. In order to improve the resolution of the measurement the concentrations of calcium and magnesium ions were made very low.



SINGLE-ROD OUTER SEGMENT of a toad is drawn into a suction pipette for recording electric currents induced by light. The outer segment in the pipette is about 50 micrometers (millionths of a meter) long and six micrometers in diameter. The rod is stimulated by the transverse slit of light, and the electrical response is amplified and recorded. To preserve the adaptation of the cell to the dark, the work is done under infrared light and is viewed on a video monitor. The piece of retina is from the toad *Bufo marinus.*



RESPONSE OF MONKEY ROD to a single photon is monitored by drawing the rod into a suction pipette. In the upper traces dim flashes activating on the average one pigment molecule were delivered to the outer segment. The response of the rod fluctuated, its amplitude ranging from zero to one or two picoamperes (trillionths of an ampere). The activation of a single pigment molecule triggers a response of about one picoampere; the fluctuations in amplitude are caused by random variations in the emission of photons from the light source. The lower trace shows the response of the same rod to a dim, steady light, which activated on the average about 10 pigment molecules per second.



WAVEFORM OF SINGLE-PHOTON EFFECT in a rod and cone from the macaque monkey retina is revealed by averaging responses to dim flashes. The change in current through the membrane is plotted as a function of elapsed time after the flash. The response of the rod (*top*) was elicited by a flash that activated an average of one pigment molecule; the response of the cone (*middle*) was triggered by a flash that activated about 200 molecules. (The response is a scaled-up version of the response to a single photon.) The bottom trace shows the response of the cone to a pulse of light one second long.

tectable in the presence of a diffuse background light. The sensitivity to the flash decreases as the background intensity is increased. A background that strongly desensitizes rod vision, however, may have little effect on the measured sensitivity of an individual rod. For example, a steady light activating 40 molecules of rhodopsin per second per rod causes a 10,000-fold drop in the sensitivity of rod vision but reduces the sensitivity of a rod outer segment by only about 20 percent. The measurements support the conclusion of the late William A. H. Rushton of the University of Cambridge that desensitization of human rod vision by background light is due to neuronal processing beyond the rod outer segments. The mechanism of the effect remains to be determined.

When the background intensity exceeds the level that corresponds approximately to the blue sky at noon, the flash sensitivity of human rod vision falls precipitously: it saturates. In such bright light, changes in the rate of photon absorption in the rods are not detected by the visual system. Electrical measurements show that single rods become unresponsive to a flash at about the same background intensity that saturates rod vision. This limit in vision therefore expresses a property of the rod transduction.

The sensitivity of a single photoreceptor to light of different wavelengths is determined by the probability that its visual pigment will absorb photons of those wavelengths. Measurements of the spectral sensitivity of single cells from the macaque, which is thought to have photoreceptors like those of man, provide a physiological basis for the spectral characteristics of human vision. The wavelengths of visible light lie roughly between 400 and 750 nanometers (billionths of a meter). Light of longer wavelengths (near-infrared) is poorly absorbed by the visual pigments; light of shorter wavelengths (near-ultraviolet) can be absorbed by the visual pigments but fails to reach the retina because it is absorbed in the cornea and lens.

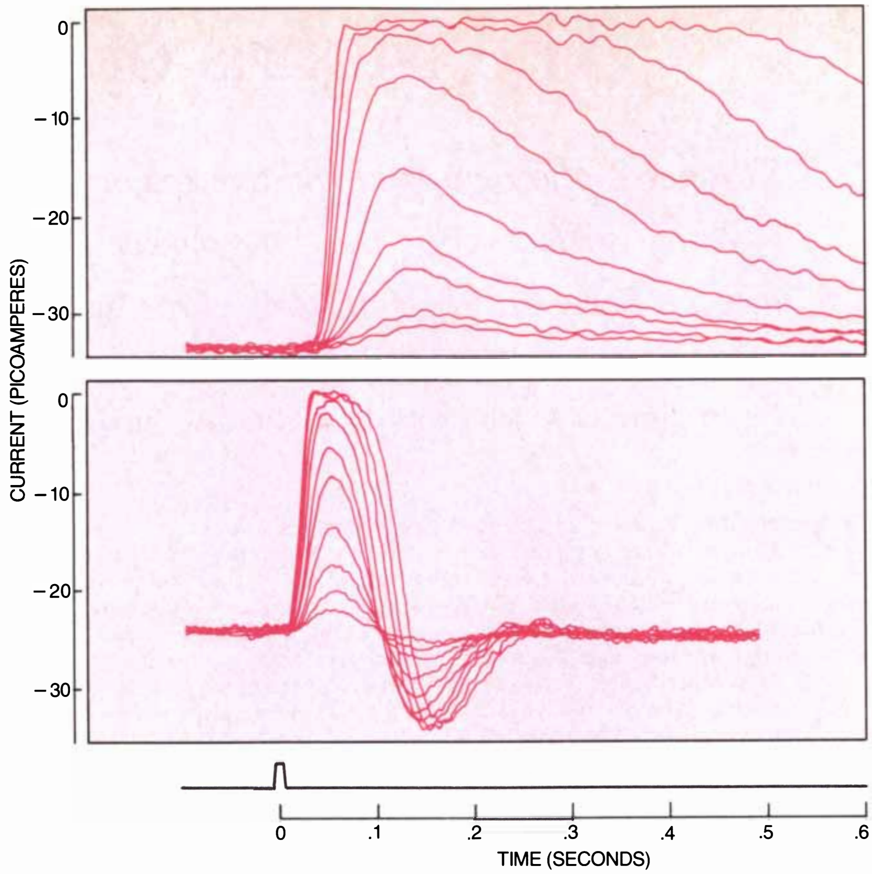
Macaque rods show a peak sensitivity in the blue-green region of the spectrum, near 490 nanometers. The measured spectral sensitivity agrees with the spectral sensitivity of human rod vision determined by psychophysical experiments.

Monkey cones fall into three groups that have peak sensitivities at roughly 430, 530 and 560 nanometers. The groups, which correspond closely to those of humans, may be called blue, green and red to indicate the relative positions of the spectral maxima. Each type of cone is sensitive to light over a broad range of wavelengths, and the sensitivities of the groups show considerable overlap. Nevertheless, the segregation of pigments into the appropriate cones appears to be quite strict. From the form of the sensitivity curves we conclude that less than one in 100,000 pigment molecules in a blue cone is of the red or green type.

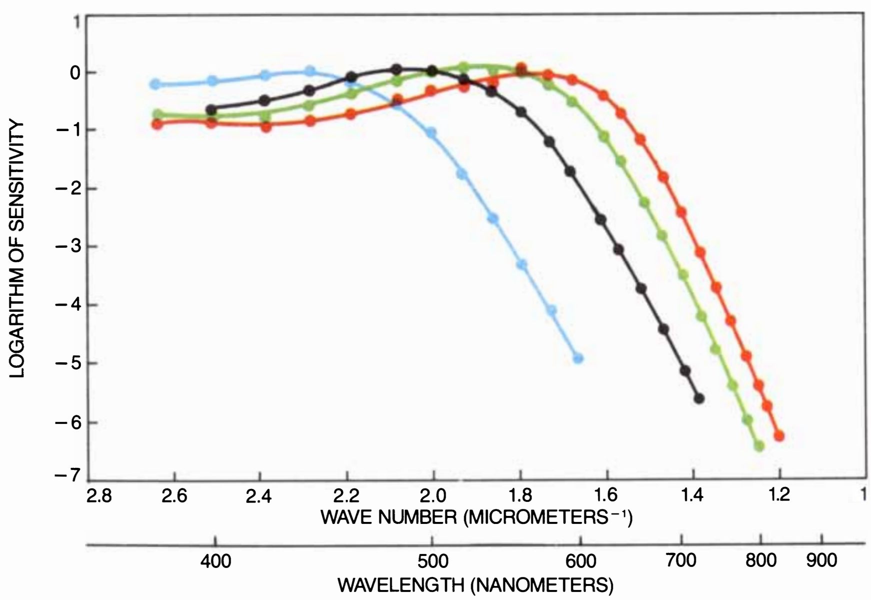
A cone's response does not depend on the wavelength of the photon that was absorbed; all stimuli that elicit identical absorptions give identical responses. By sensing the ratio of excitations in the three kinds of cone, however, the visual system is able to compute color from wavelength. It has been known for many years from psychophysical experiments that two stimuli of different wavelength composition will appear identical if both stimuli evoke, within each kind of cone, the same number of absorptions. Although the trichromacy of color vision has been well established by such observations, its exact basis has been unclear because of uncertainty about the specific cone spectral sensitivities. These sensitivities are now known for monkey cones, and it is satisfying that they predict the rules governing the light intensities a human requires to make color matches.

At long wavelengths the perception of hue is determined by the relative absorption in the red and green cones alone. As the wavelength increases beyond about 600 nanometers, the perceived hue changes from orange to a progressively deeper red. Beyond 700 nanometers a curious reversal occurs and the hue becomes more orange. This phenomenon, the "paradoxical hue shift," was discovered in 1955 by Giles S. Brindley of Cambridge. It is explained by the form of the spectral sensitivities of the red and green cones. The ratio of the red- and green-cone sensitivities has a maximum at 700 nanometers, and so this wavelength appears reddest.

The molecular mechanism of visual transduction and the central processing of photoreceptor signals are still far from being completely understood. Although the internal transmitter for visual excitation has been identified, the operation of the nucleotide cascade and the control of sodium permeability are only beginning to be characterized. Much also remains to be learned about how small signals generated by single photons are transmitted across synapses, separated from noise and processed by the visual system. The years ahead promise to be an exciting time for experiments in both areas of investigation.



MEMBRANE CURRENTS from a monkey rod (*top*) and cone (*bottom*) were recorded with a suction pipette. The outer segments were illuminated uniformly by flashes of light. The superposed recordings show the current of the outer segment as a function of time after the flash. The strength of the flashes was progressively doubled until the responses reached their maximum amplitude, and the inward current shut off completely. In the rod the response was half its maximum when 30 rhodopsin molecules were activated; in the cone the response was half its maximum when 1,200 pigment molecules were activated.



RELATIVE SENSITIVITY TO A PHOTON for rods and cones of the macaque monkey is plotted against the wavelength of the photon. The spectral sensitivities are quite similar to those of human receptors. The black curve is the spectrum of the rods, and the red, green and blue curves are respectively the spectra of the red, green and blue cones.

The Moons of Uranus

Voyager 2 photographed the five major moons at close range. All have icy surfaces, but they are darker and rockier than Saturn's moons. Early in their history three were geologically vigorous

by Torrence V. Johnson, Robert Hamilton Brown and Laurence A. Soderblom

Picture this: You are an observer at Uranus, where you have been in orbit for millions of years. It has been an uneventful epoch. Nothing much changes in your gray world; everything froze long ago. Then, at a time like any other, you spot a point of light streaking toward you out of the sun. You investigate. What you find is almost comical: a small pile of metal, a dish with several awkward, randomly gyrating protuberances, the whole contraption hurtling along at a frantic pace. In what seems to you an instant the interlude is over, and you are alone again in the serene regularity established by the planet, its circular rings and its icy, cratered moons.

Now move to California, specifically to the Jet Propulsion Laboratory: You are one of us, one of the people who launched that metal contraption, called *Voyager 2,* on its mission. Even from our comparatively limited perspective (Uranus was discovered by William Herschel only two centuries ago), the *Voyager* flyby on January 24, 1986, seemed brief. That was particularly true for those of us interested in the planet's five major satellites, which are barely visible from the earth even through large telescopes. When *Voyager 2* shot by Uranus at 72,000 kilometers per hour, we had less than a day to collect essentially all the detailed information we have ever had on the moons—and possibly ever will have: at present no further missions to Uranus are planned.

The technical obstacles were formidable. Originally intended to operate only as far as Saturn, where sunlight is four times as intense and where the moons are on the average about twice as reflective, *Voyager 2* had to contend at Uranus with targets much dimmer than those for which its cameras had been designed. To gather enough light for an image the camera shutters had to be left open for several seconds. Unless some way could be found to correct for the motions of the spacecraft and of the moons themselves, the pictures would be smeared to near uselessness. Imagine trying to photograph a speeding, charcoal-gray race car on an overcast day with very slow film and you will get an idea of the challenge that confronted *Voyager 2* and its handlers.

The challenge was fully met. While the spacecraft was proceeding from Saturn to Uranus engineers at the J.P.L. devised a method of compensating for its complex motions [see "Engineering *Voyager 2*'s Encounter with Uranus," by Richard P. Laeser, William I. McLaughlin and Donna M. Wolff; SCIENTIFIC AMERICAN, November, 1986]. The fruits of their efforts were high-quality, unsmeared images of all five major Uranian satellites: Oberon, Titania, Umbriel, Ariel and Miranda. In the case of Miranda, the innermost of the five, the images had a higher resolution than any made by *Voyager 1* and *Voyager 2* at Jupiter and Saturn. Given these results, *Voyager 2* does not appear comical or awkward to us; as a scientific experiment, it looks downright elegant.

The brevity of the Uranus encounter arose not only from the speed of the spacecraft but also from the unusual geometry of the Uranian system, which is tipped on its side. At present the planet's south pole is pointing toward the sun and the earth. Since the moons orbit roughly in the equatorial plane, *Voyager 2*'s path through the system resembled that of a bullet piercing a bull's-eye target. Instead of passing close to individual moons sequentially, as it did at Jupiter and Saturn, the probe made its closest approach to all the Uranian moons at about the same time. To ensure that each moon received high-resolution photographic coverage, the imaging sequence had to be planned to the second. Even with careful planning, though, no more than half of each moon could be photographed, because currently only the southern hemispheres are sunlit.

In the months before the closest approach a lot of camera time was devoted to scanning the planet's equatorial plane for new satellites. There was good reason to suspect their existence: Jupiter has 16 moons and Saturn 17, most of them tiny objects not visible from the earth. *Voyager 2* found 10 new satellites at Uranus, bringing that planet's total to 15. Two of the new moons are gravitational "shepherds" that patrol the inner and outer edges of the planet's largest, outermost ring. The other eight follow circular orbits between the rings and Miranda. All but one of the new satellites have diameters between 40 and 80 kilometers. The exception, designated 1985U1 pending the assignment of an official name by the International Astronomical Union, is 160 kilometers across.

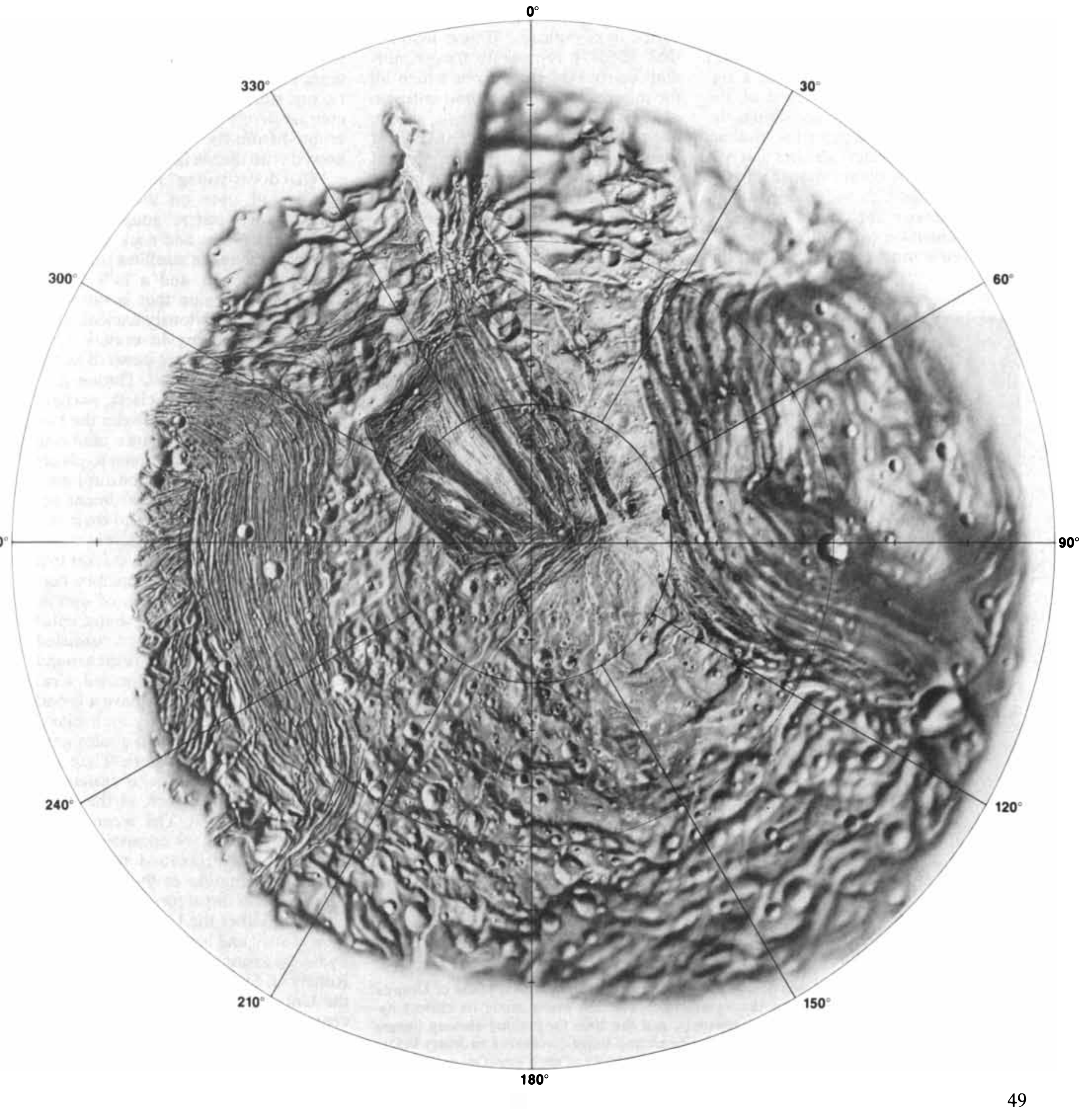
Because 1985U1 was detected more than a month before the encounter and because it happened to be on the same side of the planet as Miranda, we and

MAP OF MIRANDA based on *Voyager 2* images is a stereographic projection of the moon's southern hemisphere; the south pole is at the center and the equator is the outer ring. Only the southern hemispheres of the Uranian moons were photographed because only they are sunlit. Miranda is the smallest and most exotic of the five major moons. Superposed on its ancient cratered plains are three younger, less cratered regions of grooved and ridged terrain known as the "ovoids." The moon's surface is also crisscrossed by fracture zones; the one at the top of the map includes a cliff that is between 10 and 20 kilometers high. The Miranda map and the other maps in this article were prepared by workers at the U.S. Geological Survey in Flagstaff under the direction of R. M. Batson.

our colleagues were able to trade in a camera-time slot originally allocated to Miranda and make a closeup picture of the new moon. The image shows a somewhat irregular object with a dark, cratered surface that reflects only 7 percent of the incident sunlight. The other new satellites were not observed at close range, but they appear to be equally dark, as do the particles that make up the rings.

In that respect they are far from distinctive. Actually most of the solid material in the outer solar system that is not bright ice is very dark, with an albedo, or reflectance, of about 10 percent or less. The nature of this dark material has been the subject of considerable debate in recent years. Most investigators agree that it is dark because it is rich in carbon, like soot. The most obvious explanation is that it is essentially the same stuff as is found in carbonaceous chondrites: primitive meteorites that come from the asteroid belt between Mars and Jupiter and consist of hydrated silicates, or clays, mixed with dark, carbon-rich organic substances. Such mixtures form only at relatively low temperatures. Theoretical models suggest that conditions in the early solar nebula would have been appropriate for the formation of carbonaceous rocks from the asteroid belt outward.

Alternatively, the carbonaceous material might not be primordial rock; it might have been produced more recently, by the irradiation of methane ice or methane-contaminated water ice, which are also predicted to be present in the outer solar system. Laboratory studies have shown that ultraviolet light or energetic electrons (such as those in the solar wind or in plan-



etary magnetospheres) can induce reactions in which the methane forms a dark residue of complex organic compounds. This mechanism has been variously invoked to explain the production of dark interstellar grains, the low reflectance of cometary nuclei and the peculiar dark side of Saturn's moon Iapetus. Indeed, some workers think it accounts for the organic material in carbonaceous rocks. If they are right, then the question about the darkening of methane in the outer solar system is not so much whether it has occurred, but when: whether it was important only four and a half billion years ago, when the carbonaceous rocks formed, or whether it also occurred later on the surface of satellites.

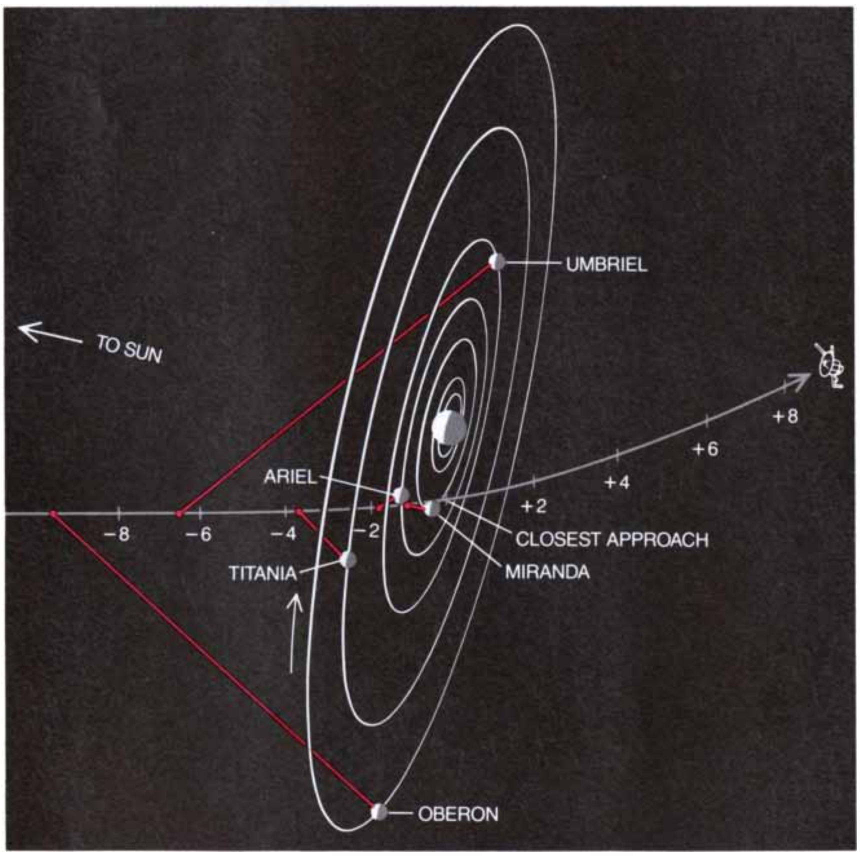
So far there is no compelling reason to believe the process has had a significant effect on the surface of the Uranian satellites. For one thing, the presence of methane on the moons, unlike that of water ice, has not yet been verified by observations from the earth. (*Voyager 2* is not equipped with a spectrometer capable of analyzing the composition of satellite surfaces.) In our view most of the dark material in the Uranian system is probably primordial carbonaceous rock, although darkened methane may have been added to the original mixture in a few locations. Experiments done by Roger N. Clark, then at the University of Hawaii at Manoa, indicate that small quantities of carbonaceous rock (less than 1 percent by volume), when they are mixed thoroughly with water ice, are enough to darken the bright ice, yielding a reflectance as low as the 7 percent level exhibited by 1985U1.

The five major Uranian moons are not nearly as dark as 1985U1; their albedos range from 20 percent (Umbriel) to 40 percent (Ariel). The observation is significant. If one assumes that 1985U1 represents the primordial ice-rock mixture from which all the moons were formed, one must explain why the larger moons should be brighter. The answer is that 1985U1 is probably a homogeneous mixture of ice and rock, whereas the major moons are not. Instead they have undergone varying degrees of differentiation: some of the dense rock has settled toward the center of the moon and some of the lighter, brighter ice has risen to the surface. The differentiation process requires internal heat, and the heat sources in all the Uranian moons probably dwindled to insignificance long ago. But judging from the *Voyager* pictures, all five moons—notably Ariel and Miranda—seem to have been geologically active early in their history.

Evidence of geologic activity on an icy satellite can take several forms. If extensive differentiation has taken place, one may see large patterns of bright and dark markings on the surface corresponding to regions where either less rock or more rock is mixed with ice. If the moon's crust has been disrupted by tectonic movements, it should have visible fault lines. And if its internal heat source has been intense enough to drive volcanic activity, one should see regions covered by comparatively young material that has erupted from the moon's interior and flowed onto the surface.

What does "young" mean? The only markers of time on the surface of satellites are craters gouged out by the chunks of ice and rock that have rained down on the satellites since they first formed four and a half billion years ago. A region that is saturated with craters is obviously ancient; a region totally lacking in craters must have had its surface re-covered in the geologically recent past. Dating geologic features more precisely, particularly if their ages fall between the two extremes, is tricky, because cratering rates have varied from planet to planet and have not remained constant over time. The heaviest bombardment occurred while the planets and their satellites were still accreting debris.

During this early period at least two distinct populations of impactors, corresponding to two sources of debris, struck the satellites of the outer solar system. The first population consisted of material remaining in orbit around the sun after the planets formed. Craters from this population have a broad size distribution; typically their diameters range from less than a kilometer to hundreds of kilometers. They are found throughout the solar system on the oldest terrains, such as the highlands of our moon. The second population of impactors consisted of debris left in orbit around the planets after the formation of their satellites. Most of these impactors were relatively small. Before the Uranus encounter their craters had been detected primarily on the moons of Saturn, notably by Robert G. Strom and his colleagues at the University of Arizona. The heliocentric and planetocentric bombardments overlapped, but the planetocentric bombardment seems to have end-



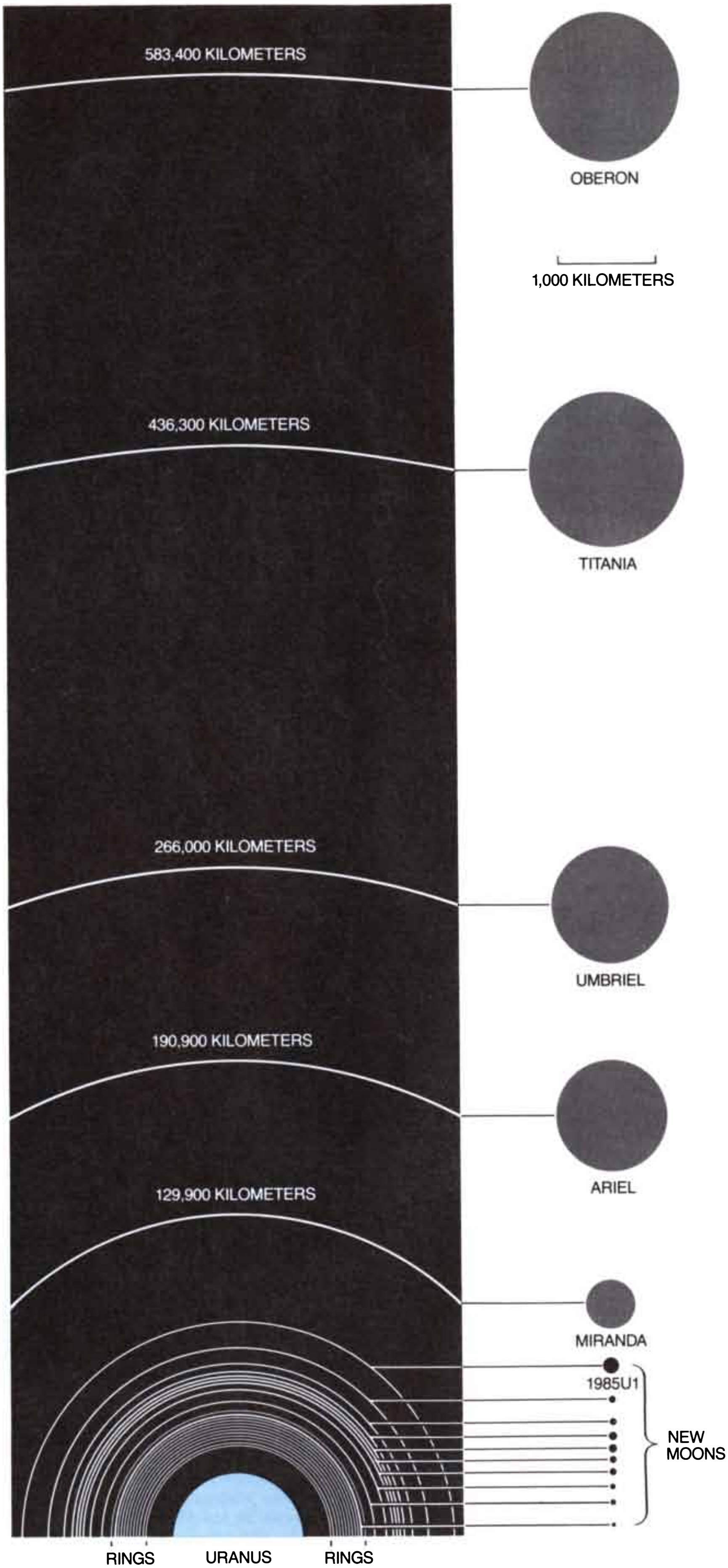
TRAJECTORY of *Voyager 2* was nearly perpendicular to the equatorial plane of Uranus. Since the moons orbit roughly in the equatorial plane, the probe made its closest approach to all the moons almost simultaneously, and the time for making closeup images was short. The colored lines indicate the positions and times (measured in hours before the closest approach to the planet) at which the best images of each moon were obtained.

ed somewhat later. This fact is a helpful clue to the early history of the Uranian moons.

Oberon and Titania, the outermost and largest of the moons, are both about 1,600 kilometers in diameter (less than half the size of the earth's moon), and they have roughly the same mass. Their surfaces, which are rich in water ice, reflect between 25 and 30 percent of the incident sunlight and are a fairly uniform gray. The gray is interrupted only by bright rays around a few craters; neither satellite displays global patterns of bright and dark markings. In spite of these similarities the two moons clearly evolved differently.

Oberon has a few features that look a bit like faults, but it shows little evidence of major tectonic disruption. Its surface is nearly saturated with large, old craters, which range in size from the limit of resolution of the images (about 12 kilometers) to more than 100 kilometers. The fact that the large craters have not been obliterated indicates that Oberon's surface has not been re-covered by fresh material since the heliocentric bombardment ended some four billion years ago.

That does not mean the satellite was completely inactive. The floor of a few of the largest craters is covered with patches of very dark material similar to that seen on 1985U1. The appearance of the craters is reminiscent of craters with dark floors, such as Tsiolkovsky, in the highlands of the earth's



URANIAN SYSTEM includes at least 11 rings and 15 moons. The orbits are measured from the center of Uranus, which has a radius of 25,600 kilometers; the moons are drawn to a common scale. The five major moons were known before the *Voyager* flyby. Of the 10 small moons discovered by *Voyager 2,* only the largest, 1985U1, was photographed at close range (*above*). Its surface is darker than that of the major moons. The large crater on its right limb is approximately 40 kilometers in diameter.

moon. On the moon the dark deposits are thought to have been laid down by volcanic eruptions following the enormous impacts that dug out the craters. Oberon too may have seen some early, localized volcanic activity. Its dark patches may consist of a mixture of ice and carbonaceous rock erupted from the interior. Alternatively, the volcanic flows could have included methane-containing ice that was darkened after it reached the surface.

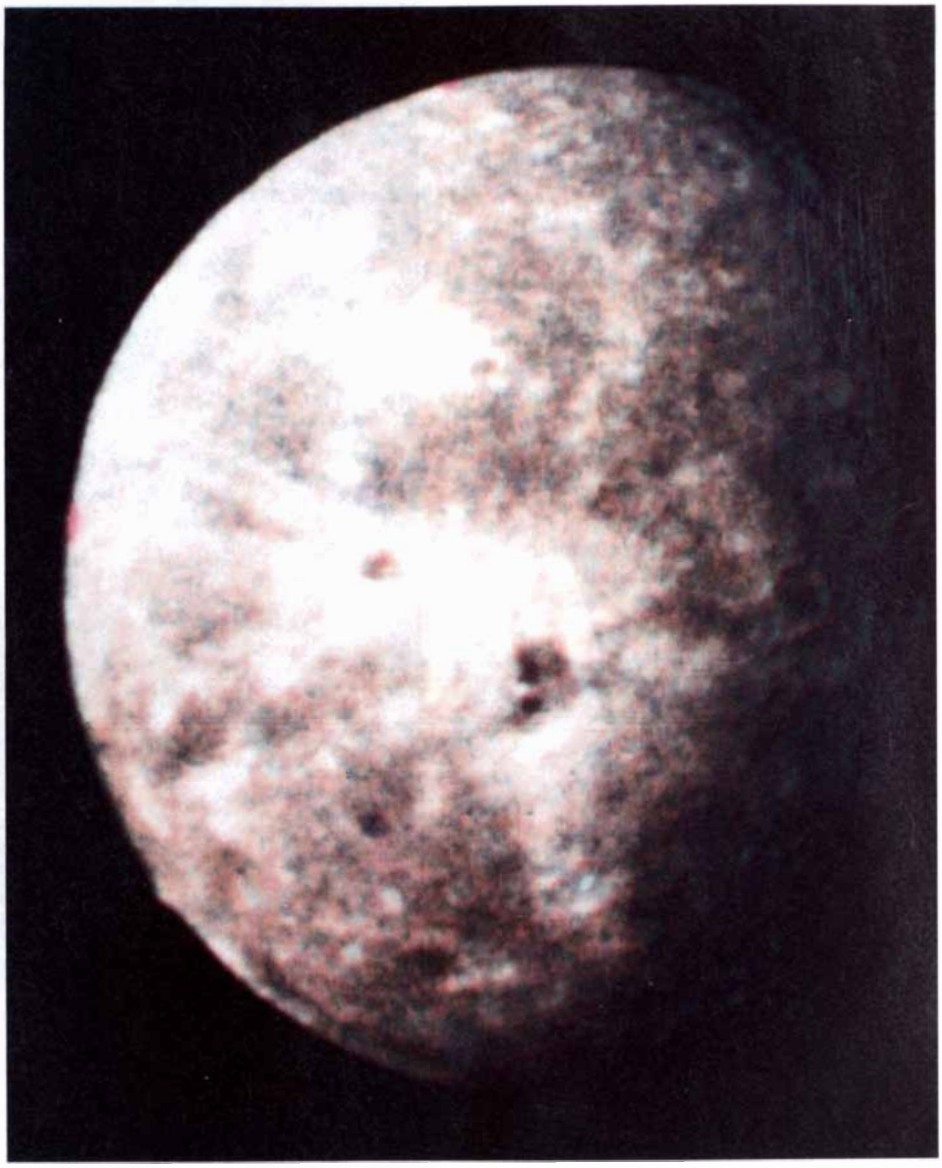
Whereas Oberon has been largely a passive target for incoming projectiles, Titania has not. Its surface bears dramatic evidence of global tectonics: a complex set of grabens, or rift valleys, bounded by extensional faults. Furthermore, although Titania's surface is heavily cratered, the evidence suggests that the craters were formed by impactors from the second, planetocentric population. Titania must have endured at least as intense a bombardment from heliocentric impactors as Oberon, and yet it has only a few large craters attributable to that population. Evidently most of the large craters were erased by some resurfacing process. Moreover, since the crater density is not uniform—there are several plains that are distinctly smoother and less cratered than the rest of the surface—the resurfacing must have lasted for a considerable time.

According to one model of Titania's evolution, the resurfacing took the form of extensive volcanic extrusion of material onto the surface. The volcanism began while the heliocentric bombardment was under way. Large craters disappeared because they were flooded or simply because the icy crust around them, which would still have been relatively warm and soft, collapsed. As the moon radiated away its internal heat, it began to freeze from the outside in. Eventually all the liquid, primarily water, froze in the interior. The water did what water always does when it freezes under low pressure: it expanded, and because there was so much of it, the entire surface of the moon was stretched. The crust ruptured along a network of extensional faults, and blocks of crust dropped down along the faults, forming the enormous grabens. The tectonic motions may have been accompanied by further extrusions of fluids; these late extrusions could have produced the smooth plains.

By the time most of the current surface had been emplaced the first bombardment had ended, and only smaller, planetocentric impactors continued to rain down on Titania. Finally, when the space around Uranus had been swept clean of debris, that bombardment also stopped. For the past three billion years or so Titania has been a quiet body, disturbed only by the impact of an occasional comet.

Eugene M. Shoemaker of the U.S. Geological Survey, an expert on cratering rates, has proposed a more drastic type of resurfacing. He thinks the early cratering of Titania by heliocentric objects was so intense that a large, late-arriving impactor could have blasted the moon apart. The shattered pieces would have remained confined to Titania's orbit and relatively soon would have reassembled into a moon. The reassembled moon would have had a new surface, one that bore no trace of the early bombardment. One virtue of Shoemaker's hypothesis is that it can explain why Oberon and Titania have similar bulk properties and yet look so different: because it is closer to Uranus, whose gravitational field must have focused and accelerated incoming debris, Titania was more likely to be shattered.

Umbriel and Ariel, the next two Uranian moons, are another interesting pair. They have nearly the same diameter (1,190 kilometers and



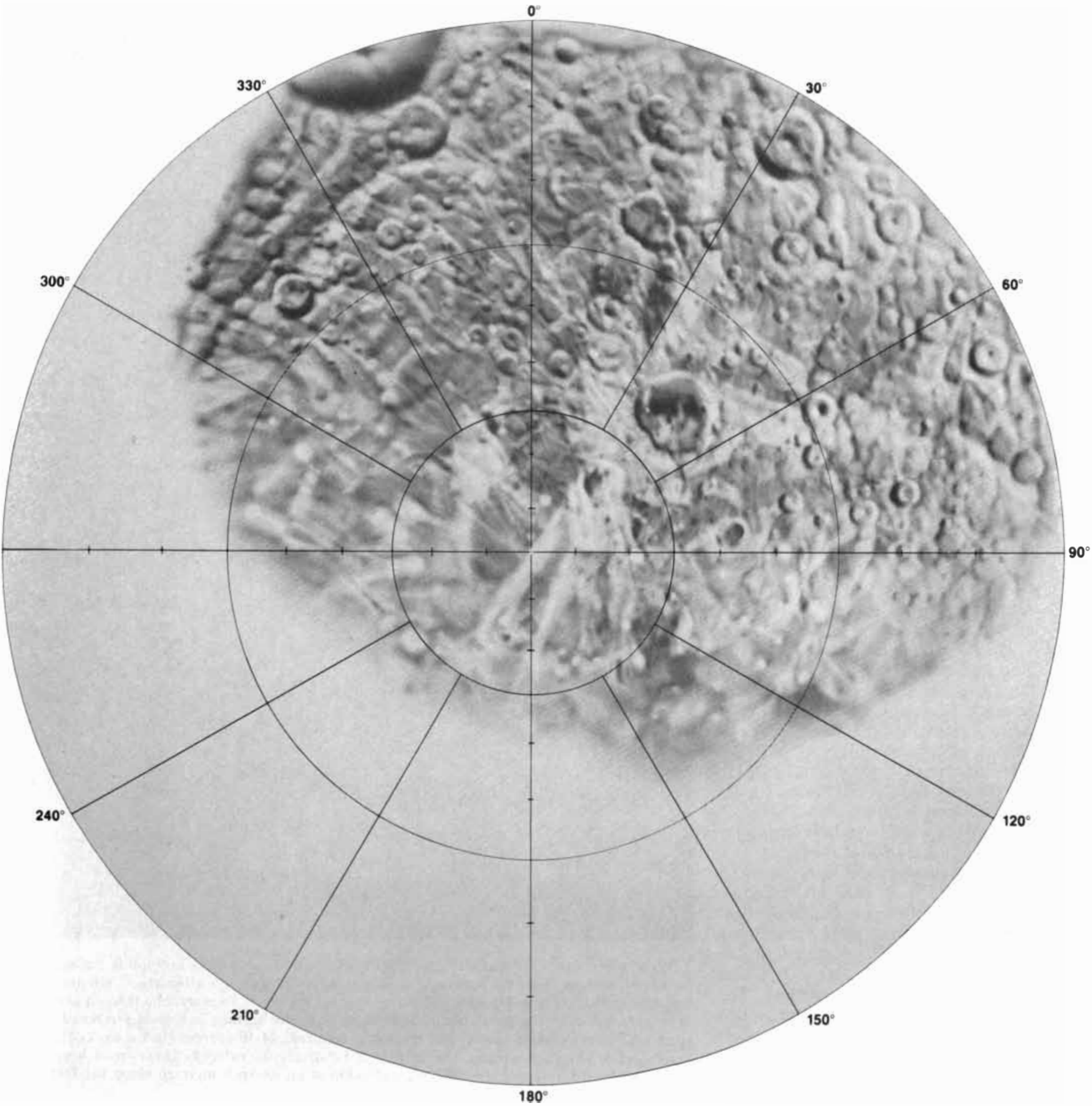
OBERON was photographed by *Voyager 2* from a distance of 660,000 kilometers. It is only slightly smaller than Titania, the largest of the Uranian moons, but its ancient, densely cratered surface shows few signs of geologic activity. Dark deposits that may be volcanic in origin are visible on the floor of a few craters, notably the large one near the center of the photograph; on the map of Oberon's southern hemisphere (*right*) that crater lies between 30 and 60 degrees longitude at a latitude of about 50 degrees. The crater is more than 100 kilometers across, and like the many other large craters on Oberon it dates from the bombardment by heliocentric debris that ended four billion years ago.

1,160 kilometers respectively) and the same mass, but the contrast between them is even stronger than the contrast between Oberon and Titania.

Umbriel is remarkable for its almost total blandness. Like Oberon and Titania it lacks global-scale variations in brightness, but unlike those two bodies its surface is not marked by bright ray craters. (Bright rays are thought to form around craters on icy satellites because the impacts eject clean, buried ice out onto the surface.) As *Voyager 2* neared the Uranian system it appeared that Umbriel might lack craters altogether, which would imply its surface was being regenerated even today. The high-resolution images made at the closest approach, however, revealed a surface dominated by large craters from the first bombardment; only the rays are absent.

Explaining Umbriel's blandness is surprisingly hard. Conceivably bright rays around its craters could have been erased by micrometeoroids that peppered the surface and mixed the ray material with the underlying dark material. (The process is called impact gardening.) A second possibility is that the rays contained methane that has been darkened by energetic radiation. Both hypotheses fail to explain why bright rays are absent on Umbriel but not on the other Uranian moons.

Perhaps the most straightforward explanation is that bright rays never formed on Umbriel. If Umbriel were covered by a blanket of dark material several kilometers thick, the ejecta from an impact would be dark; such dark-ray craters have been observed, for example, on Jupiter's moon Ganymede. The blanket might consist of a uniform, primordial mixture of



ice and rock. A primordial blanket on Umbriel is at least plausible, because Umbriel is smaller than Oberon or Titania and so might be expected to have undergone less differentiation. In addition the fact that it has the lowest reflectance (20 percent) of the Uranian moons supports the notion that it has a high concentration of dark, carbon-rich material near its surface.

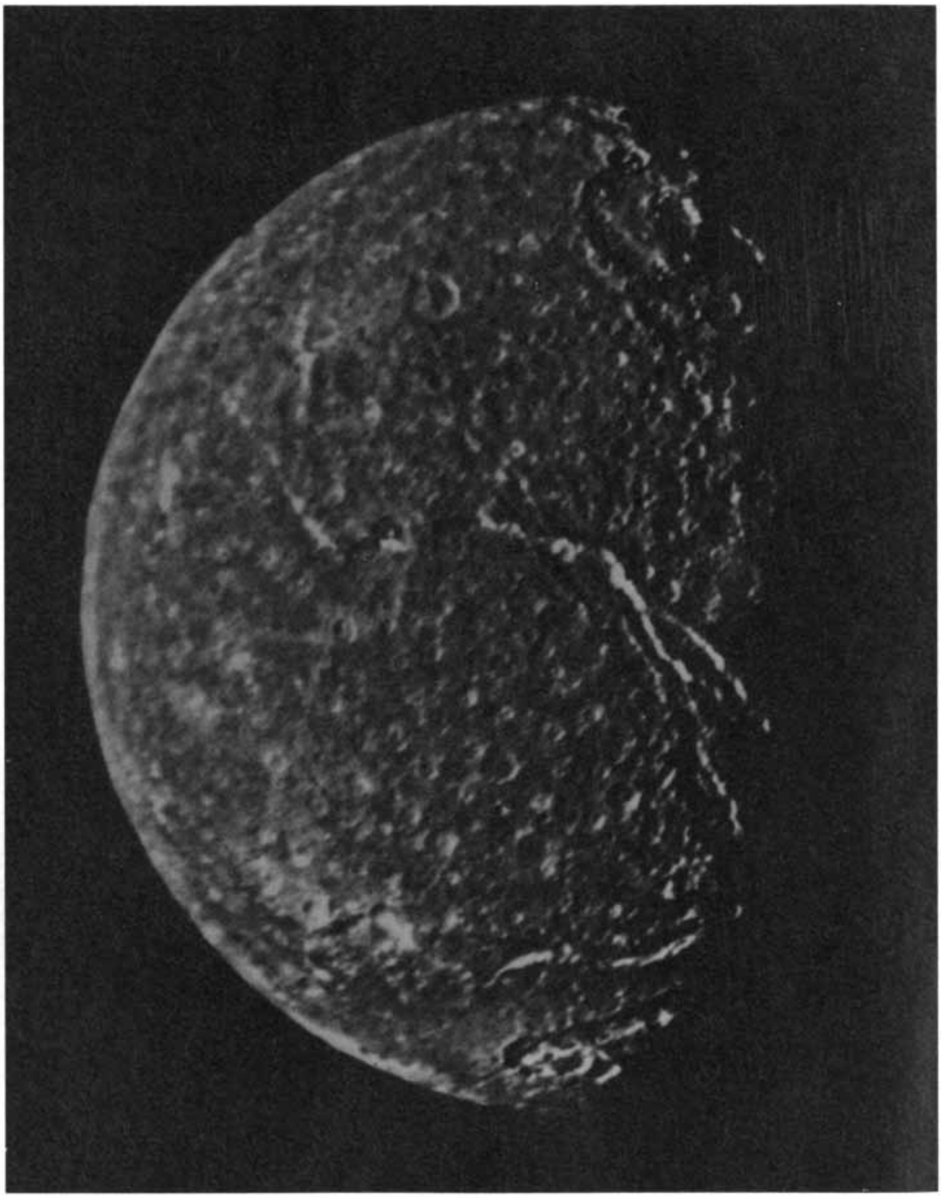
Nevertheless, there are problems with this hypothesis as well. Although most of Umbriel is bland, the moon is not altogether featureless. On the contrary, it has two rather striking bright features near the equator: a ring 80 kilometers in diameter that appears to cover the floor of an impact crater, and a spot on the central peak of another large crater. The origin of these features is a mystery, but the bright material must have come from below the surface. This means that the dark blanket, if it exists, would have to be thin or absent in those two locations and only there.

Whereas Umbriel is the darkest of the major Uranian moons, Ariel is the brightest; it has a reflectance of roughly 40 percent. And whereas Umbriel's surface is among the oldest, most cratered surfaces in the Uranian system, Ariel's is one of the youngest and least cratered. Ariel's history is reminiscent of Titania's, except that the geologic activity on Ariel was more intense, more extensive and more prolonged. Like Titania, Ariel has almost no craters from the heliocentric bombardment. In addition it has only about a third as many craters from the planetocentric bombardment as Titania, indicating that its surface was remade over a broader area and a longer period. Its global network of extensional faults is also more fully developed; in some places on Ariel the fault valleys are tens of kilometers deep.

The resurfacing of Ariel was a volcanic process. The evidence for extrusion of material onto the surface is much more conclusive than in the case of Titania. The floors of the fault valleys in the hemisphere that is always pointed toward Uranus (the Uranian moons, like the earth's moon, are locked in synchronous orbits) are covered with a smooth, relatively uncratered unit. Because the smooth unit covers previously existing geologic features—the most dramatic example is a large, half-buried crater—it must have been emplaced by volcanic flows.

In some areas the smooth unit displays grooves and ridges that parallel the axes of the fault valleys. The pattern suggests that material welled up through fractures along the axes and spread out over the valley floors, somewhat in the manner of lava erupting from rifts on the floor of the earth's oceans. The upwelling material on Ariel was almost certainly not molten rock. Instead it was probably a relatively warm, plastic mixture of ice and rock that flowed much as a terrestrial glacier does. The mixture must have been fairly viscous: in places where it has buried older features its edges form a steep scarp that is about a kilometer high.

Although Ariel has been a surprisingly active little moon, in some ways its style of activity—crustal rifting and icy volcanic flows—is reassuringly familiar, similar to what was seen on Jupiter's Europa and Ganymede and on Saturn's Enceladus. Miranda, on the other hand, is one of the strangest worlds yet observed. The innermost (and with a diameter of about 500 kilometers the smallest) of the major Uranian moons, it has barely enough gravitational strength to pull itself into a sphere. Yet its surface is a hodgepodge of complex and exotic geologic terrains that would seem more appropriate on a planet 10 times its size.

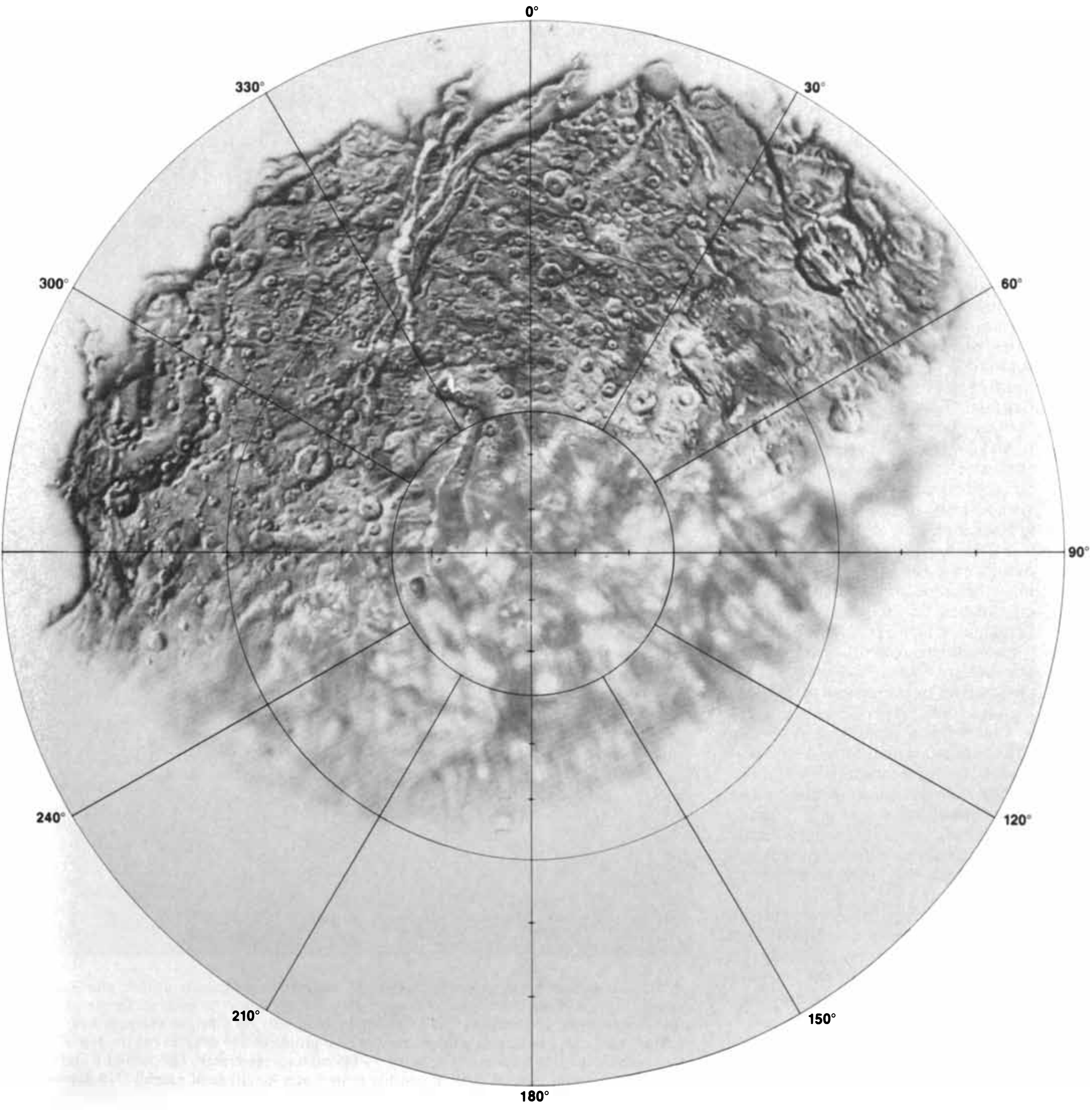


TITANIA was much more active geologically than Oberon. The photograph is a composite of two images made by *Voyager 2* from a distance of 369,000 kilometers. Titania has fewer large craters than Oberon, indicating that its surface is younger, and it has a prominent network of extensional faults, indicating that the surface has been stretched. In some areas, for example near the equator at a longitude of 30 degrees (in the lower right-hand part of the photograph), the surface is relatively uncratered. These smooth areas may have been formed by the volcanic extrusion of an ice-rock mixture along the faults.

