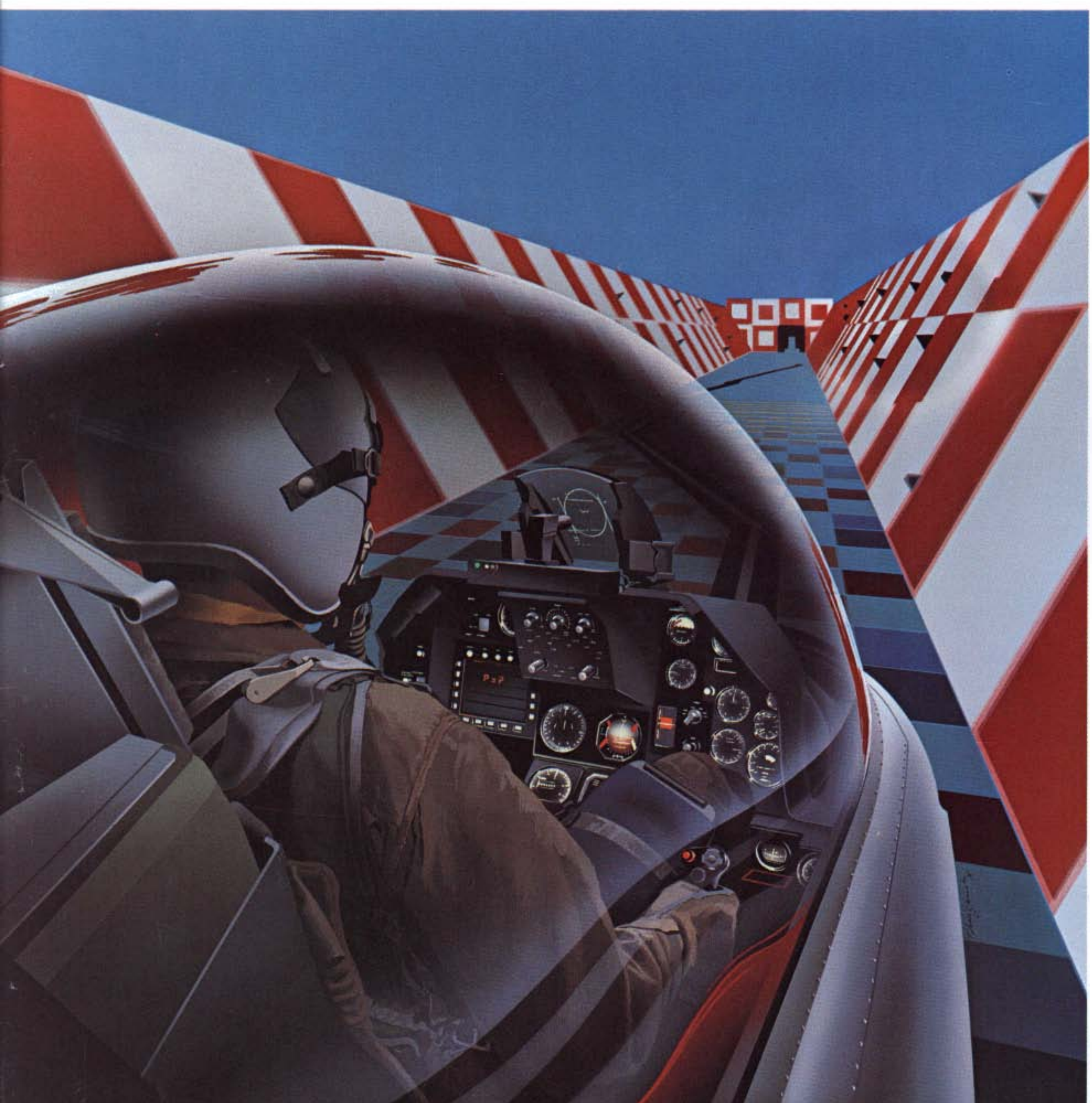