The oldest geologic unit on Miranda consists of densely cratered rolling plains pocked by the planetocentric bombardment. They cover most of the surface visible in the *Voyager* images, and they do not look particularly unusual. Superposed on the rolling plains, however, are three enormous, oval-to-trapezoidal regions that for lack of a better word we call the ovoids. The ovoids are between 200 and 300 kilometers in diameter, and judging from the number of craters on them they are substantially younger than the plains. They consist of locally parallel belts of ridges, grooves and scarps that abut one another at odd angles. Bright material as well as dark material is exposed along the scarps and in fresh craters inside the ovoids; in contrast, the rolling plains have a fairly uniform reflectance. Finally, both the ovoids and the plains are cut by huge fracture zones that circle Miranda, creating fault valleys whose steep, terraced cliffs are as much as 10 to 20 kilometers high.

The material in the ovoids may well be less dense than that in the plains. The bright areas along the scarps may consist of clean ice from below the surface, and the dark material is probably similar to whatever welled up in Oberon's craters. There is some evidence on Miranda for eruptive flows, although it is much less pervasive than on Ariel. In one of the ovoids material seems to have flowed out of a low, horseshoe-shaped cone, burying scarps and ridges and piling up behind a steep, lobe-shaped flow front.

The two models that have emerged to explain Miranda's strangeness both assume, for different reasons, that the ovoids are less dense than the roll-



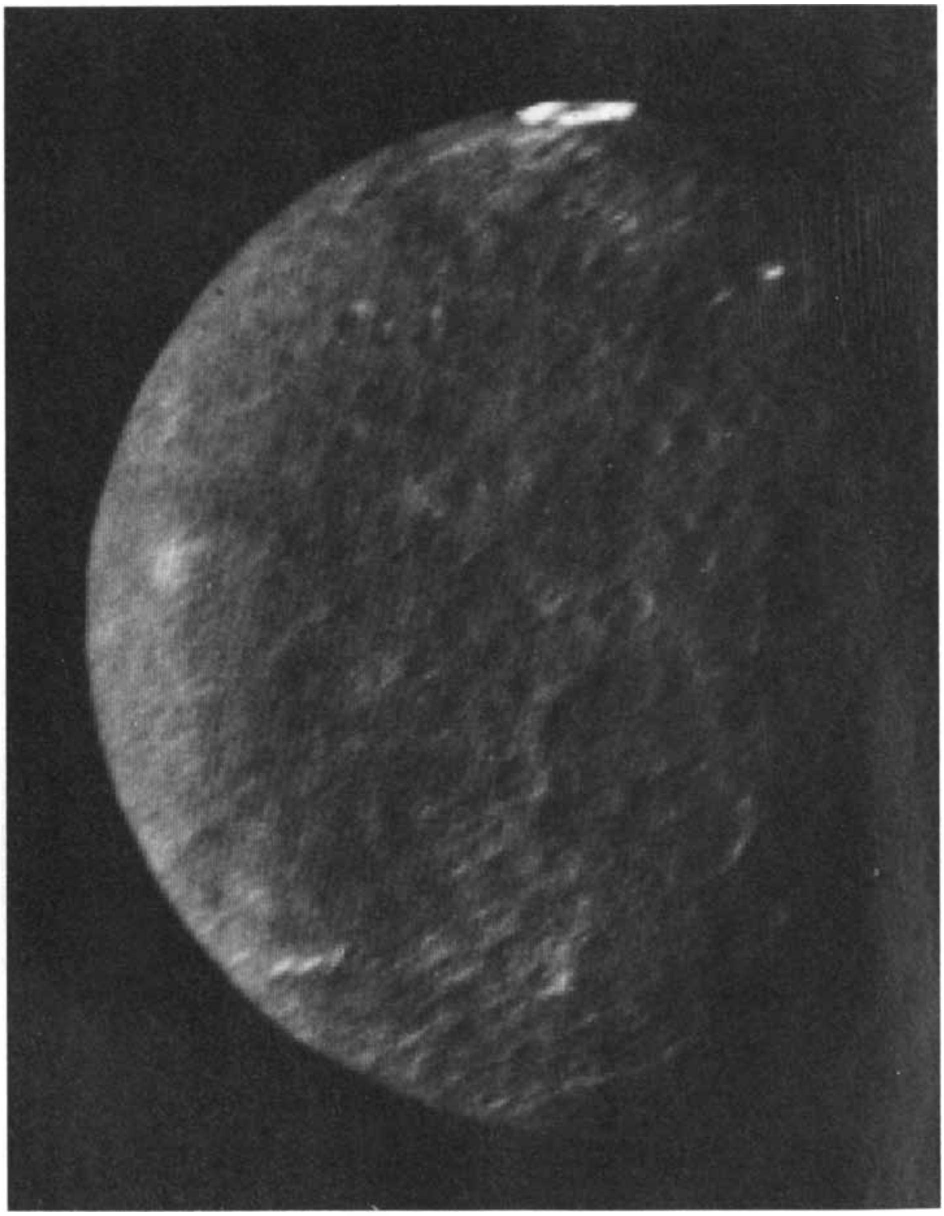
ing plains. The first model is based on Shoemaker's ideas about early cratering rates. Shoemaker thinks Miranda was blasted apart as Titania was, but five or more times rather than just once. Each time it reassembled itself. Before the last cataclysm, according to the theory, the rock and ice in the moon had time to become partially separated. Hence when Miranda was shattered again, it broke into large pieces, each of which consisted primarily of either rock or ice. When the pieces drifted randomly back together, large chunks of rock, which had been near the core of the old Miranda, became embedded in the surface of the new conglomerate. The new moon retained enough internal heat to allow viscous flow in its interior. As a result the rock masses began sinking toward the center again, enabling lighter ice to rise and flow into the space behind them and causing the surface to crack along concentric stress lines. The ovoids, in this view, are surface disturbances left by the sinking rock masses.

One argument against the theory is that it is hard to see how such large chunks of rock could remain intact through the total disruption and reaccretion of the moon. The cataclysm seems more likely to have yielded small pieces. An alternative view, put forward by one of us (Soderblom), holds that when Miranda accreted for the final time (whether it was actually blasted apart or not is unimportant), it was a fairly uniform mixture of rock and ice. Then, driven by its internal heat, it began to differentiate. As the rock sank toward the center, agglomerates of ice began rising toward the surface. In this scenario the ovoids were formed not by large sinking masses of rock but by large rising masses of ice, which eventually breached the surface. The pattern of the ovoids (they are all about the same size and are regularly spaced) suggests an internal organization to the flow, perhaps some form of convection cell.

The process of differentiation on Miranda was apparently never completed. If differentiation had continued, the ovoids would probably have been smoothed over, and Miranda would have been left with a mantle of solid ice surrounding a core of solid rock. Instead the moon was frozen at an embryonic stage of geologic development. Although some workers think the ovoids may be only a billion years old, it seems to us more likely that Miranda's evolution stopped between three and four billion years ago.

The striking thing about moons such as Ariel and Miranda is not that their geologic activity stopped long ago but that it ever began in the first place. Such a degree of activity in such small objects—Ariel is about one-third and Miranda less than one-sixth the size of the earth's moon—would have seemed extraordinary a decade ago. Then the primary source of internal heat in a solid planetary body was thought to be the gradual decay of long-lived radionuclides such as uranium, potassium and thorium. Because the quantity of radioactive elements in a planetary body is proportional to its volume, whereas the rate at which it loses heat is proportional to its surface area, a large body retains internal heat much longer than a small one and therefore undergoes a more prolonged geologic development. The terrestrial planets conform to this pattern. The earth is the largest, and it has clearly been the most active geologically; Mars is intermediate in size and has had a low level of volcanic activity throughout its history; Mercury and the earth's moon are small and had only an early, abbreviated evolution.

The moons of the outer solar system do not follow the pattern. During the past 10 years Voyager observations

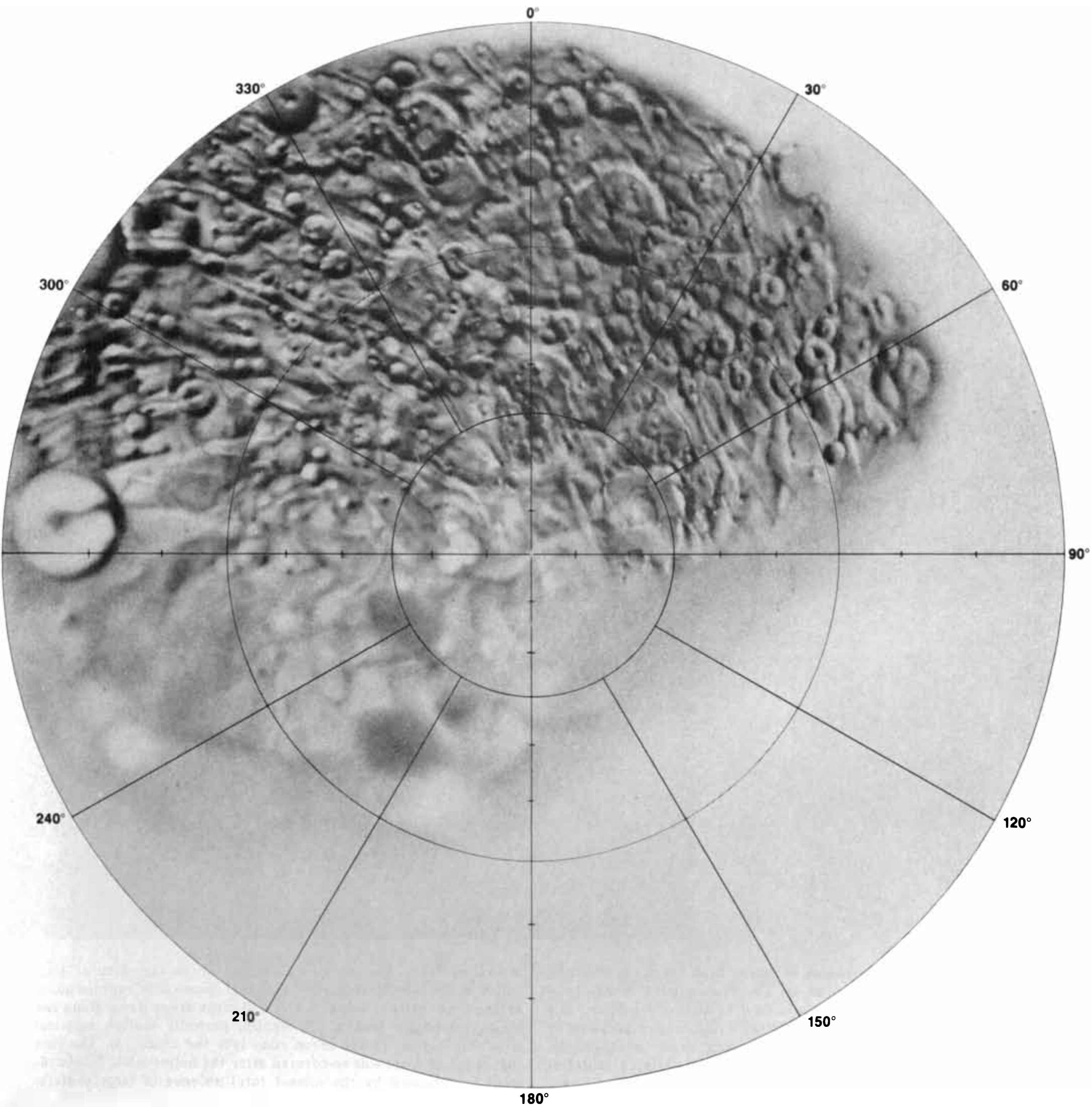


UMBRIEL is distinctive primarily for its blandness: its surface is a dark and nearly uniform gray, and the abundance of large craters shows that it is ancient. In the photograph made from a distance of 557,000 kilometers, only two bright features near the equator stand out. The first is a bright ring at a longitude of 270 degrees (at the top of the photograph); the ring appears to lie in an 80-kilometer-wide crater. The second feature is a bright spot on the central peak of another crater, at a longitude of roughly 310 degrees.

of more than 20 such moons have dramatically revised the conventional wisdom about the geologic history of small bodies. Jupiter's moon Io, where volcanoes are erupting even now, is only the most spectacular counterexample; a number of the other Jovian and Saturnian satellites also display evidence of significant geologic activity. To that list one must now add Miranda, Ariel and perhaps Titania.

The full implications of the Voyager observations have yet to be worked out, but a few guiding ideas have begun to emerge. The moons must have another energy source in addition to long-lived radionuclides. It may be that they received an early pulse of heating from short-lived isotopes such as iodine 129 or aluminum 26, both of which are thought to have been present in the primitive solar nebula. Another possible source of early heating is gravitational energy released by incoming debris as the moons accreted.

Tidal resonances are a third and potentially more prolonged source of heating. When the orbital periods of adjacent satellites differ by an integral factor, one satellite perturbs the other satellite's orbit at regular intervals that are multiples of the perturbed satellite's period. The effect is analogous to the one achieved by driving a pendulum at its resonance frequency: the perturbation is enhanced dramatically. As a result the perturbed moon's orbit becomes distinctly noncircular, or elliptical. Because of the ellipticity, the tidal bulge raised on the surface of the moon by the planet's gravity tends to shift up and down and back and forth. The shifting bulge can inject a huge amount of frictional energy into the moon, and that energy must

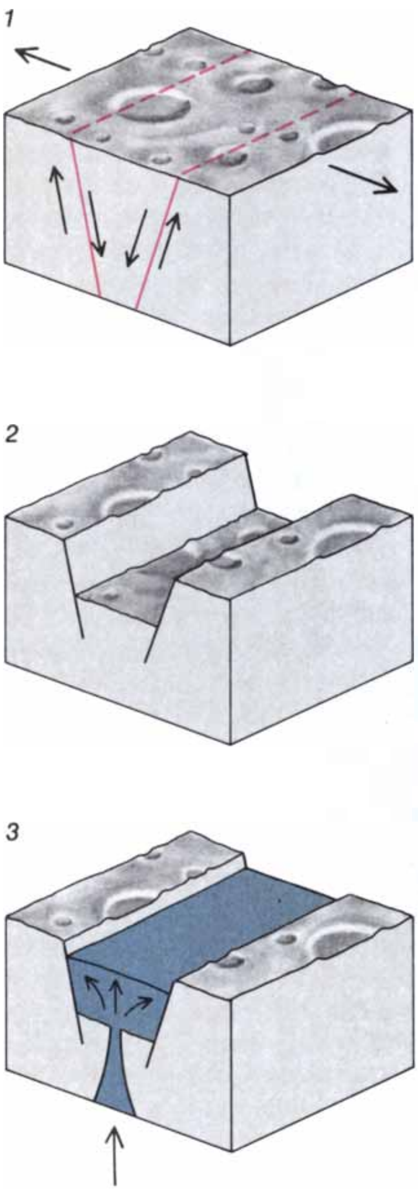


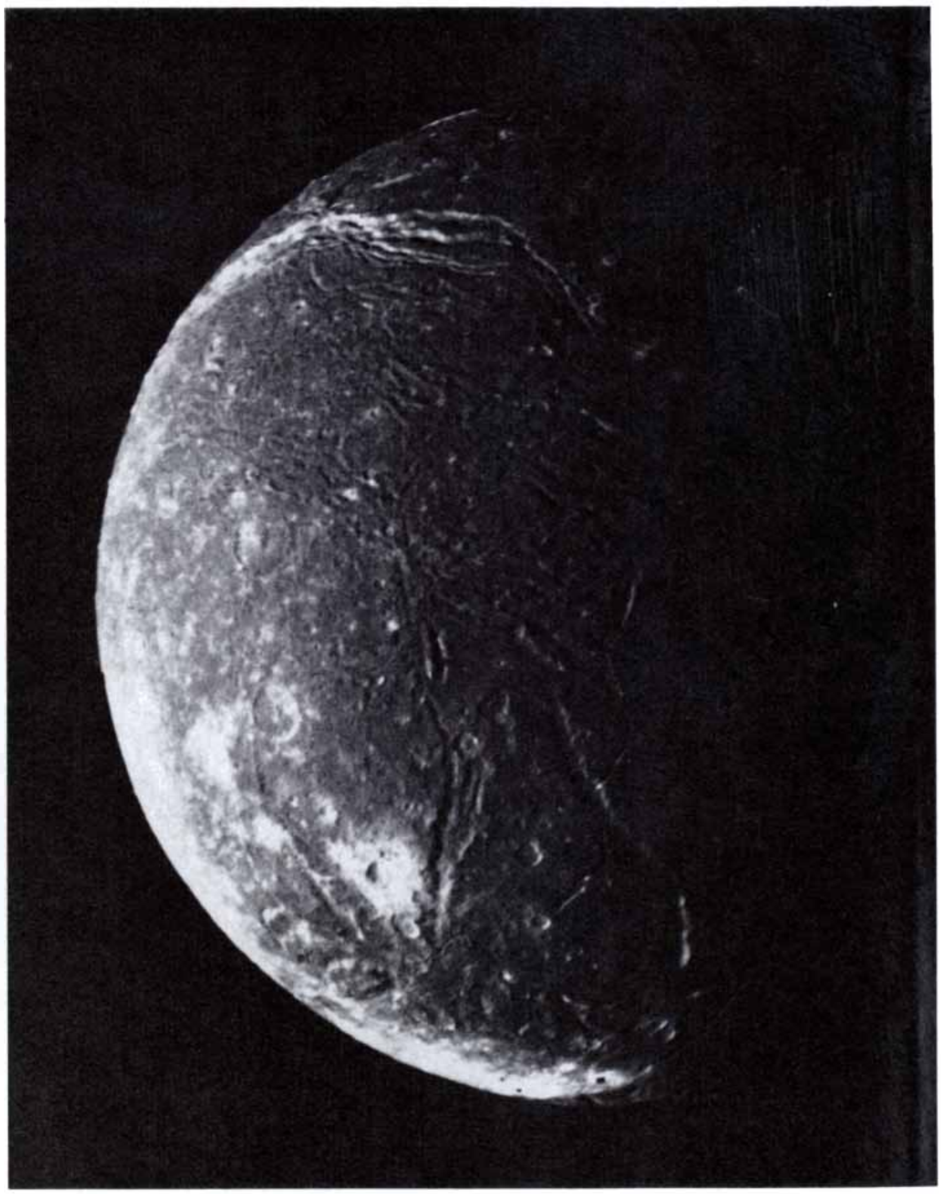
be dissipated as heat. Tidal heating is thought to account for geologic activity on Io, Europa and Enceladus.

The Uranian moons are not now in resonance. But as Steven W. Squyres, Ray T. Reynolds and Jack Lissauer of the National Aeronautics and Space Administration's Ames Research Center pointed out in a paper published before the *Voyager* flyby, Miranda, Ariel and Umbriel may have passed through a resonance in the geologically recent past. At present tidal heating seems the likeliest source of the strong geologic activity on Miranda and Ariel. The tidal effect on Umbriel would have been less significant because it lies farther from Uranus; that would explain why Umbriel and Ariel developed so differently even though they are about the same size.

The unexpected activity of moons in the outer solar system must also have something to do with their composition. To generate volcanic flows, for instance, a moon not only must have a source of internal heat but also must be made of material that can be at least partially melted. Models of the chemistry of the primitive solar nebula indicate that the moons of the outer solar system should be mixtures of carbonaceous rock and ices, primarily water ice. The surface temperature of the moons of Uranus is in the neighborhood of 80 degrees Kelvin (−193 degrees Celsius); water ice melts at 273 degrees K. Whatever the heat source of the moons was, theoretical calculations suggest that it is unlikely the internal temperature of such small bodies was ever 200 degrees higher than the surface temperature.

The moons probably contain some



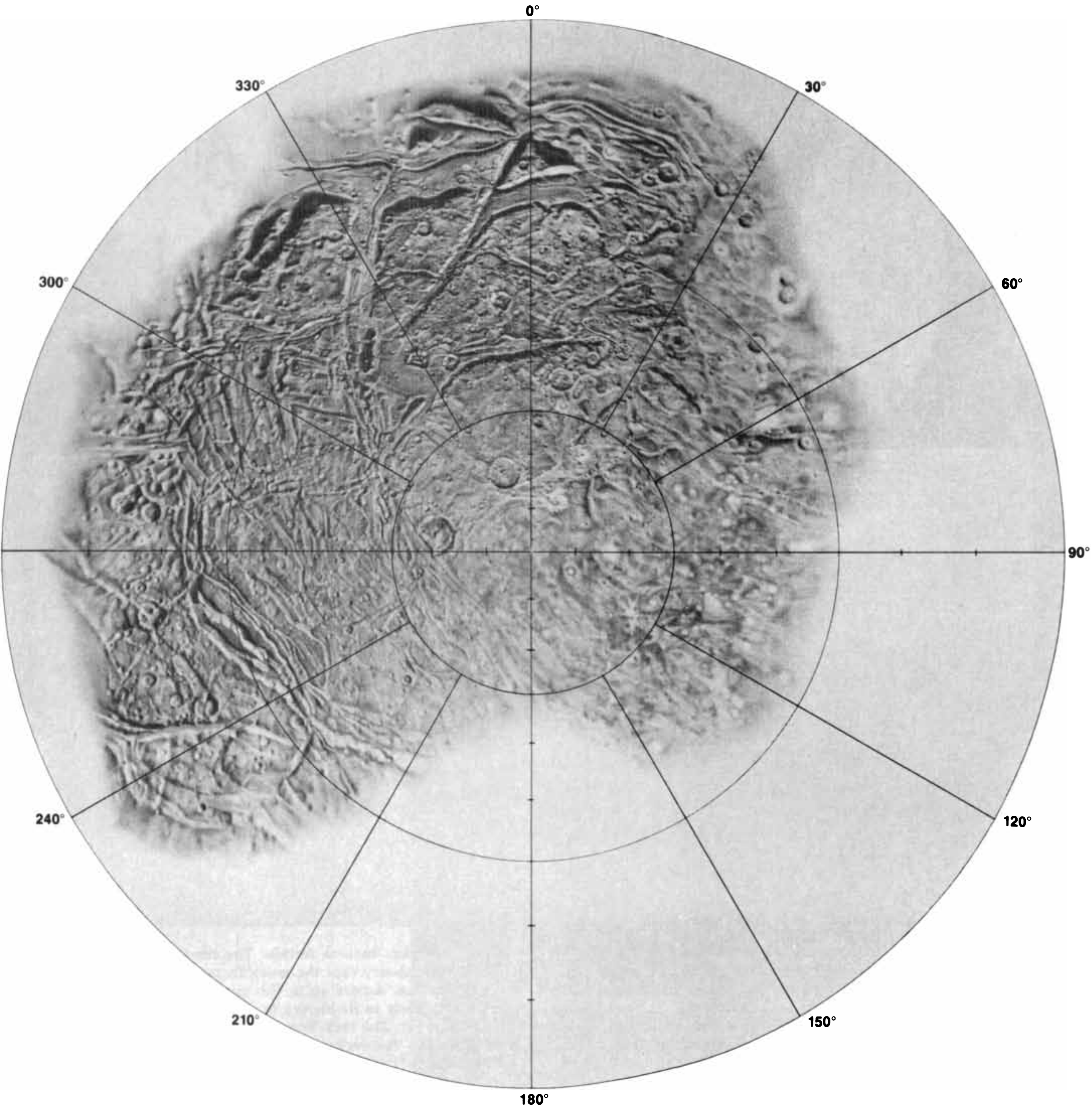


ARIEL displays more evidence of having been volcanically active than any other Uranian moons do. The photograph is a mosaic of images made from a distance of some 130,000 kilometers. The most prominent features are the broad, crisscrossing grabens, or fault valleys, along the equator (at the right in the photograph). The grabens are filled with smooth, sparsely cratered material that was extruded along the axes of the valleys and flowed over the valley floors. The process, a familiar one on the earth, is illustrated at the left. Stretching of the crust causes it to rupture along extensional faults (*1*); next a block of crust drops down along the faults, forming a graben (*2*); finally, partially molten material from the interior of the moon rises into the crack (*3*). The fact that much of Ariel was re-covered after the heliocentric bombardment is indicated by the almost total absence of large craters.

material that is more readily melted than pure water ice. The leading candidates are ammonia hydrate, methane clathrate (a form of water ice in which methane molecules are interspersed in the ice-crystal lattices) and carbon monoxide clathrate. A saturated ammonia-water mixture, for example, has a melting point about 100 degrees lower than that of pure water.

The relative proportions of ice and rock in a moon also influence its propensity for geologic activity. The more rock it has, the stronger is its radiogenic heat source, because the radioactive elements are embedded in the rock. To estimate the amount of rock in a moon one must know its average density, and to know its density one must know its mass with some precision. Estimates of the mass of the Uranian moons made before the *Voyager* flyby proved to be inaccurate, but *Voyager 2* was able to measure the masses of Oberon and Titania precisely. (The measurement procedure was an interesting one. During the weeks before the encounter the moons were continually photographed, because their exact positions were needed in order to navigate the spacecraft. The photographs revealed the tiny perturbing effects of the moons on one anothers' orbits. These perturbations, together with precision radio tracking of the spacecraft, enabled workers to calculate the satellite masses.)

When the measured masses are converted into densities, and when allowance is made for the effects of the moons' own gravitational self-compression, both Oberon and Titania turn out to consist of materials whose average density is between 1.4 and 1.7 grams per cubic centimeter. These

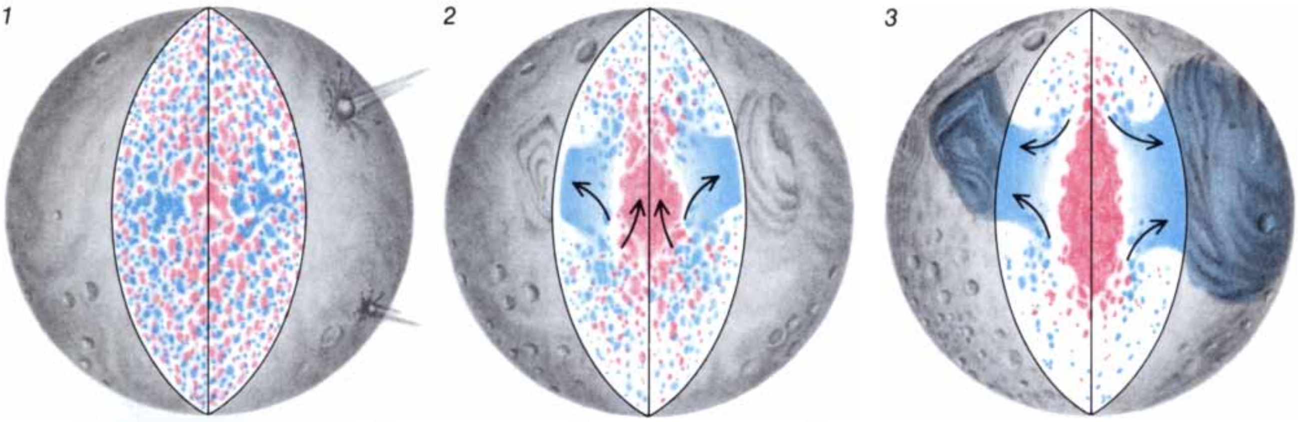


figures imply that the two moons are between 40 and 65 percent rock, which makes them considerably rockier than Iapetus and Rhea, two Saturnian moons of comparable size. The other Uranian moons may have similarly high rock fractions. (One cannot say for sure because the uncertainty about their masses is still too high.) This suggests that radiogenic heating may have been a significant contributor to their geologic activity.

The rockiness of the Uranian moons compared with the Saturnian ones is counterintuitive; simply put, one expects to find more rather than less ice as one moves farther away from the sun. Some models of the solar nebula had predicted that the Uranian moons would be no more than 35 percent rock. Apparently they were wrong.

Various revisions of the models have been proposed. One possibility is simply that the chemical reactions prevailing in the cold outer regions of the solar nebula were different from those in the inner regions. Another possibility, suggested by James B. Pollack of the Ames Research Center, is that the Uranian moons got their rockiness not from the solar nebula but from a gaseous envelope linked to Uranus. David J. Stevenson of the California Institute of Technology has proposed that an envelope rich in rocky materials could have been created by the event that is thought to have given Uranus its unusual orientation: the impact, during the final stage of the planet's accretion, of a body about the size of the earth. The moons would probably have been disrupted by such an impact, and so they must have formed afterward. In the process they may have incorporated rock from the impactor or from the planet itself.

The *Voyager* images have revealed much about the Uranian moons. But they did not reveal immediately or unambiguously why the moons are the way they are: why they are dark and rocky and geologically varied. Nor did we expect them to. Those are problems that will occupy theoreticians for some time. As for us, we are already gearing up for Neptune, for Neptune's large moon Triton and for *Voyager 2*'s next brief but intense encounter, in 1989, with another world.



MIRANDA is seen here from an angle that *Voyager 2* never saw. Because the spacecraft passed so close to Miranda (within 29,000 kilometers), the scale of its rugged topographical features could be deduced from stereoscopic views of the same region from different angles. Miranda is the only moon in the outer solar system for which such detailed topographical information is available. Using these data, workers at the Geological Survey have reprojected the *Voyager* images from new angles in order to highlight various features of Miranda's bizarre terrain. The reprocessed images lend support to the theory that the moon froze in an embryonic stage of differentiation. According to this model, Miranda had enough internal heat early in its history to allow its ice and rock to form agglomerates (*1*). The rock began settling to the core and the ice began rising to the surface (*2*). Eventually large plumes of ice breached the surface, forming the ovoid patterns (*3*). Miranda's heat dwindled before the differentiation process could be finished.

Streamline your accounting with a

Tandy® 6000 HD multiuser system

For many firms, it's hard to justify the cost of a separate computer for each individual user. That's why we designed the Tandy 6000 multiuser computer. With its XENIX® operating system, several people can perform different accounting tasks simultaneously, using low-cost display terminals.

And whether you choose a three-user or a six-user system, you'll only need one copy of software, making the system more cost effective by eliminating the need for duplicate files. All files are stored at the Tandy 6000 and can be accessed by each remote user.*

We offer the software you need, too. Increase office productivity with our integrated accounting software, including: General Ledger, Accounts Receivable, Accounts Payable, Payroll, Order Entry/Inventory Control and Sales Analysis.

Find out how your firm can benefit from a Tandy 6000 accounting system. Stop by a Radio Shack Computer Center today.

Send me an RSC-18 Software Reference Guide

Mail To: Radio Shack, Dept. 87-A-503
300 One Tandy Center, Forth Worth, TX 76102

Name ____________________

Company ____________________

Address ____________________

City ____________________

State ______________ ZIP ______________

Phone ____________________

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

TANDY COMPUTERS: In Business . . . for Business™

Radio Shack®
COMPUTER CENTERS
A DIVISION OF TANDY CORPORATION

*Files can be password protected to restrict access to certain users only.
XENIX/Registered TM Microsoft Corp.

SCIENCE AND THE CITIZEN

Germ Warfare

In 1347 Tatars besieging Kaffa, a town on the Black Sea, catapulted corpses riddled with bubonic plague over its walls. This primitive germ warfare worked: plague swept through Kaffa and its people soon surrendered. All Europe paid for the victory, however: ships sailing from Kaffa carried infested rats to other ports, sowing the seeds of the Black Death.

The recognition that the effects of even carefully engineered biological weapons can be dangerously unpredictable induced President Nixon to renounce them in 1969. For more than a decade thereafter the U.S. biological-warfare program was virtually nonexistent. Under the Reagan Administration the program has revived, its budget growing from less than $10 million to more than $60 million last year, spread among investigators in industry and academia as well as in military laboratories.

The buildup is coming under increasing criticism from the scientific and arms-control community. The U.S. efforts, critics say, could unravel the only multilateral treaty that bans weapons of mass destruction: the Biological Weapons Convention signed by 113 nations, including the U.S. and the U.S.S.R.

Barbara H. Rosenberg of the Sloan-Kettering Institute for Cancer Research, a member of the Committee for Responsible Genetics, argues that even if the U.S. buildup stays within the treaty's legal limits, it could encourage other countries to initiate their own programs and trigger a global biological arms race. Already, she observes, "there is international suspicion on all sides." These suspicions dominated a meeting of signatories of the convention in Geneva last September; a follow-up meeting is scheduled to take place this month.

Pentagon officials maintain that the convention is unverifiable and that they must counter a growing threat from the Soviet Union and possibly from other countries as well as terrorist groups. Advances in biotechnology, the officials note, have recently deepened the threat by making the design and manufacture of biological weapons far easier. The Department of Defense stresses that its buildup is purely defensive—aimed at devising medical treatment and protective gear—and is allowed by the convention.

To carry out this research, however, Defense Department investigators say they must study the "threat," which includes pathogens ranging from anthrax to Venezuelan equine encephalomyelitis. It is the work on the offensive potential of pathogens and toxins that most concerns critics. A 1986 DOD report to Congress describes investigations of the military potential of "newly discovered viruses" and of "aerosols of microbial organisms or their toxins." The Defense Department wants to expand its ability to do such research by constructing a new containment facility at the Dugway Proving Grounds in Utah. The heart of the facility would be an 18-cubic-foot chamber into which pathogens would be sprayed to test the efficacy of sensors, protective clothing, decontaminants and other gear.

The Defense Department's efforts have not gone unchecked. Two years ago Jeremy Rifkin, the antibiotechnology activist, won a court injunction that stalled construction of the Dugway facility. Congress has trimmed budget requests and has ordered the department to keep it well informed. In January the House Judiciary Committee introduced a bill that would make producing biological weapons even by Government personnel a criminal offense punishable by imprisonment for life.

Defense Department officials vehemently deny any interest in establishing an offensive capability. Their research is largely intended to protect U.S. troops stationed overseas from such common scourges as malaria and cholera, according to Thomas R. Dashiell, a Pentagon official who oversees the research programs.

The depth of the Pentagon's concern about the military rather than the natural threat from biological agents, however, is suggested by the titles of some recent Pentagon "threat assessments": "An Evaluation of Entomological Warfare as a Potential Danger to the United States and European NATO Nations," "Biological Agent Delivery by ICBM" and "Biological Vulnerability Assessment: the U.S. East Coast."

When asked whether a genuine military threat justifies such studies, Pentagon officials cite "evidence of Soviet noncompliance" with the Biological Weapons Convention. Specifically, they allege that the deadly anthrax spores that killed Soviet citizens in Sverdlovsk in 1979 had leaked from a nearby plant that was mass-producing the spores, and that the Soviets have supplied their allies in Vietnam and Afghanistan with fungal toxins known as "yellow rain." The allegations are widely disputed, however, and Pentagon officials concede privately that they have little or no hard data on Soviet biological-warfare capabilities.

Matthew Meselson of Harvard University, an authority on biological weapons, notes that the reasoning that once led the U.S. to abandon its program is still valid. Rather than showing renewed interest in exploring the potential of biological weapons, he says, the U.S. should reemphasize its renunciation of them.

Onward and Yupward

The political party that wins the allegiance of the success-bound managerial and professional cohort of the baby-boom generation could dominate American politics for many years, according to a study undertaken in 1965 by investigators at the University of Michigan. Winning their allegiance may, however, prove to be difficult because they tend to be liberal on social issues and yet conservative on economic ones.

The study is the work of M. Kent Jennings and Gregory B. Markus. Jennings started the project with a nationwide survey of 1,669 members of the high school graduating class of 1965 and at least one parent of each student. In 1965 and again in 1973 and 1982 the participants were asked about their views on such issues as domestic social spending, the Government's role in providing jobs and a minimum living standard, affirmative action for minorities, equal roles for men and women, school integration, school prayer and abortion. They were also asked about their participation in presidential elections and their choice of candidates. A final set of questions explored their attitudes toward various social and political organizations, including labor unions and big business.

Jennings and Markus define Yuppies (young, upwardly mobile professionals) as people who have graduated from a four-year college, hold professional or managerial jobs and have a family income of at least $30,000 per year in 1982 dollars. Yuppies constitute about 15 percent of the baby-boom generation (people born between 1946 and 1964). As a group, the study found, they "took the most consistently liberal positions on prayer in schools, integration, abortion and equality of gender roles."

On the other hand, their attitudes on economic issues have moved "dramatically in a conservative direction." For example, 71 percent of the Yuppies in 1982 opposed the idea that Government should take a strong role in providing jobs and a good standard of living; 46 percent said they thought the Government spent too much on domestic services and 42 percent opposed Government action to improve the social and economic position of blacks and ethnic minorities. The primary reason the Yuppies have shifted toward conservatism on economic issues, the investigators believe, is "their entrance into the professional and managerial ranks and the economic advantages of those occupations."

Non-Yuppie baby boomers were less opposed to Government action on jobs and domestic spending (respectively 57 and 35 percent) and somewhat more opposed to affirmative-action programs (49 percent). On social issues they ranked well behind Yuppies in support of an equal role for women, abortion and school integration and well ahead of Yuppies in support of school prayer.

Solving the Yuppie paradox would give a political party or movement major leverage, the workers conclude. Although Yuppies are a minority of their generation, they are "numerically important—10 to 20 million by generous definitions—and they are economically advantaged and politically active in ways that magnify their numbers."

Mars-struck

It is difficult to avoid a strong sense of déjà vu as the call for an exploration of Mars gathers strength. Prominent investigators argue that science demands the adventure; they envision a role for astronauts. Then there is the Soviet card, which plays in two suits: war and peace. A joint U.S.-Soviet program, it is argued, would promote peace; lack of a unilateral program gives the Soviets a technological edge.

As the new Congress gets to work on the budget, the Soviet theme has been gaining in intensity. In Washington word has spread that the U.S.S.R. is planning a new twin-spacecraft mission to Mars in 1992–94, and there are rumors that the Soviets will propose an international Mars sample-return mission for 1998. The 1992–94 mission would release meteorological balloons into the Martian atmosphere and send a penetrator into the soil. The new Soviet mission would expand an already impressive Mars exploration program, which also includes probes that will study Mars from orbit and land on Phobos (one of Mars's two moons) in 1989.

If they are true, rumors about the Soviet program will intensify the pressure on the National Aeronautics and Space Administration and Congress to field an ambitious U.S. planetary program. There appears to be strong grass-roots support. Indignant enthusiasts of planetary exploration swamped Congress and NASA with 12,000 letters in January and February, protesting the agency's decision to delay the *Mars Observer* until 1992. The spacecraft is planned to orbit the planet and provide data on composition, topography and climate. The letter-writing campaign successfully "bought some extra time," according to Louis Friedman, executive director of the Planetary Society, which organized the effort. In response to the protest and to concern on Capitol Hill, NASA relented.

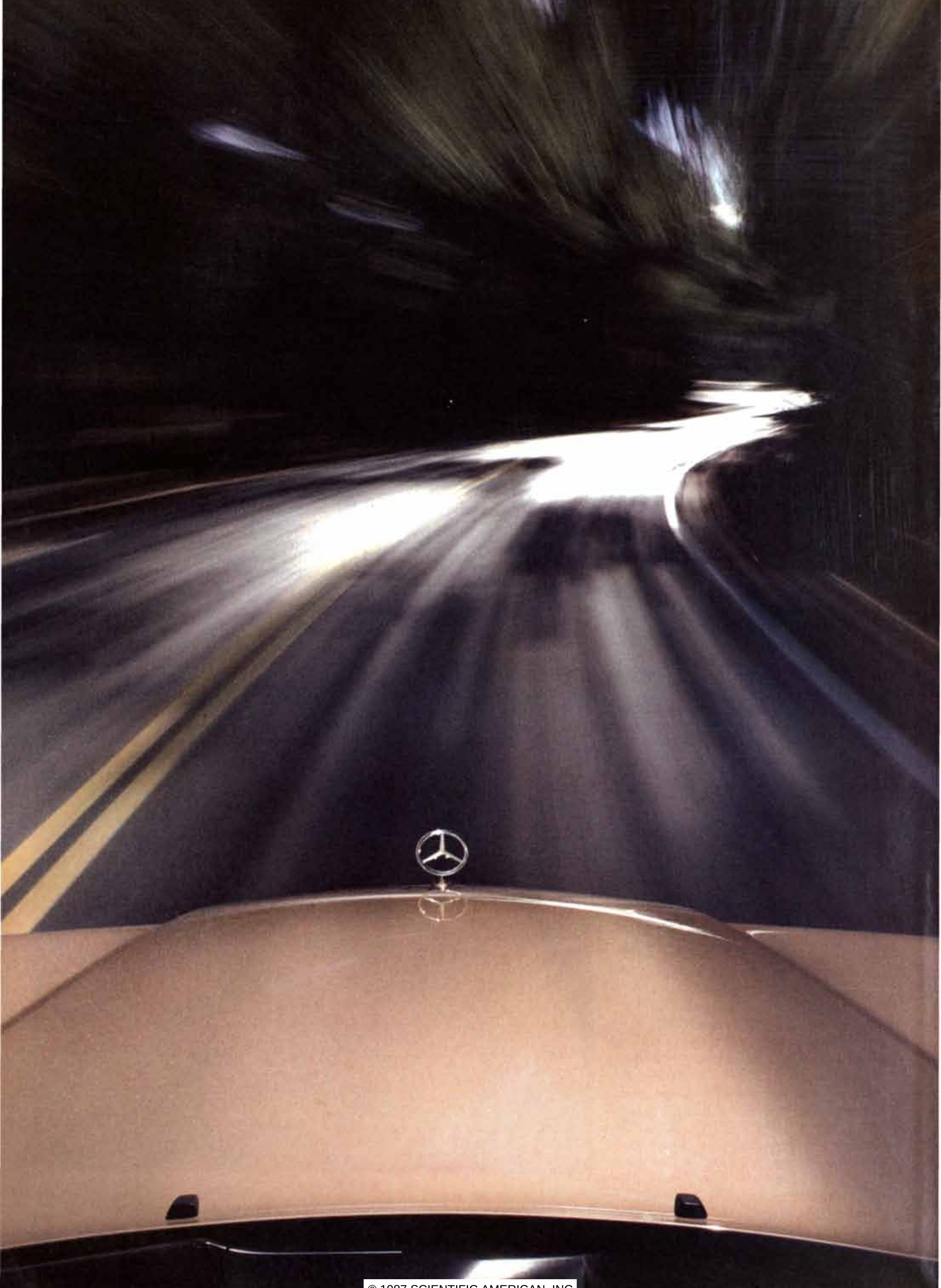
NASA's earlier plan to delay the *Mars Observer* was forced on it by a shortage of launch slots on the shuttle fleet. The problem was that by the end of 1990 NASA and its partners in the European Space Agency (ESA) will have four planetary-science payloads ready for launch but will have only three shuttle flights to launch them on. One payload had to go.

Besides the *Mars Observer* there is *Ulysses,* a joint project with the ESA to send a spacecraft out of the plane of the solar system to explore the poles of the sun; *Magellan,* formerly known as Venus Radar Mapper, and *Galileo,* which will survey Jupiter's moons and send a probe into the Jovian atmosphere. All flights have already suffered some delay, partly as a result of the *Challenger* accident. Provided the shuttle returns to service in February of next year, *Magellan* will be launched in April, 1989. *Galileo* will probably follow at the end of 1989. The question was whether *Ulysses* or the *Mars Observer* should get the 1990 shuttle slot.

Appropriations committees of both houses of Congress were disturbed when NASA proposed putting off the *Mars Observer* until 1992. Congress suggested that NASA should delay *Ulysses* for a year instead and launch it with the new, untried Titan IV rocket. NASA demurred, in part because the delay of *Ulysses* would have meant reneging on agreements previously reached with the ESA. The ESA would have been "very unhappy" with a decision to delay the mission, according to an ESA official. NASA thereupon suggested putting the *Mars Observer* on the tried and tested Titan III rocket in 1990. A Titan III launch would cost less than $100 million (compared with $250 million for a Titan IV launch).

Furthermore, launching the *Mars Observer* in 1990 would allow it to carry out joint observations with the Soviet orbiter. "Everyone here at NASA wants to do it," according to associate administrator Burton I. Edelson. A 1990 launch also has support in Congress, where Senator Spark M. Matsunaga has proposed that 1992 be designated International Space Year. NASA

SUNNY SIDE OF MARS was viewed by the *Viking 1* orbiter in 1976. This NASA image is a composite of three photographs made at a distance of 350,000 miles from the planet.



THE MERCEDES-BENZ 190 CLASS: THE SUBTLE DIFFERENCE BETWEEN MASTERING THE ROAD AND MERELY COPING WITH IT.

The road passes beneath you as always, but the sensations are markedly different. So is your state of mind. This is your first experience with a 190 Class sedan, but already you are driving with calm confidence. The car has earned your trust.

It feels resolutely stable, going precisely where you steer it, refusing to waver off course or wallow over potholes. Even the severest bumps seem only a minor disturbance as the suspension gently quells the violence underneath. Negotiating a run of switchback turns seems more routine business than high drama as the car shifts direction nimbly in response to your steering commands. Sports sedans might occasionally handle this adroitly, but they seldom feel this composed.

Suddenly the pavement deteriorates into washboard gravel, but the car tracks steadfastly ahead, curiously unfazed by the change in terrain. It occurs to you that you have yet to hear a squeak or rattle. The engine remains almost subliminally quiet, wind noise a faint whisper when you hear it at all. You normally feel an urge to stretch your legs after sitting for so long, but now you feel the urge to keep driving.

Even if you chose the automatic transmission, you still find it easy to shift manual-style when the mood strikes, locating each gear by feel without glancing downward. Your driving has become pleasurably instinctive, as driving at its best should be.

This ostensibly mystical exaltation of the driving experience springs from such technological advances as "the most sophisticated steel suspension ever put into volume production" (Britain's *Car* Magazine). And the simple fact that a 190 Class sedan is built like every Mercedes-Benz—not one ergonomic or safety principle sacrificed for the sake of cosmetic luxury or digital showmanship. Every detail of construction and assembly meeting universally envied standards.

The result is a sedan that does not "challenge" you in the macho sports-sedan tradition, but rather serves as a congenial and supremely capable ally—at once exciting and obedient, responsive and considerate. The road provides challenge enough.

Engineered like no other car in the world

does not at present have in place the formal agreements it would need to collaborate with the Soviet Union in an international sample-return mission to Mars, but the idea seems to have widespread appeal.

PHYSICAL SCIENCES

Supersite

President Reagan's decision to support construction of what will be the world's largest particle accelerator has sent a score of states or more scrambling to prepare site proposals. A technical panel will draw up a short list of finalists by the end of the year; the winner will be selected by January, 1989. Whatever state industrial-development officials may understand about the secrets of the cosmos, they do know that the Superconducting Supercollider (SSC) is guaranteed to bring significant economic benefits.

The SSC is so named because it will employ superconducting magnets and because it is very big: it will occupy a tunnel 52 miles in circumference. The SSC has to be big because even superconducting magnets are not strong enough to keep protons with the extremely high momentum imparted by the SSC on a tighter curvature. The protons, accelerated in opposite directions, will slam into one another with a total energy of 40 TeV (trillion electron volts). That is 20 times more than the energy achieved at Fermilab's Tevatron, currently the world's most powerful accelerator.

Because the tunnel will be at least 20 feet underground, stable geology is a prime consideration. Construction of an accelerator at CERN (the European laboratory for particle physics) in Geneva has been delayed for months because a tunnel hit a fracture in the Jura Mountains that filled it with two feet of water. Social factors are also considered important: the U.S. Department of Energy has asked potential sites to provide, among other things, the SAT averages of local students. Presumably prospective staff members with children will be looking carefully at the quality of local schools.

High-energy physicists believe the new particles produced by protons colliding at very high energies will illuminate the relations among the four known forces of nature and explain the origin of mass [see "The Superconducting Supercollider," by J. David Jackson, Maury Tigner and Stanley Wojcicki; SCIENTIFIC AMERICAN, March, 1986].

Opposition to the SSC focuses on the issue of cost, which is now estimated to be $4.4 billion in 1988 dollars. The SSC, opponents say, will draw funds away from other areas of science that might bring more immediate economic benefits. George C. Pimental of the University of California at Berkeley, who in 1985 studied the state of academic chemistry for the National Academy of Sciences, raises another issue. "The odds are," he says, "that the SSC will not have a great impact on the life and times of the average citizen."

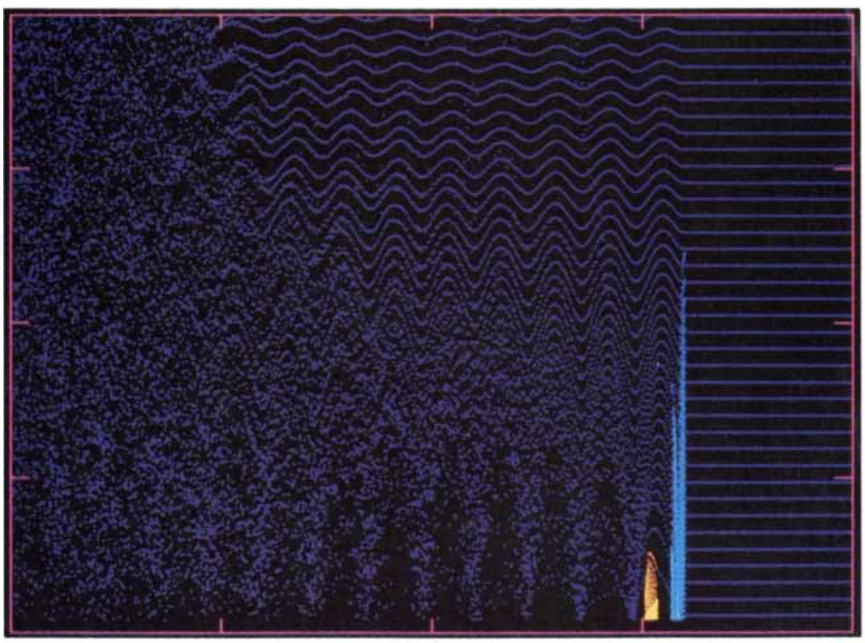
Secretary of Energy John S. Herrington assures critics that the SSC will be built with new money and so will not affect other basic research. Arno A. Penzias, a vice-president of the AT&T Bell Laboratories who is a Nobel laureate and an outspoken opponent of the SSC, says that will be true only if the Government does not in the future include SSC costs with those of other basic research. Administration officials point to their pledge to double the $1.6-billion budget of the National Science Foundation within five years as evidence that the Administration is not starving other research.

Among the states working on proposals, the front-runners seem to be Illinois, California, Texas, Washington, Colorado, Ohio and Utah. Illinois, the site of Fermilab, is particularly eager to play host and is said to be ready to spend more than $7 million to bring the SSC to the state. It has an advantage: the existing Tevatron accelerator at Fermilab near Chicago might serve to inject protons into the main SSC ring. At least two states have indicated they would compensate for whatever advantage the Tevatron is deemed to represent by putting up hard cash—up to $500 million.

At its peak SSC construction will cost $600 million per year. Congressional staff members believe the project, although popular, will receive close scrutiny on Capitol Hill. The modest $35 million the Energy Department is asking for the SSC in 1988 will be "reprogrammed" from other projects; the political strategy is apparently to get as many states as possible interested and then to ask for the big money in 1989, when a new administration will be in office. Innovative financial arrangements to bring in money from other sources would sweeten the deal. Some countries, including Japan, have already expressed interest informally.

Surfing on a Plasma Wave

After the Superconducting Supercollider is built, will physicists have to build an even bigger and costlier facility to delve deeper into the subatomic world? Not if an alternative method for accelerating particles, called plasma wake-field acceleration, lives up to its theoretical promise. The technique, in which electrons are shot through a plasma of ionized gas rather than through a vacuum, has been simulated on supercomputers and will



TWO BURSTS of electrons lasting for six picoseconds flash through a plasma-filled chamber 10 centimeters long in this simulation, by supercomputers at the Los Alamos National Laboratory, of wake-field acceleration. The first pulse (*blue*) creates a wake that boosts the smaller second pulse (*yellow*) to very high energies. The electromagnetic field created by the first pulse as it passes through the plasma "pinches" it in the middle.

soon be tested. If years of further testing prove fruitful, accelerators less than a mile long may someday boost particles to energies higher than those planned for the supercollider, which will be 52 miles in circumference.

"This may be the next step beyond the supercollider," John M. Dawson of the University of California at Los Angeles says. Dawson, a plasma physicist, proposed the wake-field concept several years ago. Now researchers at the Argonne National Laboratory, the University of Wisconsin at Madison and the Los Alamos National Laboratory, as well as at U.C.L.A., are working to validate the idea.

The technique involves sending two bursts of electrons through a plasma in rapid succession. As the initial burst speeds through the plasma it acts like a speedboat, kicking up a powerful wave in its wake. If the second, smaller burst of electrons trails just behind the first, it may be able to "surf" on the wave, absorbing as much as one billion volts of energy in the space of a meter. Conventional accelerators, in which successive voltage surges accelerate particles down an evacuated pipe, can deliver only about 20 million volts per meter; at higher voltages sparks may arc across the metal surfaces.

The wake-field concept is similar to a more mature technique, also proposed by Dawson, in which the initial "speedboat" pulse consists of laser-generated photons rather than electrons. With laser pulses workers at U.C.L.A. have generated electromagnetic fields of one billion volts per meter in a plasma; Canadian investigators have reportedly accelerated electrons with the technique. To produce short, intense pulses, however, lasers must consume enormous amounts of power; electron beams can be generated much more economically.

Workers at Argonne, in collaboration with those at Madison, have assembled equipment for a test of the wake-field concept later this year, according to James D. Simpson, head of the project. The hardware consists of an electron accelerator, a plasma drift tube 10 centimeters long and diagnostic gear. Simpson says that Argonne's powerful electron accelerator has been modified to generate pulses lasting for a scant six trillionths of a second.

The Los Alamos investigators have already simulated the machine's performance on supercomputers. Their simulations predict that the electromagnetic field generated by the passage of the first electron pulse through the plasma will "pinch" the pulse in the middle. The pinching in turn will roughly double the strength of the field, boosting it to about 170 million volts, according to Rhon Keinigs of Los Alamos.

Wake-field machines could accelerate either electrons or their antimatter twins, positrons. Applying the technique to heavier particles—in particular to protons, whose collisions will be studied in the supercollider and are being studied now at CERN and Fermilab—would be "a little trickier," Keinigs says. Although proton collisions yield more constituent particles, Dawson maintains that collisions of electrons can provide physicists with just as much important data. Because the collisions create less "debris," he adds, the results are easier to interpret.

Core Columns

Measurements of the earth's magnetic field made hundreds of years ago by Sir Edmund Halley, Captain James Cook and others have enabled a pair of investigators to sketch a picture of the events inside the core of the earth that create and sustain the magnetic field.

Writing in *Nature,* David Gubbins of the University of Cambridge and Jeremy Bloxham of Harvard University suggest that the dynamolike motions of molten metal in the core that generate the magnetic field may involve enormous, slowly rotating columns of sinking or rising fluid. The columns are parallel to the earth's axis of rotation; their positions may be determined in part by temperature variations within the inner layer of the earth's mantle (the thick shell of the earth that lies between the thin crust and the core).

Halley, Cook and the others had measured the strength and direction of the magnetic field at various points on the surface of the earth. With the aid of mathematical techniques developed for the purpose, Gubbins and Bloxham converted the historical measurements into maps of the shape and intensity of the field at the surface of the earth's core. By accumulating measurements made at several distinct times, they were able to show how the field at the surface of the core has changed.

In analyzing the changes they found several features that had not changed. Among these features are four roughly circular regions in which an intense concentration of magnetic field lines enters or exits from the surface of the core; the positions of these four regions (in relation to the positions of the continents on the surface of the earth) have been constant for more than 250 years. Two of the regions of high flux are in the Northern Hemisphere, under Siberia and western Canada. They are separated by about 120 degrees of longitude. The other two are in the Southern Hemisphere, directly "underneath" the northern pair; in other words, they are at the same longitudes as the other pair and at about the same distance from the Equator.

This arrangement suggested to Gubbins and Bloxham that the circular regions of high intensity represent the tops and bottoms of slowly rotating columns of fluid. Such columns might be expected to form naturally because of Coriolis forces (swirling forces generated by the rotation of the earth). Within each column fluid is sinking from the core's surface toward the plane of the Equator. The lines of magnetic force are "trapped" in the moving fluid, and so as it sinks it pulls flux lines with it, concentrating them within the cylinder. At the equatorial plane the fluid becomes mixed irregularly with the rest of the fluid that makes up the core.

The relative positions of the two columns—they are separated by one-third of the earth's circumference—suggests that there might be a third column, separated by 120 degrees from each of the others. That would put it under the Atlantic Ocean, a region in which the field at the surface of the core has changed rapidly over the past few centuries. Perhaps the vigorous motions in the core that have caused the field to change so rapidly have also dissipated the column of fluid that would otherwise have formed there. Equations describing the formation of the columns suggest that there should also be three columns of fluid rising from the equatorial plane toward the core's surface rather than sinking; each column of rising fluid would be between two columns of sinking fluid. Columns of rising fluid would tend to dissipate lines of magnetic force at the surface of the core, creating regions of very low flux. Gubbins and Bloxham have found evidence of such regions.

Why are the core columns stationary with respect to the earth's mantle? The core and the mantle are not tightly coupled, and rotating formations of fluid in the core would normally be expected to drift westward as the earth rotates. Gubbins and Bloxham think the answer may lie in the lower mantle, at its boundary with the core.

They note that seismic data, analyzed by Adam M. Dziewonski and John H. Woodhouse of Harvard, reveal cold spots within the mantle directly over the regions of sinking fluid. Perhaps the cold spots, by cooling some of the core fluid, are driving the downward convection of fluid in the core. Cold spots in the mantle can have only a very weak influence on the core, but if the fluid in the core is likely to establish a convective flow anyway, even a very small influence from the mantle could affect the positioning of the convective cells.

Getting Small

Pluto, the outermost known planet in the solar system, has always been recalcitrant to observation. Both its distance (about four and a half billion kilometers at its closest approach to the earth) and the fact that it reflects so little light have put a limit on knowledge about the planet. Pluto was not even discovered until 1930, and it took another half century before the planet was found to have a satellite. Until recently not even Pluto's size was precisely known; most estimates placed the planet's diameter at about 4,000 kilometers.

It is therefore understandable that when the configuration of Pluto, its satellite and the earth is propitious for gathering information on the Plutonian system, astronomers cannot let the opportunity pass unexploited—particularly since it arises only twice in the 248-year orbit of the planet. Two German astronomers have taken advantage of just such an opportunity to make telescopic measurements of the planet's size and mass as well as those of its satellite. Their results indicate that Pluto is only about half as big as it was thought to be.

Since the discovery in 1978 of Charon, Pluto's satellite, planetary astronomers have recognized that the satellite's orbital motion could be used to calculate the dimensions of both bodies of the Plutonian system: by observing how the brightness of the system varies as Charon passes in front of and behind Pluto (events respectively constituting an eclipse of Pluto and an occultation of Charon), the diameter and mass of the two bodies could be estimated. The only problem was that, from the point of view of an observer on the earth, Pluto and Charon would not be properly aligned for such measurements until the mid-1980's. Actually it took until 1986 before the celestial bodies were precisely lined up. That is when Manfred Pakull and Klaus Reinsch of the Technical University of Berlin made the appropriate light-intensity measurements at the European Southern Observatory on Mount La Silla in Chile.

Their observations and calculations have yielded the most accurate values for the size of Pluto and Charon. The astronomers report that Pluto's diameter is 2,200 kilometers and Charon's is 1,160 kilometers. Assuming that Pluto and Charon consist of the same material—probably a mixture of rock and frozen gases—Charon accounts for as much as 13 percent of the total mass of the system. In contrast, the earth's moon is about 3,470 kilometers in diameter and represents no more than 1 percent of the overall mass of the earth-moon system.

MEDICINE

Key to Alzheimer's?

A genetic defect implicated in at least one type of Alzheimer's disease has been tracked to chromosome No. 21—the chromosome associated with Down syndrome, a major cause of mental retardation.

Although it is often thought of as a disease of old age, Alzheimer's strikes members of some rare families when they are in their fifties or even younger. Two separate lines of research reported in *Science* suggest a possible underlying disorder in familial Alzheimer's disease that might also account for cases—the great majority—that are not now known to be inherited.

One of the two approaches was genetic mapping. Investigators from France, Italy and West Germany collaborated with Peter H. St. George-Hyslop and James F. Gusella of the Harvard Medical School and others to assemble four detailed pedigrees of familial Alzheimer's, one of them covering eight generations. By analyzing the pattern of inheritance of DNA in surviving family members with markers that bind at known points on a single chromosome, the workers were able to define a narrow band on chromosome 21 that probably included the site of a genetic error causing the disease. An extra copy of chromosome 21 is present in victims of Down syndrome, and older Down-syndrome patients often develop a condition like Alzheimer's. Besides the clinical effects they share, both groups have abnormal deposits in the brain of an unusual protein called amyloid beta.

The second line of research attacked the problem from a different direction. Groups headed by Dmitry Goldgaber of the National Institute of Neurological and Communicative Disorders and Stroke and by Rudolph E. Tanzi and Rachael Neve of the Harvard Medical School synthesized short strands of pseudo-DNA designed to bind to the gene for amyloid beta protein. After surmounting several obstacles both groups found that the final versions of their probes bound to chromosome 21. In other words, the gene was close to (although not necessarily coincident with) the defect mapped by St. George-Hyslop and Gusella. Similar results were reported in the *Lancet* by Nicholas Robakis of the New York State Institute for Basic Research in Mental Retardation.

Amyloid beta protein appears to derive from a larger protein that is found in many body tissues. Because parts of the precursor are strongly conserved in evolution, it seems likely that it does an important job; a paper published in *Nature* suggests that it may be a cell-surface receptor. The defect that leads to amyloid beta protein's being deposited as characteristic plaques in the brain of Alzheimer's-disease and some Down-syndrome patients might be in the gene that codes for the protein itself or in some other gene that affects it, and a major effort to separate the two possibilities will now be getting under way.

Unfortunately the new knowledge does not in itself offer hope of any cure or palliative for Alzheimer's or Down syndrome in the near future. One prospect that can be foreseen, however, is a genetic probe with which to identify individuals in affected families who are likely to develop Alzheimer's; Gusella estimates that a test might be available in from three to five years. Whether such a test could predict the disease in people outside known affected families will depend on whether some subtle genetic predisposition is found to underlie those cases too.

Doctored Data

Like the Italian, Flemish and Dutch masters of painting, distinguished research scientists at major institutions often associate their names with work in which they did not take direct part. The practice may be justifiable, but it is also risky. In the case of the painter an inferior work can debase the coinage of his posterity; in the case of the scientist it can lead to association with fraud. Although there is not even a suspicion that the lead investigators involved in such situations have knowingly participated in deceit, such incidents do raise serious questions about the atmosphere in which big-time science is done.

At the University of California at San Diego one promising junior investigator allegedly published 68 papers containing fabricated data. At the Dana Farber Cancer Institute in Boston a principal investigator withdrew two papers announcing the discovery of an interleukin when he discovered that one of his coauthors had falsified data. Trust in fellow workers, collaborators and peers "is at the core of the best science," says Gerald R. Fink of the Whitehead Institute at the Massachusetts Institute of Technology. Falsification or intentional misrepresentation of data is blatant abuse of that trust. Why do such incidents happen?

A pattern runs through the known cases of fraud, according to Efraim Racker of Cornell University. It was in his laboratory a few years ago that an investigator reportedly substituted iodine for phosphorus in experiments designed to clarify the difference between malignant cells and normal cells. Racker characterizes individuals involved in fraud as being brilliant enough to make key discoveries honestly. By choosing to falsify data bearing on important findings, they almost ensure their eventual unmasking because their experiments are certain to be subjected to repeated efforts at duplication. The cleverness with which such individuals cover their tracks can make their deceptions hard to detect. Still, notes Philip A. Sharp of M.I.T., "it is very difficult to fabricate large bodies of data," and an investigator "close enough to the experiment" can catch most tampering.

How close, though, is close enough? A primary scientist should examine at least most of the raw data and choose what fraction is suitable for publishing, says J. Stephen Shaw of the National Cancer Institute. Peer interaction within the laboratory should be encouraged, according to Timothy A. Springer of Dana Farber.

Three of the most important factors limiting the critical examination of data by leading investigators are large laboratory size, a layered structure of supervision and pressure to publish. These are also factors that tend to put stress on vulnerable junior workers. There can be no one ideal laboratory size, but many investigators think from 20 to 25 doctoral students are the most any one person can supervise. A situation in which intermediate-level scientists stand between the head of the laboratory and the doctoral students who generate the data is traditional in medical schools, according to Robert A. Weinberg of the Whitehead Institute; it is an "invitation to disaster" because it lessens the principal scientist's involvement in data collection. Kenneth L. Brigham of the Vanderbilt University School of Medicine maintains, on the other hand, that the intermediate level can be an advantage, adding an additional check on the consistency of data.

Pressure to publish can filter into any laboratory. Daniel Nathans of the Johns Hopkins University School of Medicine thinks "it is one of the responsibilities of a lab head to protect junior people from the pressure to publish and to encourage an atmosphere in which the most creative work sets the tone of the lab."

Sometimes the number of papers on which an investigator's name appears can signal a warning. In the San Diego case the rate of publication (nearly one paper every 10 days) seemed in retrospect to have merited suspicion. Yet, although it should be detrimental to a worker's reputation to "dribble papers everywhere," Sharp says, there is no magic number that signifies the limit of one individual's creativity. According to Baruj Benacerraf, head of Dana Farber, an arbitrary cap on publication would do science no good. Investigators tend to agree, though, that institutions might do well to sever the perceived link between quantity of papers and prestige or promotion.

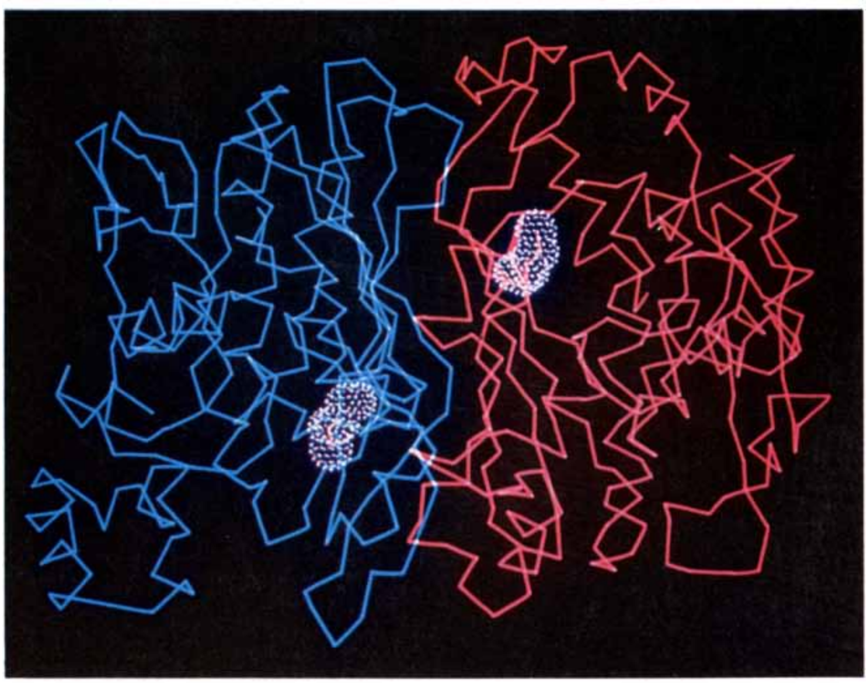
That link may have an effect on the coauthors unwittingly implicated in what turns out to be doctored work. In a recent issue of *Nature* Walter W. Stewart and Ned Feder of the National Institute of Arthritis, Diabetes, and Digestive Diseases examine the case of John C. Darsee, who was found to have altered data in papers he published while at the Emory University School of Medicine and the Harvard Medical School.

Stewart and Feder argue that some of Darsee's coauthors, including his principal investigator at Harvard, Eugene Braunwald, were not involved enough in the research to detect shoddy workmanship, let alone fabrication. It may be hard to set a precise limit on the number of papers a senior investigator can deal with, but it is clear, according to one National Institutes of Health worker who did not want to be named, "that any principal researcher who is named on over 150 papers a year does not know precisely what is going on."

Designer Therapeutics

What do malignant tumors and infectious diseases have in common, besides unpleasant consequences for their hosts? One answer is rapid DNA replication, either in the tumor cells or in the pathogen. Thwart the means by which DNA replicates, and cancers and pathogens can no longer spread. The difficulty is to find a way of doing so that affects only the disease process or at least does not disrupt other cell functions.

Two important anticancer drugs, methotrexate and 5-fluorouracil, work this way. Methotrexate interferes with the enzyme dihydrofolate reductase; 5-fluorouracil is changed in the body into a chemical that puts out of action the enzyme thymidylate synthase (TS). Both enzymes catalyze reactions that must proceed in order to make the DNA precursor deoxythymidine. Because tumor cells multiply faster than most normal cells, the agents are more damaging to malignant tissue than they are to normal tissue; they also tend to spare vital processes dependent



THYMIDYLATE SYNTHASE ENZYME is composed of two subunits, shown in red and blue. White stippling shows where the enzyme binds to its substrate. The computer image was generated by Robert M. Stroud of the University of California at San Francisco.

on RNA because RNA incorporates uridine as a building block rather than deoxythymidine.

Precise knowledge of the molecular structure of an enzyme, coupled with modeling on a supercomputer of drug-target interactions, might allow chemists to tailor-make anticancer and antimicrobial drugs, thus improving on the traditional combination of intuition and experiment. One recent group of papers in *Science* illustrates how precisely enzyme-substrate interactions can now be calculated. Using matched pairs of synthetic inhibitors for the enzyme thermolysin, investigators led by Peter A. Kollman of the University of California at San Francisco were able to calculate the energy associated with one specific hydrogen bond in a complex enzyme-substrate interaction—and then to demonstrate in the laboratory that the calculation was correct.

Exact definition of structure has also been achieved. Workers at the University of California at San Francisco have reported in *Science* the structure of the enzyme TS down to a resolution of three angstrom units, or 30 billionths of a centimeter. The technique used by Larry W. Hardy and his colleagues was X-ray diffraction, speeded by a new detector that allows diffraction patterns to be sent directly into a computer.

The TS that Hardy and his co-workers investigated was from the bacterium *Lactobacillus casei.* Hardy expects it will now be much easier to determine the structure of TS in other organisms. It may even be possible to design inhibitors that bind to a pathogen's TS but not to that of its host. To that end Hardy and his colleagues have begun to model how the enzyme changes shape when it binds to a possible inhibitor.

Joseph Kraut of the University of California at San Diego and Agouron Pharmaceuticals, Inc., a private drug-research company, professes to have already elucidated this relation. Although the details of the work are proprietary, Kraut says that Agouron is already using the structure of bound *Escherichia coli* TS to design possible drugs, exploiting recombinant-DNA techniques to produce the enzyme in large (milligram) quantities.

Knowing the structure of TS does not mean that design of an effective drug will be easy. The detailed structure of dihydrofolate reductase, for example, has been known for years, but there are still no radical improvements on its inhibitor methotrexate, discovered through the old try-it-and-see method.

Yet there are grounds for thinking TS is a better target than dihydrofolate reductase for controlling DNA replication, according to David A. Matthews of Agouron. One is that TS does not participate in other important synthetic pathways, and so blocking it is less likely to wreak unintended havoc. Another reason is that TS, unlike dihydrofolate reductase, catalyzes a rate-limiting step in DNA synthesis and hence should provide more leverage.

Several designer therapeutic drugs may be on the way. Besides TS and dihydrofolate reductase, investigators at pharmaceutical companies are studying such target proteins as renin (part of the body's blood-pressure control mechanism), phospholipase (one form of which plays an important part in arthritis) and elastase (which takes part in the destruction of lung tissue that leads to emphysema).

Immune to Pregnancy

In recent years the wider availability of vaccines against such common diseases as smallpox, cholera and polio has lowered the mortality rate of newborn infants and extended the life expectancy of adults in many Third World countries. In India health officials believe they may have found the solution to a rather different problem—birth control—in yet another vaccine: one that protects women against pregnancy as well as disease. The vaccine has proved effective and safe in animals, rendering them infertile for more than a year. Last summer India's National Institute of Immunology began tests in 88 female volunteers at five different centers. If these clinical trials are successful, the Indian Council of Medical Research may soon authorize a wider program.

The vaccine induces the secretion of antibodies that neutralize a hormone, human chorionic gonadotropin (hCG). Released soon after fertilization of an egg, hCG is vital to the process whereby the egg attaches to the wall of the uterus and begins to develop into an embryo; without hCG the egg is shed during the woman's next menstrual cycle. The vaccine is made of a subunit of hCG. By itself the hCG subunit does not induce antibodies, and so it is linked to an immunogenic "carrier" made of the tetanus and cholera toxoids. (Toxoids are attenuated toxins that serve as vaccines against specific pathogens.)

Research has already shown that when this hybrid antigen is injected into a woman, her immune system produces antibodies that bind to hCG as well as to tetanus and cholera pathogens, according to G. P. Talwar, director of the National Institute. He notes that the tetanus and cholera toxoids were chosen as carriers for three reasons: the diseases are still widespread; the vaccines have been administered to hundreds of millions of people and so their effects are well known; and the effects, while lasting for a year or more, are not permanent.

A version of the hCG vaccine in which the tetanus toxoid alone was the carrier was tested in 65 women several years ago. Four injections of the vaccine induced the production of hCG antibodies in 63 of the women for periods ranging from 300 to 500 days without interfering with their normal ovulation or menstrual cycles or affecting them adversely in any other way, Talwar says. The vaccine did produce different levels of antibodies in different women, and in some subjects the levels were not high enough to prevent pregnancy. For this reason the investigators decided to add the cholera toxoid as a carrier for the recent tests. Talwar explains that jointly the tetanus and cholera toxoids should stimulate a greater immune response.

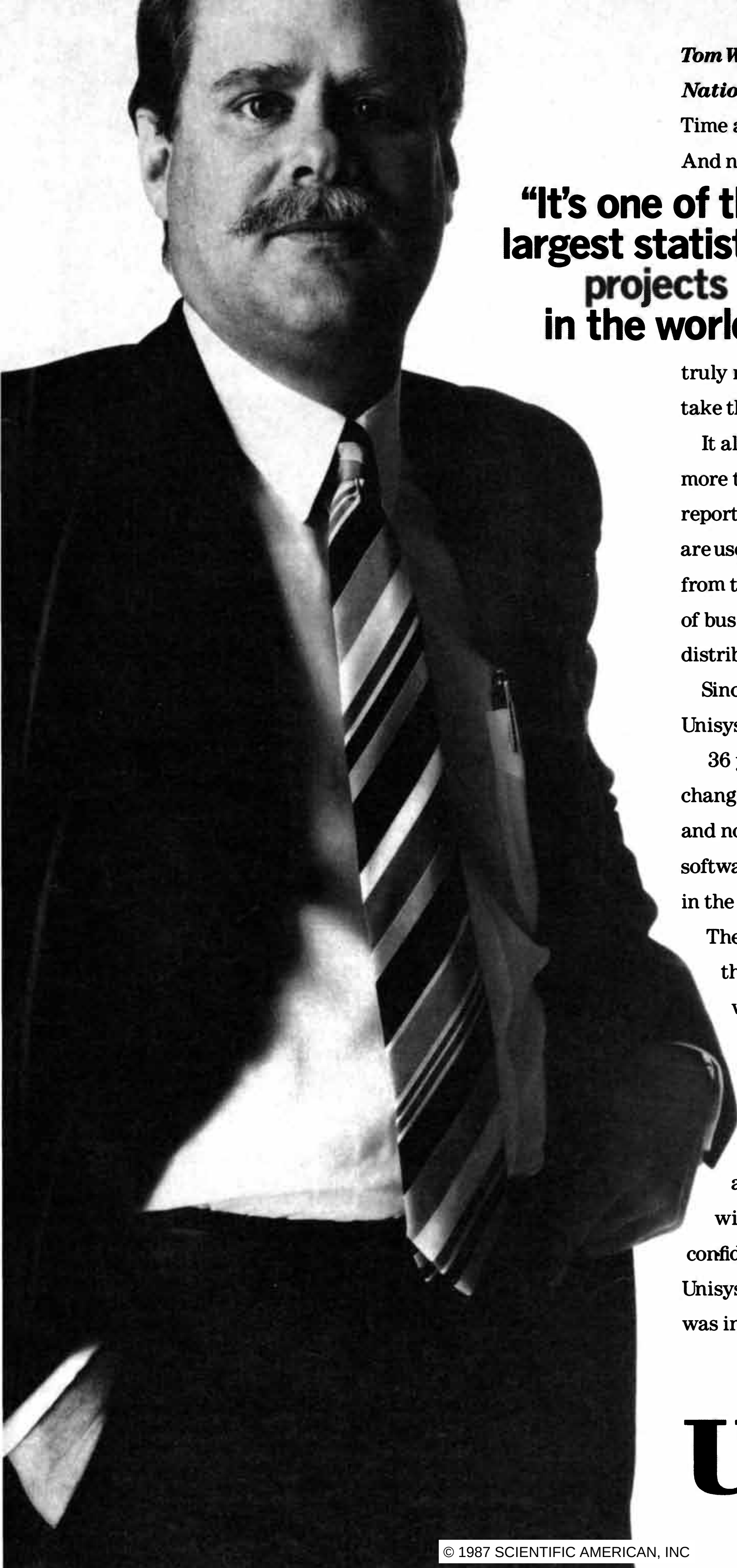
In the current trials the women receive three injections at one-month intervals and a fourth injection six months later. Ideally, Talwar says, the research will produce a vaccine powerful enough for a single injection to prevent pregnancy for up to a year.

AIDS: Breaking Out?

Heterosexual contacts of members of groups at high risk for acquiring AIDS account for 4 percent of all recorded cases of the disease. That category is the only one that is growing faster than the entire population of AIDS patients, according to Timothy J. Dondero of the Federal Centers for Disease Control (CDC) in Atlanta. Does this presage a spread through the U.S. heterosexual population comparable to the attack rate in the male homosexual population?

Because AIDS has a long latency period, the best strategy for seeing what the future holds is to analyze how newly infected individuals who have yet to show symptoms were exposed to the disease. If a growing proportion of new cases were found to be not in any of the established risk groups, it would be a sign that the virus was breaking out of its stronghold among intravenous-drug users and homosexual and bisexual men.

Unfortunately information on preclinical infection is hard to acquire. Whereas 90 percent or more of the AIDS cases that meet the definition of the disease proposed by the CDC are reported to the agency, simple infection with the AIDS virus, as revealed by the



presence of antibodies in the blood, is not recorded on a national level.

Of the scant data on infection that are available, some of the best come from blood donors who test positively for AIDS antibodies and from actual or would-be members of the military who are excluded for the same reason. Would-be blood donors who have had homosexual relations are asked not to donate, and yet some donors test positively. Roger Y. Dodd of the American Red Cross says preliminary data from a questionnaire study of such donors made a year ago suggest that "in most cases careful questioning will reveal traditional risk factors."

The same pattern is observed in studies of would-be military recruits and former members of the military in Colorado, one of five states where it is mandatory that positive antibody tests be reported. According to John B. Muth, director of the El Paso County Health Department, every member of a group of 21 men in Colorado Springs who were retired from the armed services because they were infected was found, in detailed and sympathetic counseling, to have participated in homosexual activity or, in a few cases, intravenous-drug use. Similar results emerged from a small study of would-be recruits conducted by Thomas M. Vernon, executive director of the Colorado Department of Health.

In New York City, Rand L. Stoneburner of the Health Department has established a study designed specifically to act as an "early warning" if AIDS starts to become "free-floating" among heterosexuals. Stoneburner is looking for the presence of AIDS antibody among heterosexuals who are regular patients at a clinic for sexually transmitted diseases. So far all cases among 50 patients have been traceable to known risk categories.

There are no signs so far that AIDS has established itself beyond this perimeter, and the surge of cases among heterosexual contacts of those at high risk can probably be traced to the disease's long latency. It is also true that the rate of increase in the number of cases has been declining. None of this implies an improving outlook. AIDS spreads rapidly. The declining rate of increase could mean merely that the population of susceptible high-risk individuals is becoming saturated.

No Calcium Fix

Calcium supplementation has been attracting much attention as a means of preventing osteoporosis, a bone disease of older people—women in particular—marked by the demineralization of bone and increased susceptibility to fractures. Two recent studies cast doubt on calcium therapy. They show that large doses of calcium given to women just past menopause, when bone loss is most rapid, have little or no effect on the rate at which bone density declines. On the other hand, both studies support the efficacy of estrogen therapy. Administration of the female sex hormone has long been known to slow the rate of bone loss significantly but has been resisted by some clinicians because large doses of estrogen alone can increase the risk of uterine cancer.

In the *New England Journal of Medicine* Bente Riis, Karsten Thomsen and Claus Christiansen of Glostrup Hospital in Copenhagen describe a double-blind clinical study in which 43 women in the early postmenopausal period were divided into three treatment groups. Each group received either estrogen, 2,000 milligrams of calcium a day or placebos. After two years bone-mineral content had decreased significantly in the groups receiving calcium or placebos but not in the group receiving estrogen. Additional measurements showed that estrogen prevented bone loss throughout the skeleton, whereas calcium had only a minor effect in slowing demineralization, and that effect was only in compact cortical bone: the shafts of the long bones.

A team headed by Bruce Ettinger of the Kaiser Permanente Medical Center in San Francisco has published a similar study in *Annals of Internal Medicine.* A daily calcium supplement of 1,000 milligrams (added to an average of 750 milligrams of dietary calcium) had no effect on the loss of trabecular bone, which is the spongy bone in the spine and at the ends of long bones. Nor did the team see a significant effect on the rate of cortical-bone loss.

Why not combine estrogen and calcium rather than trying to choose between the two? Ettinger's group did just that and found that giving calcium halved the amount of estrogen needed to prevent bone loss. Ironically, less estrogen may not be better, according to Ettinger. The hormone's role in preventing heart disease (by influencing the metabolism of cholesterol in the liver) may be even more vital than its effect on osteoporosis—and that role is dose-dependent.

Estrogen remains controversial. Yet combining estrogen and progesterone, Ettinger says, "completely obviates" the risk of uterine cancer and reduces the risk of fracture by 50 percent. "The vast majority of women will benefit from taking estrogen during the early postmenopausal years," he maintains. Speed is essential, because right after menopause a woman loses from 5 to 6 percent of bone mass annually from her spine.

Ultimately there may be no simple remedy for osteoporosis. Estrogen and calcium are part of the picture, but so are parathyroid hormone and perhaps such trace minerals as manganese and copper. In addition the role of the various agents depends on a person's age. For example, according to Ettinger calcium intake is less important than estrogen between the ages of 50 and 65, but it becomes important again between the ages of 65 and 80, after bone loss has slowed.

TECHNOLOGY

Great X-pectations

As molecular analysis and miniaturization become the hallmarks of modern technology, the search for tools to visualize minuscule structures becomes increasingly urgent. One of the most promising methods has finally come of age. X-ray microscopy, a technique proposed more than 90 years ago, is now proving its potential in disciplines ranging from biology to materials science.

When Wilhelm Roentgen discovered X rays in 1895, he realized the rays can "see" molecular detail because their wavelengths are extremely short—shorter by several orders of magnitude than those used in light microscopy. This powerful resolution, however, could not be harnessed without some means of focusing X-radiation. In principle a focusing device called a Fresnel lens would do the trick. The lenses are etched so that opaque ridges emanate in concentric circles from the center; the ridges are carefully spaced so that they force the radiation that slips between them to bend toward a common focal point.

Fresnel lenses for visible light were easy to construct, but in order to work with X rays the lenses had to be barely 60 millionths of a meter wide and made of gold, one of the few materials that can influence X-radiation. Until four or five years ago manufacturers could not produce Fresnel lenses to such tight specifications. Then the drive to reduce the size of semiconductor chips led to the development of microfabrication techniques accurate enough to make tiny Fresnel lenses as well.

Research in X-ray sources advanced concurrently with progress in X-ray lenses. For most applications the rays must be very intense and be confined to one wavelength. Today the standard source is a synchrotron, a type of particle accelerator, fitted with an undulator that boosts X-ray intensity and controls the wavelengths emitted. Utilizing this technology, workers at IBM and the Brookhaven National Laboratory recently achieved a resolution of 500 angstrom units (one angstrom unit is equal to one-tenth of a billionth of a meter).

An X-ray microscope produces images by beaming focused radiation at a sample; some structures absorb the X rays and others pass them on. Detectors measure either the transmission or the absorption of radiation. Transmitted X rays make "shadowgraphs," on photoresistant materials, that can be examined with an electron microscope. Alternatively, lenses can condense the rays into a small beam that scans a sample point by point; then a computer assembles and enlarges the separate transmission patterns in a composite image.

X-ray absorption prompts characteristic fluorescent emissions from a chemical element, and so the microscopes can convey the elemental composition as well as the structural features of a sample. Hence the unobtrusive probe can map the concentrations of trace elements in living tissue. It will also enable geologists to study the ancestry of rocks by revealing their mineral constituents and help engineers to locate cracks and flaws in materials under stress.

Because the microscope can accept whole, living and untreated cells as subjects, biologists can observe directly intracellular activities and responses to light, heat and chemical changes. The electronics industry that gave birth to Fresnel lenses likewise stands to benefit: X-ray microscopy can generate chemical profiles of the interfaces on semiconductor chips—junctions that have become more troublesome and more difficult to visualize as chips shrink in size.

A Polyurethane Proboscis

Given the somewhat anthropocentric outlook of *Homo sapiens,* it is not surprising that robotic arms and grippers are constructed of rigid bars with hinged or swiveling joints. Could robots be modeled after a more flexible natural manipulator—an elephant's trunk, perhaps?

To explore this possibility a team led by James F. Wilson of Duke University has built a two-foot-long, pneumatic, polyurethane trunk. Its repertory is limited: its most impressive feat involves snatching a half-pound object off a table and lifting it to a nearby shelf. Nevertheless, Wilson believes offspring of this "compliant robot" could do certain tasks, particularly those requiring heavy lifting without great precision, more cheaply than stiff-limbed robots.

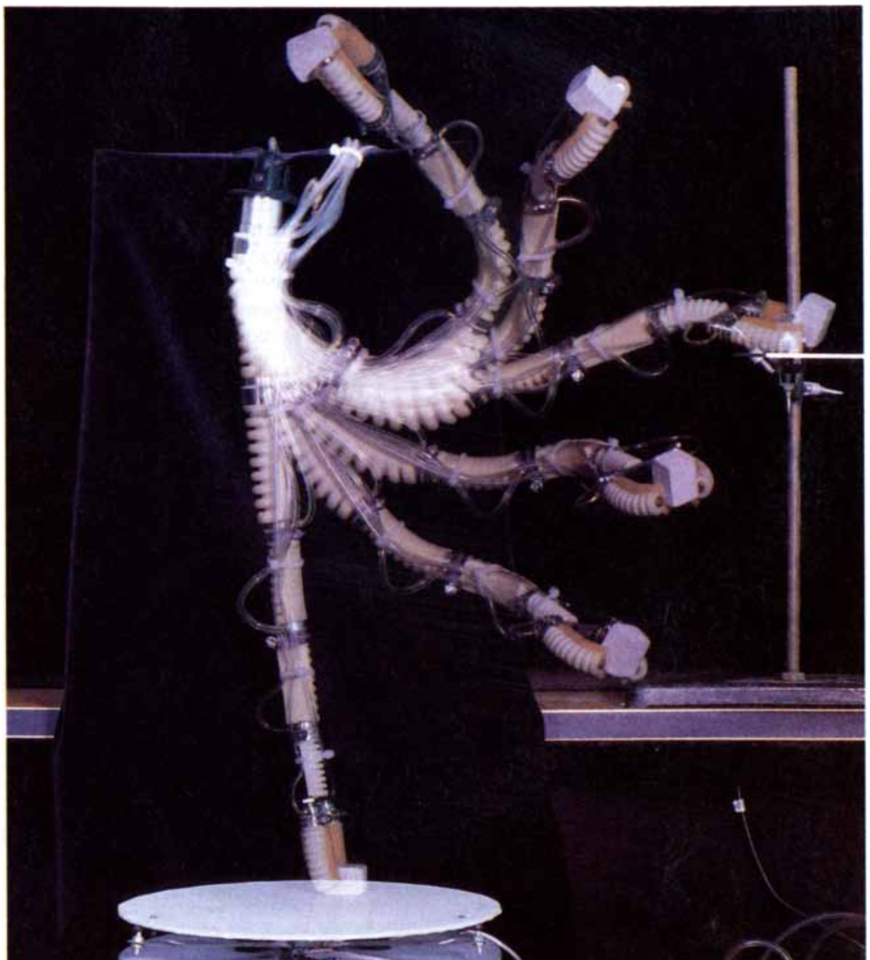
Before building the device Wilson and his colleagues spent several years studying what they call hydrostats: fixed-volume "bags" of watery tissue such as tongues, tentacles and worms, as well as trunks. Hydrostats perform a wide variety of tasks through the interaction of three types of muscles: those that traverse the length of the hydrostat and make it shorten as they contract; those that encircle the hydrostat and, as they squeeze it, make it lengthen, and those that spiral around the hydrostat and can cause it to twist.

In Wilson's first trunk the "muscles" work not by contracting but by expanding. The muscles consist of six half-bellows—flexible tubes corrugated on one side—that are joined end to end. Two more half-bellows at the end of the trunk serve as a gripper. The trunk is connected to an air compressor by means of a pneumatic controller that, on orders from a computer, regulates the flow of air into and out of each half-bellows.

Perhaps the greatest limitation of the design is that each tube, unlike a section of an elephant's trunk, can bend in only one direction. A blast of air causes the corrugated side to expand and the tube to bend in the opposite direction. Orienting the corrugation in different directions enables the artificial trunk to move in three dimensions. Wilson hopes that adding helically corrugated sections that twist under pressure will further increase the trunk's flexibility.

Wilson thinks it may be possible to build compliant robots that can lift twice their own weight. Typical steel robots, he notes, can lift only one-fourth of their weight. Since compliant robots have virtually no moving parts, they should also be more robust than conventional robots, particularly in dirty environments.

The U.S. Defense Advanced Research Projects Agency, which funds the research, declines to reveal what it might do with compliant robots. Wil-



FLEXIBLE PNEUMATIC ROBOT plucks an object off a table, lifts it high and deposits it on a shelf. The robot, built at Duke University, is powered by computer-controlled blasts of air. The sequence in this stroboscopic photograph took less than five seconds.

son suggests they could tackle back-breaking chores in a wide range of settings—for example, stacking bags of cement on a construction site, handling crates in a warehouse or loading artillery shells on a battlefield. Compliant robots might even walk on compliant legs. Wilson says he is designing a platform that can scurry about on polyurethane tubes.

Water Wings

An investigation of nature's wings and fins by an aeronautical engineer may have helped Dennis Connor's 12-meter yacht *Stars & Stripes* to sail past *Kookaburra III* in the America's Cup races off Australia. The work done by Cornelis P. van Dam of the University of California at Davis also challenges long-held notions about the design of aircraft wings; it could even spawn aircraft whose wings curve like a crescent moon.

Van Dam was inspired three years ago while doing graduate research at the National Aeronautics and Space Administration's Langley Research Center. There workers explore nature for innovative designs—asking, for example, whether the notch in a shark's tail fin or the splayed feathers of a vulture's wing tip confer some hydrodynamic or aerodynamic advantage. Van Dam began to wonder why the wings of many birds and the tail fins of many fishes curve backward.

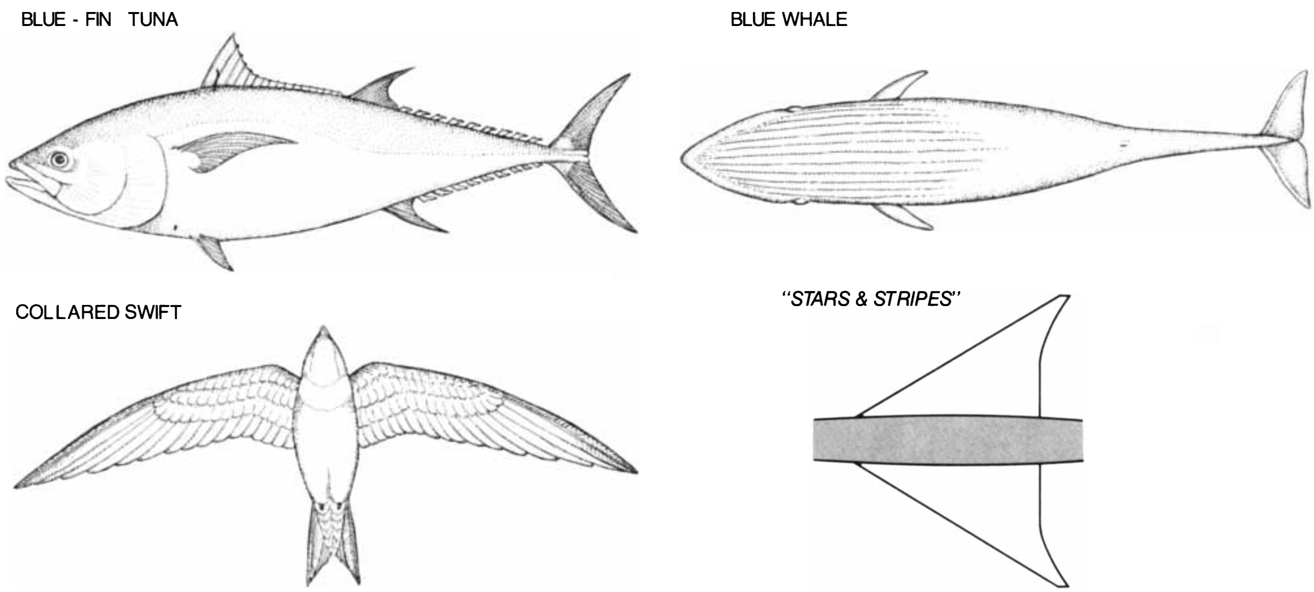
Classical aerodynamics theory, as promulgated by the German physicists Ludwig Prandtl and M. M. Munk in the 1920's, presents a very different picture of the ideal wing: one that generates the most lift (or forward thrust, in the case of a fish's tail fin) for a given amount of drag. The classical wing is roughly elliptical and juts straight out from the fuselage, at least at speeds well below Mach 1. (As a jet aircraft nears the speed of sound a shock wave forms on its wings, greatly increasing drag; it is to reduce this drag at high speeds that the wings of most jet aircraft are swept back.)

Van Dam decided he would analyze the comparative lifting efficiency of curved-back wings with an advanced computer model developed in 1982 by a NASA investigator. The technique divides a wing into panels, calculates the pressure on each panel and then the complex interaction between the localized pressures. The results were clear: a wing shaped like a half-moon or a whale's tail, with a curved front edge and a straight rear edge, can outperform the "ideal" wing; a crescent-shaped wing, whose front and rear edges both curve backward like a swift's, can work better still.

Van Dam believes the improvement is related to the way vortexes form as low-pressure air from above the wing and high-pressure air from below it swirl together behind the trailing edge. Classical theory holds that these vortexes flow back from the wing in a horizontal sheet, but van Dam's calculations suggest that the vortexes swirl vertically as well as horizontally, particularly at the tip of the wing. A wing that is crescent-shaped or one that has swept-back tips, he maintains, enhances the vertical swirling; the wing can also function like a sail, converting some of the vortexes' energy, which normally tends to induce drag, into forward thrust.

A team designing *Stars & Stripes* learned of van Dam's work two years ago, according to Charles W. Boppe of the Grumman Aerospace Corporation, a member of the group. Boppe and his colleagues were looking for ways to improve on the revolutionary winged keel that had helped *Australia II* win the previous America's Cup races off Newport in 1983; *Australia II*'s wings gave it more lift and thus a greater ability to tack upwind than Connor's boat, *Liberty*. After some reworking of van Dam's theory, the *Stars & Stripes* team designed wings with straight, swept-back front edges and with rear edges that curve back sharply at the tip. Although the effect of the wings on the boat's overall performance is small, Boppe says, "you're still talking about a factor that can win the race."

Wing shapes based on van Dam's work have been tested in wind tunnels at NASA's Langley center. "They work," according to Bruce J. Holmes, who supervised van Dam's research at Langley. Holmes notes that whereas the nature of the materials of which natural or artificial wings are built necessarily imposes constraints on their design, the advent of strong new lightweight materials such as carbon-epoxy compounds has reduced the constraints considerably. The Boeing Company, Lear Siegler, Inc., and Burt Rutan—designer of *Voyager*, the first plane to fly nonstop around the world without refueling—are now investigating the potential of wings whose tips curve back.



SWEPT-BACK WINGS AND FINS provide greater lift or thrust than those that jut straight out. A wing shaped like a whale's tail, with a curved forward edge, confers some advantage; a doubly curved wing shaped like a swift's or the tail fin of a tuna is better still. On the basis of this finding, engineers redesigned the wings on the keel (seen here from below) of the yacht *Stars & Stripes*.


NOW ONE OF YOUR MECHANIC'S TOOLS IS ANOTHER MAN'S MIND.
At GM, we think of computers in human terms. Because today, we have begun to program human knowledge—and logic—into a computer. It's called artificial intelligence.
So, even when an engine expert retires from GM, his mind can still work for you. His lifetime of experience can go into a computer. The computer can then dispense this invaluable knowledge not only to GM engineers, but to your mechanic, too. That way, he'll be able to use this computer to help him service your car.
With artificial intelligence, a lifetime of experience can now last a hundred lifetimes. Because at GM, our quest for knowledge never tires.
Or retires.
THE GM ODYSSEY: SCIENCE NOT FICTION
GM

Antiviral Therapy

How can one kill a virus and not the host cell in which it is physically and functionally incorporated? New antiviral drugs exploit the subtle molecular contrasts between virus and host

by Martin S. Hirsch and Joan C. Kaplan

In 1941 the introduction of penicillin transformed medicine, giving physicians a powerful, versatile weapon against infectious bacteria. More than 40 years later there is no comparable panacea for viral infections. The reason, paradoxically, is related to the simplicity of viruses. A drug that kills a pathogen without poisoning the patient must be able to distinguish the invader from the patient's own cells. Because bacteria are relatively complex organisms, they have many metabolic differences from mammalian cells and can reproduce outside the cells of their host. They are functionally and physically distinct, which simplifies the problem of finding substances that attack them specifically.

The simplicity of viruses, in contrast, makes them able to replicate only by physically invading a cell and coopting its biochemical mechanisms to make new viral proteins and genetic material. Because their replication cycle is intimately connected with the functions of the host cell, they present fewer unique biochemical features for selective attack. Indeed, until the beginning of the 1960's the problem of distinguishing viral functions from cellular ones in order to attack only the former was thought to be insurmountable. Instead the main strategy for controlling viral infections was (and to a large extent it still is) the development of vaccines, which do not attack a virus directly but forestall infection by stimulating the immune system in advance.

Over the past two decades, however, accumulating knowledge of viral replication has made it possible to define events that are unique to viruses and to identify compounds that might interfere with those events. Tests in the laboratory followed by carefully designed clinical trials have confirmed the promise of a few such compounds. The new antiviral drugs have already transformed the treatment of several diseases. Less than a decade ago a newborn child with widespread herpes-simplex infection or an adult with herpes encephalitis (infection of the brain) had only a 20 to 30 percent chance of surviving; today early and aggressive drug treatment of either disease can more than double the odds. Laboratory workers and clinicians are now bringing the lessons of the past 20 years to bear on a new challenge: devising therapies for AIDS (acquired immune deficiency syndrome), a deadly viral disease first recognized in 1981.

Modes of Infection

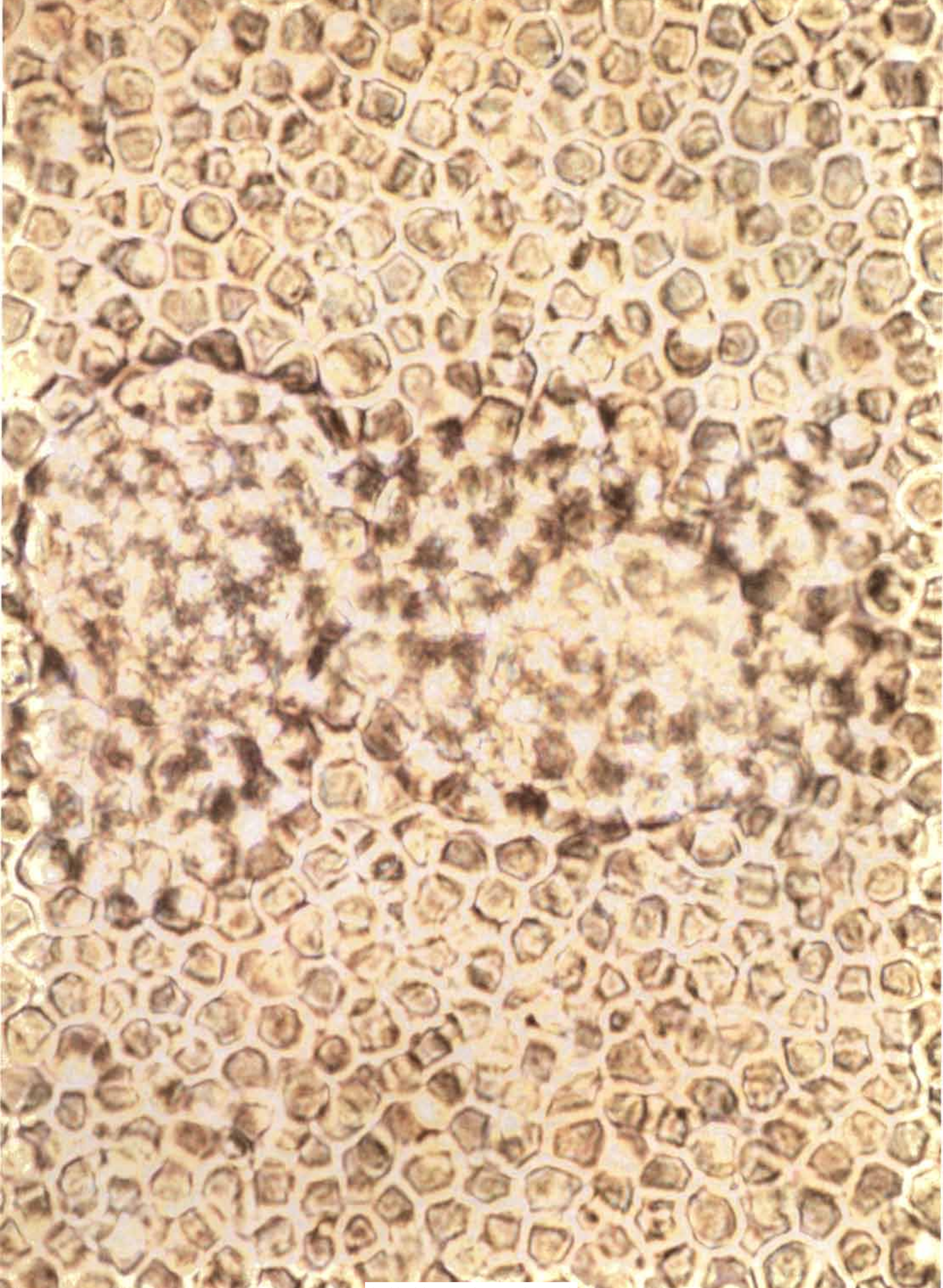
Attempts to treat viral infections must take into account the variety of ways in which viruses interact with host cells. Any virus infecting a cell binds to the surface membrane, enters the cell and sheds its protein coat. The genetic material, which may be a single- or double-strand molecule of either DNA or RNA, carries genes specifying the few enzymes and structural proteins the virus needs in order to interact with the cell and reproduce itself. The host's biochemical machinery expresses the genes during later steps in the viral replication cycle.

Those later steps determine the type of infection that ensues. In a lytic infection, which is characteristic of many DNA and RNA viruses, including those that cause the common cold and polio, the virus replicates by inducing the cell to copy the viral genetic material and make viral proteins. A multitude of new virus particles take shape, and the infected cell is lysed: it bursts, releasing the particles. Lytic infections often spread rapidly throughout the population of vulnerable cells, destroying them early in the illness. Effective treatment of most lytic infections must therefore start early and requires both prompt diagnosis and an effective drug.

Viral replication in another kind of infection, a persistent infection, does not always kill an infected cell. In a persistent infection new particles are often released gradually; the cell survives intact and continues to divide, although its function may undergo subtle changes. Persistence can take another form, in which a few cells host a lytic infection while others are not affected. In either case the infection may continue for months or years without causing overt disease. Hepatitis *B* virus, for example, can establish infections that persist indefinitely, only sometimes culminating in liver disease, such as cancer or vasculitis. Human leukemia viruses and human immunodeficiency virus (HIV), the cause of AIDS, may also produce symptoms only after a lengthy persistent infection. (HIV is now the accepted name for the virus originally called HTLV-III or LAV.)