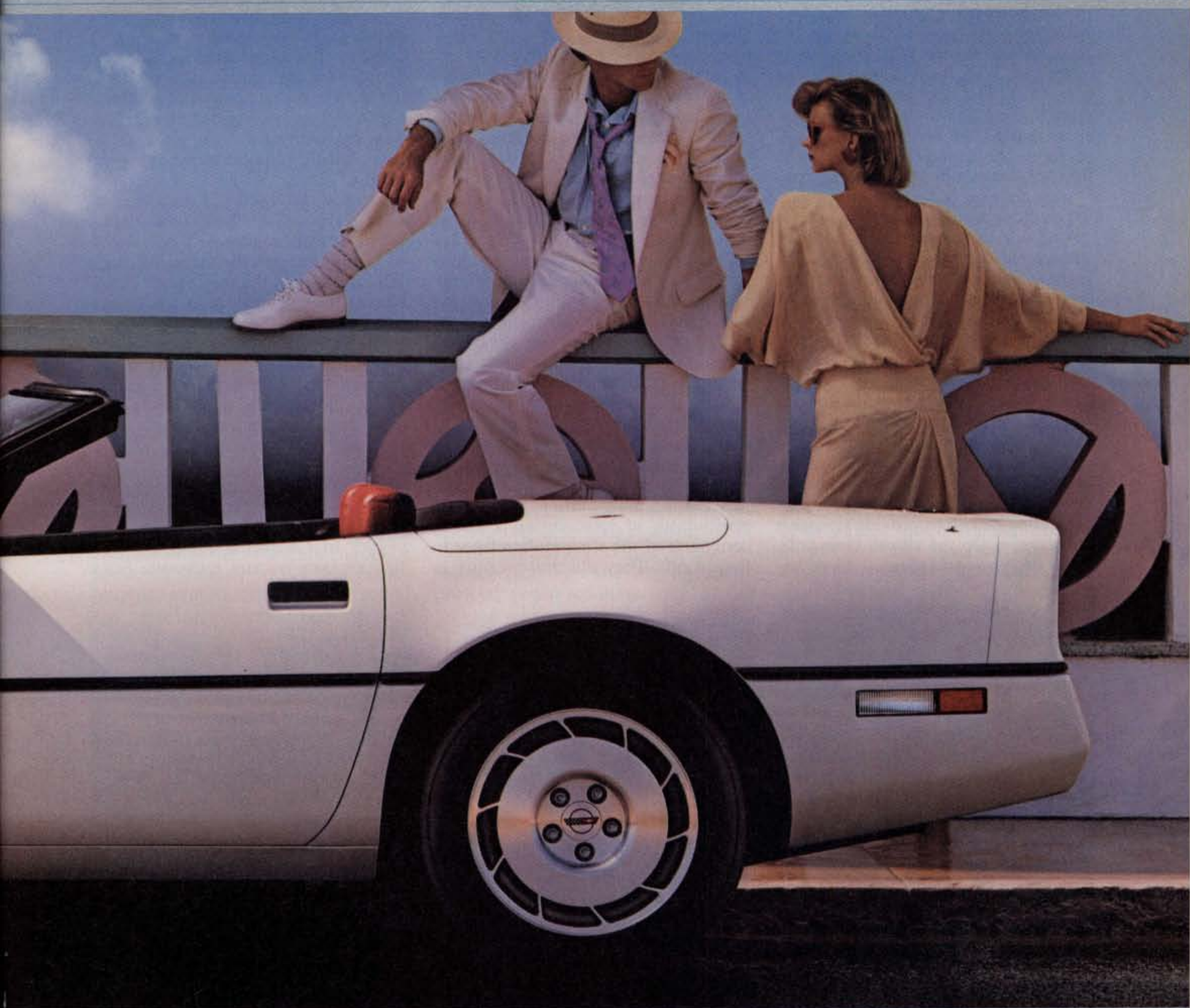
SCIENTIFIC AMERICAN