In both lytic and persistent infections the virus completes its replication cycle, in the process carrying out a number of unique steps that an antiviral drug might interrupt. A third kind

UNINHIBITED VIRAL REPLICATION has damaged human cells in a culture infected with human immunodeficiency virus (HIV), the cause of AIDS. Components of the viral envelope, synthesized as the virus multiplied, have altered the surface membrane of these *T* cells, HIV's preferred host, causing them to clump into multinucleated structures known as syncytia. In laboratory tests of antiviral drugs for AIDS one measure of a substance's efficacy is its ability to prevent syncytia from developing in an infected culture.



of infection, latent infection, offers no such opportunities. A latent virus does not reproduce itself, and the accompanying biochemical activities are absent. Indeed, such an infection can be difficult to detect, since it is not marked by the death of infected cells or the production of new virus particles. The genetic material of the virus can become integrated into the cell's chromosomes, and during cell division the genes of the virus are reproduced together with the cell's own genes and transmitted to the daughter cells. Under certain conditions latent viruses can be reactivated in a few infected cells, and an active infection develops.

Latency alternating with periods of active viral replication characterizes many herpesviruses: herpes simplex (the cause of a range of infections including cold sores and genital herpes), varicella zoster (the cause of chickenpox and shingles), cytomegalovirus and Epstein-Barr virus. The antiviral drugs that have been developed so far can affect such viruses only when they are replicating. Attacking a latent virus would entail distinguishing its genetic material from the surrounding host genetic material and selectively destroying it. That feat awaits future conceptual breakthroughs.

Developing a Drug

A knowledge of a virus's replication cycle and its interaction with host cells may suggest virus-specific functions that would be vulnerable to attack. The route from that conceptual first step to an effective antiviral drug is a long and costly one. Finding a substance that inhibits a viral function can entail laboratory screening of hundreds of compounds. Early signs of antiviral activity spur further in vitro testing of the candidate compound to gauge its effectiveness more precisely and determine its toxicity.

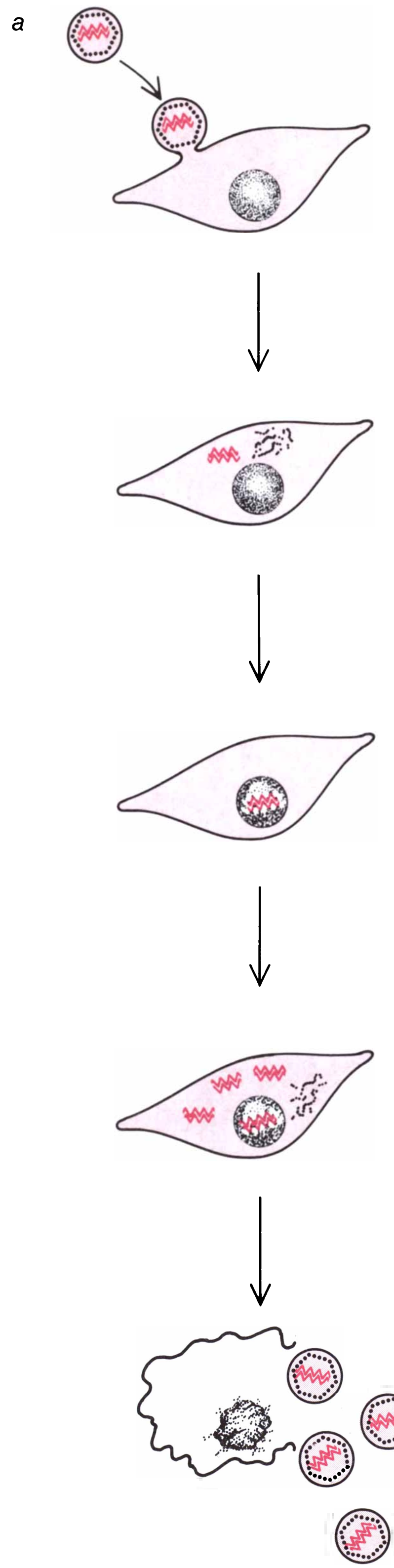
The compound is tested against a variety of virus-infected cells, including cells of human origin. The concentrations of both the drug and the virus are varied. Measurements of virus production, assays of viral proteins and observations of cell damage (such as lysis) caused by viral replication serve to indicate the drug's efficacy. At the same time the host cells are watched for signs of drug toxicity. The virus preparations against which the compound is tested should include isolates from infected patients as well as laboratory strains, which in many cases have been cultured continuously for years and may have accumulated genetic changes that could skew their response to the drug.

If the compound's initial promise is confirmed in the test tube, the pharmacokinetics of the drug—how the body transports, metabolizes and excretes the compound—must be explored, first by giving it to experimental animals. The ability of certain viruses, including HIV, to attack the brain and spinal cord and form a reservoir of potentially infectious virus there has a bearing on some animal studies. The special character of brain capillaries creates a blood-brain barrier, which prevents many compounds from entering the central nervous system. Animal studies offer a chance to observe a drug's ability to cross the blood-brain barrier. They also give additional information on drug toxicity. Often, however, tests in animals cannot demonstrate a drug's effectiveness; many human viruses either fail to infect other species or cause no signs of disease if they do.

The animal studies are followed by preliminary testing in small groups of patients. These phase-I clinical trials yield further information on the pharmacokinetics of the candidate drug and its safety. In particular they serve to indicate the maximum tolerated dose of the drug.

Large-scale clinical trials are the final test of any drug's safety and efficacy. Such trials must be carried out in a carefully chosen population of patients. It is often important, for example, that participating patients represent a similar stage of disease. Since viral replication can reach its peak soon after clinical illness develops, a candidate drug often must be tested in patients whose infection was diagnosed recently.

The variable natural history of most viral infections renders anecdotal case reports of response to a drug almost meaningless. Hence the trials must be rigorously objective and carefully controlled. If there is no established treatment for the infection, a placebo usually provides the point of comparison. The test population is divided, usually at random, into two or more groups, of which at least one is given the drug



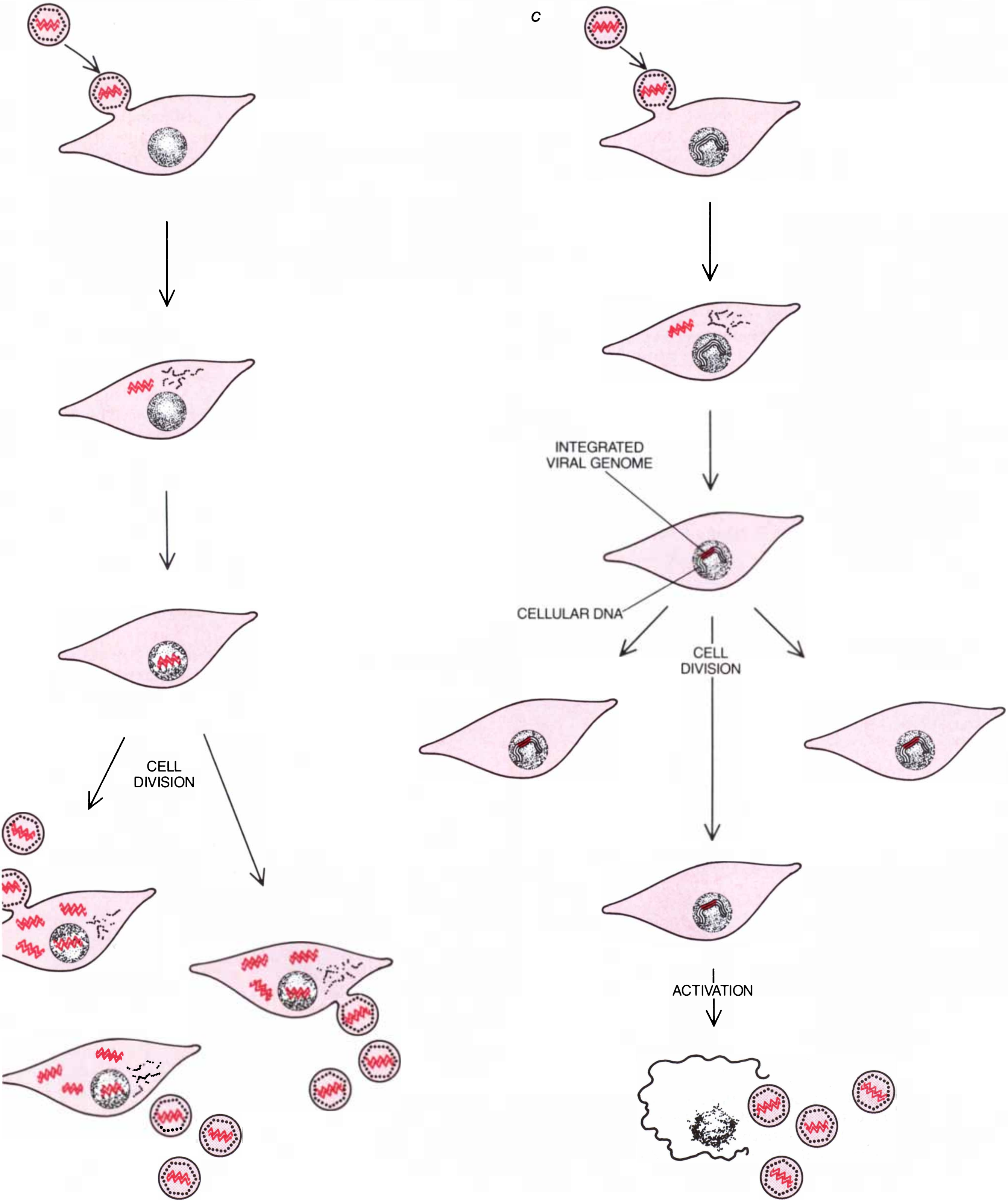
VIRUSES AND HOST CELLS can interact differently, depending on the type of virus and cell. A lytic infection (*a*) results in the death of the infected cell. After the infecting virus sheds its protein coat, its genetic material subverts the cell's biochemical machinery to make new proteins and genetic material, which are then assembled into a multitude of new virus particles. The host cell dies as the particles are released. In a persistent infection (*b*) viral replication proceeds but the cell may survive and continue to multiply, all the while releasing virus. In a latent infection (*c*) viral replication is suspended. The viral genome may become integrated into the cell's own chromosomes, and it is copied and transmitted to daughter cells in the course of cell division. Under some circumstances a virus that is latent can become reactivated in a few of the infected cells.

while another, the control group, receives a placebo. The method is usually double-blind, so that neither the patients nor the investigators know who gets the drug. Such trials are burdensome and expensive, but over the past decade they have firmly established the efficacy of the current generation of antiviral drugs.

A Model Drug

One notable success of rational and satisfactory drug development is the recently licensed compound acyclovir. Acyclovir is a highly specific inhibitor of replication by certain herpesviruses. It has largely replaced earlier generations of antiherpes drugs, which are neither as effective nor as specific.

That acyclovir might inhibit a viral



function, namely the synthesis of new DNA in the course of viral replication, was first suggested by the structure of the molecule. The drug is a nucleoside analogue: its structure mimics the structure of a nucleoside, a precursor of DNA or RNA. A nucleoside consists of one of four kinds of bases—single- or double-ring chemical groups known respectively as pyrimidines and purines—linked to a five-carbon sugar. A nucleoside analogue likewise consists of a base, or a derivative of a base, and a sugar or sugarlike group.

In the normal course of DNA synthesis free nucleoside triphosphates (nucleosides carrying three phosphate groups) react with the end of a growing DNA chain. During the reaction, which is mediated by an enzyme known as DNA polymerase, a hydroxyl (OH) group at the 3′ position on the sugar ring at the chain end forms a bond with the phosphate group at the 5′ position on the incoming nucleoside triphosphate. The other two phosphates are freed, and a nucleotide—a nucleoside carrying a single phosphate—is added to the DNA chain.

As an understanding of DNA synthesis evolved during the 1950's, it became apparent that a chemically altered nucleoside (that is, a nucleoside analogue) might interfere with the process. Nucleoside analogues were first synthesized as potential anticancer drugs; it was thought they might slow or block the accelerated DNA production taking place in rapidly dividing cancer cells. By the 1960's some viruses were known to initiate rapid DNA synthesis in infected cells, and many nucleoside analogues were tested for antiviral activity. Acyclovir, which is an analogue of the nucleoside guanosine, was among the compounds that showed early promise.

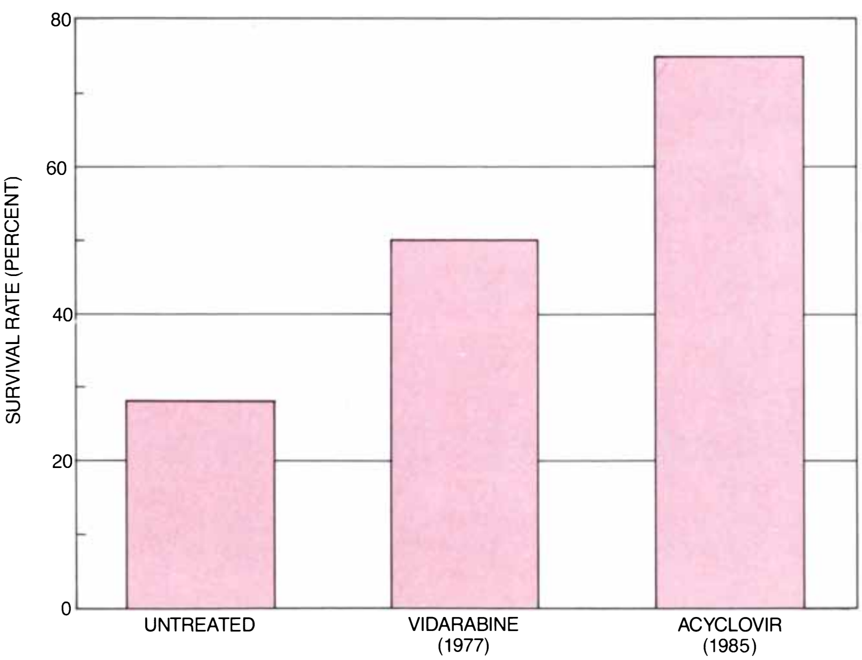
Acyclovir was soon found to have a remarkably specific antiviral effect. DNA synthesis by certain herpesviruses was found to be exquisitely sensitive to acyclovir, whereas the drug had little effect on DNA synthesis in normal cells. Some 10 years of work by Gertrude B. Elion and her colleagues at the Wellcome Research Laboratories in North Carolina revealed the mechanisms underlying the drug's specific antiviral action.

In order to interfere with the synthesis of viral DNA, the workers found, an acyclovir molecule must gain three phosphate groups, just as a nucleoside must be triply phosphorylated before it can take part in cellular DNA synthesis. Each phosphorylation is mediated by an enzyme known as a kinase.

The role of a kinase in the cell is to transfer phosphate groups from one molecule to another. Cellular kinases mediate the formation of acyclovir diphosphate and triphosphate, but the initial step, the conversion of acyclovir into acyclovir monophosphate, depends largely on a kinase specified by a herpesvirus gene. Called thymidine kinase, the enzyme is found only in infected cells. In uninfected cells a cellular enzyme phosphorylates acyclovir to a very limited extent, but the amount of acyclovir triphosphate that results is orders of magnitude smaller than the amount formed in infected cells. Hence the drug is functional primarily in infected cells.

In attacking DNA synthesis, moreover, acyclovir also singles out the DNA of the virus. Because the compound resembles a true nucleoside, a DNA polymerase can bind to activated (triply phosphorylated) acyclovir and add an acyclovir monophosphate unit to a growing DNA chain; two phosphates are released in the process. For reasons that are still not clear, the DNA polymerase of herpesviruses has a much higher affinity for the drug than the cellular polymerase; hence acyclovir monophosphate molecules are added to viral DNA in preference to cellular DNA. As the drug is added it forms a bond with a 3′ hydroxyl group at the end of the chain, just as a conventional nucleotide would. Unlike a true nucleotide, however, acyclovir monophosphate itself lacks a 3′ hydroxyl group. It offers no attachment point for the next link. Once acyclovir monophosphate has been incorporated the growth of the DNA chain is terminated.

Furthermore, whereas a DNA polymerase ordinarily binds only temporarily to the growing DNA chain, so that each molecule of enzyme can catalyze chain growth indefinitely, acyclovir monophosphate makes the association permanent. The viral polymerase is irreversibly bound to the fraudulent nucleotide it has just added to the DNA. Both the growth of the viral DNA chain and the utility of the viral polymerase are ended. The special affinity of acyclovir for viral enzymes thus leads to potent antiviral action with minimum toxicity to the host cell.



RATE OF SURVIVAL in patients with herpes-simplex encephalitis, a viral infection of the brain, has risen dramatically with the advent of effective and specific drugs for herpesvirus infections. In 1977 treatment with vidarabine was shown to increase survival as a result of the compound's ability to interfere with viral DNA synthesis and hence with replication. In 1985 acyclovir was found to have even greater efficacy: features of herpes-simplex replication make the virus singularly vulnerable to acyclovir and related drugs.

The Uses of Acyclovir

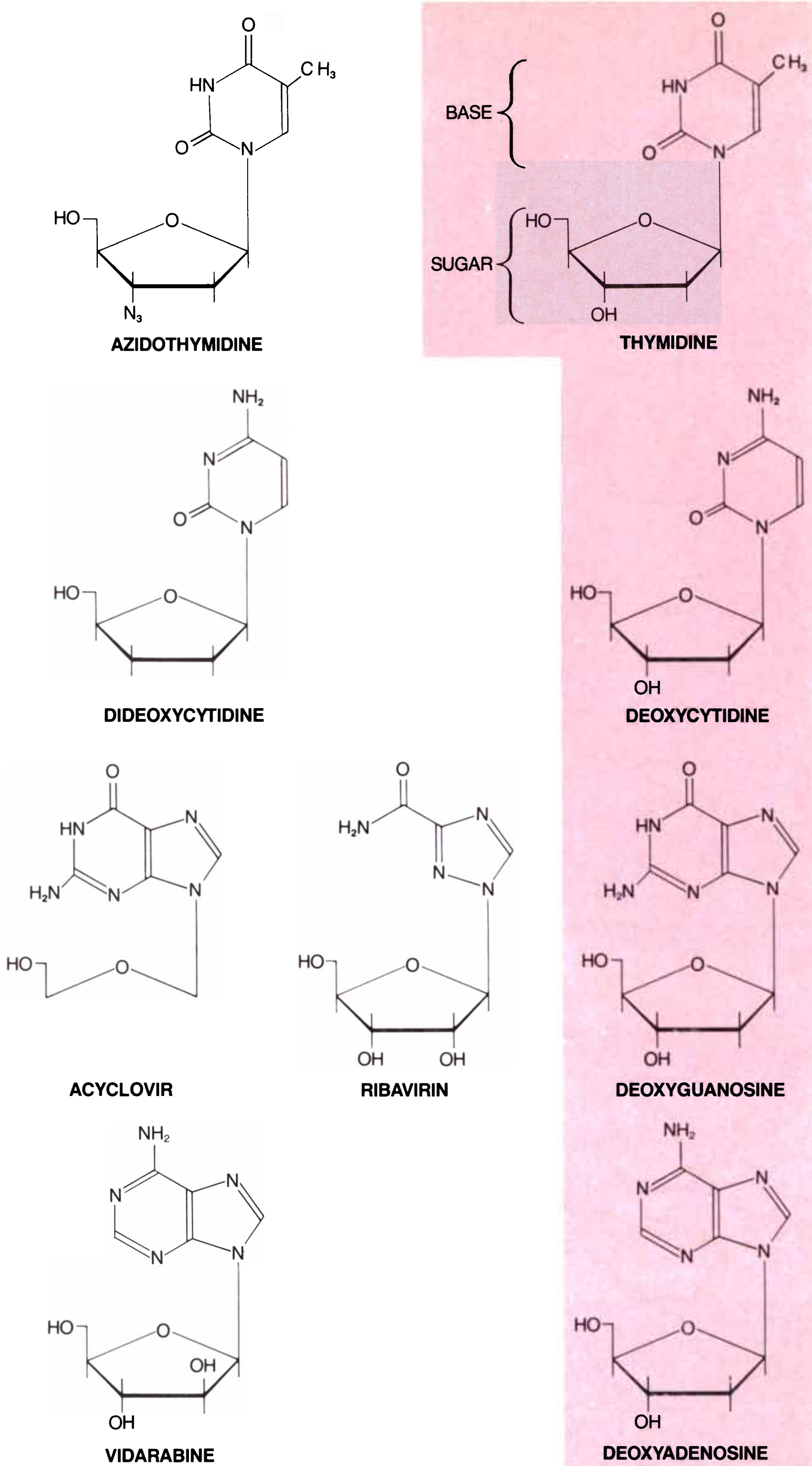
Acyclovir finds its largest application in the treatment of genital herpes-simplex infections, which are believed to affect more than 20 million people in this country alone. The initial attack, that is, the first episode of lytic infection, is usually the severest and the most prolonged and so is the most appropriate target for drug therapy. Treatment with acyclovir (which can be given intravenously, as an ointment or by mouth) relieves symptoms, reduces the amount of infectious virus released from the sores and speeds healing. The treatment does not prevent subsequent attacks or diminish their frequency or severity. Prophylactic courses of oral acyclovir can have a modest impact on recurrent infections, but the cost of the drug and its potential toxicity over the long term do not justify such regimens in most cases.

The drug has also proved to be beneficial in the infections (chickenpox and shingles) caused by varicella-zoster virus. In patients whose immune system has been suppressed for organ transplantation or impaired by disease, acyclovir can sometimes serve as a prophylactic, preventing the herpes-simplex and varicella-zoster infections to which immunocompromised patients are susceptible. The drug can also prevent the reactivation of a latent herpes infection in such patients. Acyclovir is responsible, finally, for the dramatic drop in the death rate from herpes-simplex encephalitis that was seen over the past decade.

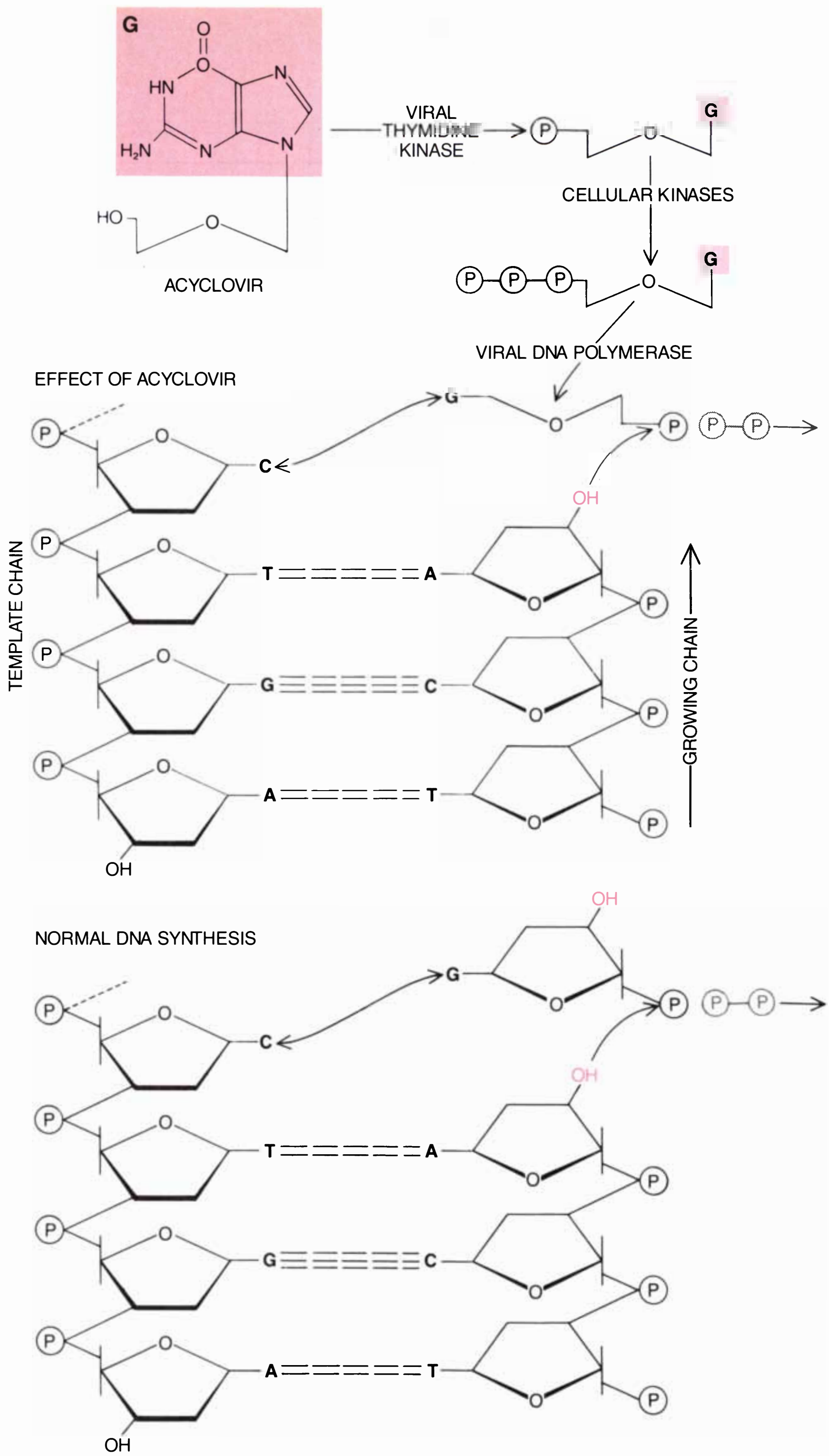
Acyclovir may eventually encounter the same clinical problem that has hampered antibacterial drugs for a number of years: the advent of drug-resistant pathogens. Drug-resistant herpesvirus mutants can be generated in the laboratory by growing the virus in the presence of acyclovir. Resistant strains have also been isolated from patients. Acyclovir resistance can have several genetic origins. One is a mutation inactivating the viral thymidine-kinase gene. Alternatively, the virus may carry a mutation resulting in an altered thymidine kinase, one that has a reduced affinity for acyclovir. In either case acyclovir is not phosphorylated in infected cells and so remains inactive. In animal models thymidine-kinase mutants seem to be less pathogenic than nonmutant strains and also less able to establish a latent infection.

Resistance can also arise from a mutational change in the viral DNA polymerase that makes it less able to interact with acyclovir triphosphate. Laboratory studies suggest that, in contrast to thymidine-kinase mutants, such resistant strains may be just as pathogenic as drug-sensitive virus is. This finding suggests that prolonged administration of acyclovir as a prophylactic or its prescription for trivial infections might favor the appearance of virus strains that are both drug-resistant and pathogenic.

For acyclovir-resistant mutants and for herpesviruses such as cytomegalovirus and Epstein-Barr virus, which are not known to carry the gene for thymidine kinase and hence are relatively insensitive to acyclovir, there



NUCLEOSIDE ANALOGUES are a class of synthetic compounds that owe their name and antiviral effects to a resemblance to nucleosides, chemical precursors of DNA and RNA. The illustration shows a variety of nucleoside analogues that are now prescribed for viral infections or are undergoing clinical trials. The analogues are shown with the corresponding nucleosides (*color*). A nucleoside is made up of a single- or double-ring base linked to a five-carbon sugar; an analogue differs from the nucleoside in large or small features of the base or sugar. An enzyme that normally acts on a nucleoside in the course of viral replication can also bind to the analogue. Because of the differences between the nucleoside and the analogue, however, binding to the analogue can incapacitate the enzyme, thereby disrupting a molecular process crucial to replication. Certain nucleoside analogues have additional features that add to their antiviral specificity and effectiveness.



EFFECT OF ACYCLOVIR in a herpesvirus infection is to inhibit the synthesis of viral DNA. First acyclovir must gain three phosphate groups (*P*). The drug relies on thymidine kinase, an enzyme specified by some herpesviruses, to add the first phosphate (*top*); cellular kinases add the next two. Viral DNA polymerase, the enzyme that catalyzes the synthesis of viral DNA, then binds to the drug as if it were an ordinary nucleoside triphosphate (a building block of DNA) and adds it to a growing DNA chain (*middle*). One of the phosphates on the acyclovir triphosphate molecule forms a bond with the 3′ hydroxyl (OH) group on the sugar at the end of the DNA chain; the other two phosphates are freed. Unlike an ordinary nucleoside (*bottom*), acyclovir itself lacks a sugar ring and a 3′ hydroxyl group. Hence it cannot support further chain growth. Moreover, the viral DNA polymerase, which would normally go on to catalyze further additions to the chain, remains permanently bound in a complex with the DNA and the drug and is incapacitated.

are alternative strategies. A newly synthesized nucleoside analogue, DHPG (dihydroxy-propoxymethyl guanine), shows considerable activity against these herpesviruses both in the test tube and in patients, although the drug's precise mode of action is not known. Vidarabine, a compound that was the mainstay of treatment for life-threatening herpesvirus infections before the advent of acyclovir, can also function in the absence of thymidine kinase. Like acyclovir, vidarabine is a nucleoside analogue that is activated in the cell by triple phosphorylation and then interferes with the synthesis of viral DNA.

Other Antiviral Successes

Respiratory viral infections remain a major cause of illness, but drug development has led to advances in treating influenza and diseases caused by respiratory syncytial virus: pneumonia and bronchiolitis in infants and young children. Two compounds responsible for progress against influenza, known as amantadine and rimantadine, predate the recent conceptual advances in antiviral therapy. Their effectiveness against influenza *A* was discovered in the early 1960's through laboratory screening of a large number of compounds. Even now their precise mode of action is not clear. Some workers have proposed that the drugs may prevent the influenza *A* virus from shedding its coat as it infects a cell or may interfere with early stages in the transcription of the viral genetic material into messenger RNA (mRNA), which then serves as a template for protein synthesis. Only amantadine is licensed in this country; it is most often given as a prophylactic against infection during an influenza outbreak.

The more recent effort to block specific steps in viral replication by means of nucleoside analogues has yielded a drug effective against respiratory syncytial virus. The compound is ribavirin, a guanosine analogue that acts against a broad range of viruses in vitro and has also shown effects in clinical trials against Lassa fever, a viral disease that is endemic in Africa.

Ribavirin's efficacy is thought to reflect at least two biochemical activities. Phosphorylated ribavirin interferes with an enzyme that is crucial to the synthesis of the guanosine precursors of DNA and RNA. The drug's resemblance to guanosine also enables it to inhibit the "capping" of viral mRNA, in which a modified guanosine molecule is added to the end of newly transcribed mRNA. The cap seems to enable mRNA to bind to ribosomes, the structures in the cell on which proteins are made. The capping process is universal, but ribavirin is selective in its inhibitory effect. It inhibits the capping of viral mRNA—and hence the production of viral proteins—far more than it affects the corresponding functions of human cells.

Even so, ribavirin may be toxic to blood cells and can lead to anemia when it is given by mouth. Recent trials have shown, however, that ribavirin can act against respiratory syncytial virus when it is administered as an inhaled aerosol, a tactic that carries the drug directly to the infected tissues but minimizes its absorption and therefore its toxic effects.

Interferons

Antiviral therapy has progressed not only by searching for drugs that inhibit viral replication but also by appropriating one of the body's own defenses: a class of proteins known as interferons. The name was coined in the late 1950's, when it was found that cells exposed to a virus secrete proteins that enable other cells to resist infection. Advances in molecular biology have helped to bring interferons into clinical use, just as they have figured in developing synthetic antiviral drugs.

Interferons were purified in the late 1970's, making it possible to study their activities in detail. Cells produce interferons not only when they are exposed to virus directly but also when they encounter double-strand RNA (a form of genetic material found in many viruses) or certain compounds that stimulate the immune response. The secreted interferon molecules bind to surface receptors on other cells, inducing the target cells to synthesize several new proteins.

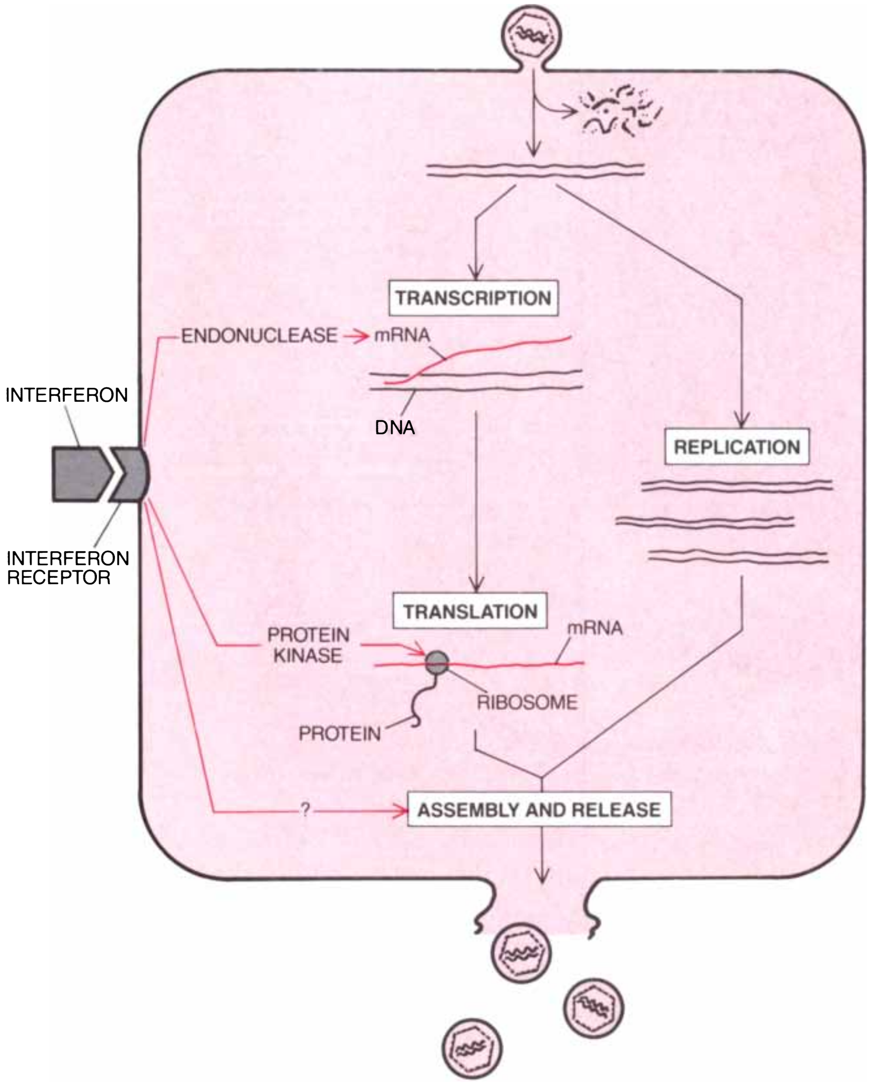
One protein initiates a series of biochemical steps that leads to the activation of a previously dormant enzyme, which then breaks down the mRNA of virus replicating within the target cell. A second protein is a kinase that inactivates the cell's machinery for protein synthesis. Through one or both mechanisms (they vary depending on the kind of cell and the virus infecting it) the virus is prevented from expressing its genes as proteins. In certain infections interferons also have other effects that have not yet been explained at a molecular level. Electron micrographs show a buildup of defective virus in some interferon-stimulated cells, which suggests that interferons can inhibit the assembly and release of new virus particles.

The therapeutic promise of interferons could not be tested until large quantities became available. Several years after the proteins were first purified the techniques of recombinant DNA enabled workers to clone human interferon genes and insert them into bacteria and yeast cells. The genetically modified bacteria or yeast cells were then capable of producing human interferons in virtually limitless quantities [see "The Purification and Manufacture of Human Interferons," by Sidney Pestka; SCIENTIFIC AMERICAN, August, 1983].

The ensuing clinical studies suggested that interferons can serve in preventing or treating certain herpesvirus infections and perhaps also in treating some warts, caused by papilloma virus. In patients with chronic hepatitis *B* and viral respiratory infections the utility of interferons remains unclear. Because of the inconclusive results of many clinical trials and the variety of toxic side effects they have shown, interferons remain in something of a clinical limbo, their value as antiviral agents still uncertain.

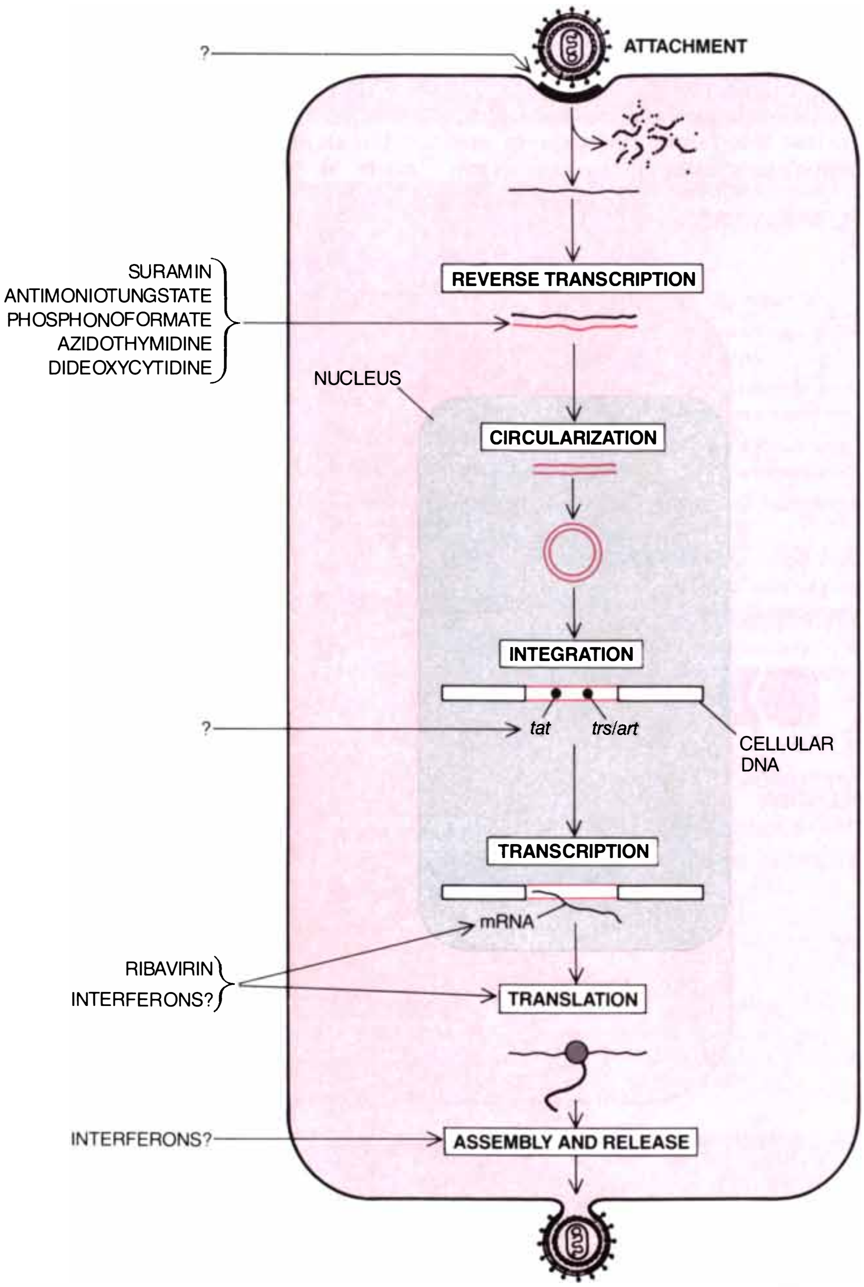
The Current Challenge: AIDS

For all of the progress that has been made in developing antiviral drugs



INTERFERONS bind to receptor molecules on the surface membrane of an infected cell (*left*) and may then exert their antiviral effect at several junctures in the viral replication cycle. One is the transcription of the viral DNA: the process by which viral genetic information is copied into messenger RNA (mRNA). Interferons can initiate biochemical events that activate an endonuclease, an enzyme that breaks down the viral mRNA. Another possible target of interferons is translation: the production of proteins from the mRNA template, which occurs on organelles known as ribosomes. Interferons can stimulate the cell to make a protein kinase that inactivates a factor taking part in translation. A third target may be the final step in viral replication: the assembly of new virus particles from viral proteins and replicated genetic material and the release of the particles. The mechanism by which interferons may block virus assembly and release is not known.

and delivering them to the site of infection with a minimum of toxic side effects, antiviral therapy is still in its infancy. Some viral diseases, such as rabies and brain infections caused by the arboviruses, have so far proved untreatable. Progress against other infections, such as the common cold and viral diarrheas, has been meager. The greatest challenge that antiviral therapy faces as it develops into a mature science is the urgent need for a treatment that will have lasting effectiveness against AIDS.



MOLECULAR TARGETS for existing and hypothetical AIDS drugs punctuate the replication cycle of HIV, the causative virus. The virus infects its host cell by binding to specific cell-surface molecules; an effective antiviral drug might work by blocking attachment. Within the cell the virus sheds its protein coat and converts its genetic material, RNA, into DNA in a process known as reverse transcription—the target of drugs now in use or under study. The DNA then enters the cell nucleus and adopts a circular form, and some of it is incorporated into the cell's chromosomes. The integrated DNA is later transcribed into mRNA, which is translated into viral proteins. Genes known as *tat* and *trs/art* specify proteins that stimulate transcription and translation and might serve as targets for future drugs. An existing drug that interferes with transcription and later processing of mRNA is ribavirin. Interferons too can inhibit transcription, and perhaps also translation and the last event in the cycle: assembly and release of new HIV particles.

AIDS was recognized just six years ago; its causative virus, HIV, was not identified until 1983 [see "The AIDS Virus," by Robert C. Gallo; SCIENTIFIC AMERICAN, January]. The disease has spread at a frightening pace, and more than 30,000 patients have now been diagnosed in the U.S. alone. The number of people in this country who are infected but have not yet developed symptoms is much larger, probably more than a million. Several features of AIDS militate against therapy. The virus's main target is the *T*4 lymphocyte, a white blood cell that marshals the immune defenses. The depletion of *T*4 cells in AIDS causes a severe depression of the immune response, and so an AIDS drug may have to fight the virus without much help from host immunity. Furthermore, the virus also infects cells in the central nervous system, where it is shielded from many drugs that might otherwise be effective by the blood-brain barrier.

Yet the crucial basis for an antiviral strategy in AIDS, an understanding of the viral replication cycle, is now developing. In infecting its host cells HIV binds to specific cell-surface receptor molecules. It penetrates the cell cytoplasm and sheds its protein coat, baring its genetic material: a single strand of RNA. A viral enzyme known as reverse transcriptase accompanies the RNA. In a step characteristic of retroviruses, the class of viruses to which HIV belongs, the enzyme "reverse-transcribes" the RNA into DNA. Ultimately some DNA copies of the HIV genome become integrated into the chromosomes of the host cell.

The integrated viral genome, known as a provirus, may remain latent until the host cell is stimulated, possibly by another infection. The proviral DNA is then transcribed into mRNA, which directs the synthesis of viral proteins. The provirus also gives rise to other RNA copies that will serve as the genetic material of viral progeny. The proteins and the genomic RNA congregate at the cell membrane and assemble to form new HIV particles, which then bud from the cell. Two HIV genes, designated *tat* and *trs/art*, seem to orchestrate this burst of replication, which destroys the cell. They code for small proteins that boost the transcription of proviral DNA and the synthesis of viral proteins.

The most tempting drug target in the HIV replication cycle is reverse transcription, a step that is both crucial to viral replication and irrelevant to host cells. Several drugs have already been shown to reduce the activity of reverse transcriptase in vitro. HIV replication is demonstrably slower in the presence of those compounds, which include suramin (an antiparasitic agent), antimoniotungstate (also called HPA-23), phosphonoformate and a new class of nucleoside analogues called dideoxynucleotides. All are now undergoing clinical trials in patients with AIDS or related disorders.

Another inhibitor of reverse transcription, azidothymidine (AZT), has already proved its clinical efficacy. Like acyclovir, AZT is a nucleoside analogue—it is modeled on the nucleoside thymidine—that originated in the anticancer efforts of the 1960's. AZT interferes with the synthesis of new DNA during reverse transcription by terminating the growth of the DNA chain. Triple phosphorylation by cellular kinases activates the compound, which then binds to reverse transcriptase (for which it has an affinity 100 times greater than its affinity for cellular DNA polymerases) and is incorporated into the DNA chain. At the sugar ring's 3′ position AZT has an azido (N_3) group instead of a hydroxyl group. Incoming nucleotides find no attachment point, and so reverse transcription comes to a halt.

AZT's antiviral effects in vitro are selective as well as potent: the concentration at which it inhibits viral replication is well below the level at which, for example, it begins to impair the immune functions of *T* cells. In the light of encouraging laboratory results a group led by Samuel Broder at the National Cancer Institute carried out preliminary trials of AZT in 1985–86. The workers studied the compound's pharmacokinetics and safety in a small group of patients with AIDS or the cluster of symptoms known as AIDS-related complex (ARC). Large-scale placebo-controlled trials, meant to continue for six months, then began in 282 patients with AIDS or ARC.

By the 16th week of the study an independent review board saw striking differences in death rate between the patients getting a placebo and those receiving the drug; the drug recipients had also suffered fewer infections, had gained weight and had shown an improvement in immune function. As a result the trials were called to a halt and all the placebo recipients were given AZT. More must be learned about the drug, however, before it can be prescribed on a broad scale. It is not clear how long-lasting its effects will prove to be, or whether the drug will be as effective in patients at earlier stages of infection. The prospect of long-term use has also heightened concerns about AZT's considerable toxicity to bone-marrow cells, the precursors of blood cells.

Steps in the HIV replication cycle other than reverse transcription may also be fruitful targets for antiviral attack. Human interferon alpha, which like other interferons can inhibit the production of viral mRNA, the synthesis of viral proteins or the assembly and release of new particles, slows HIV replication in vitro and is now undergoing clinical trials. Ribavirin, which affects protein synthesis as well as other events in the viral replication cycle, has also been tested in patients, and there are preliminary indications that it may delay or prevent the development of AIDS in people who are infected but do not show symptoms.

One can also conceive of antiviral agents that would interfere with the initial step in infection—the attachment of an HIV particle to specific cell-surface receptor molecules—or would attack the protein products of the *tat* and *trs/art* genes, thereby interfering with the expression of other HIV genes. Such agents (for the moment largely hypothetical) might be exquisitely specific and hence less toxic than the antiviral drugs now being developed for AIDS.

Future Antiviral Strategies

The rapid emergence of promising treatments for AIDS is testimony to the new molecular understanding of viral replication and to two decades of clinical experience with other antiviral drugs. How will antiviral therapy advance in the years to come? Better ways to use existing drugs will undoubtedly account for much progress. Assay systems for earlier diagnosis of viral infections will enable the physician to begin drug treatment before extensive cellular damage is done. Means of sending drugs to infected tissues that are now difficult to reach, such as the central nervous system, will bring other infections within range of treatment.

In AIDS and in other severe and prolonged viral illnesses, antiviral therapy may come to rely increasingly on combinations of drugs, each of which would attack a different viral function. An antiviral synergy is one possible advantage of the strategy. Our group, for example, has shown that the combination of a reverse-transcriptase inhibitor such as AZT or phosphonoformate with interferon alpha has a synergistic effect on HIV in vitro.

Such combinations are now being tested in patients, where one might expect additional benefits. Because the dose of each compound is lower, toxic side effects are less likely. The chance that a resistant strain of virus will emerge is also much reduced: the virus would have to develop ways to cope with several drugs at once. Similar advantages might also be had from a regimen in which two or more compounds are given in succession rather than together. In AIDS, combination therapy may come to mean not only prescribing an assortment of antiviral drugs but also administering drugs together with natural immune-regulating proteins in an effort to restore patients' immune responses.

Designing a synergistic combination of antiviral agents, like identifying a promising compound in the first place, calls for close attention to the drugs' molecular mechanisms. In combination some drugs may in fact have antagonistic effects. Markus Vogt in our laboratory has shown, for example, that ribavirin has an affinity for the same cellular enzymes that phosphorylate AZT to its active form. In vitro AZT loses much of its effectiveness when it is combined with ribavirin. Clinical trials of this combination of drugs should therefore be approached cautiously and carried out under controlled conditions.

Molecular studies of viral replication cycles and the mechanisms of experimental drugs will surely remain the major source of progress. As the molecular vulnerabilities of additional viruses are identified, currently untreatable infections will yield to highly specific drugs. Molecular studies of latent viruses may even give rise to drugs selective enough to attack those sequences of silent viral DNA, which are distinguished only by their foreign origin from the host genetic material surrounding them.

A strategy for future drug development may emerge from the study of viruses by X-ray crystallography, which can reveal their structure in atomic detail [see "The Structure of Poliovirus," by James M. Hogle, Marie Chow and David J. Filman; SCIENTIFIC AMERICAN, March]. Computer-graphics programs based on such structural knowledge could yield models to predict how and where a given experimental compound would bind to a virus particle; the information might guide the chemical modification of the original compound or the synthesis of a new one to achieve a molecule that would fit precisely into a particular niche in the virus architecture. The tailormade drug would bind to a virus with high specificity, preventing it from attaching to its host cell or shedding its protein coat. Such a strategy for rational drug design would powerfully augment current approaches to antiviral therapy.

ROOFING TILES made of amorphous-silicon solar cells can be substituted for conventional roofing tiles, thereby providing a means of supplying the building with solar-generated electricity. These tiles were made by the Sanyo Electric Co., Ltd., in Japan.

SOLAR-POWERED HOUSE has been operating for five years in Osaka. The blue panels on the roof, made of amorphous-silicon solar cells, supply the electricity; the overhanging red panels below the roof take care of the heating and cooling of the structure.

Photovoltaic Power

The technology and economics of converting energy from the sun directly into electricity have improved rapidly. Within 15 years megawatt power plants based on solar cells could be in service

by Yoshihiro Hamakawa

It is virtually a certainty that the worldwide demand for energy will increase steadily in the coming decades. One could also foresee, although with a little less certainty because of technological variables, that the suppliers of energy will have to contend with shortages of fossil fuel and with concerns about the polluting effects of those fuels and the possible dangers of nuclear energy. Clearly a power source based on an inexhaustible and nonpolluting fuel could be expected to find a role. It now appears that such a source is at hand in the solar cell, which converts sunlight into electricity.

The technology is not new; solar cells have provided the electricity for space vehicles since the advent of the space age and are now appearing in such consumer items as calculators and watches. What is new is that the technology is approaching the stage at which million-watt power plants can be built. This progress stems from improvements in the efficiency of solar cells at converting sunlight into electricity and reductions in the cost of making them. As a result the cost of photovoltaic power has dropped from $50 to less than $5 per peak watt in the past decade. With further advances it seems probable that large power plants based on solar cells will be in place by the turn of the century.

The formal name for the device at the root of this power-producing technology is the photovoltaic cell. Up to now virtually all such cells have been made of crystalline silicon, but other materials are both possible and promising. They include semiconductors made of such alloys as copper indium diselenide, cadmium telluride and gallium arsenide.

Whatever the material is, the steps whereby the cell converts sunlight into electricity are essentially the same. The focus of interest is the movement of charge carriers in a semiconductor, since electricity—an electric current—consists of a flow of these charge carriers through a circuit. The process can be observed in a solar cell made of crystalline silicon.

The silicon atom normally has 14 electrons associated with it. Four of them are valence electrons, meaning that they are available to interact with other silicon atoms or with atoms of other elements. In a crystal of pure silicon each atom is in valence with four other atoms: it shares two electrons with each of them. One of the electrons is its own and the other belongs to the neighbor.

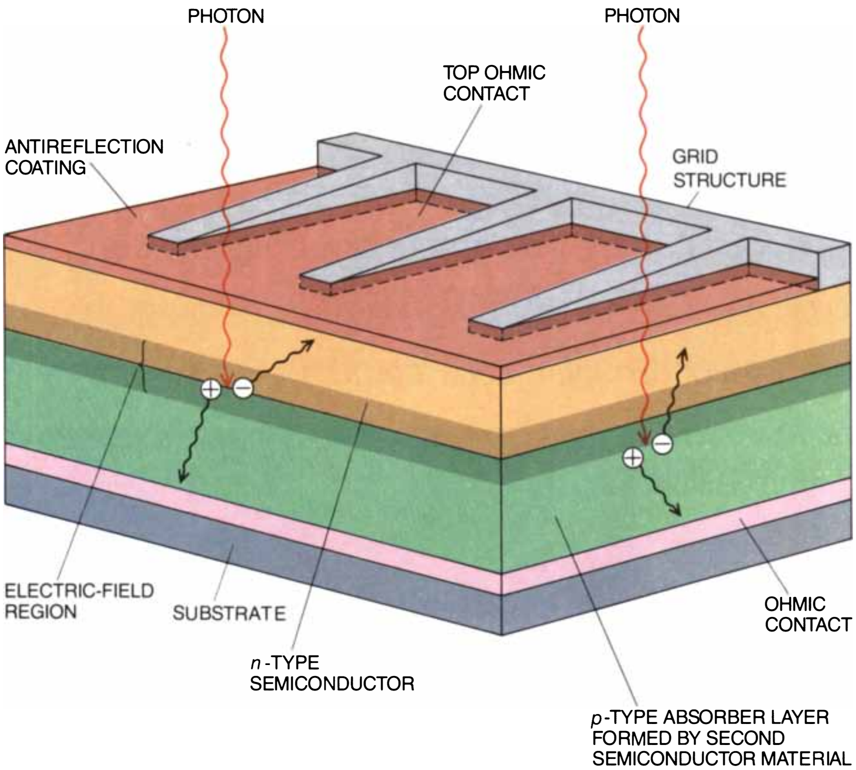
This fairly strong electrostatic bond between an electron and the two atoms it is helping to hold together can be broken by an input of energy. With an appropriate input the electron is lifted to an energy level called the conduction band, where it can move freely and participate in the conduction of electricity. In making such a shift it leaves behind a "hole," which is a place lacking an electron. A nearby electron can be expected to move into the hole, exchanging places with it. Hence both electrons and holes can move within a crystal.