SIMULATING FLIGHT

$2.50

July 1986



Today's Chevrolet

You get perfectly focused pictures. Because the N2020 focuses automatically, whether your subject is moving or still.

What happens when you don't set the exposure? You get perfectly exposed pictures. Because the N2020 offers three programmed modes for auto-exposure, too.

What happens when you don't advance the film? You get perfect pictures anyway. Because the N2020's built-in motor drive advances it for you.

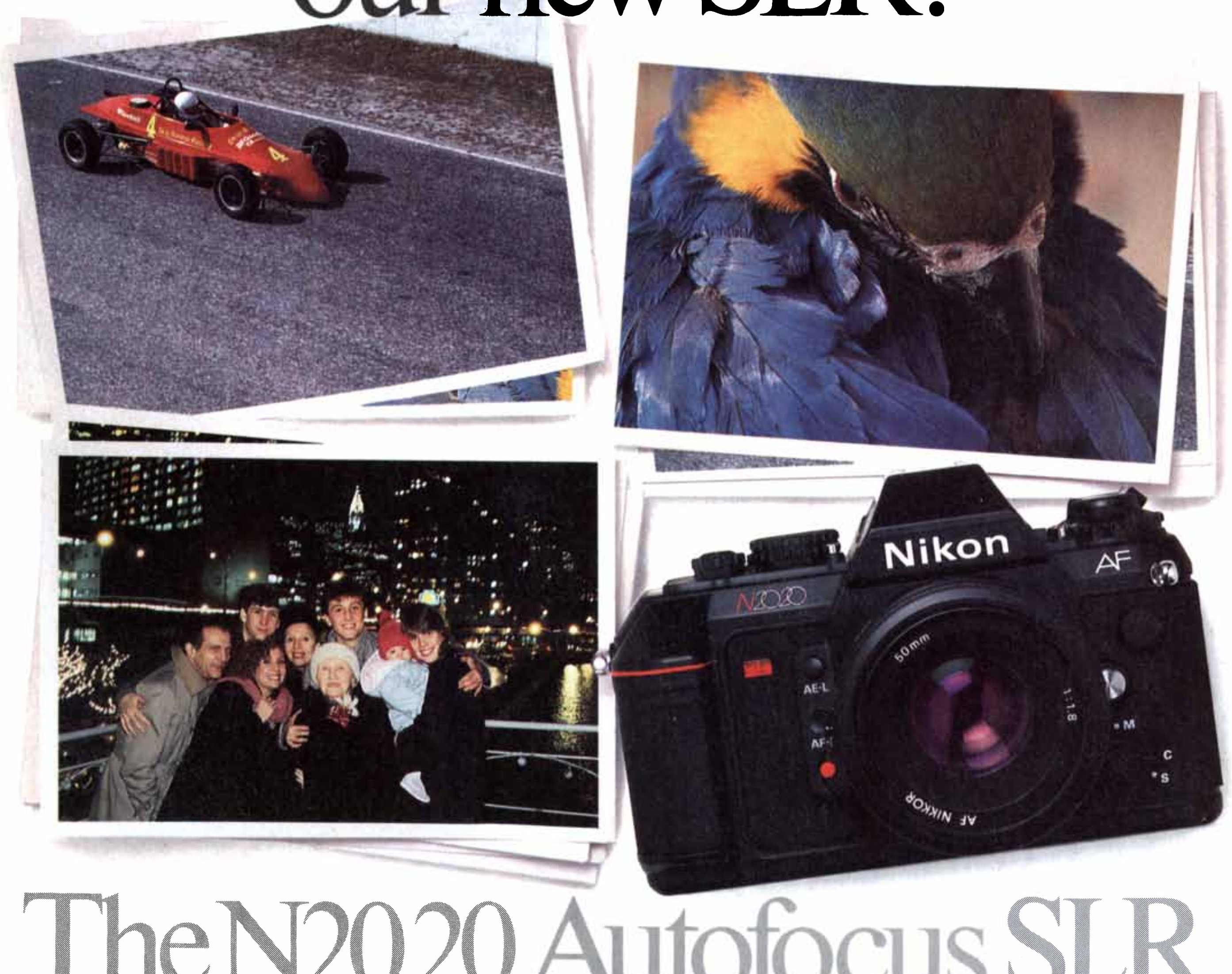
What happens when you don't have much light? You get perfect pictures even in total darkness. Because the new SB-20 Autofocus Speedlight helps the N2020 "see" in the dark.

You also get a choice of unparalleled Nikon lenses including our popular new zooms. And lots of manual options, too.

So what happens when you buy an N2020? Your pictures will be perfect whether you are or not.

Nikon
We take the world's greatest pictures.

What happens when you don't focus our new SLR?



The N2020 Autofocus SLR.

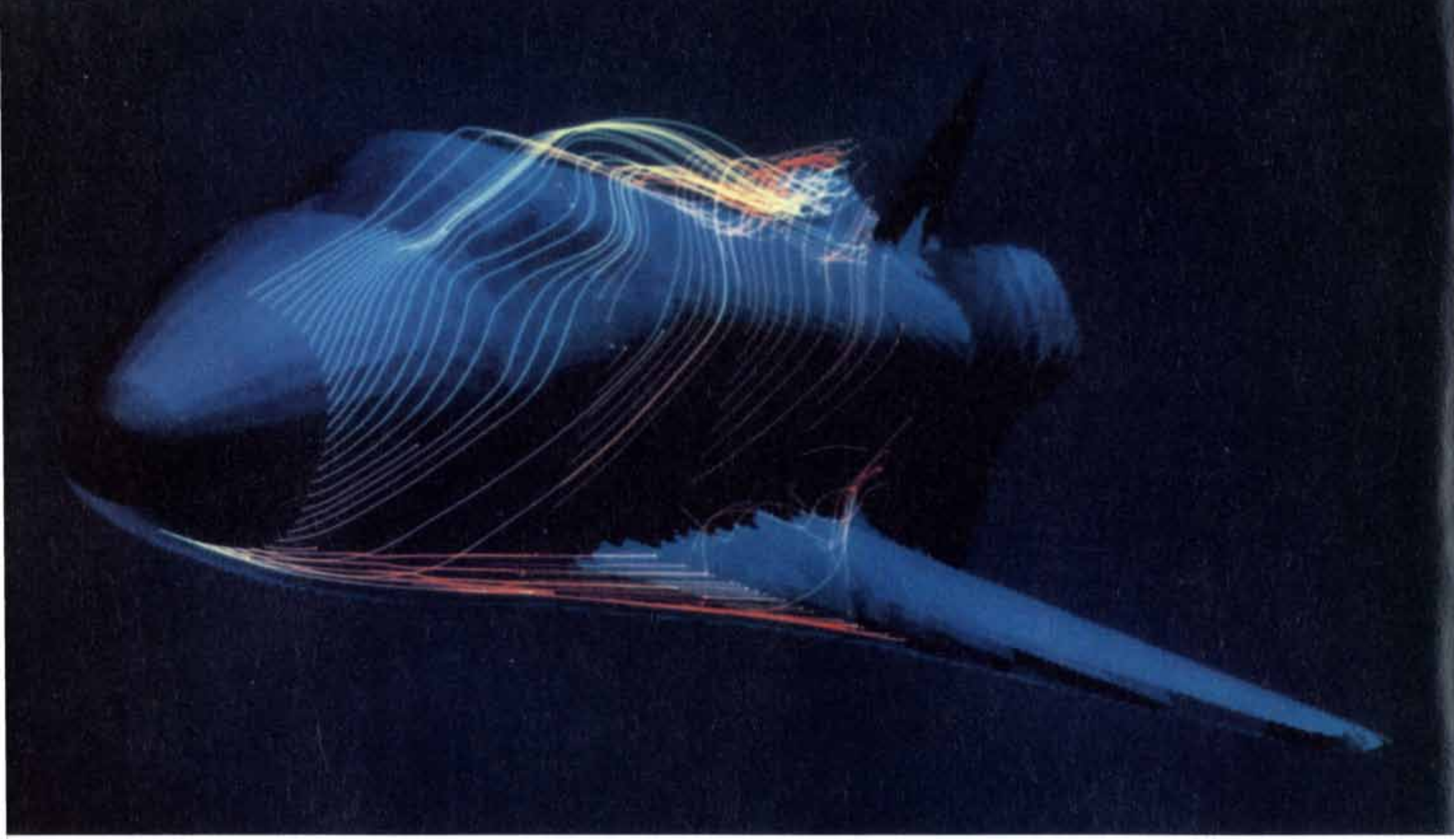
Nikon USA CLUB One year membership in the Nikon USA Club is free with every purchase when you submit the Nikon USA Club application. For further information write: Dept. N9, Nikon Inc., 19601 Hamilton Ave., Torrance, CA 90502-1309. © Nikon Inc., 1986.