A solar cell is basically a diode of large area. In a diode the passage of current, in the form of free charge carriers such as electrons, is impeded in one direction and facilitated in the other. This result is achieved by the presence of a fixed electric field in the diode, resulting from the fact that the diode is made with two semiconductors of different materials. Such an electric field will propel carriers of a given charge through it and repel carriers of the opposite charge.

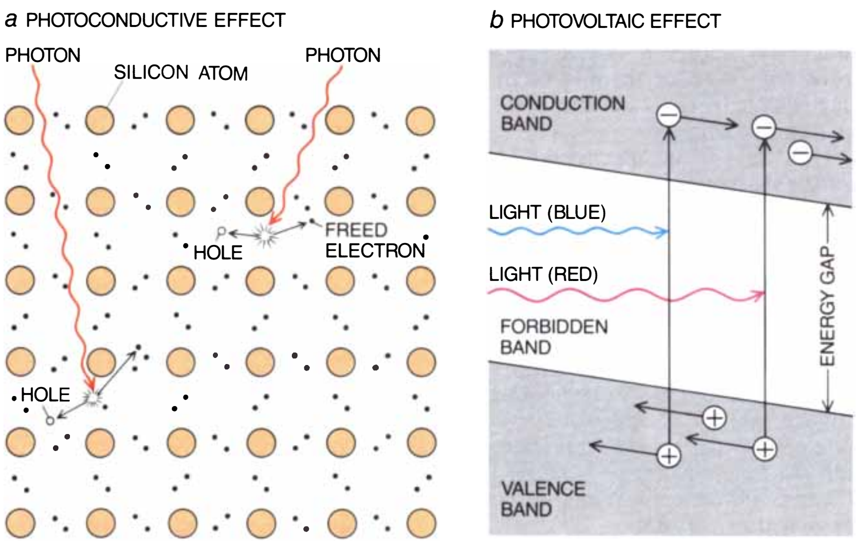
A solar cell takes advantage of this phenomenon. The cell is made with different semiconductor layers to create a diodelike field fixed within the cell. This built-in electric field is positioned as closely as possible to the region where sunlight is absorbed in the device. Sunlight carries energy (about 1,000 watts per square meter of exposed surface). When solar photons strike a semiconductor in a solar cell, they may be reflected, go through or be absorbed. The photons that are absorbed are the ones of potential use.

A photon that is absorbed in a semiconductor may produce heat by jostling the semiconductor material or it may have enough energy to remove an electron from a bound state and lift it to a free state in the material's conduction band. For example, a photon with a wavelength of less than about 1.2 micrometers has enough energy to produce a free electron when it is absorbed in silicon.

The promotion of an electron to the conduction band creates a hole. The hole represents a positive free charge that acts analogously but oppositely to a negative free charge. If a hole is present in a silicon semiconductor, it can move freely because a nearby bound electron can easily jump into the hole's position, effectively moving the hole to the place vacated by the jumping electron.



SOLAR CELL based on crystalline silicon consists of several layers. From the bottom up they are a substrate made of glass or plastic; a metallic layer that serves as a contact in the external electrical circuit; a *p*-type silicon semiconductor that tends to collect "holes," or positively charged particles; an *n*-type silicon layer made in a different way to collect negatively charged electrons; an antireflection coating, and a top-contact grid. The plus and minus symbols represent a hole-electron pair created by the absorption of a solar photon, or quantum of light. Near the *p-n* junction an electric field resulting from the different materials in the semiconductors pushes the electrons upward and the holes downward. This flow of charge makes an electric current through the external circuit.



PHOTOVOLTAIC EFFECT of photons in a crystal of silicon (*a*) results when an incoming photon hits an electron (*black dots*) bound to an atom of silicon. The electron is released and becomes a negative free charge, leaving in its place a positive hole. This phenomenon is the photoconductive effect. In quantum-mechanical terms (*b*) the electrons freed by photons of appropriate energy become more energetic and move from the valence band, or bound state, to the conduction band. If the photoconductive effect occurs near the electric field of a cell, the photogenerated hole-and-electron pair are separated and pushed to opposite sides of the cell. There they can be part of an electric current.

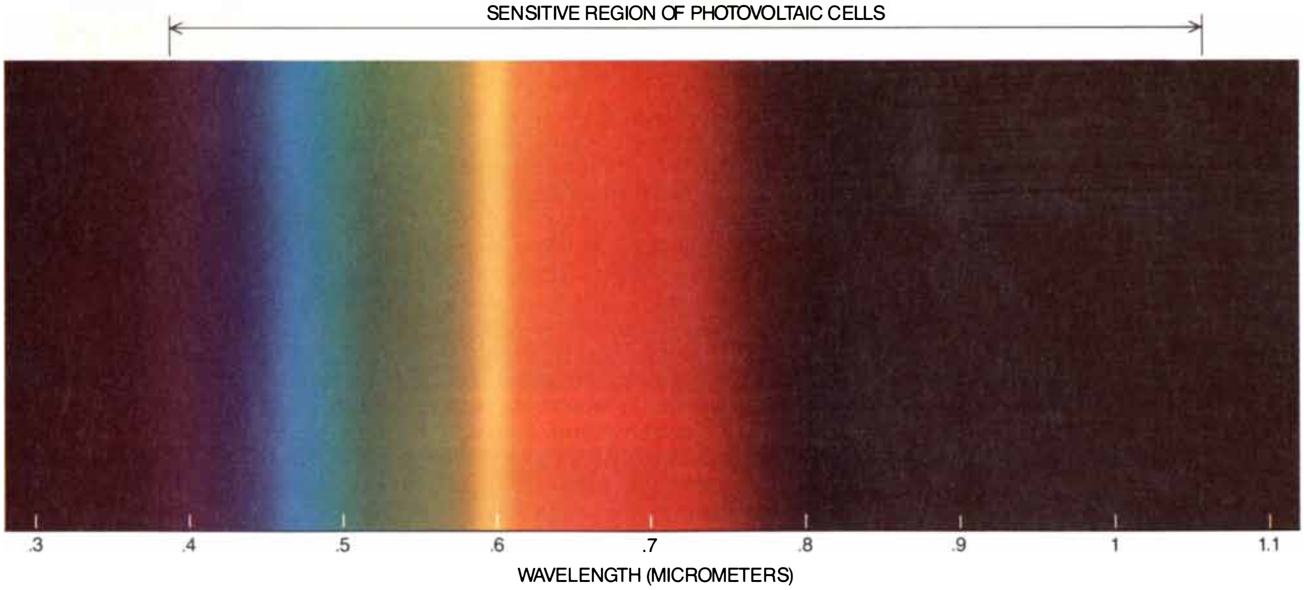
When light is absorbed in a semiconductor, it produces two free carriers: the free electron in the conduction band and the free hole in the valence band. The key process in the solar cell's conversion of sunlight into electricity is that these light-generated free carriers are then driven in opposite directions by the force of the built-in electric field. The holes and electrons near the field provide an example. Suppose the field points in such a way (in relation to these light-generated carriers) as to facilitate the passage of electrons and repel holes. The electrons will be accelerated through the field and the holes will stay near the place where they were generated. Once the free electrons have gone through the electric field they will not return, because the field, acting the way a diode field does, will keep them from coming back.

Hence, as light shines on the cell, an increasing number of positive charges are pushed to the top of the cell and negative ones to the bottom, or vice versa depending on the type of cell. If a conductor is attached to both the top and the bottom, the free charges will flow through it as electricity. Electricity will flow continuously, as direct current, for as long as the cell is exposed to light.

In a typical solar cell a material such as glass or plastic serves to support the cell and is called the substrate. On it a conducting layer such as a metal is deposited as a back contact. Next the light-absorbing semiconductor is deposited. On top of it goes either a different or a modified semiconductor. The interface between the two semiconductors is the location of the critical built-in field. The top semiconductor is often transparent in order for sunlight to pass through it and be absorbed as close to the electric field as possible. This arrangement improves the probability that electrons or holes generated by light will reach the field region and be separated.

The final component of a typical cell is the top contact grid. When the top and bottom contacts are connected in a circuit, useful electricity will flow while the cell is illuminated.

Several other phenomena are of interest in understanding the performance of a well-designed solar cell. Two processes, drift and diffusion, influence the effectiveness with which light-generated free carriers contribute to the flow of current. Free carriers can be generated by light absorbed in the cell, either in the narrow field region or outside it. Carriers generat-



RESPONSE REGION of crystalline-silicon photovoltaic cells begins inside the violet part of the visible spectrum (near .35 micrometer in wavelength) and goes beyond the visible spectrum into the infrared. Solar energy peaks between .5 and .55 micrometer.

ed within the region, which is typically about .5 micrometer wide, are quite effectively separated by the strong field. They are propelled at high speed to the top and bottom of the cell. This is the drift process.

Diffusion applies to free carriers generated outside the field region. It refers to their tendency to bounce about randomly for a time. Any free electron has a certain probability of encountering the built-in electric field. If it does, it will be propelled to the other side of the cell and will be able to contribute to the electric current. High-performance solar cells are designed to maximize this probability.

The thickness of a solar cell, and to a great extent its cost in materials and processing, is determined by its ability to absorb the available sunlight. Two critical factors, the band gap and the absorption coefficient, influence the absorption of light. The band gap of a semiconductor is the characteristic energy level at which it begins to absorb sunlight. Materials such as silicon, gallium arsenide and others employed in solar cells are chosen because they characteristically begin absorbing light at long wavelengths. Hence they are able to absorb a large fraction of the available sunlight.

In addition different semiconductor materials absorb sunlight best at different thicknesses, which can range from 100 micrometers to less than one micrometer. So-called thin-film solar cells are made of materials that absorb in the one-micrometer range, allowing for great reductions in material and fabrication costs. (It takes about 50 micrometers of crystalline silicon to absorb as much sunlight as amorphous silicon can absorb in one micrometer and copper indium diselenide can absorb in .1 micrometer.)

Variations in absorption are due to differences in the atomic structure of different semiconductor materials. Some materials, called direct-band-gap materials, are strongly absorbing; others (indirect-band-gap materials) absorb weakly. Crystalline silicon is an indirect-band-gap material; thin-film materials such as amorphous silicon and copper indium diselenide are direct-band-gap materials. In the indirect type two coincident phenomena are required to promote a valence electron to the conduction band: the light absorption itself, and the favorable interaction of a lattice vibration (called a phonon). In the direct-band-gap materials the light energy alone is sufficient. The difference in probability between the single event and the two events accounts for the order-of-magnitude difference in the absorption of light by the materials.

In working with solar cells one is also concerned about the amount of sunlight reaching the earth. Just beyond the earth's atmosphere the solar constant, as it is called, is 1,358 watts per square meter. By the time sunlight reaches sea level the absorption of solar energy by water vapor, ozone (O_3) and carbon dioxide has reduced the input to about 1,000 watts per square meter. At the earth's surface this solar energy lies mainly in the range of wavelengths from .4 micrometer to 2.5 micrometers; the energy peaks between .5 and .55 micrometer, which is approximately the wavelength of the green color of leaves.

The output of a solar cell is determined by the energy spectrum of the optical absorption in the active area of the semiconductor and by the depth of the *p-n* junction (the interface between the layer that tends to collect positive charges, or holes, and the layer that tends to collect negative charges, or electrons) as measured from the side of the cell facing the sun. An example of the kind of efficiency to be expected from a solar cell is seen in the crystalline silicon cell. Such a cell is sensitive to light in the range from about .35 micrometer to about 1.1 micrometers, encompassing nearly all the visible spectrum and extending into the infrared. Because of losses arising from such factors as heat, reflection and photons that are not absorbed by the semiconductor, the theoretical limit of a silicon cell's efficiency in converting sunlight into electricity is about 28 percent. Further losses within the cell mean that the actual efficiency falls somewhat short of the theoretical maximum.

It is often pointed out that the conversion efficiency of a solar cell is less than half that of generating systems based on nuclear power or fossil fuel. The comparison is nonsensical. At a conversion efficiency of 38 percent a steam turbine fueled by oil is wasting 62 percent of the expensive fuel. (Indeed, the fuel came from sunlight-associated processes far less efficient than a solar cell.) Moreover, the generated energy causes thermal pollution, acid rain and atmospheric contamination. On the other hand, even if a solar-cell installation is only 10 percent efficient in converting sunlight into electricity, it is effectively utilizing an otherwise neglected energy source that costs nothing.

This is not to say that there is no need to try to improve the efficiency of solar cells. Any gains in efficiency contribute to reducing the cost of the output power of solar-cell installations. Hence a good deal of research and development is directed at improving the conversion efficiency (sunlight into electricity) of candidate technologies so that they will produce more power.

An obvious but important step is to apply an antireflective coating to the front surface of the cell in order to increase the amount of nonreflected light and thus the number of photons penetrating to the *p-n* junction. Reflection losses in an uncoated cell can reach 30 percent. Another effort is to employ what is called back-surface-field technology, whereby an electric field is created at the back surface of a semiconductor to push randomly moving free carriers back toward the *p-n* junction in order to enhance the separation of electrons and holes. Through such efforts the efficiency of mass-produced crystalline solar cells has now reached about 14 percent.

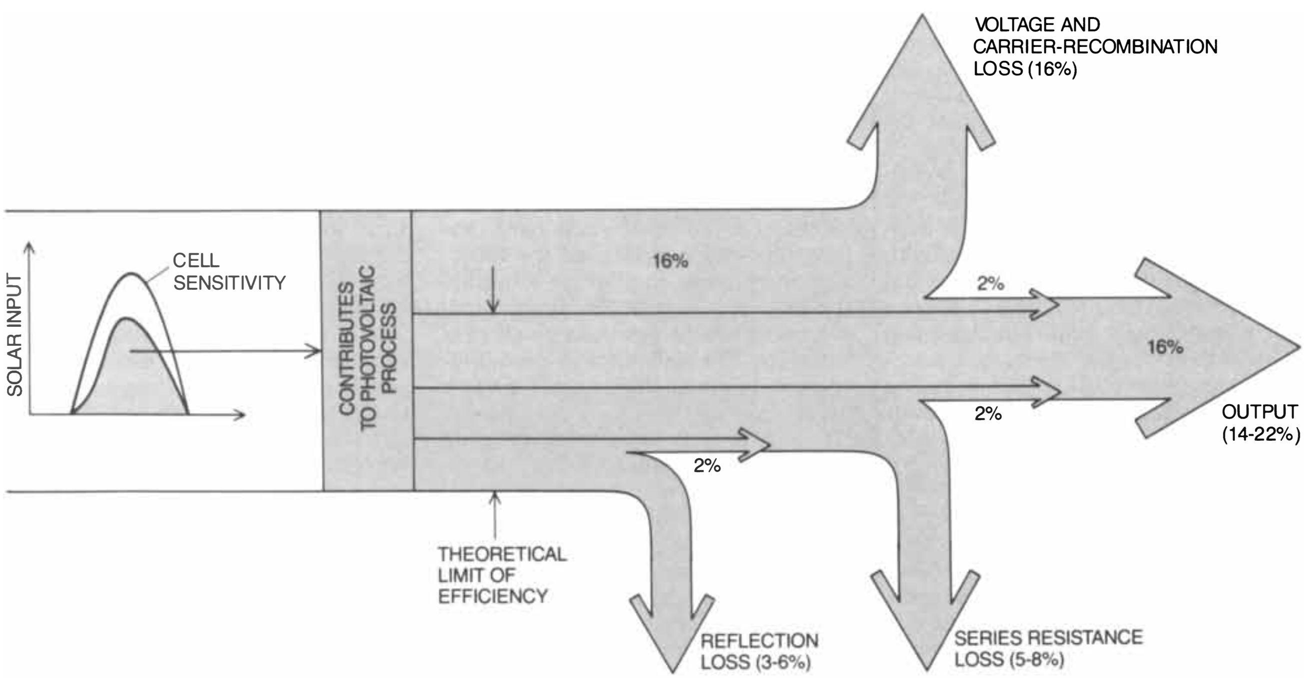
Recently several innovative materials have been developed for photovoltaic cells. One of the most important is amorphous silicon (a-Si), which does not have the regular lattice structure of crystalline silicon. The amorphous structure increases considerably the probability that light will be absorbed and that an electron will be lifted to the conduction band. The material is therefore considerably better than crystalline silicon at absorbing sunlight. Hence it can be manufactured in thin-film form for semiconductors (.5 micrometer thick compared with up to 300 micrometers for crystalline silicon). Because they contain less material, thin-film cells cost less.

A related development is the invention of amorphous silicon carbide (a-SiC). With this alloy the manufacturer can control not only the valency (by doping the material with impurities to supply either more or fewer electrons) but also the size of the energy gap—the amount of incident energy needed to lift an electron to the conduction band—to which the cell responds (by varying the amount of carbon). This discovery opens up an amorphous-silicon-alloy age featuring such new materials as amorphous silicon germanium (a-SiGe), amorphous silicon nitride (a-SiN) and others. Such alloys change the band gap of amorphous silicon so that it can be adapted to different functions in thin-film cell designs.

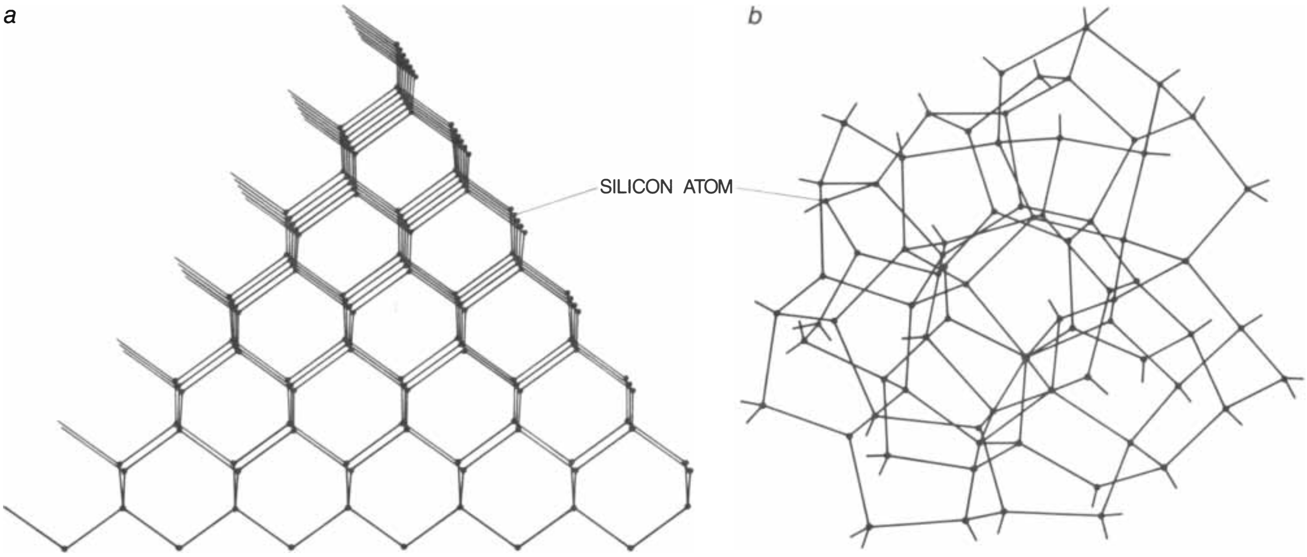
Another advance in thin-film solar-cell technology is to stack cells in order to make more use of the solar spectrum. For example, the top cell might collect blue light, allowing red light to pass on to a cell that is more efficient in that part of the spectrum.

A sealed panel containing a number of cells interconnected at the factory is called a photovoltaic module. Several modules form an array. To collect sunlight more efficiently a solar array can be built into what is called a flat-plate installation, which is mounted on a support structure that can be designed to track the path of the sun throughout the daylight hours. Another technique for collecting sunlight is a concentrator installation, which employs reflectors or lenses to focus sunlight on the solar cells. The two types of installation differ operationally in that a flat plate will work satisfactorily even in the diffuse light of a cloudy day, whereas a concentrator works well only in direct sunlight. The disadvantage of the concentrator installation may be partly overcome by equipping the system with high-efficiency solar cells, including single-crystal silicon and gallium arsenide cells, which are now in the efficiency range from 20 to 26 percent.

The use of a concentrator installa-



ENERGY FLOW in a crystalline-silicon solar cell utilizes in principle about 44 percent of the solar energy to which the cell is sensitive. About 16 percent of the energy is lost by various processes within the cell, leaving a theoretical limit of efficiency of 28 percent. Further losses reduce the actual efficiency of the cell in converting sunlight into electricity to between 14 and 22 percent. The arrows labeled "2 percent" indicate lost energy, as from losses to reflection, that can be re-collected by a well-designed cell.



STRUCTURAL DIFFERENCE between crystalline silicon (*a*) and amorphous silicon (*b*) is indicated. Until recently most solar cells were made of crystalline silicon, which has the regular lattice structure of a typical crystal. The amorphous, or irregular, structure increases considerably the possibility that light will be absorbed, because the photons interact more with the amorphous structure. Hence a cell can be made with less material by means of what is known as thin-film technology. Amorphous silicon is one of several materials recently developed for thin-film cells; other materials include gallium arsenide and cadmium telluride.

tion bears on the question of reducing the cost of solar-cell power plants. Optical devices such as lenses and reflectors can be cheaper than solar cells, and so a power plant employing them may require less of an investment in photovoltaic devices. Moreover, in a concentrator installation the small-area solar cell at the focus of concentration is not the cost driver of the system; therefore it is possible to utilize highly efficient photovoltaic devices that would be much too expensive for a flat-plate installation.

Cost is what has delayed the advent of large-scale generating plants based on photovoltaic power. In 1974 the price of a solar-cell module was $50 per watt of peak power. The cost of power-grid electricity from such a module would have been extremely high: $3 per kilowatt-hour, assuming a 20-year life and adding another $50 per peak watt to cover the costs of other components of the system. At that time the cost of electrical energy generated by fossil fuels and nuclear energy was from five to 10 cents per kilowatt-hour. The sharp difference necessitated a tremendous research effort aimed at bringing down the cost of photovoltaic electricity. The work has encompassed cell materials, cell structure and mass-production processes. As a result of those efforts the cost of a module has come down by one order of magnitude, to about $5 per peak watt. Further reductions, perhaps by another order of magnitude to less than 40 cents per peak watt for thin-film cells such as amorphous silicon, can be expected in the next decade.

Thin-film technology serves to illustrate the possibilities. Cells made of three of the thin-film materials (amorphous silicon, copper indium diselenide and cadmium telluride) have achieved 10 percent efficiency in the laboratory. Several companies in the U.S. and Japan are making commercial modules of amorphous silicon, in some cases from one square foot to four square feet in area. The efficiency of the modules is about 8 or 9 percent. A few companies are commercializing the other two materials, which match amorphous silicon in their potential to achieve the long-term efficiency and cost level required for thin-film technology. Manufacturers of copper indium diselenide have been able to achieve an efficiency of almost 10 percent for modules 100 square centimeters in area. Several companies working with cadmium telluride have been experimenting with electrodeposition as a means of making solar cells at low cost. Some companies have made cadmium telluride modules that have an area of one square foot or more.

The thrust of the work aimed at improving these materials so that they will meet the requirements for producing power at a competitive price is to increase the efficiency to about 15 percent for modules, to develop fabrication techniques that will turn out modules at a cost below $50 per square meter and to ensure that the modules are durable, operating efficiently for up to 30 years in outdoor installations. The fact that thin-film devices have reached a conversion efficiency of 10 percent (from zero only a decade ago) indicates the rate of progress. Several companies profess to have achieved efficiencies of 11 or 12 percent in laboratory cells. In the end the efficiency of working cells will have to be 16 or 17 percent in order to achieve an average of 15 percent in large arrays, because of inherent losses.

Progress toward these efficiencies is steady, and new strategies to achieve them (such as by stacking cells) are being developed systematically. The attainment of large-area fabrication seems to be clearly within sight. Significant cost reductions will result when the new materials go into large-scale production.

As the price of solar cells has come down over the past 10 years, several new applications for the technology have appeared. In addition to the variety of solar-powered electronic devices now available on the retail market, they include water pumps and signals for automobile and railway traffic. Generators that are light in weight and powered by solar cells provide electricity for boats and for land vehicles such as golf carts. As Kenneth Zweibel of the Solar Energy Research Institute has pointed out, the technology should find a number of applications in developing countries.

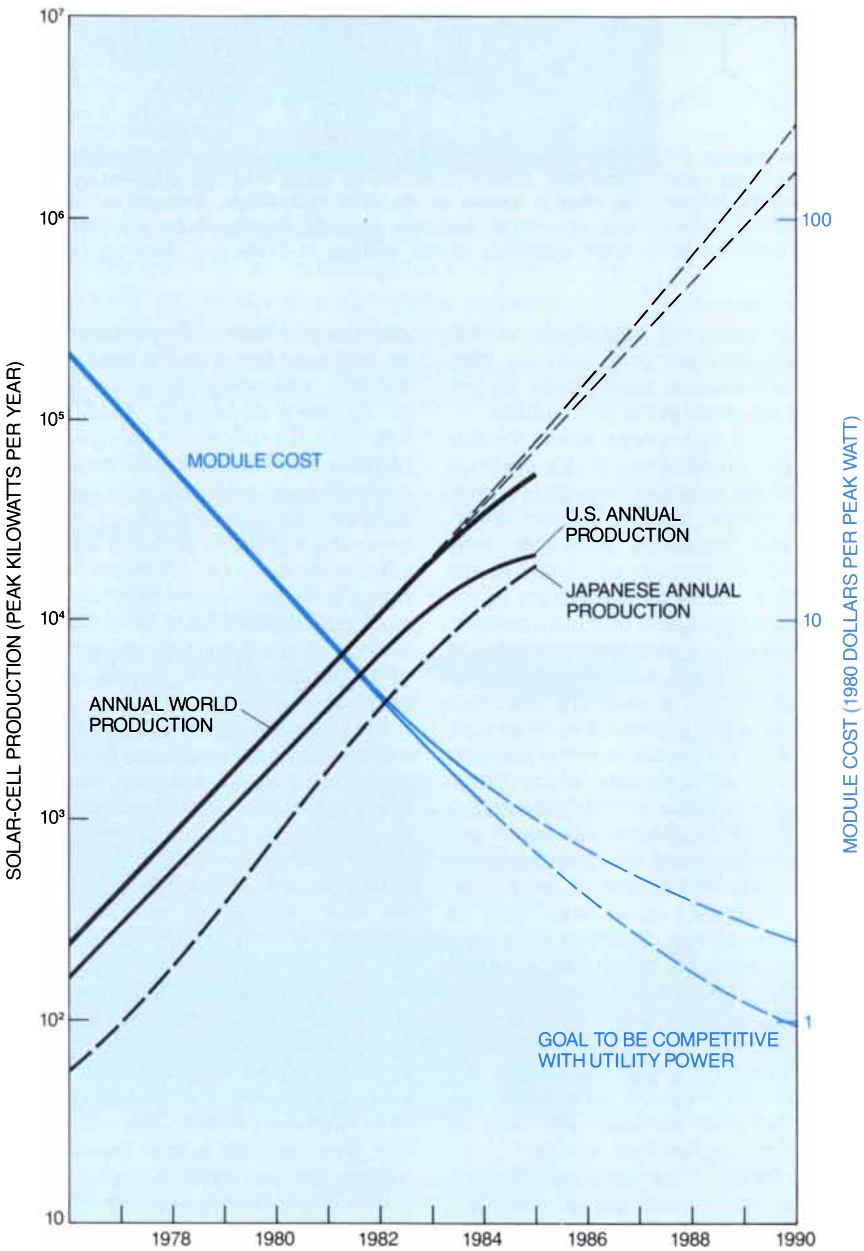
The next step is a large-scale photovoltaic system hooked up to a utility grid. For the purpose of tests and demonstrations leading toward such a capability, the U.S. Department of Energy in the late 1970's provided money for nine intermediate plants, ranging in size from 10 kilowatts to 100 kilowatts. These plants have proved the feasibility of intermediate-size plants and larger ones. As of 1984 the number of fairly large photovoltaic power plants built or planned was 14. They range in size from 200 kilowatts to 120 megawatts. The locations include the U.S., Italy, Japan, Saudi Arabia and West Germany. The European Community and some individual countries are developing 20 photovoltaic demonstration plants in the range from 15 to 300 kilowatts. Japan has in operation more than 10 plants in the range from three to 200 kilowatts, and a one-megawatt plant built by the New Energy Development Organization began operation in 1985. Japan's Sunshine project, started in 1974, aims at developing renewable-energy ideas to meet a considerable part of the nation's energy demand by the year 2000. The U.S. Department of Energy has had a photovoltaic program with similar aims for several years.

The largest solar plant now operating is at Carrisa Plains in California. It was built in less than a year by ARCO Solar, Inc. and Pacific Gas & Electric Company and provides 7.2 megawatts of peak power to the utility's grid.

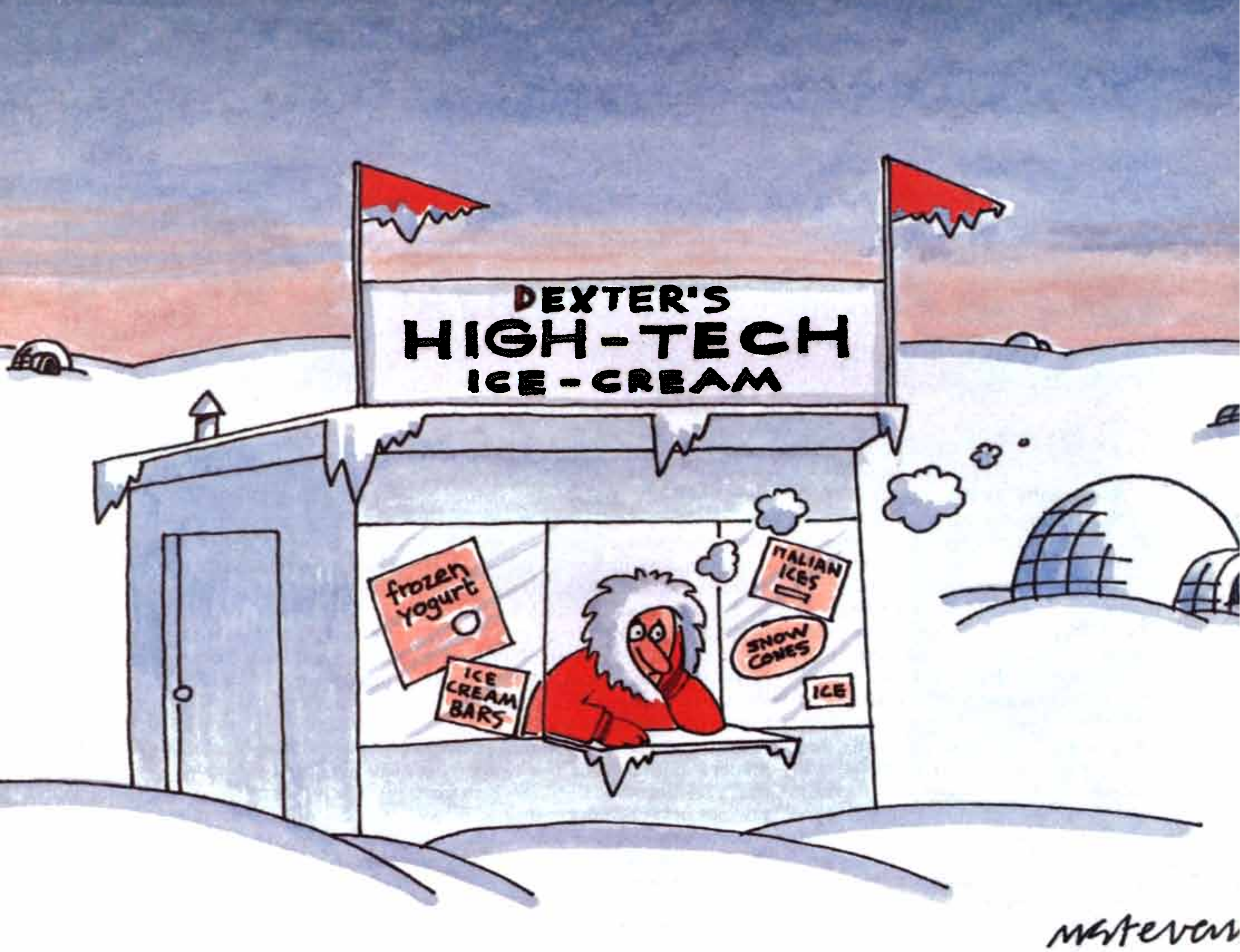
Taking into account the improvements that can be expected in materials and performance, the cost of electricity from photovoltaic systems could be competitive with conventional power plants by 1995. Stated in terms of cost per kilowatt-hour, the conventional plants will be producing electricity for between three cents (hydroelectric stations) and 35 cents (coal-fired steam turbines). Photovoltaic plants should by then be at about eight cents. Future improvements in photovoltaic technology should make it possible to sell the power at less than four cents per kilowatt-hour.

A solar photovoltaic system has a number of advantages in addition to its reliance on a clean and inexhaustible resource. It has no moving parts, so that a solar power station can be virtually maintenance-free—a phenomenon that has been demonstrated in remotely operated lighthouses, telecommunication relay stations and space vehicles. Solar cells can easily be mass-produced, as transistors and semiconductor integrated circuits are. In an expanding market this attribute can be expected to bring about cost reductions resulting from economies of scale. Solar arrays consist of a number of solar-cell modules and allow a wide range of application sizes and types with the same conversion efficiency and technology. A plant of megawatt size can be built in less than a year, in sharp contrast to conventional power plants and particularly to nuclear plants. As a result power companies can have more flexibility in planning for projected changes in the demand for electricity.

In the light of all those considerations one can expect that the era of economically feasible photovoltaics will arrive in the near future, sooner than anyone would have supposed a decade ago. The achievement will require sustained research and development work, international cooperation and perhaps some initial subsidies by individual governments. Given the right conditions, the photovoltaic industry could eventually rival the electronic industry in size.



PRODUCTION AND COST of solar cells have moved in opposite directions in the past decade, and the trend is projected to continue. A module is a sealed panel containing a number of solar cells interconnected at the factory; several modules constitute an array.



Better Check The Climate Before You Move Your Business.

While most of the country may be cooling down, New Jersey business just keeps getting hotter.

In the first half of 1986, business starts in the nation were up only 8.4% over the previous year. New Jersey, on the other hand, boasted a 30.3% increase.

And although New Jersey is one of the leading states for hi-tech industries, we are not interested in just being the Silicon Valley of the East. We're looking for those emerging growth industries that are best suited to our region. We call this strategy, *niche-finding*.

Once we find a niche, we support it. New Jersey has created five university-based Centers of Excellence — telematics, biotechnology, materials science, food technology, and waste management.

New Jersey is the right place. Geographically, socially, and politically. Twenty percent of our current employment is provided by firms which didn't even exist ten years ago.

If you're an entrepreneur with vision, take a look at New Jersey. We warm up to innovation and creativity.

For more information, write: Ben Ferrara, Director, Division of Economic Development, CN 823, Trenton, New Jersey 08625. Or call 609-292-7757.

We won't leave you out in the cold.

Department of Commerce and Economic Development
I plan to locate or expand my business in New Jersey.
Please have a relocation specialist contact me.

Name__________
Title__________
Company__________
Address__________
City_____ State_____ Zip_____
Tel__________

SANE4/87

New Jersey And You, Perfect Together.

Thomas H. Kean, *Governor* Borden R. Putnam, *Commissioner,* Department of Commerce & Economic Development

The Emergence of Animals

Some 570 million years ago animals diversified at an unprecedented rate. New body plans and ways of living emerged in concert with a new kind of community, one characterized by complex food chains

by Mark A. S. McMenamin

As recently as the middle of this century the earliest known fossils had all come from geologic formations in a stratigraphic interval known as the Cambrian, which contains rocks formed between about 570 and 505 million years ago. The Cambrian fossils included animals with body plans similar to those of a number of living animals. The fossil record therefore posed a perplexing question: Where were the ancestral forms that had given rise to these plentiful, advanced and diverse early sea animals? The sudden appearance of animal fossils in the lowest Cambrian strata, and the absence of animal fossils in Precambrian strata, made the boundary between the Precambrian and Cambrian periods the cardinal division of the geologic time scale.

In recent years a number of areas throughout the world have indeed yielded animal fossils older than the Cambrian. Still, the fact remains that a tremendous explosion of life took place in the Cambrian period. The start of the Cambrian marks the first appearance of a large number of major groups of animals. Many of these have survived until today. Many others are so unusual that they cannot be assigned to any known phylum. (A phylum is a group of animals that share the same general body plan.) Although most kinds of body plan found in present-day creatures first appeared during the Cambrian period, only a tiny fraction of the animals inhabiting the Cambrian seas gave rise to anything living today. Most of the rest can be viewed, with the benefit of hindsight, as short-lived experiments. The Cambrian period saw a higher percentage of such experimental groups of animals than any other interval in the history of the earth.

The emergence, during the transition from the Precambrian to the Cambrian, of so many kinds of creatures radically changed the nature of the relations among animals. Creatures evolved during the Cambrian to fill ecological niches that had never before been filled. Animals that fed on living matter, rather than scavenging dead organic matter or relying on symbiotic relationships with photosynthesizing algae, became much commoner. Predators emerged. The animals of the Cambrian coexisted in a web of relations not unlike those that connect modern animals. The transition from the Precambrian to the Cambrian thus saw not only the appearance of many modern kinds of animals but also the development of the modern kind of animal community.

The transition from the Precambrian to the Cambrian can be divided into four main stages. The first stage is marked by the appearance of the very first shelly fossils (only one or two shelly species from this stage have been found) and the so-called Ediacaran fauna: flat, soft-bodied creatures named for the Ediacara Hills of South Australia, where many of the first specimens were found. In a very rough approximation, this stage took place in the period between about 700 and 570 million years ago. The second stage, which began about 570 million years ago (with an uncertainty of plus or minus 40 million years), saw the disappearance of the Ediacaran faunas and the first assemblages of shelly faunas of low diversity (in which about five species of shelled animals are found together). This stage lasted for about five to 15 million years.

The third stage, which lasted for about 10 to 20 million years, is characterized by shelly faunas of moderate diversity (more than five but fewer than 15 species of shelled animals found together) and the first appearances of a group of unusual, vase-shaped creatures known as archaeocyathans. The fourth stage, which lasted for some 15 to 30 million years, saw the first shelly faunas of high diversity and the earliest appearance of trilobites. Trilobites are extinct arthropods (animals having jointed appendages) that were made up of three sections (head, thorax and tail) and covered by a shieldlike carapace, or exoskeleton; like modern arthropods, trilobites grew by shedding their carapace, and cast-off carapaces are commonly found as fossils.

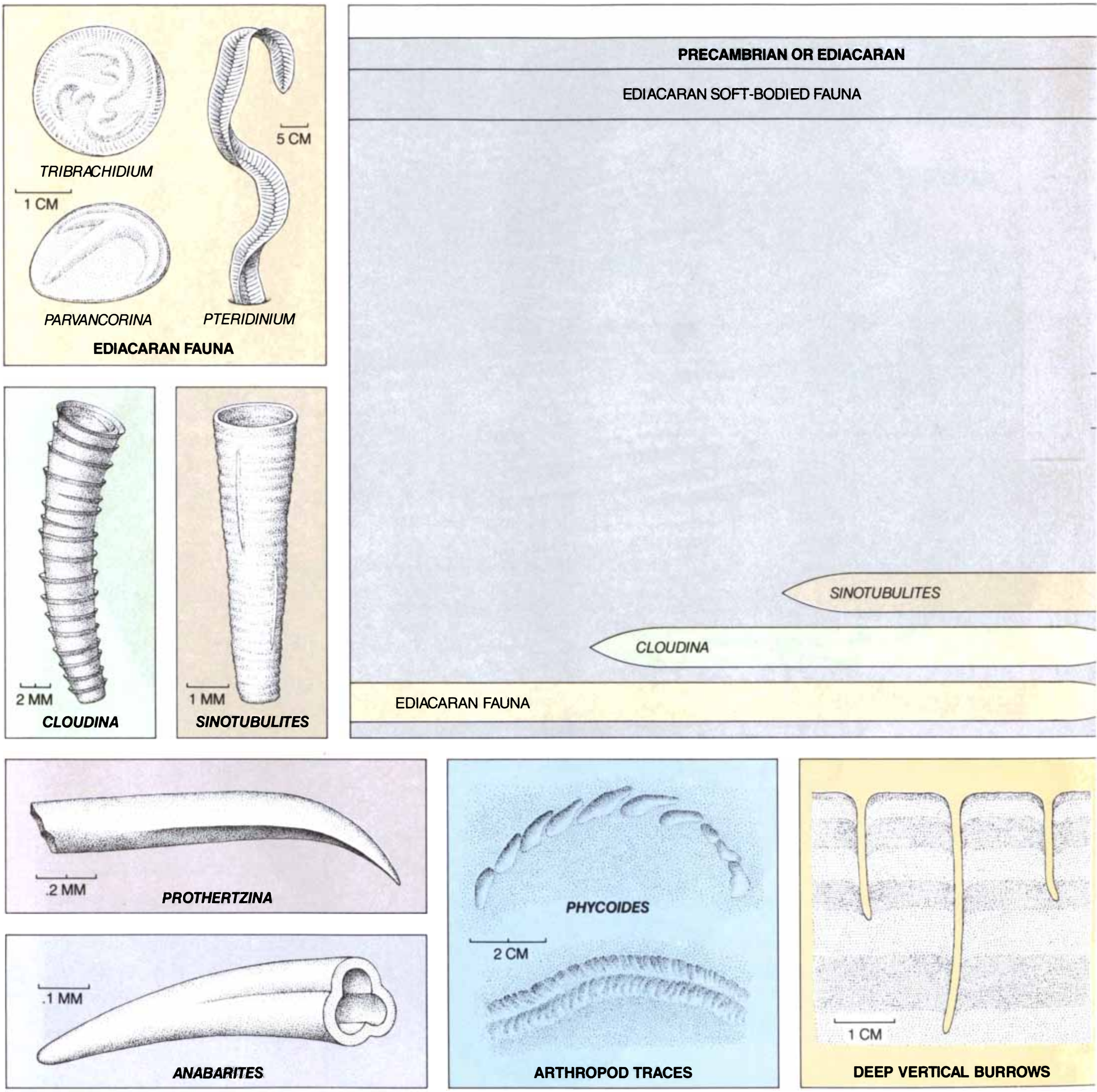
The appearance of the Ediacaran fauna, which indicates the first of these four stages, is found in strata above the Precambrian occurrences of glacier-generated marine sediments known as tillites. These tillites record drawn-out episodes of worldwide glaciation. The Ediacaran fauna must therefore have come into being not long after the last major episode of late Precambrian glaciation. Fossils of Ediacaran soft-bodied animals are found with fossil traces (animal tracks and shallow burrows) formed along the surface of what was then the sea floor. There are many fewer fossils at this stage than at later stages, but this can be accounted for partly by the fact that soft-bodied creatures (which decay rapidly) tend not to be preserved as fossils; it does not necessarily indicate that Ediacaran species were not abundant and widespread. A handful of shelled fossils are also found at this stage, including *Cloudina,* which is a tube-shaped fossil from Namibia that has a shell made of calcium carbonate, and *Sinotubulites,* another tubular fossil of calcium carbonate, which is found in Precambrian rocks of southern China.

This interval also yields tube-shaped fossils known as sabelliditids and vendotaenids. Sabelliditids are typically a few centimeters long and from one to several millimeters in diameter. They are fossils of tubular sheaths original-

HELICOPLACUS, a creature that became extinct about 510 million years ago, only 20 million years after its first appearance, had what might be called an experimental body plan: its parts were organized in a way that is not found in any living creature. *Helicoplacus* was shaped like a cylindrical spindle and was covered with a spiraling system of armor plates. It emerged during the transition from the Precambrian to the Cambrian period. During that transition more kinds of body plan arose than at any time before or since. Most, like the plan of *Helicoplacus,* proved to be unsuccessful. The specimen shown here is about five centimeters long.

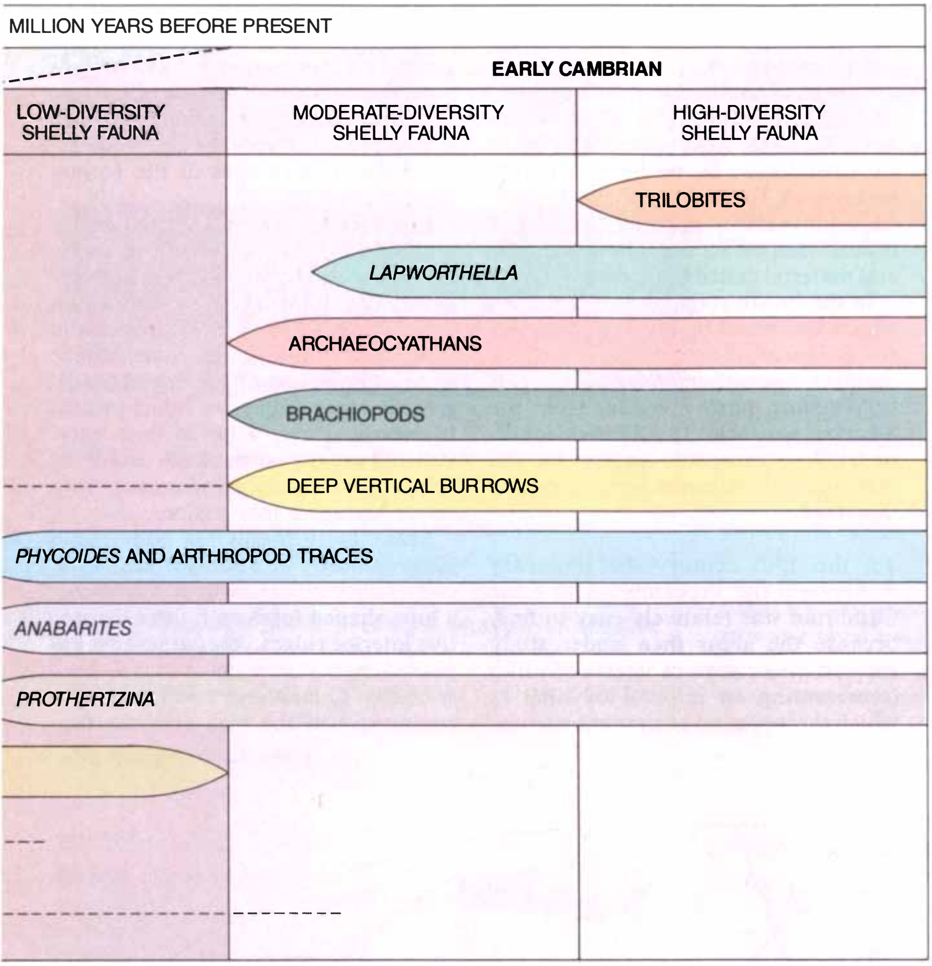
ly composed of a flexible organic substance that probably housed wormlike, filter-feeding animals. Sabelliditid tubes, along with *Cloudina* and *Sinotubulites,* give evidence of early tube-dwelling animals that had taken up a sessile (stationary), filter-feeding way of life. Vendotaenids are also tubular fossils originally made of a flexible, organic substance, but they are much smaller than sabelliditids (less than a centimeter long and about a tenth of a millimeter wide). Vendotaenids may represent sheaths formed as external secretions by colonies of Precambrian bacteria.

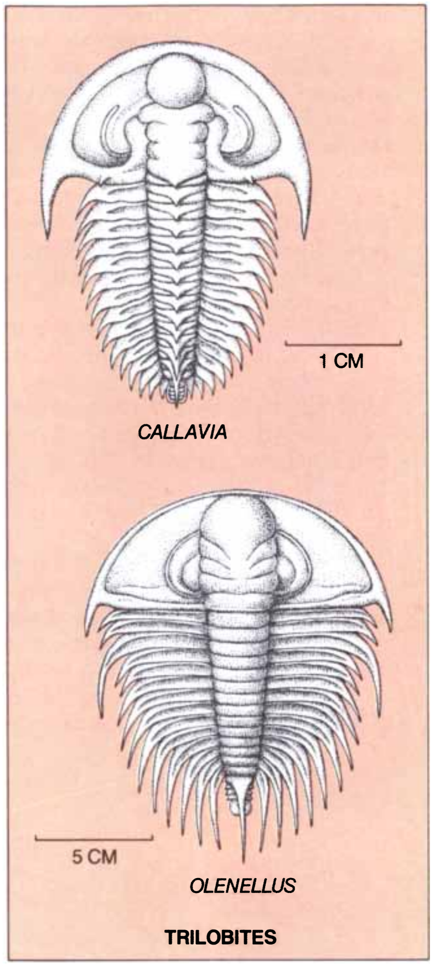
At the beginning of the second stage of the transition period, fossils appear—of animals and pieces of animals—whose hard parts are based on

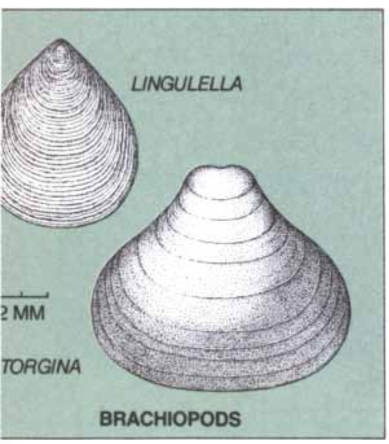


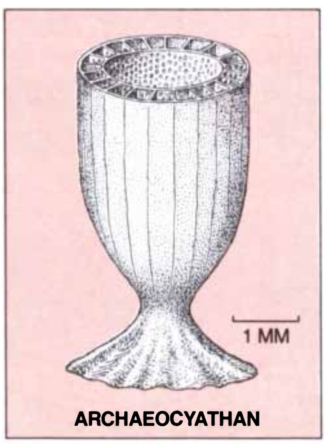
ANIMAL LIFE became much more complex and diverse during the transition from the Precambrian to the Cambrian period. (All the creatures shown here lived in the sea.) The so-called Ediacaran fauna, which were soft-bodied and tended to be flat, lived between 700 and 570 million years ago, during the first stage of the transition. This stage also saw the appearance of *Cloudina* and *Sinotubulites,* tube-shaped fossils of calcium carbonate shells. A characteristic fossil of the second stage of the transition is *Protohertzina,* a relic made of calcium phosphate that resembles the spines with which some modern predators grasp prey; *Protohertzina* was probably a grasping spine of an early predator. Another shelly fossil from the same period is *Anabarites.* It and *Protohertzina* are found in rock formations over a wide range, including sites in Asia, Australia, the Middle East and North America. In rocks from this period the diversity and number of trace fossils (fossils of the tracks and shallow burrows left by animals) in-

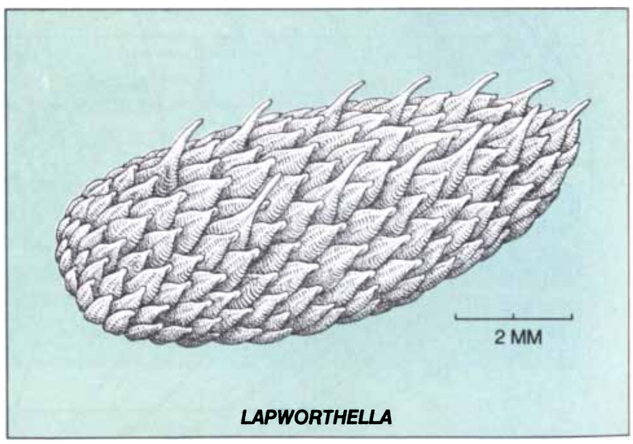
calcium phosphate. The earliest of these phosphatic fossils are a few millimeters or less in length, and they make up a low-diversity grouping of early phosphatic shelly fauna. One example is the tusk-shaped fossil called *Protohertzina.* It has strong microstructural similarities to the sharp spines that a modern phylum of tiny but voracious predators called chaetognaths, or arrow worms, use for grasping. *Protohertzina* itself probably served as a grasping spine of some similar creature, and it is the first fossil that can be identified with any confidence as having been part of a predatory animal. Along with the phosphatic fossils appear some shelly ones consisting of calcium carbonate. The original mineralogy of these early shelly fossils is sometimes difficult to ascertain, be-











crease markedly. There are some tracks of great complexity; examples are *Phycoides pedum,* which records a series of feeding and burrowing motions, and certain tracks that have chevron-shaped grooves, which were probably made by some early arthropod. (Arthropods, such as the modern lobster, are creatures that have jointed appendages.) Rock formations from this stage also show the earliest deep vertical burrows. During the next period the brachiopods (two-valved creatures that resemble clams) emerged, as did a strange group of creatures known as archaeocyathans, which had double-walled, porous skeletons made of calcium carbonate. Toward the middle of this third period *Lapworthella,* an armored animal, first appeared. (The rendering of *Lapworthella* shown here is a reconstruction based on fossils of individual sclerites, or pieces of the armor coat.) The fourth phase of the transition marks the appearance of the trilobites, arthropods that were covered by a shieldlike exoskeleton, which they shed as they grew.

cause calcium phosphate can replace shell matter made of calcium carbonate with such fidelity that structural features only a few thousandths of a millimeter wide are preserved.

At about the same stratigraphic level there is a tremendous increase in the number and diversity of fossilized tracks and traces on the sea-floor sediment. Deep vertical burrows appear for the first time and there is an abundance of complex trace fossils such as *Phycoides pedum,* a track that records a series of feeding and burrowing motions made by a bottom-dwelling animal. There are also animal tracks that have chevron-shaped grooved markings, formed as the appendages of a crawling or burrowing arthropod scraped the sediment.

The Ediacaran fauna, with one possible exception, seems to have become extinct at this level, although some paleontologists believe the lack of fossilized corpses of Ediacaran-style soft-bodied animals may be due to scavenging by burrowing animals, whose number and activity increased markedly at this time.

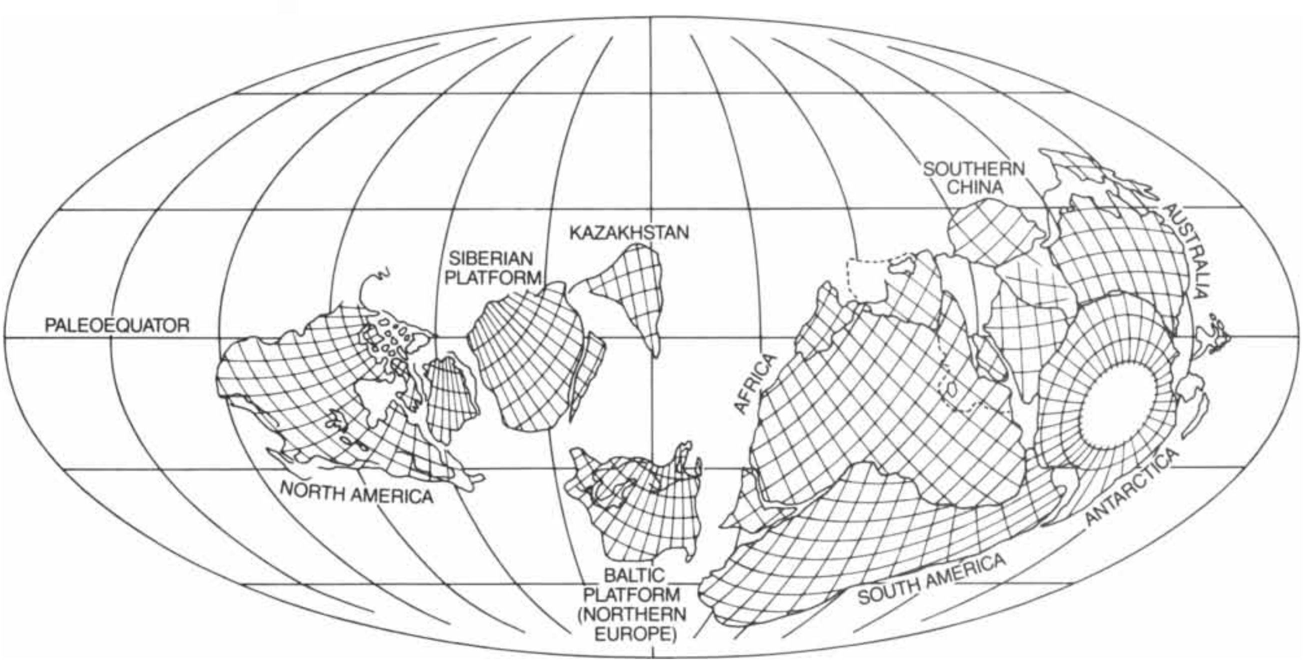
The third stage of the transition from the Precambrian to the Cambrian is indicated by moderate-diversity shelly faunas. In most areas of the world these appear in strata immediately above those containing the low-diversity faunas. In a region called the Siberian platform, which is in central and western Siberia, the moderate-diversity faunas are accompanied by the earliest archaeocyathans: vase-shaped fossils that had double-walled, porous skeletons made of calcium carbonate. The archaeocyathans bear a slight resemblance to corals and sponges, but in fact they have no close relationship to any living group and are currently placed in their own phylum. Archaeocyathans, in combination with calcareous algae (multicellular algae whose branches and light-gathering organs were reinforced by needles of calcium carbonate), formed the earliest wave-resistant reefs by growing together in mound-shaped accumulations of skeletal material called bioherms.

In the fourth stage of the transition, which is marked by the high-diversity shelly faunas, the geographic range of the archaeocyathans expanded greatly, reaching many areas far from the Siberian platform. In addition, fossils of trilobite carapaces appear for the first time in sediments formed during this stage.

In the 19th century the boundary between the Precambrian and the Cambrian was relatively easy to find, because the areas then under study showed major gaps or unconformities (representing an interval of time in which there was no preservation or deposition of rocks) between Precambrian and Cambrian sedimentary layers. The expanded data base available to 20th-century paleontologists has actually complicated the task of locating the boundary precisely, because many regions are now known where Precambrian and Cambrian formations grade continuously into each other. Determining where the boundary lies in any given rock formation is necessary if paleontologists are to correlate the information gathered at one site with that gathered at another and to determine the relative ages of the formations in question.

Precambrian and Cambrian stratigraphic sections are difficult to correlate from site to site for two reasons. One reason is that there are only a limited number of early fossil species to study. The other is that many of the species that do exist have long stratigraphic ranges (they are found in strata deposited over a broad time interval) and are not particularly useful for splitting stratigraphic sequences into biostratigraphic subdivisions.

Many early fossils are widespread geographically as well as stratigraphically. One such organism is *Anabarites,* a tube-shaped fossil with three distinctive interior ridges. *Anabarites* first appeared during the stage of low-diversity shelly faunas and continued as a component of the high-diversity fau-

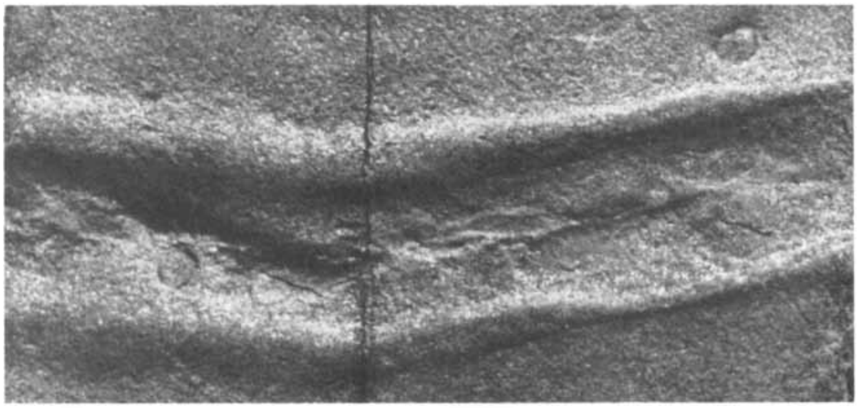


GEOGRAPHY of the world during the transition from the Precambrian to the Cambrian period presented the conditions in which the emerging life of the transition period could flourish. In the late Precambrian much of the earth's landmass was concentrated in a single supercontinent. The breakup of the supercontinent, which probably took place near the beginning of the Cambrian period, provided large regions of new coastline for colonization. Many of the newly formed continents were at or near the equator, and so they offered a warm, equable climate. The increasing distances between the early Cambrian continents encouraged provinciality (differences among separated faunas). Grid lines drawn on the continents mark present-day lines of latitude and longitude.

nas. It is found in Australia, China, India, Iran, Kazakhstan, Mongolia, western and eastern North America and Siberia. The widespread distribution of some elements of the Ediacaran fauna and the early shelly faunas stems from the positions of the continents in the late Precambrian and early Cambrian. Most of the continents were near the equator, and there is good reason to suppose that in the late Precambrian many of the present-day continents were part of a single supercontinent.

Indeed, Gerard C. Bond and his colleagues at the Lamont-Doherty Geological Observatory have found evidence that continental rifting occurred around the edges of North America during the transition from the Precambrian to the Cambrian. The proximity of the newly separated continents in the early stages of continental drift, and the fact that the continents sat at roughly the same latitude, made it possible for animals to spread widely: there were no broad oceans or extreme differences in temperature to prevent faunas from migrating from one continental shelf to another.

Provinciality (that is, differences among species in different locales) began to develop throughout the world when the first widespread archaeocyathans and trilobites appeared. The tendency toward provinciality was no doubt intensified by the growth of progressively larger oceans between continents throughout the early Cambrian period. Also, Allison R. Palmer of the Geological Society of America has shown that at this time groups of formations called carbonate belts began to form along the margins of several continents. Carbonate belts are shallow marine areas formed by the accumulation of shells made of calcium carbonate. Because they limit access to the open sea, carbonate belts make it possible for animals living in areas between a belt and the shoreline to evolve in isolation from similar animals living on other continents. The effect is particularly evident in some trilobite groups.



FOSSILIZED TRACKS give an indication of the complexity of the creatures that made them. Some tracks formed in the Precambrian, as is shown in the top photograph, are the result of shallow, horizontal burrowings in the sediment on the sea floor. They are about a millimeter wide. The broad track shown in the bottom photograph was made in the Cambrian period by a crawling mollusklike organism. It is about 1.5 centimeters wide.

Because most continents in the late Precambrian were near the equator, the climate after the last Precambrian glacial episode was probably quite equable. As the global climate became warmer, food supplies in shallow marine waters began to stabilize at relatively low levels. A decrease in worldwide temperature extremes would have contributed to the stability of marine food resources. In general, a smaller gradient in temperature between the poles and the equator leads to a smaller seasonal overturn of the ocean, causing fewer deep-ocean nutrients to reach shallow marine waters. A stable, nonfluctuating supply of food is crucial for many marine animals, and particularly for those that live in the Tropics and are more accustomed to stable conditions than creatures that live in areas having more pronounced seasonal variations.

As a matter of fact, the limited supply of food available during the Precambrian helps to explain the unusual, flattened bodies of the Ediacaran creatures. The flatness and thinness of the Ediacaran body plan (the pancake-shaped *Dickinsonia,* for example, had a maximum thickness of no more than about six millimeters although its diameter could exceed one meter!) maximized the ratio of surface area to volume. A high ratio of surface area to volume is ideal for the feeding strategies known as photoautotrophy and chemoautotrophy, which are favored in environments containing small amounts of nutrients.

Photoautotrophy involves a symbiotic relationship with photosynthesizing algae. Within the tissues of the host animal the algae are protected from animals that might feed on them. In return they release nutrients to the host and remove waste products. For such symbiosis to work, a substantial part of the host's body must be exposed to sunlight so that the algae can photosynthesize efficiently. Many reef corals and a few tropical clams harbor symbiotic photosynthetic algae.

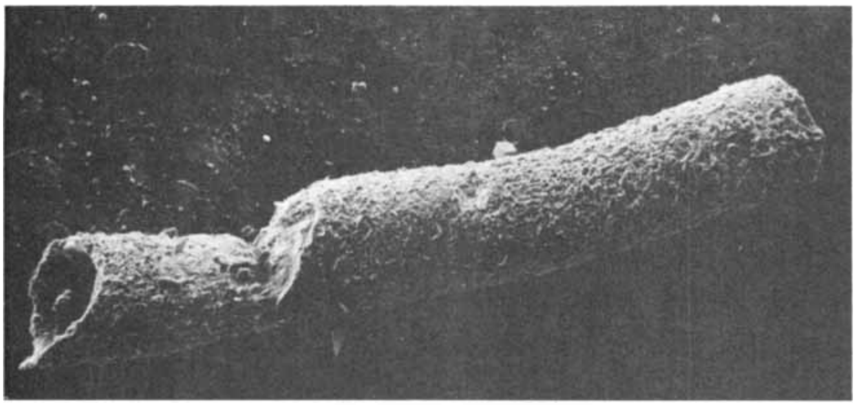
The other strategy, chemoautotrophy, is the direct uptake of energy-supplying nutrients from seawater. It too can sometimes involve internal symbiosis, in which the animal harbors chemosynthetic bacteria. The strategy is common, for example, among modern animals living near deep-sea hydrothermal vents. Some, however, seem to absorb dissolved nutrients directly, without the aid of bacteria.

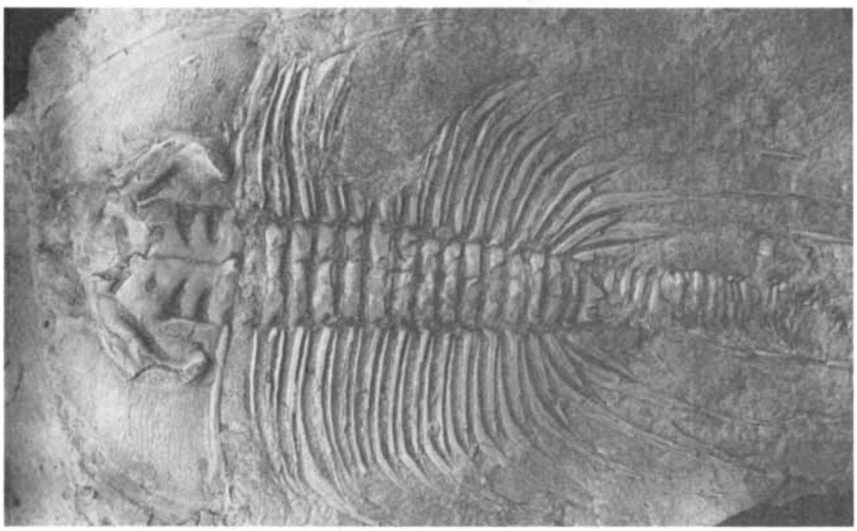
The flattened bodies of the Ediacaran fauna would have made possible efficient uptake of nutrients in seawater or efficient absorption of light by photosynthesizing algae. Recent studies by Pamela Hallock-Muller of the University of South Florida in St. Petersburg have shown that in low-nutrient waters the host-symbiont association is particularly favored, because the host can recycle its waste products directly through the symbiont rather than releasing them into the environment. The Ediacaran fauna may therefore have been well adapted to the marine conditions that prevailed throughout the late Precambrian, when the waters of shallow seas are thought to have been low in nutrients.

Toward the end of the Precambrian and during the transition to the Cambrian the food supply changed, probably in response to changes in the chemistry of the oceans, and other feeding strategies became more important. Heterotrophy—the consumption of other organisms (either animals or plants)—became increasingly important near the end of the Precambrian. Evidence for increased heterotrophy is found, for example, in the fossil record of stromatolites: structures shaped like domes, columns or cones that were built up in layers by seafloor carpets of algae. Stromatolites grew upward toward the sun, successive layers of algal mats rich in organic matter alternating with layers of sedimentary particles that were trapped by the mats. Although each layer was much less than a millimeter thick, the layers gradually accumulated to form large columns or domes; some Precambrian stromatolites were more than 10 meters high.

About 800 million years ago stromatolites underwent a marked decline in diversity, which Stanley M. Awramik of the University of California at Santa Barbara has attributed to the appearance of algae-eating animals: stromatolites are easily overgrazed, and disturbances in the algal mat can cause the stromatolite to stop forming.

Some other possible evidence for increased heterotrophy is found in microfossiliferous cherts, which are sedimentary rocks made up of microcrystalline quartz. When thin sections of these cherts are illuminated and viewed under a microscope, microbial fossils can be seen embedded in the translucent slices of rock. The sheaths of such fossils became more robust between 800 and 700 million years ago, perhaps to provide protection from early grazing animals. At about the same stratigraphic level, fossils of traces made by scavenging and deposit-feeding animals appear. The animals that made these traces were almost certainly the ancestors of the Cambrian shelly organisms.





WOUNDED ANIMALS provide striking evidence of predation during the transition period. The image at the top shows a damaged and healed shell of *Hyolithellus* enlarged some 40 diameters. The photograph at the bottom shows a wounded and healed carapace (exoskeleton) of the trilobite *Olenellus robsonensis* seen at three-fourths actual size. That the wounds have healed is important, because it demonstrates that the damage was done during the life of the animal and was not inflicted later on a corpse or an empty shell.

Perhaps the most astonishing aspect of the Cambrian faunas is that so many radically different types of animals appeared in such a short interval. What gave rise to this sudden diversification? James W. Valentine of Santa Barbara and Douglas Erwin of Michigan State University have proposed that a high order of genetic repatterning was possible because the genome (the full complement of genes) of multicellular animals was less complex in the transition period than it is today. Fewer kinds of mutation would have been fatal because the linkages among various parts of developmental genetic programs were not as intricate in Precambrian and early Cambrian animals as they are in today's fauna.

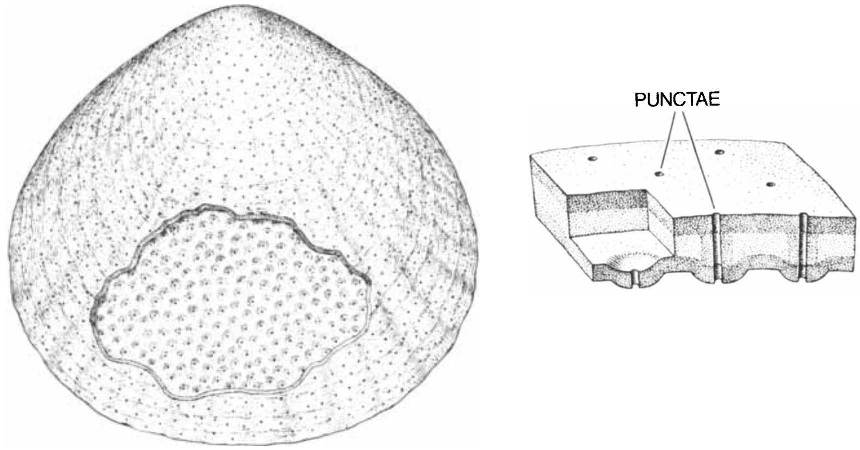
This genetic flexibility, Valentine and Erwin argue, is one of two reasons so many high-level taxa (such as classes—which are the largest divisions within phyla—and phyla) came into being so suddenly. The other reason is that there were so many niches available that had never before been occupied. The Cambrian could therefore see the appearance of new phyla and classes at a rate that has not been matched since: radically new animals, without competition, could become the founders of new classes or phyla.

The Cambrian diversification, while rapidly establishing new phyla and classes, also initiated the first complex communities of animals linked by food chains. The existence of new types of communities in turn created niches for new types of animals. A key element in the establishment of animal communities is predation, which establishes a hierarchic chain of food-transfer connections among animals.

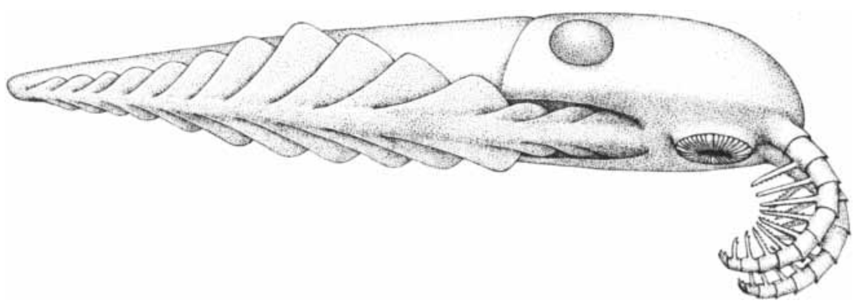
Previous assumptions that predators were not important in Cambrian communities have been overturned by new indications that the impact of predation was great. There are essentially three kinds of evidence: actual fossils of predators, specimens of damaged (and sometimes partially healed) prey and antipredatory adaptations in some animals.

Protohertzina, the fossil that resembles a modern arrow worm's grasping spine, is a relic of one early Cambrian predator. Another predator is *Anomalocaris,* which has recently been reconstructed from nearly complete specimens by Derek Briggs of the University of Bristol and Harry Whittington of the University of Cambridge. *Anomalocaris* is gigantic by Cambrian standards (about 45 centimeters long) and resembles no living animal. It had a body shaped like a flattened teardrop, both sides of which were flanked by swimming fins. At its broad head were a pair of jointed appendages for drawing prey into its hideous mouth, which was shaped like a pineapple ring lined with teeth. *Anomalocaris* is probably an example of a short-lived, experimental phylum. Briggs and Whittington suggest that *Anomalocaris* is largely responsible for one of the other major signs of predation in the Cambrian: fossils of wounded trilobites.

There are numerous fossilized trilobite specimens that have had bites taken out of the carapace. In most cases the wounds are partially healed, indicating that the carapace was still attached to the trilobite when it was damaged, rather than having been damaged by a scavenger after being shed by the trilobite. There are other examples of predatory damage, such as small shelly fossils with boreholes in them. The holes resemble those made by certain modern predators that drill through shells to eat the soft meat inside.



ANTIPREDATORY ADAPTATION may explain the numerous punctae, or small holes, found in the shells of a group of animals known as mickwitziid brachiopods. Holes may have acted as channels for transporting chemical deterrents to the outside of the shell.



PREDATOR from the early and middle Cambrian period may have fed largely on trilobites. This creature, *Anomalocaris,* was much larger than most animals of its day (it grew to be about 45 centimeters long). It had gripping appendages to carry food to its mouth; it swam by generating a wavelike motion along the finlike membranes on its underside.

The third indication that predation was important in the early Cambrian is the number of features of early Cambrian animals that could have had antipredatory functions. For example, the shells and exoskeletons that appear for the first time in a number of the new phyla probably served as a defense against predators. The deep vertical burrows that proliferate at the time of the low- and moderate-diversity shelly faunas could have given protection from predators unable to move through sediment. Several trilobite species evolved long spines, which may have made them more difficult for such contemporary predators as *Anomalocaris* to attack.

In addition, Stefan Bengtson of the Institute of Paleontology in Uppsala, Ed Landing of the New York State Geological Survey and Simon Conway Morris of Cambridge have shown that many small shelly fossils are actually sclerites: disarticulated parts of a spiny, maillike coat of armor that probably protected the upper surface of slow-crawling animals. Such animals may have resembled small marine porcupines or hedgehogs. Another adaptation is found in the class of brachiopods (creatures with a two-valved shell roughly resembling that of clams) referred to as mickwitziids. These animals had numerous pores, called punctae, running through their shell walls. The punctae may have delivered chemical deterrents to the outer surface of the shell, discouraging predators and parasites.

Predators, then, were an important part of the marine environment during the Cambrian. Exoskeletons, which may originally have evolved as a means of protection, were also key factors in the evolution of some strikingly new body plans. For example, without a bivalved shell brachiopods would not be able to produce the internal currents that make their filter-feeding mechanism efficient.

Even though the mechanics of the establishment of Cambrian communities are becoming clearer, one central question remains unanswered: Why did the Cambrian revolution occur when it did and not tens or even hundreds of millions of years earlier? This question is particularly puzzling because scavenging and grazing animals (known from their trace fossils) predate the Cambrian boundary by as much as 200 million years.

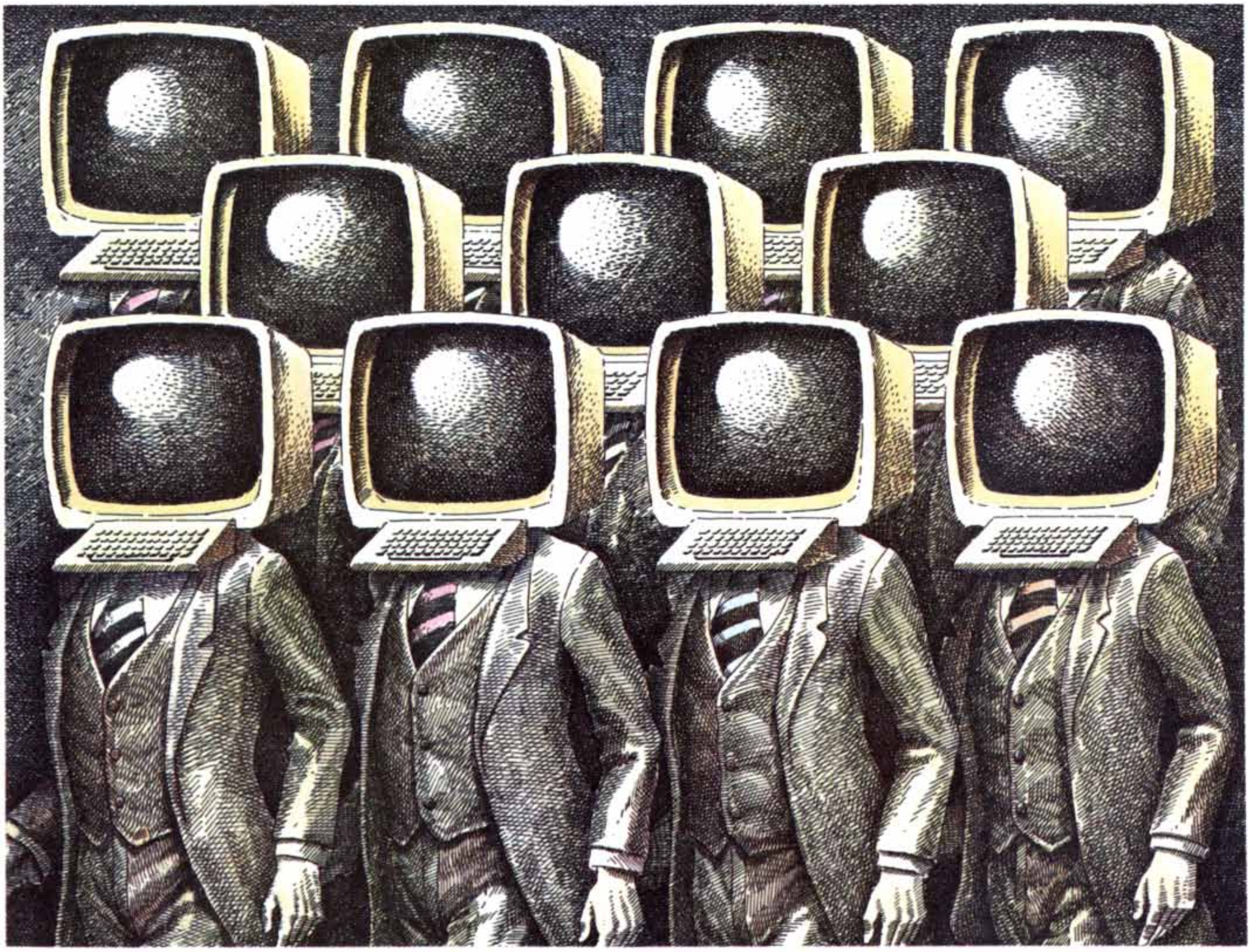
One answer may have to do with the chemistry of the oceans. The concentrations in seawater of phosphate, as well as of isotopes of sulfur and strontium, underwent dramatic (if poorly understood) fluctuations during the transition from the Precambrian to the Cambrian. As Peter Cook and John Shergold of the Bureau of Mineral Resources, Geology and Geophysics in Canberra suggest, the large phosphate deposits found in Precambrian-Cambrian sediments in many parts of the world may represent an episode of global phosphogenesis, when increases in the availability of phosphate and other nutrients made it easier for animals to form phosphatic skeletons.

A stumbling block to this hypothesis is that calcareous shells seem to be at least as prevalent as phosphatic shells at the boundary between the Precambrian and the Cambrian. It may be more accurate, then, to see the episode of phosphate deposition as part of a larger event in which the oceans saw a sudden increase in the amount of available nutrients. An environment rich in nutrients would no longer have favored the bioenergetics of the host-symbiont relationship and so it would have increased the number of scavengers and deposit feeders, leading in turn to an increase in the number of predators. As such heterotrophic organisms reached what might be called a critical biomass, an "ecological chain reaction" (in the words of Martin Brasier of the University of Hull) could be initiated: newly evolved animals might create niches that could be filled by other, newer animals, eventually resulting in complex communities in which there were abundant shell-bearing animals.

The Ediacaran fauna (and the other Precambrian animals) had originated in a world characterized by a massive supercontinent, waning glaciation and relatively low supplies of marine nutrients. The remarkably numerous and diverse Cambrian animals appeared in a world marked by the breakup of the supercontinent (which made extensive tropical shorelines available) and abundant marine food supplies. It is not yet known with certainty whether the Cambrian animals appeared because of such changes in the global environment, because of a series of fortuitous changes in the genetic programming of animals or because of some combination of these causes and other, unrelated factors.

Whatever the reason for their origin, the earliest Cambrian innovations (such as shelly organisms, predators and deep burrowers) went on to colonize the world rapidly. The combination of environmental influences (such as nutrient-rich waters) and biotic changes (such as the emergence of predators) caused a major change in the nature of animal communities. Modern animals, including human beings, are the direct descendants of the animals that first appeared during the Cambrian explosion, and the style of ecological interaction these early animals brought about has characterized nearly all animal communities of the past 570 million years.

All too often, you can become just another face in the crowd simply by following it.



Following the crowd is no way for any hi-tech company to select a new location. You'd overlook more than you can imagine.

A site in Maryland, for example, gives you access to a pool of over 107,000 scientists and engineers. Where else can you find such a concentration of talent? Nowhere in the country.

And that's just one of the little-known facts that qualify Maryland for virtually any company's list of options.

Let us tell you more. Contact Michael Lofton, Dept. of Economic and Community Development, Dept. 152, 45 Calvert Street, Annapolis, Maryland 21401. (301) 974-3514.

Maryland

More than you can imagine.

Acoustics of Ancient Chinese Bells

Bronze bell chimes were important orchestral instruments until they vanished from history 2,000 years ago. A chime recovered by archaeologists has revealed their sophisticated acoustical design

by Sinyan Shen

In 1978 a set of Chinese bronze chime bells large enough to occupy the entire stage of a modern recital hall was unearthed in Hubei Province in southern China. The chime, which dates from the fifth century B.C., consists of 65 bells encompassing five octaves, a range greater than that of most contemporary instruments. A filigree of gold-inlaid inscriptions on the bells and their frame documents the existence of an elaborate theory of music that specified the design, scales and instrumentation of ancient orchestras. This record and recent investigations of the chime itself have prompted a complete rewriting of the history of acoustics.

The ancient inscriptions confirmed what modern scholars were only beginning to suspect: that the bells were constructed in such a way that each could produce two separate pitches. This property sets the Chinese chime bells apart from Western church bells, which are known for their single, lingering tones; unlike church bells, the chime bells could perform complex, rapidly metered music. For reasons that are not clear, the principles and practices surrounding the unique bell design were never passed down. Consequently the way to play chime bells remained a mystery for more than 2,000 years.

The bells incorporate many unusual features whose sophistication and precision can be appreciated only in the light of the dual-pitch design. Since the chime was discovered, studies of vibrational properties and tuning methods have revealed the depth of understanding possessed by Chinese metallurgists and musicians. The design of the bells requires a theoretical grasp of physics and engineering formerly thought to have evolved only in the late 18th century. Indeed, the acoustical principles exploited in the Chinese bronze bells have astonished even 20th-century acousticians.

The study of acoustics in the Western world is relatively young. In 1787 a German physicist named Ernst F. Chladni sprinkled sand on vibrating plates to show that some regions of the plates remained stationary during vibration. These motionless regions were termed nodal lines; their distribution describes the modes a vibrating body assumes. Each so-called normal mode is associated with a characteristic frequency of vibration, and the frequency of vibration determines the perceived pitch.

Vibrating bodies move in many different modes simultaneously and generate many different frequency components called partials. The partial with the lowest frequency is called the fundamental; there are many higher frequencies called overtones. When a bell is struck or a string is plucked, all these frequencies come into play, but some are stronger (louder) than others. The relative strength of partials in a musical sound constitutes its tonal quality, just as a combination of wavelengths determines the color of light.

In 1890 Lord Rayleigh studied the bells in the tower of his church in Terling, England, and performed experiments on several bells in his laboratory. He identified six partials in the bells. Rayleigh, who laid the foundation for subsequent work on bell acoustics, thought that a bell could produce just one fundamental pitch. Since his experience was confined to Western bells, he could not have foreseen the lesson that Chinese bells thousands of years old would teach.

About 80 years ago single chime bells began to appear among other archaeological finds in China. Later sets and entire ensembles of the bells turned up; today thousands of bells and more than 50 chimes have been recovered. Although the bells were thoroughly scrutinized, investigators did not recognize the dual-pitch potential until 1977. Some doubts persisted until the 1978 discovery of the chime in Hubei Province.

There were earlier clues that had escaped notice. While studying the Jing-li bell chime, which had been un-

BELL CHIME of Zenghou Yi, ruler of a fifth-century B.C. Chinese principality, is

earthed in Henan Province in 1957, investigators at the National Music Research Institute played "The East Is Red" using pitches obtained by striking the bells at their center. An $\#E_5$ was missing. Anxious to complete the piece, the team found the note by striking a $\#C_5$ bell on its side. Their success was regarded as accidental.

In 1977 Huang Xiang-peng, Lu Ji, Wang Xiang, Gu Bo-bao and their colleagues at the institute examined a bell chime discovered in Shanxi Province and found that every bell, when struck on a side, produced a pitch higher than the one sounded from the center. The interval between the pitches was always a minor or major third, a difference in frequency equivalent to that between four or five consecutive keys on a piano. The team's observation sparked animated discussions about whether the phenomenon was accidental or deliberate, and whether the second pitch was a fundamental or an overtone.

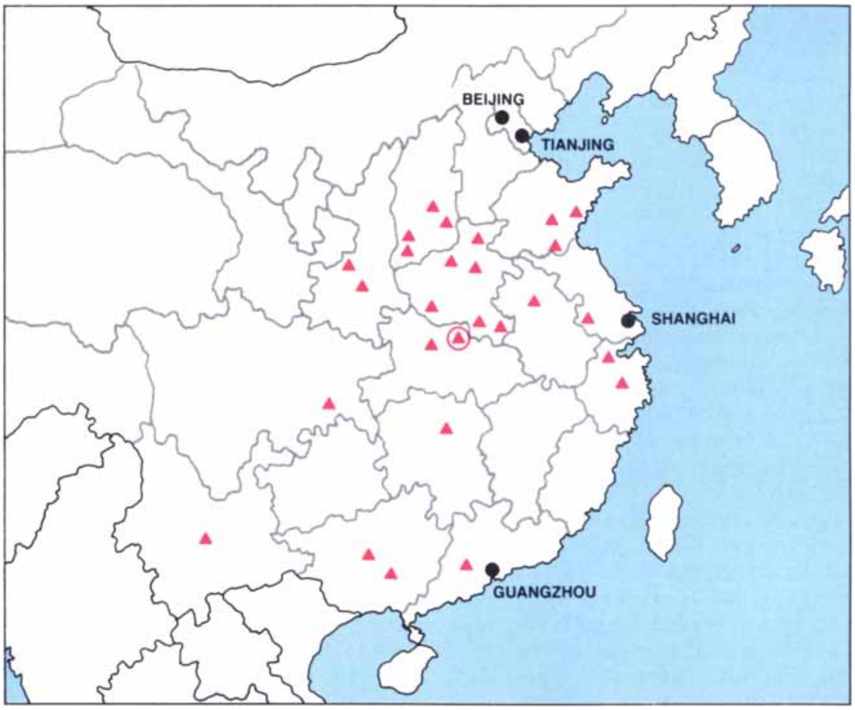
The investigators then studied more than 200 bells from the Shang (16th to 11th centuries B.C.) and Zhou (11th century to 221 B.C.) periods, spanning the historical lifetime of the chimes. They concluded that the bells were deliberately constructed to produce two pitches. In one chime the side striking positions were decorated with glyphs of the phoenix, a practice presumably linked to ancient legends in which the singing of the phoenix connotes music.

When, a year later, the magnificent 65-bell chime was found in Hubei Province, the researchers' conclusion was borne out. The chime had been interred in the tomb of Zenghou Yi, marquis of an ancient principality known as Zeng. It was part of two large orchestras also preserved in the marquis's tomb. The bronze *zhong* bells, which combined make up a type of chime called *bian-zhong,* were intact and almost perfectly tuned. The musical treasures of Marquis Yi confirmed the dual-pitch design of chime bells through detailed records inscribed on the bells themselves.

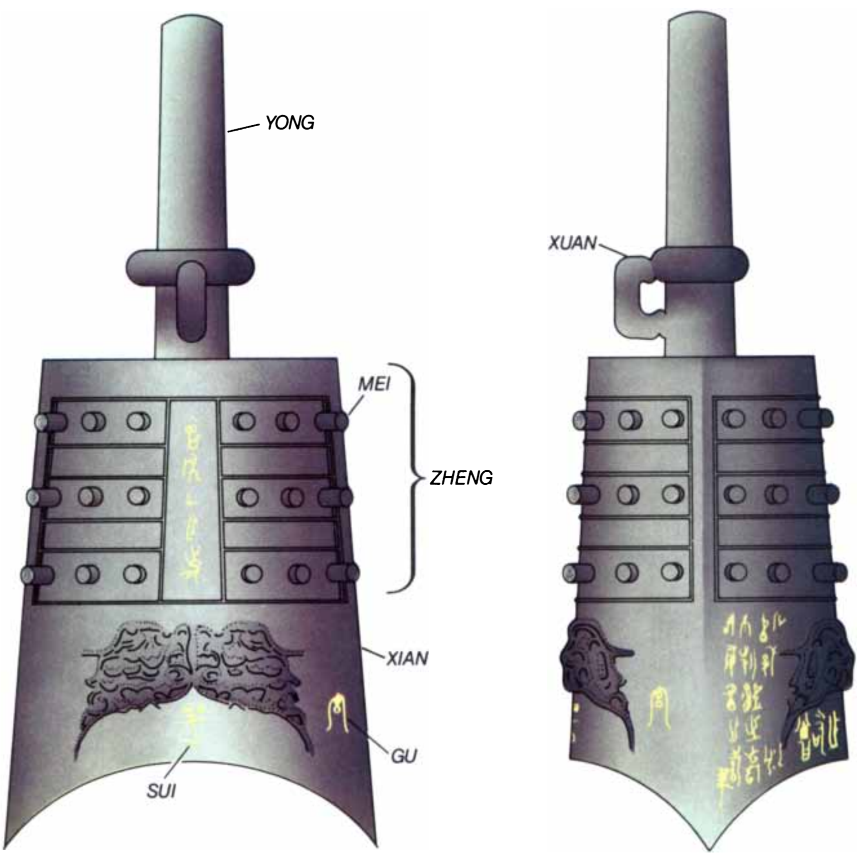
Because chime bells are intended to perform music in the company of orchestras, their desirable qualities are somewhat different from those of a ceremonial bell or a church bell. Each member of the chime should have a broad dynamic range and, to allow the performance of complex melodies, its tones must please the ear and attenuate quickly, without a prolonged echo.

The geometric configuration of the chime bells is crucial to their achieving these acoustical properties. The *zhong* bell has an asymmetrical construction. Unlike the church bell, which is circular in cross section, a *zhong* bell is oblate: its horizontal cross section is a flattened oval. The lip of the bell does not lie on a plane but arches upward in the front and back, then downward into hornlike feet on the left and right sides. The front and back faces meet at a ridge called the *xian.* Covering four

the most impressive set of *zhong* bells found to date. The chime consists of three tiers of bells mounted on an L-shaped frame. Exhaustive acoustical studies have shown how unique structural features enable each bell to sound two pristine fundamental pitches.



MAP OF BELL-CHIME DISCOVERY in China includes only the eastern and southwestern provinces because all the chimes were found there. More than 50 sets of bells have been unearthed since 1900. A circle marks the origin of Zenghou Yi's collection.



ZHONG BELL FEATURES are intimately related to the tonal quality and performance of the bell, shown here in front (*left*) and side views. The bell can produce two pitches because its cross section is asymmetrical; the tones are refined by bronze *mei* nipples. Some bells also carry inscriptions that indicate exact striking positions for each pitch.

regions of the upper bell body are 36 bronze nipples known as *mei*.

In a *bian-zhong* chime each *zhong* bell is suspended by a collar called the *xuan* from a hook mounted on the beam of the chime frame. The clapperless bells are arranged in tiers and played with different kinds of strikers. Usually bells of the high and middle registers are hung at or above eye level; the performers use hammerlike mallets. Bells of the low register are hung in the bottom row and are played using rods wielded in a near-horizontal trajectory.

A bell chime consisting of several dozen bells requires from five to seven performers. Performers of different registers stand on opposite sides of the chime. The bell mouth hangs downward at an angle of about 30 degrees from the vertical, balanced by a long, heavy nose called a *yong* that tilts toward the back so that the side to be played faces the performer. Depending on their pitch, the *zhong* bells range from a few inches to several feet in height: these last are relatively large instruments by modern standards.

The two fundamental pitches that give the bells their notoriety arise from two distinct regions on the bell face. One pitch comes from the lower center of the bell, a position called *sui,* and the other from the areas to the left and right of the *sui,* called the *gu* positions. The term *sui* is said to mean "mirror" and probably refers to the similarity between the curve of the bell lip and the concave burning mirror used at the time to start fires. *Gu* means "drum" or "music-producing." Because of a strong cultural emphasis on right-handedness, the right *gu* was played much more often than the left.

When played in a set, dual-pitch bells allow efficient performance and reduce the overall size of the bell chime. The design of Western bells is not nearly as practical. The "strike" tone of a Western round bell (the pitch heard by a listener) is not the "hum," or fundamental tone, but the note an octave above the hum tone. For instance, a modern American bell with a fundamental pitch of middle C (C_4 at 256 cycles per second) has a perceived pitch of C_5 (512 cycles per second). It weighs 800 pounds. According to the approximate rule that the pitch of a round bell is inversely proportional to the cubic root of its mass, a decrease in frequency by a factor of two would require eight times the original mass. An American bell with a middle-C strike tone would therefore weigh more than three tons—nearly 1,000 pounds more than all the bells and racks in the Zenghou Yi chime combined.

ACOUSTICAL ANALYSIS revealed the precision of the *zhong* design. Zhou engineers learned to manipulate the overtones in the bell's "voice" as well as the two fundamentals. Wang Xiang (*seated*) and Huang Xiang-peng of China's National Music Research Institute are shown here determining the pitches of Zenghou Yi's chime. In spite of their age, the bells were almost perfectly tuned.

The inability to produce a bell that can sound its acoustical fundamental, then, has had serious implications for the material requirements and casting constraints of bell manufacture. Yet the church-bell design is the result of centuries of careful experimentation. The Chinese design, the result of thousands of years of labor, was lost after the Han period (206 B.C. to A.D. 220). What acoustical secrets does it embody that eluded Western designers?

A bell is a very complex acoustical body. Its partials cannot be expressed as simple arithmetic ratios, in contrast to the completely elastic string or a vibrating air column whose frequencies correspond to the ratio 1 : 2 : 3 : 4 : 5 : 6 and so on. Both the oblate *zhong* bell and the round church bell are special adaptations of the acoustical system known as vibrating plates. For plates and bells alike, increasing the thickness and elasticity of the vibrating material increases vibration frequency, whereas increasing diameter and density decreases the vibration frequency.

Mary D. Waller of the London School of Medicine for Women studied the normal modes of vibrating circular plates and published her results in 1937. Her nodal figures consist basically of radii (designated m) distributed symmetrically around the plate's center, and circles (designated n) concentric with the perimeter of the plate. The simplest mode, the one corresponding to the fundamental tone, exhibits four nodal lines that divide the plate into four vibrating sections shaped like pieces of pie. Adjacent segments are always moving in opposite directions at any given instant. The next mode, which generates a frequency 1.7 times that of the fundamental, has only one nodal circle, which delineates an inner circular segment and an outer ring. Other modes arise from additional combinations of radii and circles, creating a rich interplay of partial frequencies.

Acoustically, a round bell behaves like a plate stretched into a flared bell shape and suspended at the center. The normal modes of a round bell, viewed from above, are quite similar to the patterns seen on circular plates. Vibrational motion is most intense at the rim. When the bell is given a blow, the struck side is forced inward and adjacent regions are pushed outward. The rim then passes through its initial circular shape to form another elongated circle at a right angle to the first.

In the course of these vibrations certain parts of the bell remain relatively still compared with others, namely the points at which the different circular distortions intersect. These points represent the nodal lines of the bell, called nodal meridians. They are not nodes in the strict sense, however, because the intersections are not exact, and so some motion continues to occur in the plane of the surface. This motion is what makes a glass goblet ring when one rubs a wet finger around the rim.

Nodal meridians on a round bell are spaced evenly, as are the nodal radii of a plate. Because of this symmetry, a clapper can be used with a round bell, since striking any point on the rim produces the same vibrational effects. The meridians of the asymmetrical *zhong* bell, on the other hand, are not evenly distributed. The consequence of this asymmetry is that, for any given number of nodal lines, more than one spatial arrangement is possible.

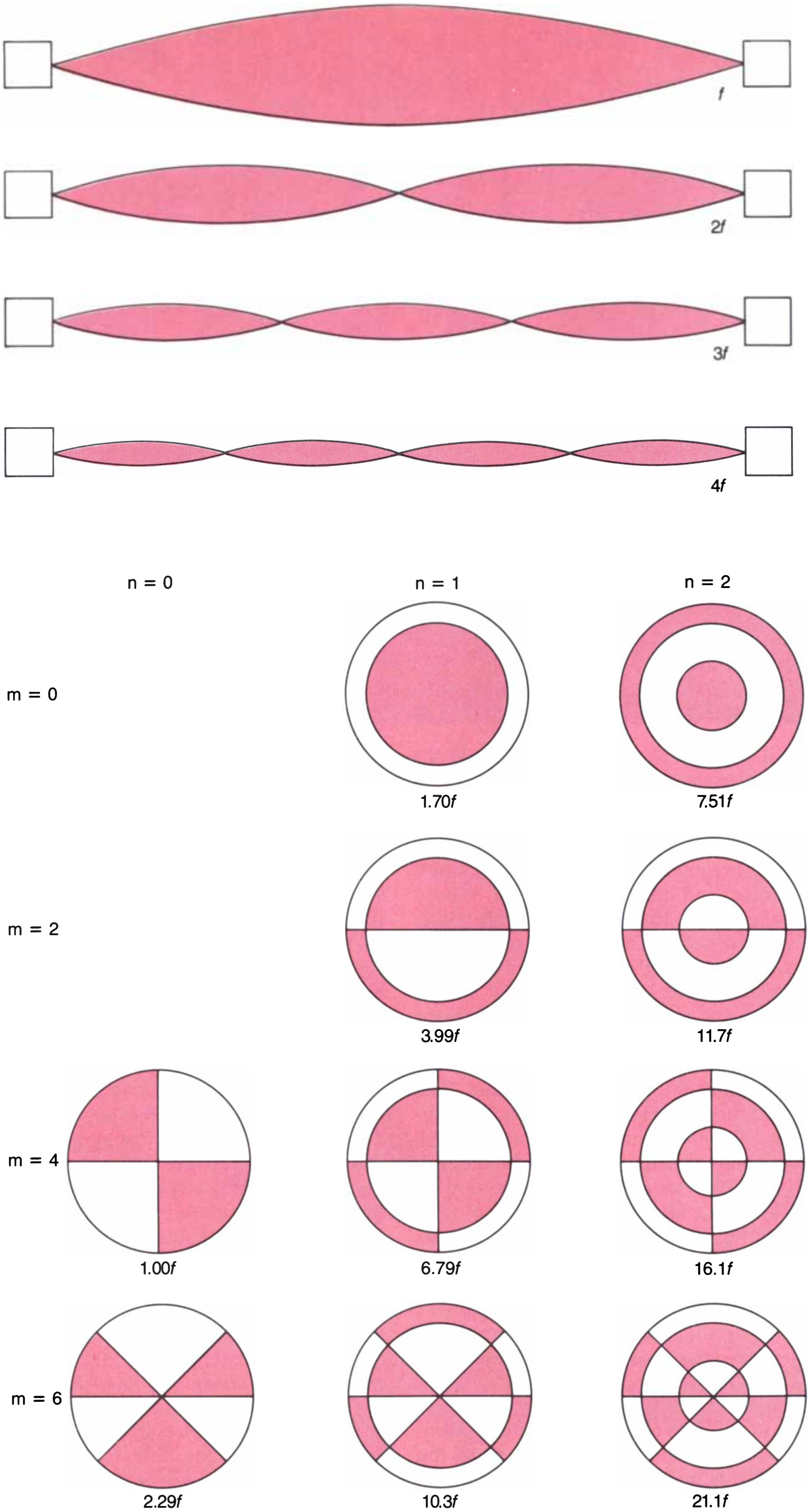
Indeed, the *zhong* bell has two well-defined sets of modes that can be selectively activated by striking different positions on the bell: the *sui* and *gu* positions. Although at the two fundamental pitches the modes have the same number of nodal lines ($m = 4$ and $n = 0$), the placement of the lines is indeed different, so that the frequencies generated are different. This "degeneracy" of vibrational modes accounts for the chime bells' extraordinary acoustical properties. In general the *gu* mode produces higher frequencies than the *sui* mode, but the two are not mutually exclusive: they share certain high-frequency partials that issue from the more complex modes.

The ancient Chinese refined their dual-pitch design to make the two sets of modes distinct in acoustical character yet comparable in musical function. They worked to separate the two pitches of each bell by excluding common qualities. For instance, when the *sui* position is struck, the faces and sides of the bell experience the greatest movement and the *gu* areas in between represent the silent, motionless nodes. The *sui* areas become nodes when the bell receives a blow at the *gu* position. Hence each striking position falls at the point that is least disturbed when the other position is struck, which is also the area least involved in producing the alternate tone.

In order to locate so precisely the nodal meridians of the two fundamental modes, the ancient Chinese must have possessed a theoretical grasp of the physics of music far beyond historians' initial estimates. Given such an understanding, it would be fairly straightforward, although not easy, to find the strike positions that isolate the fundamentals. To achieve the best resolution between pitches, however, overtones as well as fundamentals should be separated. Nodal meridians for the important *gu* overtone modes converge naturally at the *sui* strike position; therefore they do not contribute to the *sui* tone. The meridians of *sui* overtone modes, however, do not congregate at the *gu* position, so that traces of *sui* partials could interfere when the *gu* tone is struck. Herein lies the rationale for the arched lip of the *zhong* bell.

When we examined the two sets of normal modes, we found that the arch in the bell lip, by changing the shape of the vibrating "plate," also alters the nodal patterns of the most important *sui* overtones. Because of the arched lip, the *sui* overtone meridians converge at the point designated, not coincidentally, as the *gu* strike position. It is typically three-fifths of the way from the *sui* position to the *xian* ridge. The striking position is so critical to obtaining the proper pitch and tone quality that the ancients made inscriptions on the bells to indicate *sui* and *gu* placement unequivocally. Neither the concave rim design nor the precision in identifying the convergence of nodal lines could have been accidental.