Established 1845 SCIENTIFIC AMERICAN July 1986 Volume 255 Number 1

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1986 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 painting on the cover shows the cockpit of a simulator for an F-16 jet fighter surrounded by a computer-generated scene. In such a simulator a pilot can practice many flying tasks while sitting safely on the ground. Consequently flight simulators are increasingly employed in the training of both military and civilian pilots (see "Flight Simulation," by Ralph Norman Haber, page 96). The painting is based on a simulator called ESPRIT, which is now under construction at the Link Flight Simulation Division of the Singer Company. The system is designed to conserve computational power by tracking the direction of the pilot's gaze at each moment and displaying in great detail only the area that is in the line of sight (*circular area on right bank of image*). Programs that simulate flight are also available for personal computers (see "Computer Recreations," by A. K. Dewdney, page 16).

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
18–22	James Kilkelly
23	Andrew Christie
24	Gabor Kiss
32	Barrie Rokeach
34–35	Patricia J. Wynne
36–37	Patricia J. Wynne (*top*), Ilil Arbel (*bottom*)
39–40	George V. Kelvin
41	National Optical Astronomy Observatories
42–46	George V. Kelvin
47	Jack O. Burns and J. Ward Moody
49	Daniel Pauletti, Southwest Foundation for Biomedical Research
50–56	Alan D. Iselin
71	National Aeronautics and Space Administration
72–74	Ian Worpole
76	Andrew Tomko
77	Ian Worpole
78	Peter Molnar
79	Servicio Aerofotográfico Nacional del Perú
81	James Kilkelly
82–86	George Retseck
89	Peter Möller and J. Rayford Nix
90	James Kilkelly
92–95	Carol Donner
97	Bob Foster, McDonnell Aircraft Company
98–100	Ian Worpole
101	Link Flight Simulation Division of the Singer Company
102	Ian Worpole
103	Link Flight Simulation Division of the Singer Company
105	Neil F. Hadley and Greg Hendricks
106	Tom Prentiss and Neil F. Hadley (*top*), Tom Prentiss (*bottom*)
107	Elizabeth McClain (*top*), Neil F. Hadley (*bottom*)
108	Tom Prentiss
109	Barry K. Filshie (*top*), Tom Prentiss and Neil F. Hadley (*bottom*)
110	Neil F. Hadley (*top*), Tom Prentiss (*bottom*)
112	Tom Prentiss
115	Charles A. Knight, National Center for Atmospheric Research
116–119	Michael Goodman

SCIENTIFIC AMERICAN

In Other Languages

LE SCIENZE
L. 3,500/copy L. 35,000/year L. 45,000/(abroad)
Editorial, subscription correspondence:
Le Scienze S.p.A., Via del Lauro, 14, 20121 Milano, Italy.
Advertising correspondence:
Publietas, S.p.A., Via Cino de Duca, 5, 20122 Milano, Italy

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

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

POUR LA SCIENCE
25FF/copy 245FF/year 290FF/year (abroad)
Editorial, subscription, advertising correspondence:
Pour la Science S.A.R.L.,
8, rue Férou,
75006 Paris, France

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

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

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

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

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

Advertising correspondence all editions:
SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue
New York, NY 10017
Telephone: (212) 754-0262 Telex: 236115

LETTERS

To the Editors:

In the middle of the 19th century Michael Faraday discovered many fundamental laws of nature using homemade equipment, which he paid for with money saved from his modest salary. Later James Clerk Maxwell was an amateur scientist, and Albert Einstein earned his living as a postal clerk while he formulated the theory of relativity. Only one generation ago pioneers in high-energy physics were timidly hoping to receive subsidies of tens of thousands of dollars to enable them to construct the early accelerators. The accelerator escalation that has taken place since then is breathtaking; to recognize the trend is not as complex as solving the equations of unified field theory.