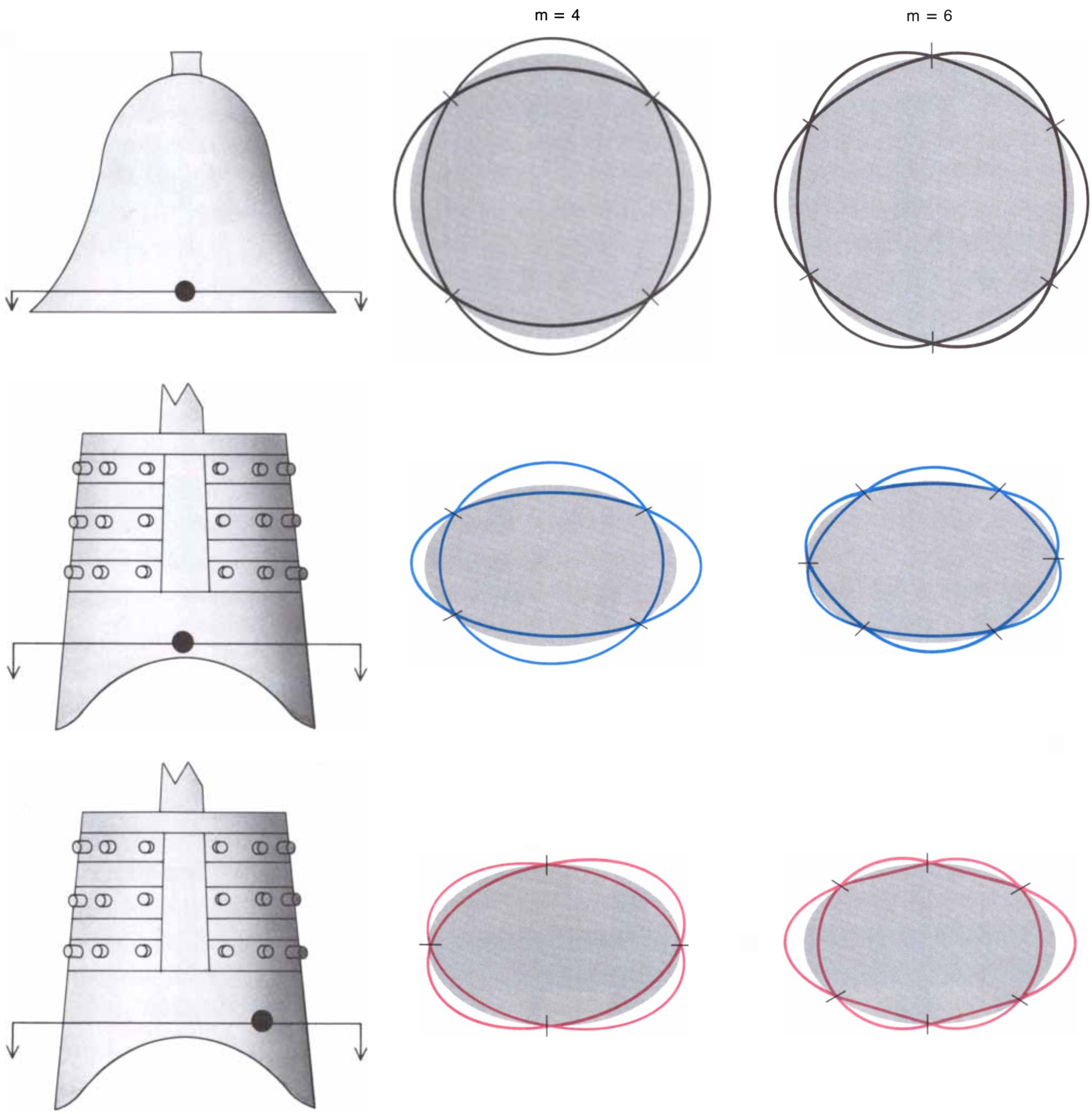
The *mei* nipples clustered at the top of all but the highest-register bells are also more than ornamental. They help to balance the strength of the two fundamentals so that their volumes are



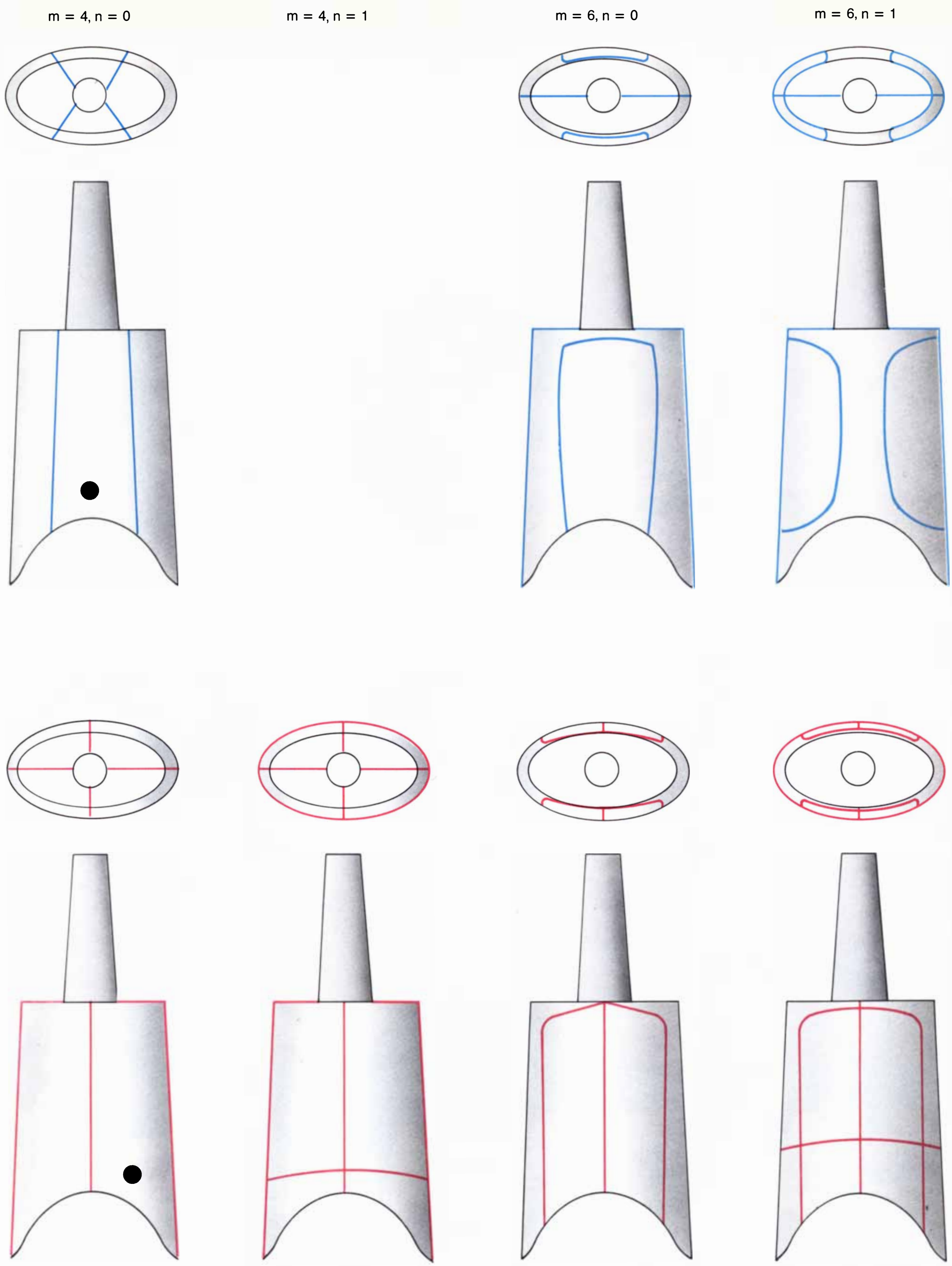
PARTIAL FREQUENCIES, which make up the tonal quality of a sound, result from the different modes of a vibrating body. The lowest frequency is called the fundamental; other frequencies are overtones. When a taut string is plucked (*top*), it produces partials that are whole-number multiples of the fundamental frequency *f*. A vibrating plate (*bottom*), however, gives rise to partials that cannot be related by simple arithmetic ratios. Instead frequencies depend on the combination of nodal radii (*m*) and nodal circles (*n*): regions of the plate that remain still during vibration. Red and white represent movement in opposite directions. A bell, whether it is round or oblate, is a special kind of vibrating plate.

comparable. More important, the *mei* act as another device to separate the two bell tones. Recent laboratory studies have found that the nipples change the complete overtone structures, or frequency spectra, of *sui* and *gu* tones. The *mei* provide extra weight around the bell shoulders, altering nodal patterns in the upper part of the bell. Accordingly they are most pronounced on large bells. Without the *mei*, the *sui* and *gu* fundamentals are easily distinguished, but they have certain high-frequency overtones in common. With the nipples, overtone frequencies shift so that little overlap between the two sets of partials occurs.

The interval between the two pitches on a *zhong* bell is selected by casting and tuning. The choice of the interval is arbitrary, but it should suit the melodic progression of the compositions it will perform. In addition the interval should not be a discord since, in spite of the efforts of the designers, traces of the secondary tone may persist after the primary tone has died down. Zhou engineers tuned their bells so that the intervals of the overtones, as well as those of the fundamentals, were harmonic. The second partial of the *gu* tone, for example, is always an



MOTION AT THE RIM when a bell has been struck illustrates the distribution of nodal lines for a round bell (*top*) and for the two pitches of a Chinese *zhong* bell, called *sui* (*middle*) and *gu* (*bottom*). Gray indicates resting positions; lines show changes in the bell's shape after it has been struck. Modes in which *m* is equal to 4 and 6 are represented. At these modes the nodal radii of a round bell are evenly spaced; hence only one pattern of distribution is possible. Because a chime bell is oblate, however, a given number of nodal lines can be arranged many different ways, and different nodal patterns give rise to different pitches. As is shown here, the pattern of nodal radii is determined by the point at which the bell is struck. Chinese designers managed to maximize the separation between the two *zhong* pitches by having the nodal lines of one pitch serve as the striking position for the other.



NODAL PATTERNS for the *sui* (*top*) and *gu* (*bottom*) pitches describe the modes of the fundamentals and the most conspicuous overtones. Top-view renderings above each bell recall the vibrating-plate system for enumerating radii and circles. Horizontal lines count as circles even when they do not encircle the bell, but near the bell shoulders they do not contribute to the bell's sound and so are not counted. As the top illustration on page 110 shows, the *sui* tone has two dominant overtones and the *gu* tone three.

octave plus a major or minor third above the *sui* tone; the partials of a *zhong* bell with a minor third separating its two pitches are in the ratio 1 : 1.2 : 2.4 : 2.81 : 3 and those of a bell with an interval of a major third are 1 : 1.25 : 2.5 : 2.81 : 3.

When bell-chime intervals from several periods are compared, a historical trend toward the major- and minor-third intervals represented in Marquis Yi's bell collection becomes apparent. Lacking examples of ancient music, one can only assume that this preference matches a taste for major and minor thirds in the musical composition of the time. In Europe these intervals were not recognized as harmonic until the 12th century.

When Marquis Yi's *bian-zhong* was found, its bells were almost perfectly tuned in spite of their prolonged burial. A second set of 36 bells discovered in the same principality of Zeng in 1981 were even better tuned. Ordinarily a vibrating plate is tuned by the addition or removal of material, but with this method it would seem impossible to tune one bell pitch without altering the other, since both are contained on one continuous body. How did the Chinese tune two pitches on a single *zhong* bell?

Again the ancients made use of their remarkable expertise in pinpointing nodal lines. They could tune just one pitch by paring bronze off a bell's inner surface if, in paring, they carefully followed the nodal lines of the other. Hence tuning the *sui* pitch entailed removing metal from the *gu* nodal lines, and vice versa. At the same time, many bells were cast so accurately that they required no modification.

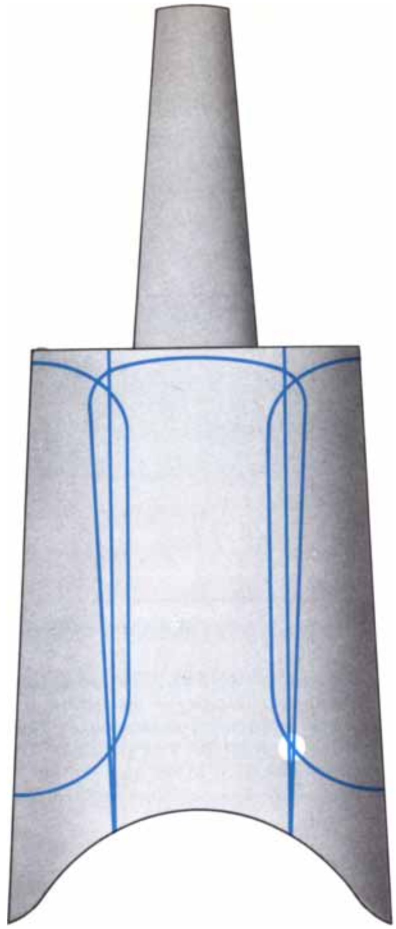
The methodology of bronze casting with pottery sectional molds was fully developed in ancient China. Even so, manufacturing a low-register *zhong* bell presented quite a formidable task. Large, complicated objects were often made by casting separate parts, then uniting the components in a final mold. But the bronze chime bell, no matter how large, was always cast as a single piece.

In the modern world the closest approximation of a Chinese bell chime is the carillon. A carillon consists of a group of bells carefully selected to yield equal-tempered chromatic intervals. The bells are played from a keyboard. Because some of the partials are in discord, only a limited number of chords are within the carillon's scope, and ordinarily only a single melody is played. Considerable knowledge and skill are needed to render satisfactory effects; even expensive installations can give disappointing performances. If the carillon is the best that modernity has to offer, one might wonder how the Chinese arrived at the *bian-zhong* design so long ago.

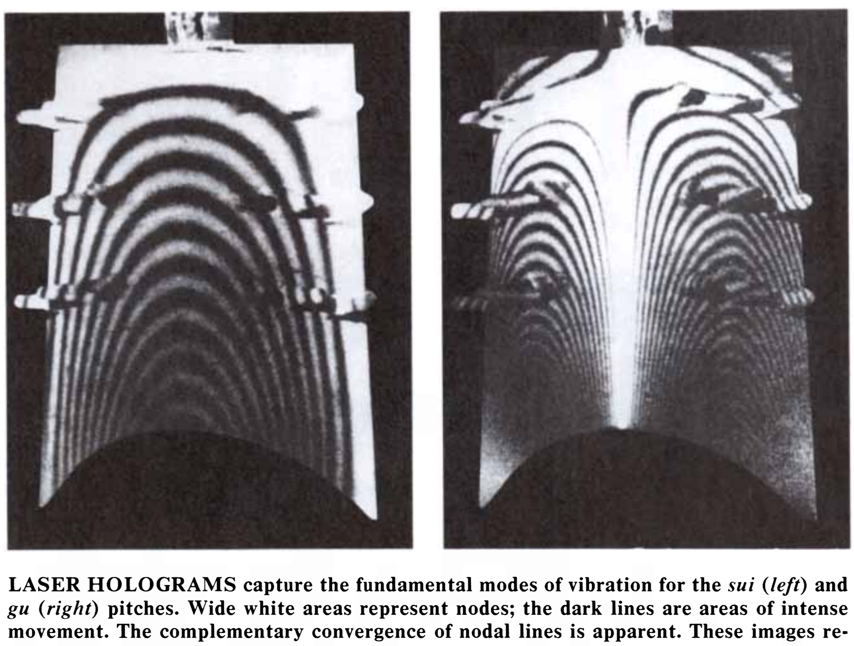
Since ancient times Chinese musicians have been sensitive to subtle differences in tonal quality. This sensitivity is manifested in many of their instruments. For example, I once asked a Chinese wind player about the design rationale of the *shuang-guan,* an oboe consisting of two apparently identical cylindrical oboes. He replied, "No two reeded tubes have the same tonal spectrum. With two tubes on the *shuang-guan,* the player has control over a broader range of tonality and can be more selective of the timbre."

Perhaps this heightened sensivity to tonal structure led the ancient Chinese to experiment with a slightly elliptical bell design, which gives rise to a fuller set of vibrational modes and a broader range of overtone possibilities. *Ling* bells, relics of the early Shang period, were oblate but still had clappers and produced only one tone; their musical character can be inferred from the fact that they were used as bells for dogs and cows.

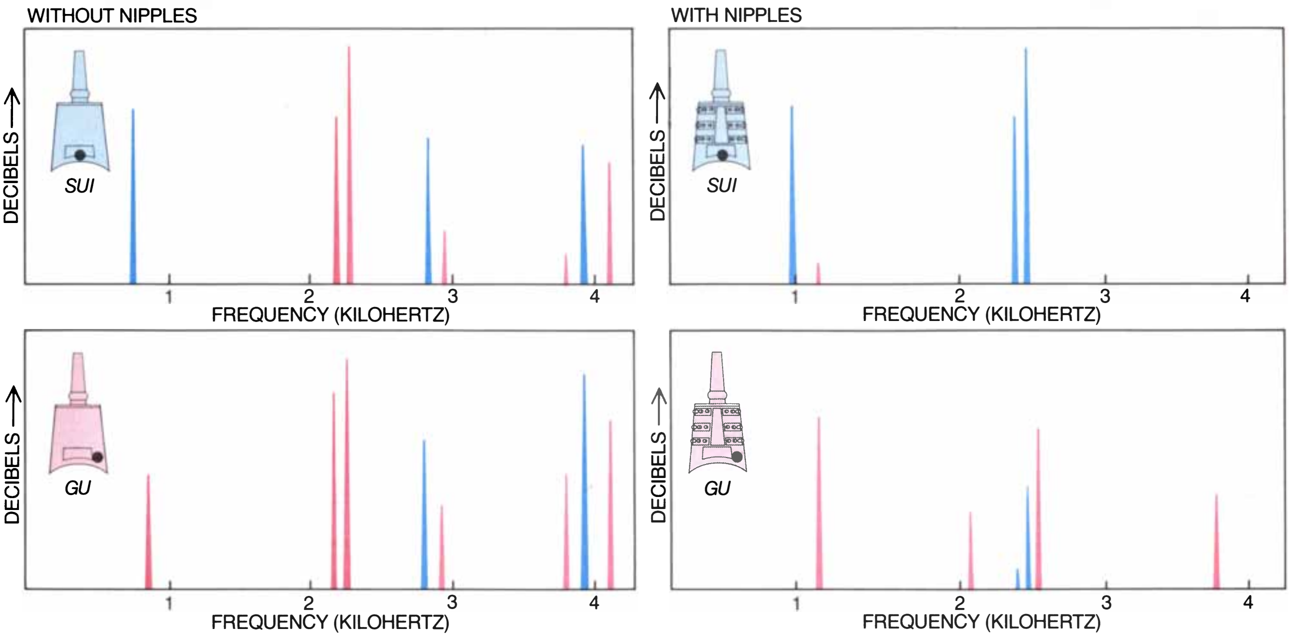
In subsequent efforts the Chinese may have struggled with oblate bells that sounded two distinctly different pitches but produced only awkward or muddy tones. The large, oblate *zheng* handbell, which predates the chime bell by hundreds of years, exemplifies this stage of development. Indeed,



NODAL-LINE CONVERGENCE for the three dominant partials of the *sui* tone is an important consequence of the chime-bell design. If the perimeter of the bell were flat, *sui* nodes would never intersect; hence a blow to any part of the bell would excite at least one *sui* tone. The arched lip of the *zhong* bell, however, rearranges nodal lines so that they congregate at the *gu* position (*circle*). This convergence helps to clear the *gu* tone of "muddy" *sui* echoes.



LASER HOLOGRAMS capture the fundamental modes of vibration for the *sui* (*left*) and *gu* (*right*) pitches. Wide white areas represent nodes; the dark lines are areas of intense movement. The complementary convergence of nodal lines is apparent. These images result from the differential reflection of light by moving and stationary surfaces of the bell.

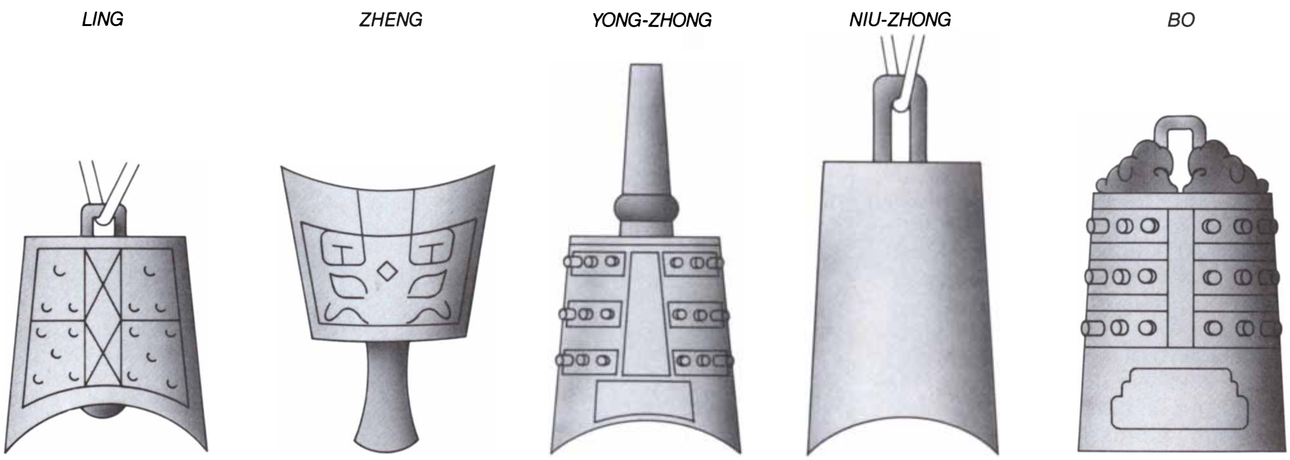


GRAPHS OF PARTIAL FREQUENCIES emphasize the importance of the *mei* nipples in separating the *sui* (*blue*) and *gu* (*red*) tones. The frequencies generated by bells without *mei* (*left*) are almost identical for the two positions. Only the fundamentals are noticeably different. When nipples are added (*right*), nodal patterns in the upper part of the bell are rearranged, changing some frequencies and eliminating others. Almost all traces of *gu* tones disappear from the *sui* profile, and only hints of *sui* partials remain in the *gu* tone. *Mei* nipples also serve to increase the volume of the *gu* fundamental so that it matches the volume of the *sui*.

zheng bells lend their name to the upper half of the *zhong* bell, and the *yong* nose that holds the *zhong* at a 30-degree tilt is derived from the *zheng* handle. The clapperless *zheng* bell, however, has a muddy *gu* tone because its flattened mouth does not force the convergence of *sui* nodal lines.

With modification of the bell's surface, cross section and thickness, the two pitches were brought into a well-defined harmonic relation and their individual tones were sharpened; eventually the sophisticated *zhong* design emerged. Contemporaries of the *zhong* bells include the *niu-zhong,* which is also dual-pitch, and the ceremonial *bo,* which lacks the musical agility of the *zhong*. Both these bells are played in the vertical position.

The art of music has a long and distinguished history that parallels the rise of human civilization. The principles embodied in the Zenghou Yi bell chime suggest that the science of music may have a history almost as long, and just as distinguished. China in the Shang and Zhou periods possessed a level of acoustical science that was essential in supporting the elaborate musical art of those periods. Physics and engineering worked hand in hand to perfect wind, string and percussion instruments and to arrive at orchestration. These are accomplishments that have modern counterparts. In contrast, the overall design of a large set of dual-pitch bells for the performance of music is an achievement that has no equal in the modern physics of music.



BELL ANCESTRY suggests the route the ancient Chinese followed to arrive at the *zhong* design. The *ling,* oldest of the five bells drawn here, goes back more than 3,600 years. It had a clapper and produced just one unremarkable note. The *zheng* handbell is an early example of the dual-pitch design; it was used by Chinese soldiers, who carried it mouth up. The *niu-zhong* also produces two pitches, but it lacks the clarity of the popular *yong-zhong* at low registers. *Niu-zhong* are visible in the top tier of the chime on page 105. Another *zhong* contemporary, the *bo,* served as a monotone ceremonial bell. The bells are not drawn to scale.

SCIENCE/SCOPE®

The world's fastest digital integrated circuit, a gallium arsenide (GaAs) chip that runs at a clock rate of 18 gigahertz (GHz), or 18 billion cycles per second has been built by Hughes Aircraft Company scientists. The ultra high-speed circuit operates as a divide-by-two frequency counter and is five times faster than currently available GaAs integrated circuits and ten times faster than commercial silicon circuits. Fastest frequency reported previously for static frequency dividers was 13 GHz for a laboratory device requiring cryogenic temperatures; the Hughes circuit operates at room temperature. Operation of digital circuits at multi-GHz frequencies opens new areas of digital communications and signal processing, promising better noise immunity, a wider range of functions, and less complexity than their analog counterparts. Applications are foreseen in fiber optic communication links, supercomputers, advanced radars, and satellite communications.

Ships at sea will be able to determine their positions via satellite. A maritime navigational system is one of the new services proposed for the existing system of Marisat satellites, launched in 1976. For the past 4 years, the trio of Marisats has been providing telecommunications services for the International Maritime Satellite Organization (INMARSAT), a cooperative of 47 countries that operate a worldwide system for maritime communications. Leases with INMARSAT have been renewed for three years by Comsat General Corporation, owner of the satellites, enabling the Hughes-built satellites to continue providing communications services to the military, shipping, and offshore industries.

A unique computerized visual system helps military forces simulate battlefield terrain. The system provides unusual realism and flexibility to help with a wide range of training and mission planning requirements. It can generate lifelike three-dimensional scenes from a computer database created with aerial photography. Pilots can use the system for nap-of-the-earth flight training, even to the point of seeing simulated radar and infrared displays. The Hughes system also can be used for intelligence analysis and team tactics training.

A night vision system for helicopters significantly reduces pilot workload by eliminating wasted movements, simplifying controls, and providing excellent video images and object detection in reduced visibility. The Hughes Night Vision System (HNVS) is a low-cost, forward-looking infrared (FLIR) system that provides a pilot with automatic tracking and digital video processing. It superimposes FLIR video, flight symbology, and navigational data on a single display, which can be mounted on the flight panel or in a helmet visor. The helmet visor display projects a FLIR image onto a biocular holographic combiner on a see-through visor. A helmet linkage, which moves the FLIR as the pilot's head moves, reduces the pilot's workload further and improves flight safety.

Hughes Missile Systems Group, in Canoga Park, California, an attractive suburb of Los Angeles, is advancing every phase of research, development, and manufacturing technology as it applies to tactical guided missile systems and strategic defense. Opportunities for engineers and scientists are in analog/digital circuit design, software engineering; high-voltage power supply design; microwave process, transmitter and radome design; electro-optical design; IR imaging sensors; focal plane arrays; quality assurance engineering; operations analysis; mechanical engineering; and production design and systems engineering. Send resume to Hughes Missile Systems, Engineering Employment, Dept. S2, 8433 Fallbrook Ave., Canoga Park, CA 91304. Equal opportunity employer. U.S. citizenship required.

For more information write to: P.O. Box 45068, Los Angeles, CA 90045-0068

The First Technology

A young archaeologist re-creates a prehistoric flaked-stone technology in order to understand how our ancestors made and used early stone tools more than two million years ago

by Nicholas Toth

Human beings are not the only animals that use tools. A variety of mammalian and nonmammalian species employ materials such as stones or twigs in obtaining food or in other activities. Our nearest relatives, the chimpanzees, occasionally modify raw materials to shape simple implements for collecting edible insects and water and for grooming. What is more, the behavior is learned: young chimpanzees become proficient by imitating their elders. Yet even among chimpanzees the use of tools does not appear to be indispensable, whereas among modern human beings it is essential for survival. Few people could survive in the wild without implements of some kind. How did this come to be?

The answer lies between two and three million years ago in our African past. At some time in that span our ancestors embarked on a way of life that included a critical new element never observed in any other species: the manufacture of tools by means of a flaked-stone technology. This new technology seems to have developed in association with critical changes in diet and in biological form. How the early stone tools fitted into the lifeways of their makers is not altogether clear. Traditional studies have been hampered by a lack of comparative information, because early hominids (representatives of the family that includes modern human beings) were neither exactly like modern humans nor like chimpanzees or baboons.

Partly to overcome this lack, some archaeologists have begun to try to put themselves in the place of early hominids to see how the first tools might have been made and employed. At a site called Koobi Fora in northern Kenya, where the oldest archaeological remains are 1.9 million years old, I did a series of replicative and functional experiments in which I made and used the earliest types of tools. The results of such "experimental archaeology" are intriguing. My research suggests that the earliest stone technology placed great emphasis on stone flakes and not solely on the flaked core, or parent rock, as has traditionally been thought. It also offers evidence that by about two million years ago our ancestors had become right-handed, a finding that has important implications for the development of the human brain.

Experimental archaeology has a long although somewhat erratic history. Beginning in the 19th century a number of antiquarians attempted to flake stone in order to find out how stone tools had originally been made. Some of the most enterprising "experimental archaeologists" of the past century, however, were forgers, who made stone tools in order to sell them (a practice that tended to give such experimentation a bad name). In general neither the legitimate nor the illegitimate efforts were particularly systematic, and with a few exceptions the idea of in-depth functional experiments in archaeology lay fallow until the second half of the 20th century. At that time pioneer scholars of African prehistory such as J. Desmond Clark of the University of California at Berkeley and Louis Leakey began to put the experimental approach to early stone tools on a more rigorous, problem-oriented footing.

Yet much of the recent work only skimmed the surface of what seemed possible using the experimental approach. Most experiments concentrated on, say, the manufacture of a single type of tool to find out how it was done; many of these tools were relatively sophisticated compared with the

AUTHOR BUTCHERS ELEPHANT with stone tools he made himself; an associate loads the meat for weighing. (The large African elephant died of natural causes; none of the animals employed in the author's research were killed for experimental purposes.) The combination of stone flakes and larger implements employed to butcher the elephant is characteristic of the earliest stone-tool technologies, which developed in Africa about 2.5 million years ago.

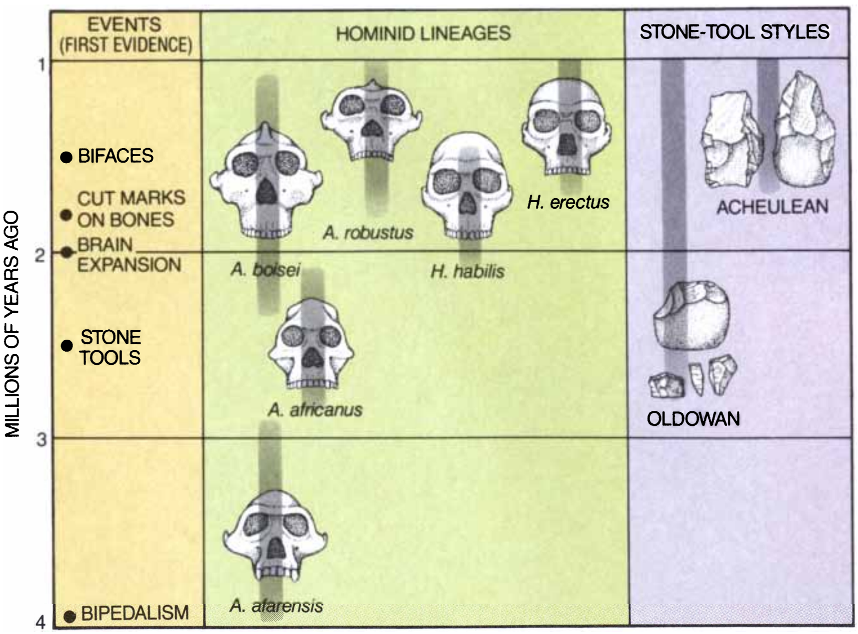
earliest types. Alternatively, Clark and others gave prehistoric tools to contemporary African hunter-gatherers in an attempt to find out how the ancient specimens might have been employed originally. In contrast to these approaches, I wanted to make and use the entire range of artifacts from the early stages of technology. In that way I hoped to reconstruct not just a single technique but, insofar as possible, the entire universe of technical possibilities available to our early ancestors.

Koobi Fora proved to be an excellent place for such an approach. This area, which lies on the northeast side of Lake Turkana in northern Kenya, was the site of a research program headed by the late Glynn Isaac, then at the University of California at Berkeley, and Richard E. F. Leakey of the National Museums of Kenya. Under their joint direction an international team of scientists, including paleontologists, anatomists, geologists, geochemists and archaeologists, had been carrying out an integrated program of research since 1969. In a project directed by Isaac and by John W. Harris of the University of Wisconsin at Milwaukee, the prehistoric sites in the Koobi Fora area had been excavated by means of modern techniques, resulting in excellent recovery of all archaeological materials [see "The Food-sharing Behavior of Protohuman Hominids," by Glynn Isaac; SCIENTIFIC AMERICAN, April, 1978].

There was much to be learned from this trove of material, because Koobi Fora includes 20 excavated sites from the early Stone Age, ranging in age from about 1.4 to 1.9 million years. Those sites are found in a wide variety of geographic and sedimentary contexts, providing much comparative information. In addition, all the types of rock utilized in prehistory for making stone tools are still accessible in modern stream beds, and so they were available for my experiments. The chief material available to early hominids was a medium-grained lava; other rock types that were less commonly used included chert, quartz and ignimbrite (a volcanic rock that is formed from hot ash flows).

The Koobi Fora sites are today situated in the midst of a semiarid badland bounded by volcanic highlands on the east and Lake Turkana on the west. Between some 1.4 and 1.9 million years ago, however, the area was quite different in an ecological sense. For one thing, it was considerably richer. The plain was covered by a savanna grassland. Along the streams that flowed down from the highlands toward the forerunner of Lake Turkana were gallery forests: thin strips of green running parallel to the stream bed on each side. The landscape was populated by such animals as ancient forms of elephant, rhinoceros, hippopotamus, buffalo, giraffe, zebra, wildebeest, lion, leopard, crocodile, hyena, warthog and porcupine, as well as a range of antelopes, large saber-

AFRICA was the cradle of humanity: most of the important early evolutionary events took place there. The map shows Stone Age sites that are thought to be more than one million years old. The author did much of his "experimental archaeology" at Koobi Fora.



BRAIN EXPANSION accompanied the emergence of technology; there may well have been feedback between the processes. For more than a million years, while stone tools arose, there were two hominid lineages. One lineage—the later australopithecines, among them *Australopithecus boisei*—probably had a tough vegetable diet. A larger-brained lineage, probably with a diet of higher quality, included an ancestor of human beings: *Homo habilis*. The first stone technology, the Oldowan, included tools made from cobbles. Beginning about 1.5 million years ago a new technological mode, the Acheulean, appeared. Many Acheulean tools were made from large flakes, some more than 20 centimeters long.

toothed cats and primates such as baboons and colobus monkeys.

The hominid fossils of the area suggest that there were two different hominid forms in the same landscape. Both species were experiments in bipedal locomotion (walking upright on the hind limbs) among apes. The first was a robust (large-muscled), small-brained form known as *Australopithecus boisei.* The later australopithecines, the group that includes *boisei* and related small-brained hominids, are thought to have been specialized for eating a diet of relatively tough vegetable foods—as is suggested by their large cheek teeth and jaw muscles, along with patterns of tooth wear.

The other bipedal ape, known as *Homo habilis,* was in all probability the direct ancestor of human beings. *H. habilis* was, compared with its australopithecine cousins, more slender (gracile is the technical term) and larger-brained. Roughly 1.6 million years ago this species appears to have given rise to a descendant called *Homo erectus,* which was the immediate predecessor of *Homo sapiens.* Anatomical features suggest both early *Homo* forms were dietary generalists, probably eating a wider range of foodstuffs than the australopithecines, including a fairly high proportion of nutrient-rich foods such as animal products.

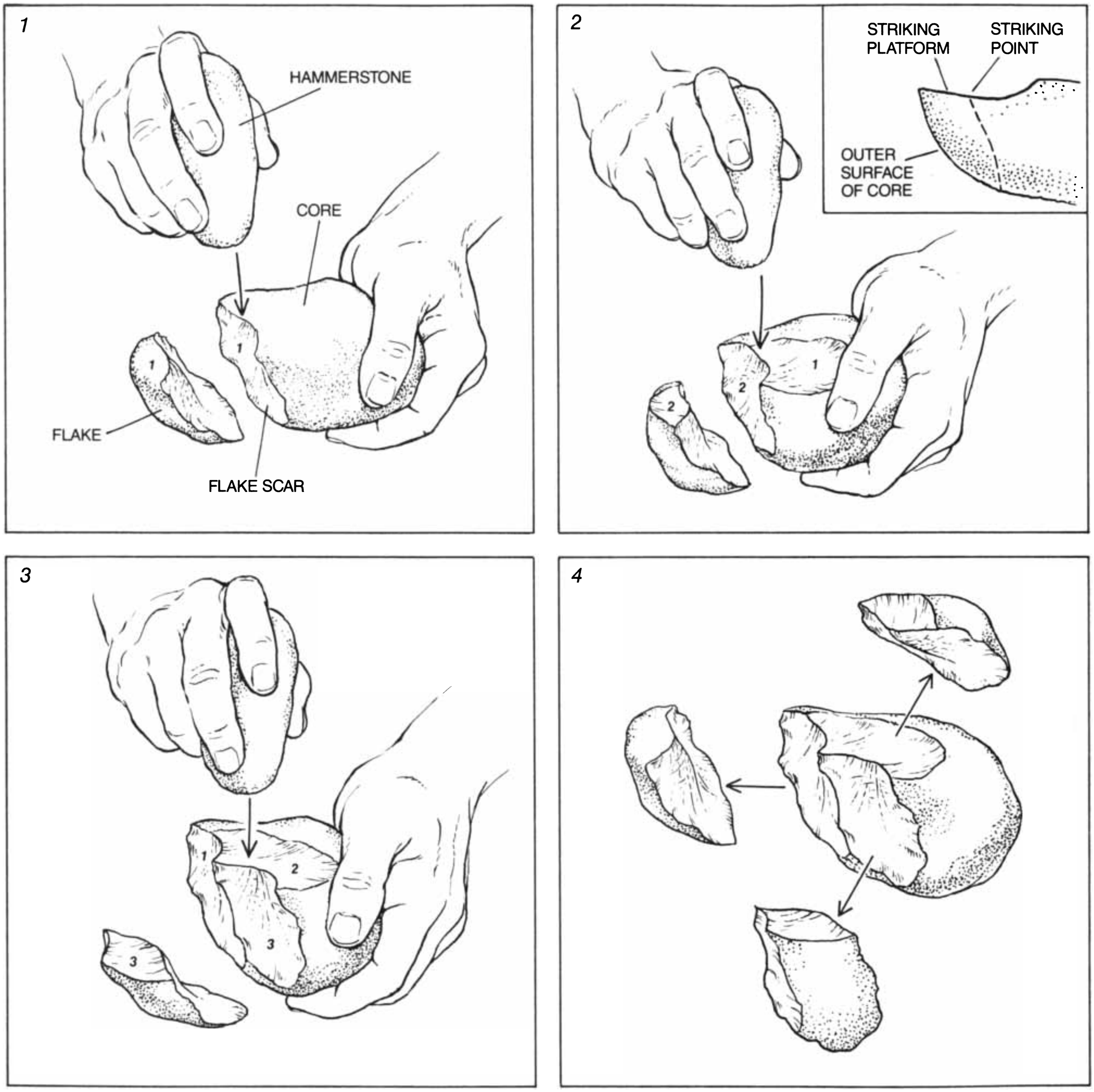
For about a million years, until approximately 1.2 million years ago, the australopithecines lived in the same African landscape as the *Homo* lineage. (Indeed, their remains are sometimes found at the same sites.) The period of their coexistence is, from the point of view of technology and evolutionary adaptation, a critical one. The earliest reliably dated tools, about 2.4 million years old, have been found in the Omo Valley in Ethiopia. At the Omo sites archaeologists have found fragments of quartz pebbles and other rocks that were deliberately shattered. Unfortunately no animal bones were preserved with these tools, and so the uses they served are not known.

At sites such as Koobi Fora and Olduvai Gorge in Tanzania the deficiency is remedied. These sites have yielded not only simple stone artifacts as old as 1.8 to 1.9 million years but also well-preserved fossilized bone. The technology of this period is generally referred to as Oldowan, in reference to the Tanzanian site, where such artifacts were first described by Louis and Mary Leakey. Most of the Oldowan implements are lava cobbles or quartz and quartzite blocks that have been struck with a cobble hammer, resulting in a core and flakes or fragments. Sometimes the flakes were retouched or even made into cores themselves.

After about 1.5 million years ago a new technology was added to the Oldowan repertory. This technology is known as the Acheulean. Its hallmark is that the raw material for the implement was often not a small river cobble but a large flake struck from a boulder. Such flakes, some more than 20 centimeters long, offered excellent material for making the large bifacial forms that scholars of prehistory call picks, hand axes or cleavers.

Since the *Homo* lineage and the australopithecines were both present for much of this technological development, it is not yet possible to demonstrate conclusively who made the majority of early stone tools. Circumstantial evidence seems to point to *Homo* as the main toolmaker, however. Stone tools are not directly associated with the australopithecines that lived before the emergence of the genus *Homo*, and stone-tool manufacture continued to flourish after the *Australopithecus* line became extinct, by about a million years ago. Furthermore, when hominid fossils and stone artifacts are both present in the same level, at least some of the hominids tend to be *Homo*. Finally, the representatives of the genus *Homo* were larger-brained and therefore perhaps more likely to have been the habitual toolmakers.

I set out to understand the nature of early stone technologies by means of a holistic experimental approach. In



BIFACIAL FLAKING was one of the chief methods employed by early toolmakers. It was generally carried out by means of the technique called hard-hammer percussion. A stone serving as a hammer was held in one hand and the core, or rock to be flaked, was held in the other (*1*). After a flake had been struck off, the core was turned so that the flake scar (here numbered to match the flake) could serve as a platform for striking off the next flake (*2*). For effective flaking the angle between the striking platform and the core's outer surface must be less than 90 degrees (*inset*). Striking off more flakes (*3, 4*) yields a core with a bifacial edge.

OLDOWAN-STYLE TOOLS were made by the author, who has replicated the entire range of Oldowan core forms, along with a variety of flakes. In traditional terminology the core forms in the upper row would be called (*left to right*) hammerstone, unifacial chopper, bifacial chopper, polyhedron, core scraper and discoid. In the bottom row are a flake scraper (*left*) and six unretouched flakes. The author concludes that simple flakes may have been just as important in early stone technology as the core forms.

ACHEULEAN-STYLE TOOLS, made by the author, include an ovate hand ax, a pointed hand ax, a cleaver and a pick (*upper row*). Below are a spheroid, a flake scraper and three flakes. The author found Acheulean tools very effective for butchering larger animals.

my work, stone technology was studied as a system, one that included several stages. The first stage was the initial acquisition of raw material. Acquisition was followed in turn by manufacture, use and discard—and ultimately by the geologic burial of the artifacts. Another step, transport of the stone, could have taken place at any point between the acquisition of the raw material and the discard of the flaked stone. I wanted to examine each step in the system. Because the great majority of the sites at Koobi Fora have yielded artifacts in the Oldowan style, I focused much of my attention on that style, but without ignoring the Acheulean.

In thinking about the acquisition of stone for toolmaking, one important consideration was how far the hominids had carried the raw material before fashioning the tool itself. At the Koobi Fora sites that were adjacent to streams at the time of occupation, the primary source of material could have been nearby gravel bars. At those sites transport distances might have been as little as a few meters. Geologic studies have shown, however, that at other sites, particularly those in river deltas or near the margins of the ancient Lake Turkana, the nearest sources of rock suitable for flaking would have been at least several kilometers away. Many contemporaneous sites at Olduvai Gorge give evidence of similar transport distances.

For the purposes of comparison I examined both the gravels in modern stream beds and those that would have been on the surface at the time of hominid occupation. My aim was to find out what range of rock types had been available to the early toolmakers. In both modern and prehistoric samples the predominant type of rock was a dark lava. Generally more than 90 percent of the cobbles in each stream bed were of this material, the remainder consisting of ignimbrite, chert and quartz. Although the rarer rocks tend to be easier to flake than lava and also yield sharper edges, the early hominids at Koobi Fora appear to have made no special attempt to pick them out: the rock types chosen for artifacts are in direct proportion to their abundance in the prehistoric gravels. (At Olduvai Gorge there are indications of greater selectivity.)

Yet the protohuman toolmakers at Koobi Fora were well aware of other characteristics of their material. They had a keen eye for avoiding lava cobbles with defects that render them unsuitable for flaking. The interior of a heavily weathered cobble can become oxidized, so that the stone will flake unpredictably when it is worked. The sign of such excessive oxidation is often a hairline fracture on the surface, known as a "weathering flaw." Rocks with weathering flaws are common in the stream gravels at Koobi Fora, but the characteristic fragments they give rise to are rare among the artifacts uncovered there. It seems clear the early hominids had already learned to reject such inferior material.

The second stage of the system is tool manufacture. In the course of thousands of experiments I tested the same raw materials and methods that early hominids might have employed. I tried a range of flaking techniques to find out which of them was used at Koobi Fora. Among them were hard-hammer percussion (striking a core with a sharp, glancing blow from a stone hammer), anvil technique (striking the core on a stationary anvilstone) and bipolar technique (striking the core with the hammerstone while it rests on the anvil). The fracture patterns on cores, flakes and fragments resulting from each technique were analyzed along with the overall efficiency of the technique. The comparative data showed that the only major technique employed at Koobi Fora was hard-hammer percussion, which for lava cobbles is the most efficient of the three techniques.

The cores that result from hard-hammer percussion are conventionally classified in a variety of morphological types, which archaeologists have given names such as choppers, discoids, polyhedrons and core scrapers. Most traditional studies have accepted the tacit premise that among early toolmakers these cores were the primary tools, the flakes and fragments being largely waste material. Indeed, it has been thought that the shapes of these early cores correspond to the "mental templates" of the early toolmakers: consistent geometric shapes on which they drew in manufacturing their implements.

My experimental findings suggest that far too much emphasis has been put on cores at the expense of flakes. It seems possible that the traditional relationship might be reversed: the flakes may have been the primary tools and the cores often (although not always) simply the by-product of manufacture. My experiments show that many of the early core shapes can emerge without any conscious intent in the process of removing flakes and fragments from a cobble. Thus the shape of many early cores may have been incidental to the process of manufacture and therefore indicative of neither the maker's purpose nor the artifact's function. Conversely, the sharp-edged flakes often turn out to be extremely effective cutting tools, as I shall show in describing my functional experiments.

The notion that many of the Oldowan core forms did not necessarily result from mental templates was confirmed by work with untrained stone knappers (toolmakers). In order to understand better the process by which one learns to flake stone, I supplemented my own experiments with investigations of toolmaking among selected samples of people with no formal training in stone-tool studies. Initially most of them produced heavily flaked and battered cores that bore little resemblance to early Stone Age tools. In only a few hours, however, most novices can master the basic mechanics of stone flaking. Thereafter they can make the entire range of Oldowan core forms. What is more, many of these forms emerge spontaneously in the process of learning to flake stone.

Although the early stone artifacts from Koobi Fora do not display much technical finesse, it is clear to me that the early hominids had also mastered the basic skills. In order to flake stone by hard-hammer percussion in a controlled way, three conditions must be met at the same time. First, the core must have an acute edge (one with an angle of less than 90 degrees) near which the hammer can strike. Second, the core should be struck a glancing blow at a point generally about a centimeter from the acute edge. Third, the blow should be directed through an area of high mass, such as a ridge or a bulge [*see illustration on page 115*]. Examination of the Koobi Fora artifacts reveals that their makers had gained the cognitive and motor skills needed to fulfill the three conditions and strike off a reliable sequence of flakes.

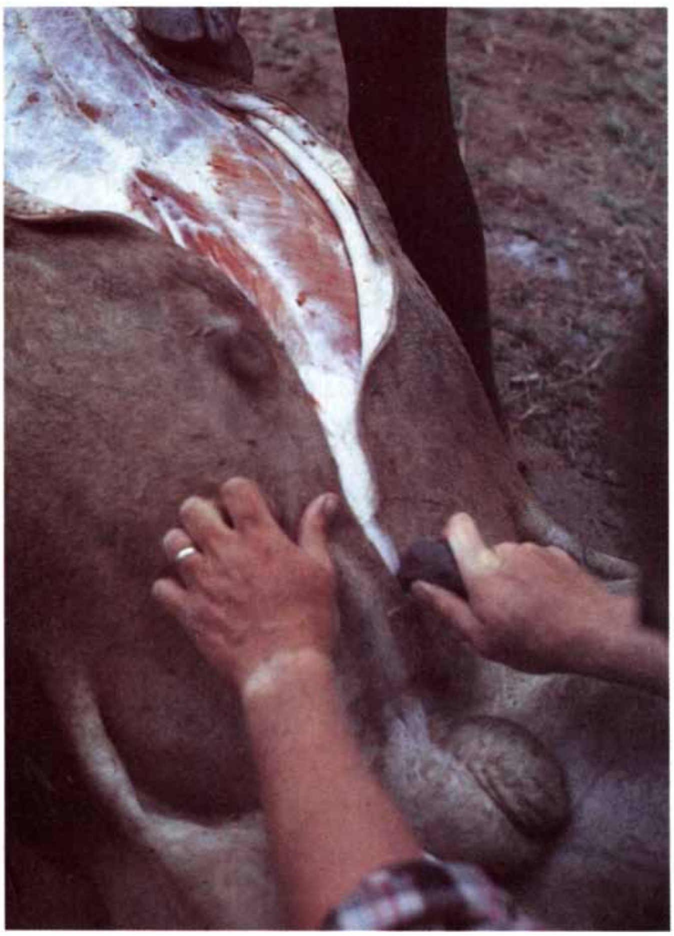
Now, these principles can be applied in at least two different types of flaking: alternate bifacial flaking and unifacial flaking. In alternate bifacial flaking, flakes are struck from opposite sides of the core in alternation to produce two flaked faces. In unifacial flaking a series of flakes are removed from a single face of the core. The flakes resulting from alternate bifacial flaking cannot reveal whether the knapper was right- or left-handed. In unifacial flaking, however, the pattern of flakes does depend on the handedness of the maker: "right-oriented" flakes, consistent with a right-handed knapper, can be distinguished from "left-oriented" ones by the pattern of cortex, or weathered rind, on their surface [*see illustration on page 120*].

That difference has provided the basis for some intriguing findings about the original development of handedness among human beings. I analyzed the proportion of right- to left-oriented flakes resulting from my own flaking and found that it was 56 : 44. Since I am right-handed, it was not surprising that the right-oriented flakes predominated. (The ratio is not higher because much flaking is bifacial, which yields comparable numbers of right- and left-oriented flakes.) When I measured the ratio among flakes gathered at Koobi Fora, however, I found almost exactly the same ratio—57 : 43. Other early Stone Age artifact assemblages also show this ratio. This surprising finding suggests that the preference for the right hand had already developed among human beings by between 1.9 and 1.4 million years ago.

Handedness is by no means a trivial matter in an evolutionary sense. The widespread dominance of one hand seems to be unique to human beings. Animals do show individual preferences for one paw or hand over the other, but these balance out, and the overall ratio among any population is likely to be about 50 : 50. Among human beings, however, right-handers predominate by about 90 : 10, and the preference appears to be under genetic control. What is more, handedness is associated with a profound lateralization of the human brain. In most modern human beings the left cerebral hemisphere (which controls the dominant right hand) is the center of time-sequencing skills, including language ability; the right hemisphere is more specialized for spatial patterning. The appearance of a preference for the right hand among the Koobi Fora hominids may well indicate the evolution of lateralization.

Lateralization of the brain may in turn have brought with it increased capacities for conceptualizing the future, a conclusion supported by my work on tool transport. Transport was a subject I was particularly interested in. I had touched on it in analyzing the transport distances from sources of raw material, but I wanted to know more. In analyzing this problem the cortex of the lava cobbles was again most helpful. Clearly, when a stone is flaked, the pieces with cortex come off first. I examined the distribution of cortical and noncortical flakes that resulted when I flaked various kinds of core myself. If the hominids had carried out the entire sequence of manufacturing steps at one site, predictable patterns of cortical and noncortical flakes would be represented there.

If, on the other hand, cores had been transported from place to place and worked on a bit at a time, the distribution would be skewed: some sites would have fewer of the cortical flakes that are made in the early phases of flaking. That is precisely what I found. Compared with my computer-generated models of the entire population of flakes resulting from tool manufacture, the Koobi Fora sites tended to include anomalously high proportions of noncortical flakes (the kind produced in the later stages of flaking). It would appear that early hominids carried partially flaked cores with them, perhaps in simple containers. At stopping points the cores may have been worked again, and at places of prolonged or frequent occupation much of the flaked material dropped out to form the artifact concentrations that archaeologists call sites.



EXPERIMENTAL MEAT PROCESSING revealed that different stone implements were best for various tasks. For slitting the hide of a wildebeest (*left*) an unretouched flake worked well. For removing the leg of a bovid (a gazellelike animal) a bifacial hand ax

This spatial and chronological pattern suggests that the hominids living around Koobi Fora had considerably greater powers of organization and planning than modern chimpanzees have. Although some chimpanzees do make tools for future use, the time between manufacture and employment is usually less than a few minutes. Moreover, only rarely do chimpanzees transport objects farther than 100 meters. In contrast, among the early hominids raw materials may have been carried more or less constantly in preparation for potential uses in the future. The development of such mental skills may have been part of the groundwork for tool use. Increased tool use in turn may have contributed to new selective forces that favored further brain development, as Sherwood L. Washburn of the University of California at Berkeley has suggested.

In all probability one aspect of this feedback cycle—a cycle of major adaptive importance—was an application of implements to an increasing range of tasks. Based on current knowledge of contemporary hunter-gatherers as well as nonhuman primates, I assembled a list of functions that tools could have performed among early African hominids and then experimented to find out what types of tools were best for carrying out each function. The list included such jobs as butchering animals, woodworking, hideworking, cracking hard-shelled nuts and digging. On the basis of hundreds of feasibility experiments, I have drawn the following generalizations.

All my butchering experiments were carried out on animals that had died of natural causes or been slaughtered by traditional methods for food. The subjects included domestic animals such as goats, sheep, pigs, cows and horses along with wild species such as deer, zebras, oryxes, wildebeests and even elephants. The most effective tool depended on the specific task within the overall butchering. A sharp, unmodified flake was most efficient for opening the carcass. For dismembering and defleshing an unmodified flake or a retouched flake with an acute edge was best. For chopping dried, residual meat from a scavenged carcass, which Lewis Binford of the University of New Mexico has proposed was an important hominid behavior, a large flake or a sharp-edged core proved most suitable. For cracking bones to extract marrow or breaking open the skull to get the brains, an unmodified cobble or a heavy core did the best job.