The cement is barely dry at the SuperCERN accelerator here in Geneva, but the carefully reasoned budget request for the next round is already being prepared. Next in line, apparently, is the accelerator described in "The Superconducting Supercollider," by J. David Jackson, Maury Tigner and Stanley Wojcicki [SCIENTIFIC AMERICAN, March]. This machine will be 52 miles in circumference and is projected to cost $4 billion.

At each stage of the accelerator escalation virtually identical arguments have been advanced to justify the leap to the next quantum. These arguments have almost always included, in one form or another, the assertion that a grasp of the fundamental nature of physical reality is just around the budgetary corner. In the words of the authors, "less energy or less luminosity begins to compromise the discovery potential." And who is willing to take such a risk? (Certainly no politician.)

The Superconducting Supercollider (SSC) may indeed be the ultimate solution. Or it may not. Based on past experience, a fairly probable scenario is that after the $4 billion (plus cost overruns and operating expenses) is spent a number of obscure results will be obtained (and reported on the front page of the *New York Times*), the usual quota of Nobel prizes will be awarded and physics will move on to the next level.

What if, as in the past, the SSC raises as many questions as it resolves, if not more? Shall we then be asked to build a 200-mile accelerator at a cost of $20 billion? Is there a limit and, if there is, how shall we recognize it?

KENT GORDIS

Geneva

To the Editors:

It is intellectual curiosity and the accompanying ability to investigate the laws of nature and benefit from investigation that distinguish the human species from other forms of life. Surely the drive to constantly improve our understanding of the fundamental nature of matter and energy has plenty of historical precedents in all areas of human endeavor.

The continuing exploration of the moon and the solar system is a natural extension of the spirit that led the early American pioneers to explore the Western frontier. The telescopes of Galileo and Kepler, which provided our earliest detailed understanding of planetary motion, have led to giant optical and radio telescopes that enable us today to probe the outer reaches of the universe. The understanding of human anatomy and of the workings of the body led naturally to a quest to find means of fighting disease and prolonging human life.

Gordis asks a legitimate question about whether there is an end in sight to our need for ever larger accelerators. The simple answer is that we do not know, but some elaboration might be worthwhile.

There are cases in the history of science in which a discovery or a theory effectively terminated a certain line of investigation. Newton's law of gravitational attraction solved the riddle of planetary motion; the breaking of the genetic code concluded a certain phase of molecular biology. There are strong reasons to believe the SSC might provide a similar breakthrough in particle physics, but only time and experiments will resolve this question.

Gordis raises an additional question: What is a feasible or desirable level of Federal support for accelerator construction and hence research in high-energy physics? Before answering we must point out that even though the SSC is a very expensive scientific instrument, its construction would not represent a radical break with past levels of funding for high-energy physics. As accelerators have become bigger (and therefore more expensive) their number has become smaller. Thus the total funding level (after correcting for inflation) has not changed very much over the past quarter of a century. Except for a significant increase in the high-energy-physics budget during the peak years of its construction, the SSC should not require major changes in the funding level.

We conclude with a personal opinion regarding Gordis' last question. It is true that the Large Electron-Positron Collider (SuperCERN in Gordis' terminology) is still unfinished, but it is a complementary machine to the SSC and will not address the same questions. The field of high-energy physics is complex enough so that a single accelerator cannot address all the important problems. Often a multiprong approach is necessary. We do not know whether fundamental questions about the nature of matter and energy will remain after the exploitation of the SSC, but if they do, curiosity will undoubtedly drive us to explore them. Because a $20-billion accelerator does not appear economically feasible, this drive will provide an additional strong impetus to a search for new, more economical and more compact acceleration techniques that would open a still higher energy regime to exploration.

There are already significant efforts to develop such techniques, but the construction of the SSC cannot await them: it appears that it will be close to a quarter of a century before they could result in a proposal for a practical scientific instrument. That time scale is simply too long to preserve the vitality of the field, particularly when one considers that it may take more than a decade to design and build a modern accelerator. If we want to pursue the age-old quest to understand the fundamental building blocks of matter, we have no choice but to build the Superconducting Supercollider.

J. DAVID JACKSON

MAURY TIGNER

STANLEY WOJCICKI

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1936: "Why does alcohol mix freely with water while gasoline merely floats on top? Why does kerosene mix freely with gasoline but not with water? Why do salt and sugar disappear in water with the stir of a spoon while powdered sulfur, iodine, or mothballs merely settle to the bottom? Why will soap mixed with water remove dirt and grease? The answers to questions such as these have come from a study of the sizes, shapes, and charges of the molecules in the new science of molecular structure."