The making of digging sticks and spears also entailed different implements: heavy, acute-edged cores for severing a straight branch from a tree, flakes and flake scrapers for the finer shaping of the point and a rough stone surface for the final grinding. A variety of simple and retouched flakes were the most efficient tools for scraping fat and meat off a hide or removing hair from the outer surface. Unmodified cobble hammers and anvils worked best for cracking hard-shelled nuts. For digging to reach buried water or underground vegetation such as roots and tubers, however, stone artifacts were usually not as efficient as other materials, which included unmodified antelope horns, broken limb bones from large animals and sharpened digging sticks.

The technological repertory of the

was best (*second from left*). For cracking bovid limb bones to obtain the marrow an unmodified cobble or, as here, a heavy core was most effective (*second from right*). A similar core was well suited for breaking open a bovid skull to get at the brains (*right*).

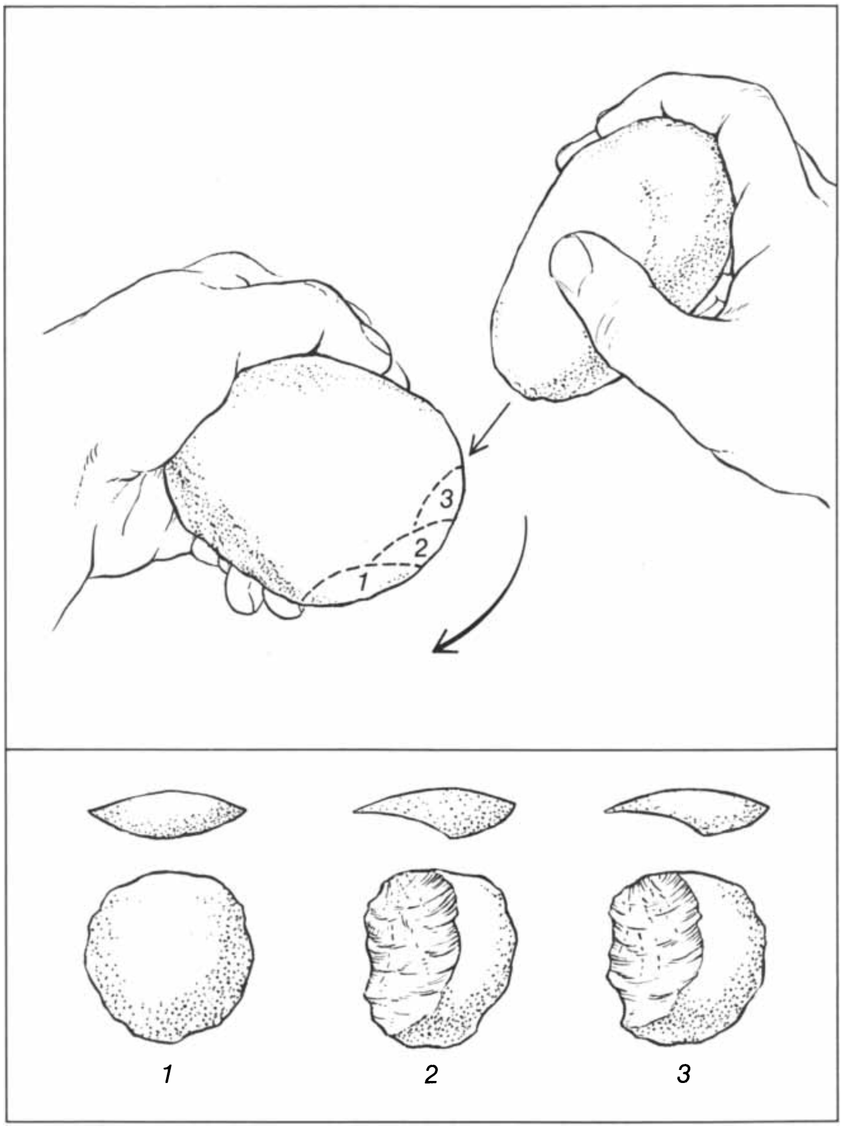
early hominid groups may well have included weapons, among them wood spears, clubs, throwing sticks and cobble missiles. In this study I made no attempt to kill animals with primitive weapons. I did, however, carry out some experiments that were aimed at clarifying the problem of killing. Wood and horn-tipped spears thrust into the chest cavities of dead cows and goats were found to penetrate as much as 30 centimeters unless they hit a rib. Such spears penetrated the chest cavity of an adult African elephant, in contrast, only three or four centimeters, but it must be kept in mind that early hominids were considerably stronger than modern human beings of average height and build, and these simple weapons may have been far more effective in their hands.

Some of the humblest implements I experimented with were various kinds of containers. Simple devices for carrying water, food or partially worked cores could have been made out of the shells of ostrich eggs or tortoises, pieces of bark or animal skins. A branch employed as a skewer and slung over the shoulder could have provided a valuable tool for carrying meat. Although such containers are very modest implements, they could have been quite important in the course of human evolution. Indeed, the first tools of all may have been carrying implements that enabled a group of hominids to transport food, thereby reducing their dependence on the immediate environment.

Functional studies are valuable for understanding how early tools might have been used, but they cannot prove how the tools were actually employed. Two recent advances make it possible to bridge that gap. The first is microscopic study of the characteristic polishes formed on the edges of stone tools in the process of working various materials. Lawrence H. Keeley of the University of Illinois at Chicago Circle is a pioneer in this approach [see "The Functions of Paleolithic Flint Tools," by Lawrence H. Keeley; SCIENTIFIC AMERICAN, November, 1977], and I collaborated with him in examining the microwear polishes on a set of implements from Koobi Fora.

Unfortunately microwear analysis is not suitable for examining the medium-grained lavas of which the majority of Koobi Fora tools were made; that stone is too heterogeneous and too coarse to yield good data under the microscope, and chemical weathering can alter the surface of a lava artifact over time. The minority of finer-grained stones (particularly cherts) yielded helpful data, however. Of the 59 fresh, unweathered artifacts I submitted to Keeley, wear polishes were found on nine. Four had polishes consistent with animal butchering, three with woodworking and two with cutting softer vegetation. Although this is a small sample, it suggests that by 1.5 million years ago meat processing may have been an important technological activity, complemented by the manufacture of simple wood tools and the cutting of grasses (perhaps for laying meat on to protect it from dirt while a carcass was being butchered, or, as Keeley has suggested, for bedding material).

The second advance in methodology is based on the converse modification: of the bone by the tool. Striations on fossil bone from Koobi Fora and Olduvai Gorge have been shown to be cut marks left by hominids employing stone tools for skinning, dismembering and defleshing the carcass. Such work, done by Henry T. Bunn of the University of Wisconsin at Madison, Richard Potts of the Smithsonian Institution and Patricia Shipman of Johns Hop-



UNIFACIAL FLAKING yields flakes that can indicate which hand was favored by the toolmaker. A right-handed individual holds the hammerstone in the dominant right hand and the core in the left hand; the core tends to be rotated clockwise as a sequence of flakes is struck off one face. Left-handers reverse the arrangement and generally turn the core counterclockwise. The pattern of cortex, or weathered rind, on the flakes can show which hand held the hammerstone. The flakes numbered *2* and *3* in the lower panel are "right-oriented," suggesting a right-handed toolmaker. (On left-oriented flakes, made by left-handers, the crescent of cortex has the opposite orientation.) Analysis of large populations of prehistoric flakes enabled the author to determine that early hominids appear to have become preferentially right-handed by between 1.9 and 1.4 million years ago.

kins University, corroborates the finding that meat processing was a significant early activity. In addition some marrow-yielding long bones from sites in both areas show fracture patterns that could have been made by stone hammers. How such animal resources were obtained is controversial—the relative contributions of hunting and scavenging are debated—but it seems clear that technology based on the flaking of stone was integral to processing the carcasses.

Such evidence, combined with my own extensive experimental studies, has enabled me to begin to formulate a picture of the emergence of flaked-stone technology. I think the major impetus for the development of flaked-stone tools was a shift toward exploitation of animal-food resources (by scavenging or small-scale hunting) in an opportunistic foraging strategy that was still dominated by plant foods. This dietary shift probably took place more than two million years ago, and it may have been associated with major climatic changes. About 2.5 million years ago the African climate appears to have become significantly drier. Elisabeth S. Vrba of Yale University has suggested that the climatic shift brought about large-scale animal extinctions along with a burst of speciation marked by the appearance of new animal species.

A technology of flaked stone may have offered the early hominids a means of adapting to the new conditions by moving opportunistically into niches that had previously been the domain of carnivores or scavengers. Hominids are not well equipped biologically for such niches: our small canine teeth, flat cheek teeth and lack of claws make us inefficient carnivores. Furthermore, as Keeley has pointed out, sharp-edged rocks or other materials can rarely be found in nature. Yet the early hominids, like modern chimpanzees, may already have possessed a technology based on unmodified stone hammers for plant processing. Accidental fracture of the stone hammers or anvils could have given rise to sharp fragments whose potential for cutting was discovered by trial and error.

In this way our ancestors could supplement their biological equipment: sharp-edged stones became the equivalent of canines and carnassials (meat-cutting teeth) and heavier rocks served as bone-crunching "jaws." That flakes were an integral part of this process is suggested by a finding that emerged from my work with Keeley. All nine of the artifacts from Koobi Fora that showed traces of wear were simple flakes. Not one of them would have been considered a "tool" in most traditional classifications of Oldowan forms. Furthermore, many of the fine, parallel striations on bones from Oldowan sites seem to be cut marks made by unmodified flakes. Once the notion of flaking stone had been assimilated, a crucial feedback loop between tool use and brain expansion may have begun.

LARGE FLAKE detached from a boulder is one hallmark of Acheulean technology. About 1.5 million years ago hominids began to quarry such large flakes and make them into hand axes, picks and cleavers of the type shown in the bottom illustration on page 116.

Then, about 1.5 million years ago, there came about a second technological leap: the development of the Acheulean-style implements. My work has shown that many of the Acheulean "hand axes," "picks" and "cleavers" are in fact excellent tools for the heavy-duty butchering of larger mammalian species. It would seem likely, then, that this second advance was associated with more systematic butchering of large carcasses, acquired either through scavenging or through hunting, as investigators such as Clark have suggested in the past.

Both the work of other investigators and my own research suggest that changes in diet and biology were strongly associated with the development of early technology. Further work will undoubtedly add to our knowledge and modify our interpretations. What will remain is a certain inherent fascination with the simple stone tools found at the early African sites. They represent the first evidence of culture (the transmission of learned behavior) in the prehistoric record. Culture and technology, in their complex contemporary manifestations, are often taken for granted, their origins forgotten. Yet every modern technology—including those for exploring other parts of the solar system, splitting the atom, engineering genetic material and performing organ transplants—is linked in an unbroken line of culture with the inception of flaked-stone tools between two and three million years ago.

THE AMATEUR SCIENTIST

Making a barometer that works with water in place of mercury

by Jearl Walker

What might be described as the first barometer figured in a curious experiment done in 1641 by the Italian scientist Gasparo Berti. He closed the lower end of a long tube with a stopcock and immersed it vertically in a tub of water. Then he filled the tube from the top with water, sealed the top and opened the stopcock. Only a small amount of water flowed out of the tube.

Berti contended that the slight fall of the water column left a vacuum in the tube above the column. The possibility that a vacuum can exist in nature had already been proposed by other investigators, but the idea had yet to be accepted universally. Berti also maintained that water remained in the tube because the atmospheric pressure on the exposed water surface in the tub balanced the pressure from the weight of the water in the tube.

Berti's tube extended up the side of a multistory building and therefore was too big to serve as a useful barometer. A few years later his compatriot Evangelista Torricelli proposed that the water be replaced with mercury, which is 13.6 times denser, so that the instrument's height could be reduced. Torricelli's associate Vincenzo Viviani constructed such an apparatus. He sealed one end of a glass tube about 30 inches long and filled it with mercury. Pressing a finger over the open end, he inverted the tube and lowered its end into an open container of mercury. When he released his finger, mercury flowed from the tube until the pressure from the weight of the mercury left in the tube matched the atmospheric pressure. This instrument was the first true barometer. By noting the height of the mercury column one could measure the atmospheric pressure. High pressure pushed mercury into the tube, increasing the column's height. Low pressure allowed mercury to leave the tube, decreasing the column's height.

Because mercury is so dense, the variations in column height are usually small and sometimes must be recorded in measurements of less than a millimeter. Why not stick with water as the working fluid? Although the barometer would be quite tall, the variations in atmospheric pressure should yield changes in column height 13.6 times greater than those in a mercury barometer. Unfortunately the water barometer falls quite a bit short of the barometric ideal.

The explanation for this failure lies in the fact that the space above the water column is not a full vacuum as Berti conjectured. Some of the water vaporizes, filling the space with vapor. Whenever the temperature changes, the amount of vapor varies. Hence the pressure the vapor exerts on the water column also varies. For example, in warm weather the increased vapor pressure reduces the height of the column. The device is therefore not accurate as a barometer.

Sam Epstein of Los Angeles has found a way around this difficulty. His barometer contains common antifreeze (ethylene glycol) diluted with water. The antifreeze reduces vaporization at the top of the barometric column. The barometer's residual dependence on temperature is accounted for by a table that Epstein formulates when he calibrates the instrument.

Epstein has made two such barometers. One of them is described as a barometer of the cistern type and the other is a barometer of the siphon type. The cistern barometer is similar in design to Berti's apparatus. To accommodate the low density of the antifreeze mixture (1.07 times the density of water), the device stands about 35 feet high. Since monitoring the fluid level can be a nuisance if one must climb several flights of stairs, Epstein designed a remote sensor to read the column height from the base.

If you want to build a similar apparatus, your first move should be to find a suitable place for it to stand. It should be fastened to a wall by brackets. It should not be in direct sunlight because of the heating effect. Although the freezing point of the antifreeze mixture is low (about −34 degrees Fahrenheit), you must be careful not to mount the device where the mixture might freeze in winter.

The barometer is made out of six sections of polyvinyl chloride (PVC) tubing, each section six feet long and one inch in internal diameter. If you decide to do without a remote sensor, make the top section of transparent plastic. Join the sections with union pieces sealed with primer and cement. Be certain the seals are tight; after the cement sets, the sections can be separated only by sawing them apart. For the working fluid mix four gallons of good-quality antifreeze with an equal amount of distilled water.

The sensing elements are 24 platinum electrodes spaced along the top section. Each electrode is numbered according to its height. An additional electrode, mounted on the opposite side of the section, serves as a common electrode. When a voltage is applied between the common electrode and a numbered one, a small current is conducted through the antifreeze solution. The numbered electrodes are connected to a rotary switch at the base of the barometer. When you turn the switch, you put a different electrode in the circuit. By rotating the switch to find the highest electrode that conducts electricity you can ascertain the approximate height of the fluid.

In order to make the electrodes, buy from a laboratory-supply house a nine-inch length of platinum wire (No. 24 gauge). Also get 25 glass tubes, each tube six inches long and five millimeters in internal diameter. Cut the wire into nine-millimeter sections and bend each section into the shape of a Z. Its two outer arms should each be four millimeters long. Heat the end of a glass tube in a gas-and-compressed-air flame from a Meker burner. Hold the end just above the tip of the blue part of the flame, rotating it slowly to provide uniform heating. The end soon softens and closes. When the opening is only slightly wider than the arms of the wire, insert a wire section into the tube and continue heating the glass so that it seals around the middle section of the Z.

When the tube closes around the wire, gradually diminish the air supply to the burner while rotating the tube in the flame. After another minute remove the tube from the flame and let it cool. Be careful not to overheat the wire or it will melt, ruining the device.

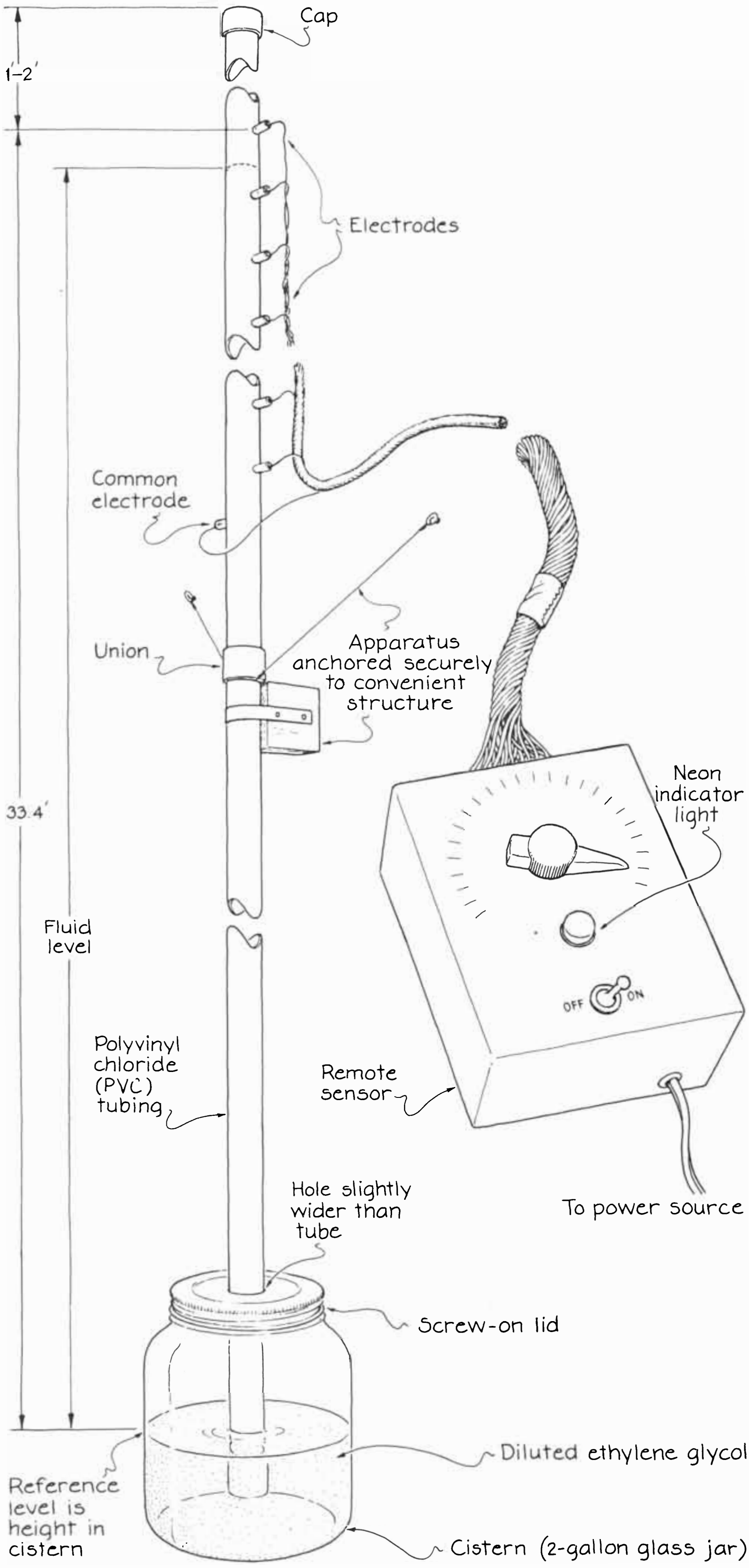
After the tube has cooled cut the end containing the wire to a length of one inch. Then fire-polish the ragged end left by the cut.

Each electrode is connected to the rest of the sensor circuit by hookup wire, which has strands and is insulated. Have a six-inch length of this hookup wire handy with one end stripped of its insulation. Put approximately half an inch of 50/50 rosin-core solder in the electrode. Hold the section in a gas flame until the solder melts. Then quickly remove the section from the heat and push the stripped end of the hookup wire into the molten solder. After the device has cooled make sure the wire is firmly embedded; if it is not, repeat the procedure. When the wire is properly set, seal the platinum end of the tube with PVC cement.

Prepare the electrodes and drill holes for them in the top section of PVC tubing. Make each hole just large enough to accept an electrode. Space the centers of the holes for the numbered electrodes at 2.4-inch intervals, starting about a foot below the top of the tubing. Then seal each electrode with PVC cement. When the barometer is later assembled, put the highest electrode 33.4 feet from the level of the fluid in the cistern.

From the stub of hookup wire at each electrode run enough additional wire to reach the base of the barometer. Each wire is attached to a resistor connected to one of 24 positions on a rotary switch. Number the electrodes, beginning with the lowest one. The resistors connected to the lowest six electrodes are rated at 90 kilohms. The higher electrodes, which are farther from the common electrode, require resistors with less resistance. Group the 24 electrodes into four sets of six. Decrease the resistance by 20 kilohms for each successively higher set. Each of the resistors has a power rating of .5 watt.

The rotary contact of the switch is connected to a double-pole, single-throw toggle switch. A wire runs from the toggle switch to the common electrode. Along the route install a small neon lamp with no internal resistor; it signals when current is conducted through the antifreeze solution. The toggle switch is connected to an isolation transformer that is plugged into a wall outlet. When you want to take a measurement, turn on the toggle switch. (It should be off at other times.) Select a position for the rotary switch. If the neon lamp lights, turn the switch counterclockwise until the lamp goes out. Otherwise turn the switch clockwise until the lamp goes on. The last position of the switch indicates the height of the fluid in the



Sam Epstein's water barometer

barometer to an accuracy of approximately two inches.

The rest of the construction does not depend on the remote sensor. Mount the PVC tubing on a wall and plug its lower end with a rubber stopper taped on or with a threaded cap. The cistern is a widemouthed two-gallon jar with a screw-on lid. Drill in the lid a hole slightly larger than the tubing. Half fill the jar with the antifreeze mixture. Slip the lid over the lower end of the tubing and have someone hold it well above that end.

Raise the support of the jar until the lower end of the tubing is in the antifreeze mixture. Put a funnel into the top of the tubing and slowly pour in the antifreeze mixture, trying to avoid air bubbles. After the tube is full allow 30 minutes for any air bubbles to dissipate. Then cap and seal the top of the pipe. Unplug the lower end and screw the lid of the jar into place. Allow one day for the vapors of ethylene glycol and water to saturate the region above the fluid column.

If you do not build the remote sensor, add a scale alongside the upper end of the fluid column by which you can measure the atmospheric pressure. The height of the fluid column is measured from the fluid level in the cistern. The scale should extend in height from 28.8 to 33.4 feet, the same range covered by the placement of electrodes for the remote sensor. If the cistern is wide enough, the fluid level in it hardly changes as atmospheric pressure varies, and so you can attach the scale to the barometer.

The scale can be marked in pressure units called inches of mercury. The full length of the scale, 4.6 feet, represents a pressure change of 4.6 inches of mercury. Mark the scale in .1-foot sections. Each section then represents a pressure change of .1 inch of mercury. For example, if the atmospheric pressure increases so that the column of antifreeze mixture climbs by .1 foot, the column in a mercury barometer would climb by .1 inch. (You could mark the scale in any other units of pressure.)

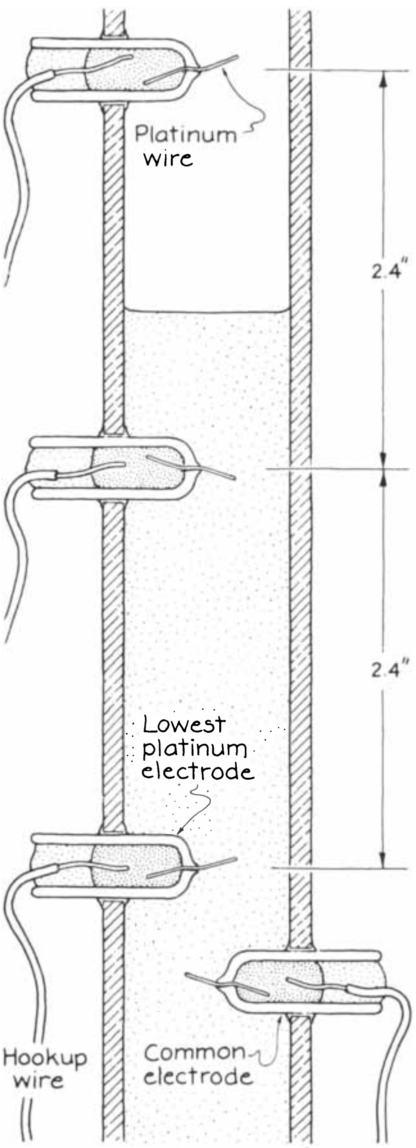
If the barometer is exposed to a considerable range of variations in temperature, you will need to record the height of the fluid throughout the entire range of temperatures. Compare the reading with the atmospheric pressure measured by means of some other device, such as a mercury barometer, that is not appreciably influenced by temperature. At very low temperatures the vapor pressure above the fluid hardly alters the height of the column. At higher temperatures the vapor pressure lowers the height.

After you have constructed a table of corrections for the temperature, you no longer have need for the second barometer. Each time you make a reading on the antifreeze barometer note the temperature on a thermometer placed next to the base. Use the table to determine the correction that must be added to your reading to get the true atmospheric pressure.

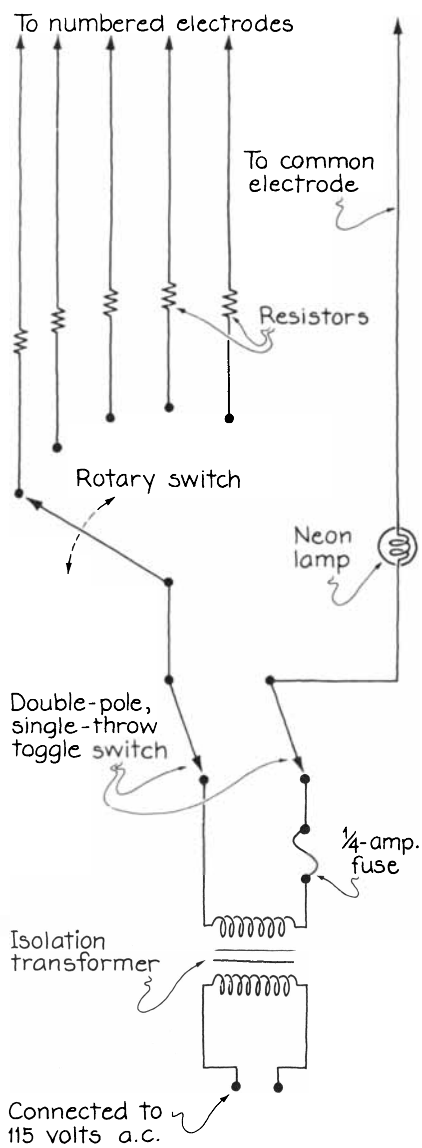
Epstein's second device, the siphon barometer, is read at the base without a remote sensor and does not require a cistern. It is in the shape of a U but with one side shorter than the other. The short side is made of transparent plastic pipe that is six feet long with an internal diameter of one inch. The long side consists of seven sections of PVC tubing. The sides are connected by PVC elbows and a short length of PVC tubing. When you have cemented the tubing, pipe and elbows together, mount the device on a wall, insert a funnel in the short side and pour in enough antifreeze mixture to fill both sides to a height of five feet.

Now a calculation must be made to ascertain how much additional fluid will be needed to balance the normal atmospheric pressure at your altitude. At sea level the normal atmospheric pressure will balance the weight of 29.92 inches of mercury. To find the equivalent height of the antifreeze mixture, multiply the height of the mercury by the ratio of the density of mercury to the density of the antifreeze. Dividing the result by 12 to convert it into units of feet, you will find that the equivalent height of the antifreeze mixture is 31.7 feet.

Less fluid is needed above sea level, more below it. Determine your elevation with respect to sea level. For every



The placement of electrodes



The circuitry of the remote sensor

100 feet above sea level subtract 1.3 inches from the benchmark height of 31.7 feet. For every 100 feet below sea level add 1.3 inches to the benchmark height. When you have determined the appropriate column height for your elevation, calculate the volume of fluid you will need to fill the PVC tubing to that height.

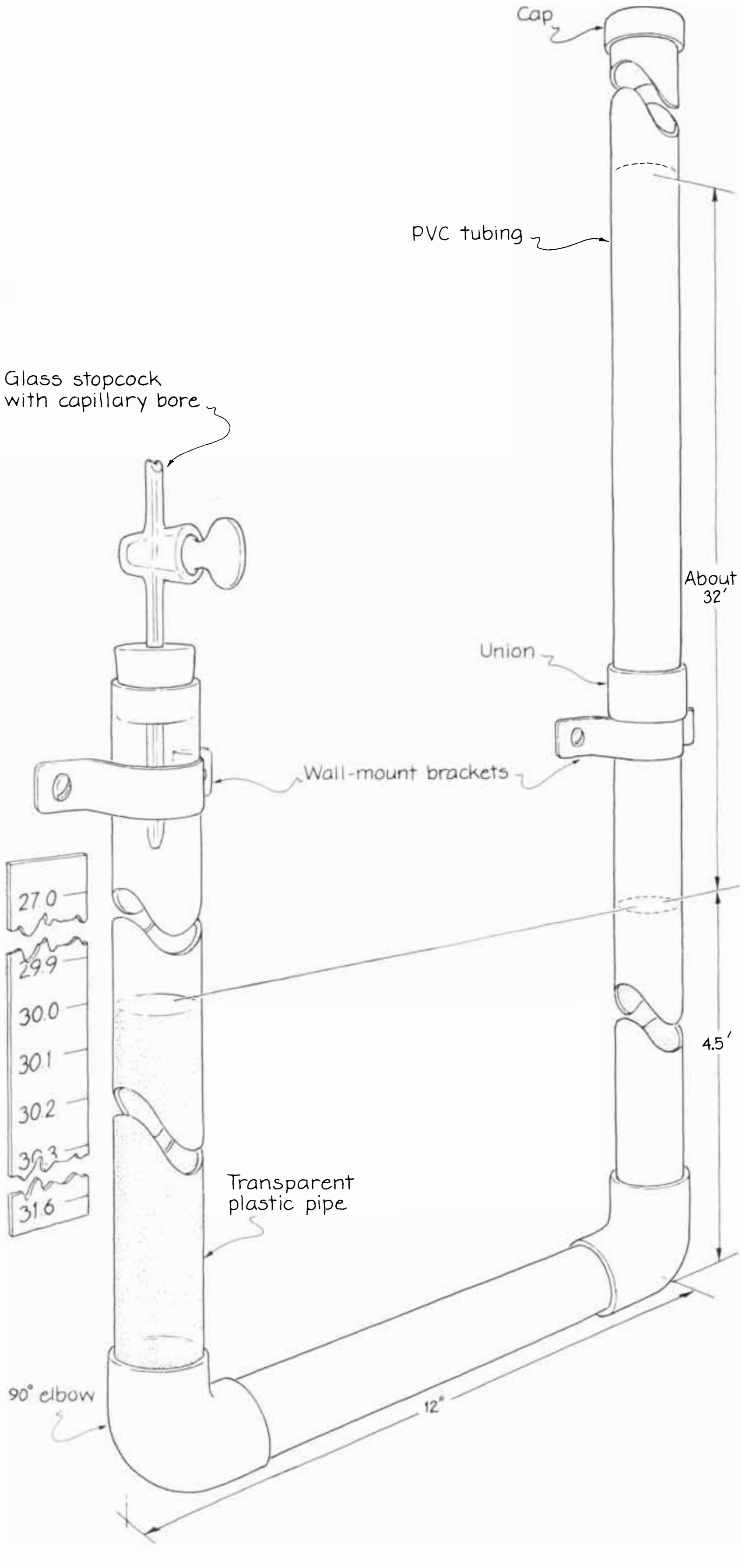
To add fluid to the barometer begin at the short side. Bring the fluid level up to the top of the pipe and close that side with a rubber stopper. Insert a glass stopcock through a hole in the stopper; the bore of the stopcock should be two or three millimeters in diameter. Close the stopcock, push the stopper into place, seal it with electrical tape and add the rest of the fluid to the long side of the barometer.

Allow 30 minutes for air bubbles to escape from the fluid and then attach a bicycle air pump to the protruding end of the stopcock. As you open the stopcock, pump air into the barometer, forcing the fluid level in the long side to the top. Immediately close the stopcock. Cement a cap to the top of the long side. When the cement has set, slowly open the stopcock. Wait one day for the vapors to saturate the space above the fluid in the long side. Replace the stopcock and the rubber stopper with another stopper through which you have inserted a two-inch glass tube with an internal diameter of one millimeter. Fasten the stopper with tape and cover it with a cap that is loose enough to pass air.

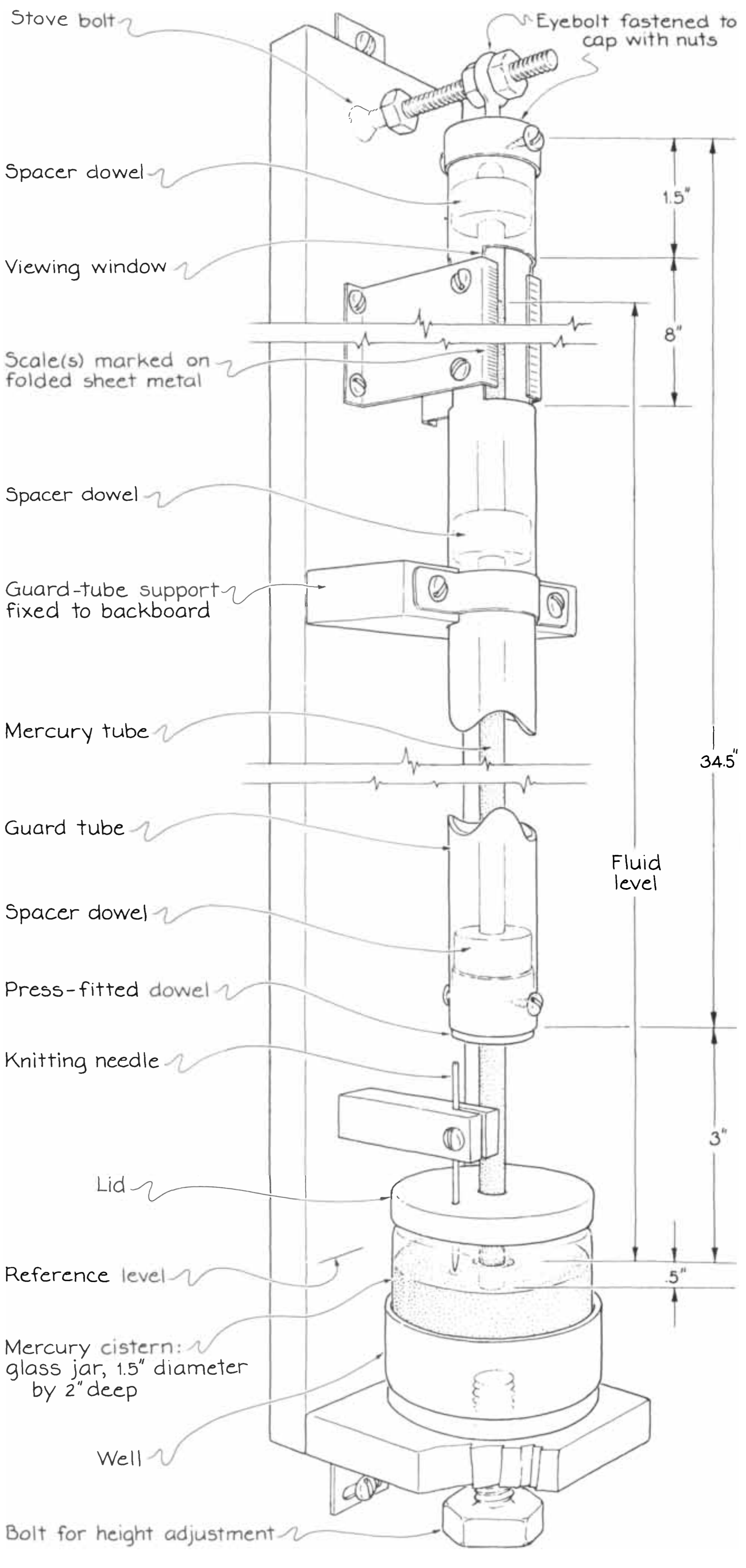
The scale for measuring the atmospheric pressure is mounted along the short side. It should run in the opposite direction from the scale on the cistern barometer. The reason is that an increase in atmospheric pressure drives the fluid down in the short side. Since the pressure variations are split evenly between the two sides of the barometer, the distance between marks on the scale is half the distance on the scale for the cistern barometer. Hence a change in height of .05 foot now represents a pressure change of .1 inch of mercury.

To calibrate his barometers Epstein built a mercury barometer. Since he is a chemist, he is well acquainted with the precautions that should be taken in work with mercury. If you decide to build a mercury barometer, you should get advice from someone similarly knowledgeable about mercury. Do the work in a well-ventilated area. Do not touch the mercury or get any of it in your mouth.

Epstein's barometer holds about two pounds of mercury, which he bought from a dental-supply house. The backboard support for the barometer is half-inch wood mounted on a



Epstein's siphon barometer



Epstein's mercury barometer for calibration

wall with two picture hooks, each one strong enough to support 50 pounds. The barometer must be exactly vertical. The cistern, a jar with a metal screw-on lid, sits in a well that is supported by a wood extension from the backboard. Drill the extension so that a round steel stock one inch in diameter can be press-fitted through it. Drill and thread the stock to accept a 3/8-by-16-inch steel bolt, which should be greased before it is put in place. The end of the bolt fits securely in a threaded hole at the bottom of the well.

Two holes are drilled in the lid of the cistern. One hole, exactly centered, is slightly larger than the mercury tube that fits through it. The other hole is slightly larger than a plastic knitting needle that extends into the cistern. A rubber liner cemented inside the lid has similar holes but fits tightly around the mercury tube and the needle. The holes are punched with a laboratory cork borer. Add a barely visible V notch to the hole through the liner for the mercury tube. The notch passes air so that the air pressure inside the cistern matches the external air pressure. The notch must be small to keep the escape of mercury vapor to a negligible amount.

A support for the needle extends from the backboard. The tip of the needle indicates the height of a reference level scratched on the backboard. Whenever the atmospheric pressure is read, the height of the cistern is adjusted until the level of the mercury in the cistern is at the tip of the needle.

The tube holding mercury rests inside a guard tube made from thin-walled electrical conduit with an internal diameter of three-quarters of an inch. Make the guard tube 34.5 inches long. It has two viewing slots near the top through which the mercury level can be monitored. A scale marked in inches or millimeters is positioned along the front side of the guard tube. The wall behind the slots should be white to provide contrast with the mercury. A cap for the guard tube has a No. 10 eyebolt mounted in its center and is fastened to the tube by three sheet-metal screws.

Two supports for the barometer extend from the back wall. One of them is made of wood; its outer end is cut to fit around the lower end of the guard tube. A clamp made out of sheet metal holds the tube snug against this wood support. The other support is a 3/8-inch bolt that goes through the backboard, extends over the barometer and passes through the eyebolt. The eyebolt is held in position by nuts on the support bolt.

The mercury tube, made of flint glass, is 35 inches long and five millimeters in internal diameter. For safety Epstein decided that the wall thickness should be at least one millimeter. He sealed one end of the tube with a Meker burner. Three half-inch wood dowels were slipped over the tube to serve as spacers in the guard tube. They have a centered hole slightly larger than the mercury tube. Two of them are glued in place with epoxy at the middle and bottom of the tube. The third dowel is glued 4.5 inches below the open end of the tube.

In filling the mercury tube, Epstein held it by clamps so that its sealed end was at the bottom. To catch any mercury that might spill he placed a wide glass container holding about half an inch of water under the tube. He also took the precaution of filling the tube in a carpetless room; if mercury spilled, he could amalgamate it with dust or filings of copper or zinc moistened with vinegar.

Epstein filled the tube by slowly pouring the mercury into a funnel held in place by a piece of rubber tubing in the open end of the tube. To facilitate the flow of mercury he inserted a four-foot length of piano wire into the tube through the funnel. Tapping the tube with a pencil, he worked the wire up and down, being careful not to scratch the glass. He continued the process until the tube was filled to within an inch of the top and the air bubbles in the mercury were eliminated.

With the help of an assistant he inverted the guard tube, placed its lower end on a soft surface, slipped the tube of mercury into it and pressed a wood dowel into the upper end of the guard tube. After sliding the inverted cistern lid over the protruding open end of the mercury tube, Epstein completed filling the tube with mercury. Wearing a surgical glove, he closed the open end of the mercury tube with a finger as he and his assistant inverted the assembly and lowered the tube into the mercury in the cistern. He removed the finger when the mercury tube was below the fluid level in the cistern. Being careful not to let the lower end of the mercury tube move above the fluid level in the cistern or touch the bottom of the cistern, he clamped the guard tube to the support extending from the back wall. Then he placed the cap on the guard tube and adjusted the height of the tubes until he could run the support bolt through the eyebolt.

After screwing the cistern lid to the cistern, Epstein positioned the knitting needle. To make a reading he rotated the bolt below the cistern until the fluid level in the cistern was at the tip of the needle. He then read the height of the mercury column to an accuracy of about one millimeter.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

POLYPHONIC COMPOSITION. Owen Swindale. Oxford University Press, 1962.

STUDIES FOR THE CONTEMPORARY DRUMMER. John Xepoleas and Warren Nunes. Hansen House, 1981.

COMPUTERS AND MUSIC. Special issue of *Byte,* Vol. 11, No. 5; June, 1986.

THIRD-GENERATION NUCLEAR WEAPONS

THE EFFECTS OF NUCLEAR WEAPONS. Edited by Samuel Glasstone and Philip J. Dolan. Government Printing Office, 1983.

DIRECTED ENERGY MISSILE DEFENSE IN SPACE. Ashton B. Carter. Office of Technology Assessment, U.S. Government Printing Office, 1984.

NUCLEAR WEAPONS DATABOOK, VOL. 1: U.S. NUCLEAR FORCES AND CAPABILITIES. Thomas B. Cochran, William M. Arkin and Milton M. Hoenig. Ballinger Publishing Co., 1984.

ENDLESS GENERATIONS OF NUCLEAR WEAPONS. Theodore B. Taylor in *Bulletin of the Atomic Scientists,* Vol. 42, No. 9, pages 12–15; November, 1986.

NUCLEAR TESTING IS A PANDORA'S BOX. Theodore B. Taylor in *FAS Public Interest Report,* Vol. 39, No. 10, pages 1–11; December, 1986.

HOW PHOTORECEPTOR CELLS RESPOND TO LIGHT

ENERGY, QUANTA, AND VISION. Selig Hecht, Simon Shlaer and Maurice Henri Pirenne in *The Journal of General Physiology,* Vol. 25, No. 6, pages 819–840; July 20, 1942.

RESPONSES OF RETINAL RODS TO SINGLE PHOTONS. D. A. Baylor, T. D. Lamb and K.-W. Yau in *The Journal of Physiology,* Vol. 288, pages 613–634; March, 1979.

SPECTRAL SENSITIVITY OF SINGLE CONES IN THE RETINA OF *MACACA FASCICULARIS.* B. J. Nunn, J. L. Schnapf and D. A. Baylor in *Nature,* Vol. 309, No. 5965, pages 264–266; May 17, 1984.

THE PHOTOCURRENT, NOISE AND SPECTRAL SENSITIVITY OF RODS OF THE MONKEY *MACACA FASCICULARIS.* D. A. Baylor, B. J. Nunn and J. L. Schnapf in *The Journal of Physiology,* Vol. 357, pages 575–607; December, 1984.

THE MOONS OF URANUS

THE MOONS OF URANUS, NEPTUNE AND PLUTO. Robert Hamilton Brown and Dale P. Cruikshank in *Scientific American,* Vol. 253, No. 1, pages 38–47; July, 1985.

SATELLITES. Joseph A. Burns and Mildred S. Matthews. University of Arizona Press, 1986.

VOYAGER 2 IN THE URANIAN SYSTEM: IMAGING SCIENCE RESULTS. B. A. Smith et al. in *Science,* Vol. 233, No. 4759, pages 43–64; July 4, 1986.

ANTIVIRAL THERAPY

ANTIVIRAL AGENTS AND VIRAL DISEASES OF MAN. Edited by George J. Galasso, Thomas C. Merigan and Robert A. Buchanan. Raven Press, 1984.

ANTIVIRAL CHEMOTHERAPY AND CHEMOPROPHYLAXIS. Raphael Dolin in *Science,* Vol. 227, No. 4692, pages 1296–1303; March 15, 1985.

PROSPECTS OF THERAPY FOR INFECTIONS WITH HUMAN T-LYMPHOTROPIC VIRUS TYPE III. M. S. Hirsch and J. C. Kaplan in *Annals of Internal Medicine,* Vol. 103, No. 5, pages 750–755; November, 1985.

PHOTOVOLTAIC POWER

PHOTOVOLTAICS: SUNLIGHT TO ELECTRICITY IN ONE STEP. Paul D. Maycock and Edward N. Stirewalt. Brick House Publishing Co., 1981.

AMORPHOUS SEMICONDUCTOR TECHNOLOGIES AND DEVICES (VOLS. 2, 6 AND 16 OF JAPAN ANNUAL REVIEWS IN ELECTRONICS, COMPUTERS AND TELECOMMUNICATIONS). Edited by Y. Hamakawa. Ohmsha, Ltd., Tokyo, 1982–84.

BASIC PHOTOVOLTAIC PRINCIPLES AND METHOD. Kenneth Zweibel and Paul Hersch, Solar Energy Research Institute. Van Nostrand Reinhold Company, 1984.

THE EMERGENCE OF ANIMALS

THE CAMBRIAN RADIATION EVENT. M. F. Brasier in *The Origin of Major Invertebrate Groups,* edited by M. R. House. Academic Press, 1979.

A CASE FOR TWO LATE PROTEROZOIC–EARLIEST CAMBRIAN FAUNAL PROVINCE LOCI. Mark A. S. McMenamin in *Geology,* Vol. 10, No. 6, pages 290–292; June, 1982.

A COMPARATIVE STUDY OF LOWER CAMBRIAN *HALKIERIA* AND MIDDLE CAMBRIAN *WIWAXIA.* Stefan Bengtson and Morris Simon Conway in *Lethaia,* Vol. 17, No. 4, pages 307–329; 1984.

THE DAWN OF ANIMAL LIFE: A BIOHISTORICAL STUDY. Martin Glaessner. Cambridge University Press, 1985.

CONTINUING WORK ON THE PRECAMBRIAN-CAMBRIAN BOUNDARY. John W. Cowie in *Episodes,* Vol. 8, No. 2, pages 93–97; June, 1985.

ACOUSTICS OF ANCIENT CHINESE BELLS

AN AMAZING DISCOVERY IN CHINESE MUSIC. Lee Yuan-Yuan in *Chinese Music,* Vol. 2, No. 2, pages 16–17; June, 1979.

THE MUSIC OF THE ZENGHOU ZHONG. Lee Yuan-Yuan in *Chinese Music,* Vol. 3, No. 1, pages 3–15; March, 1980.

ANCIENT CHINESE TWO-PITCH BRONZE BELLS. Ma Cheng-yuan in *Chinese Music,* Vol 3, No. 4, pages 81–86; Vol. 4, No. 1, pages 18–20; Vol. 4, No. 2, pages 31–36; December, 1980–June, 1981.

THE FIRST TECHNOLOGY

HUMAN EVOLUTION: AN ILLUSTRATED INTRODUCTION. Roger Lewin. W. H. Freeman and Company, 1984.

THE ARCHAEOLOGY OF HUMAN ORIGINS: STUDIES OF THE LOWER PLEISTOCENE IN EAST AFRICA 1971–1981. Glynn L. Isaac in *Advances in World Archaeology,* Vol. 3, pages 1–87; 1984.

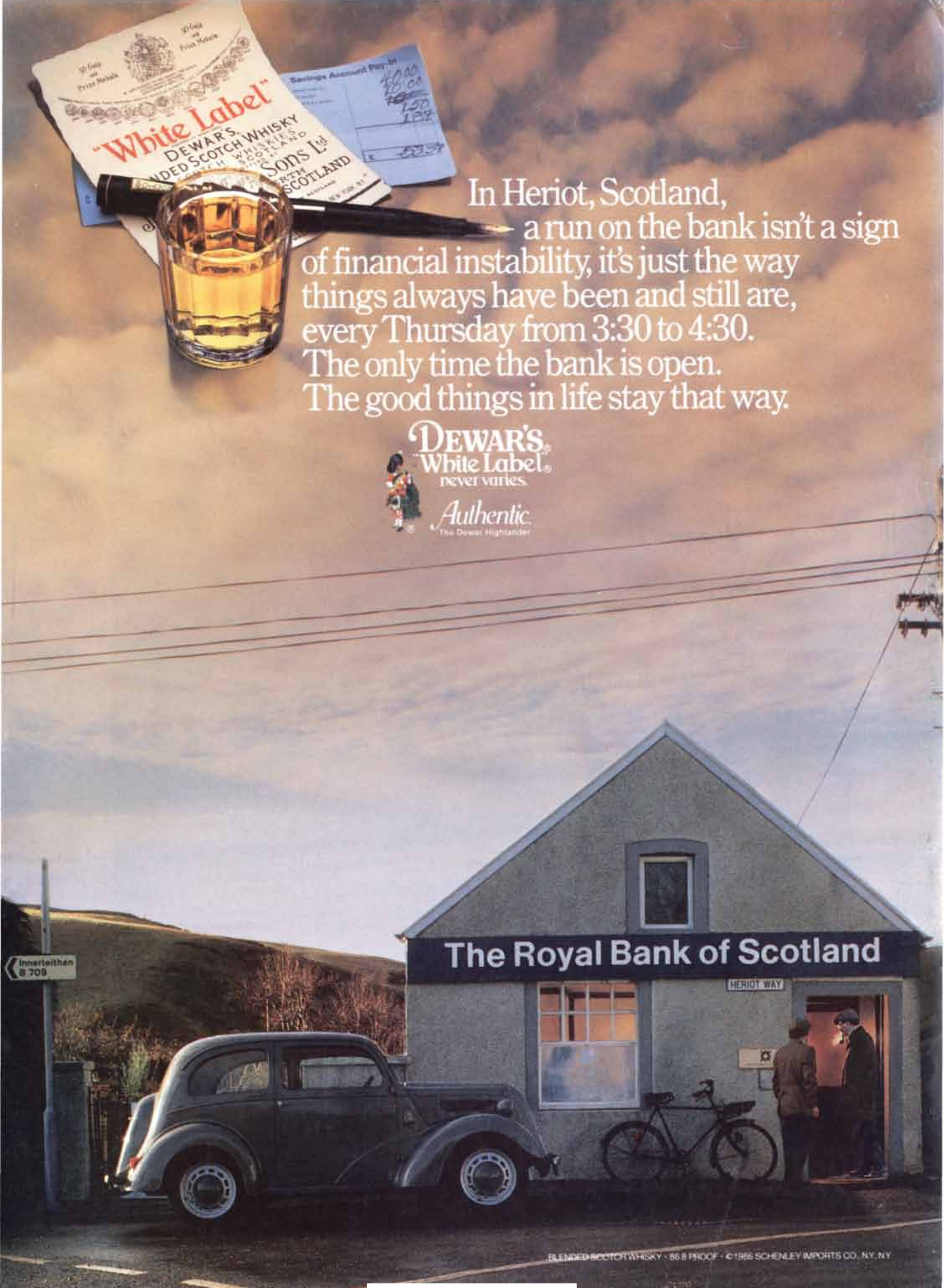
THE OLDOWAN REASSESSED: A CLOSE LOOK AT EARLY STONE ARTIFACTS. Nicholas Toth in *Journal of Archaeological Science,* Vol. 12, No. 2, pages 101–120; March, 1985.

STONE AGE SITES IN THE MAKING: EXPERIMENTS IN THE FORMATION AND TRANSFORMATION OF ARCHAEOLOGICAL OCCURRENCES (B.A.R. INTERNATIONAL SERIES 319). Kathy D. Schick. B.A.R. (British Archaeological Reports), 1986.

THE FIRST MILLION YEARS: THE ARCHAEOLOGY OF PROTOHUMAN CULTURE. Nicholas Toth and Kathy D. Schick in *Advances in Archaeological Method and Theory,* Vol. 9, pages 1–96; 1986.

THE AMATEUR SCIENTIST

THE HISTORY OF THE BAROMETER AND ITS USE IN METEOROLOGY. W. E. Knowles Middleton. The Johns Hopkins Press, 1964.



In Heriot, Scotland,
a run on the bank isn't a sign of financial instability, it's just the way things always have been and still are, every Thursday from 3:30 to 4:30.
The only time the bank is open.
The good things in life stay that way.
DEWAR'S
"White Label"
never varies.
Authentic.
The Dewar Highlander
"White Label"
DEWAR'S
SCOTCH WHISKY
SONS Ld
SCOTLAND
Savings Account Pay-in
The Royal Bank of Scotland
HERIOT WAY
Innerleithen
B 709
BLENDED SCOTCH WHISKY · 86.8 PROOF · ©1986 SCHENLEY IMPORTS CO., NY, NY