"Cutting metal bodies under water by burning them with an electric arc is being done extensively. This means of dealing with difficult situations is saving time and money and achieving results that would be exceedingly hard to effect in any other way. With the current switched on, the diver has only to touch with the tip of the tool the metal to be burned and then retract the electrode a fraction of an inch. Since the metal body acts as the 'ground,' or return circuit, an incandescent arc is produced, and that arc, with jets of high-pressure oxygen issuing from the tip of the electrode, melts the opposing metal and blows the fused particles away. Another valuable application of the device is to weld metals under water."

"A new process, applied in Germany and the Netherlands, recovers sulfur dioxide, the objectionable gas produced as a by-product of many industrial operations, in pure form directly from the dilute waste gases. The new 'sulphidine' process dissolves the gas in a mixture of xylidine (similar to aniline) and water. When the mixture is saturated, it readily parts with the gas on heating. There is an important market for the gas in pure form."

"A few years ago bridge engineers were startled by a suggestion made by Walter E. Irving that open gratings be used as decking on bridges. The traditional idea was that the floor of a bridge should be of solid construction. The new construction has been used with such complete satisfaction in a number of large bridges that it has taken its place as an important contribution to bridge building. Its main advantage is its light weight. Another important advantage is that at no time does ice or snow accumulate."

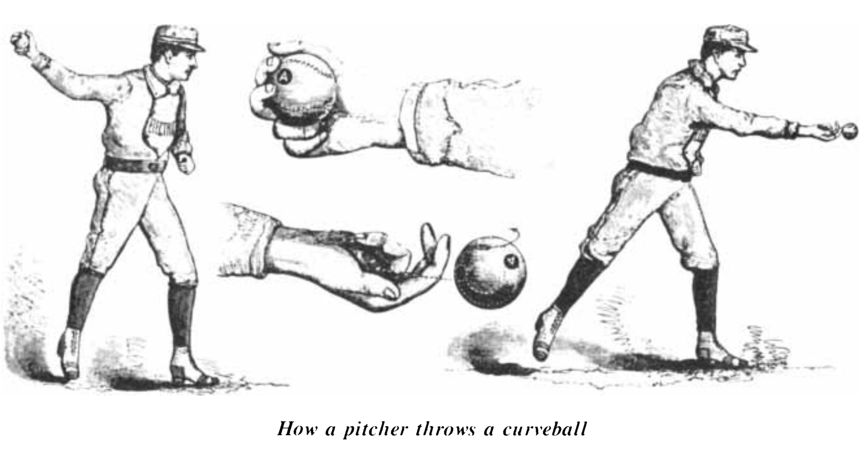
"A possible revolution in lighting practice may be approaching as a result of developmental work now in progress on two new forms of mercury lamps known as fluorescent and capillary types. The first appears to have great potentialities in providing a wide range of remarkable colored lighting effects; the second produces, in some cases, a brilliance exceeding that of the sun and emits a wealth of photo-chemical rays useful in commercial, industrial and therapeutic applications."

JULY, 1886: "The American Telephone and Telegraph Company of New York has recently been organized for the purpose of establishing direct telephonic communication between the large cities of the country. The first line has been constructed between New York and Philadelphia, the length of the route adopted being about 100 miles. The line will probably be open for business within a few weeks, and it is expected that it will prove a great convenience."

"It is a fact so curious as to be worthy of remark that the canal projects of Panama and Nicaragua have, almost simultaneously, met with an ill-fortune which can scarcely help but weaken the confidence of the public in their practicability. At Panama the $120,000,000 that was to have been sufficient to pay for the construction of a tide-level canal has been expended. Even if another $120,000,000 should be put into the project, there is abundant evidence to prove that it would be insufficient. As to the project for a lifting lock canal at Nicaragua, the recent earthquake in the vicinity of its route sweeps away in a moment a principal argument put forward in favor of its selection. Nicaragua, we have been told, is outside the zone of earthquakes."

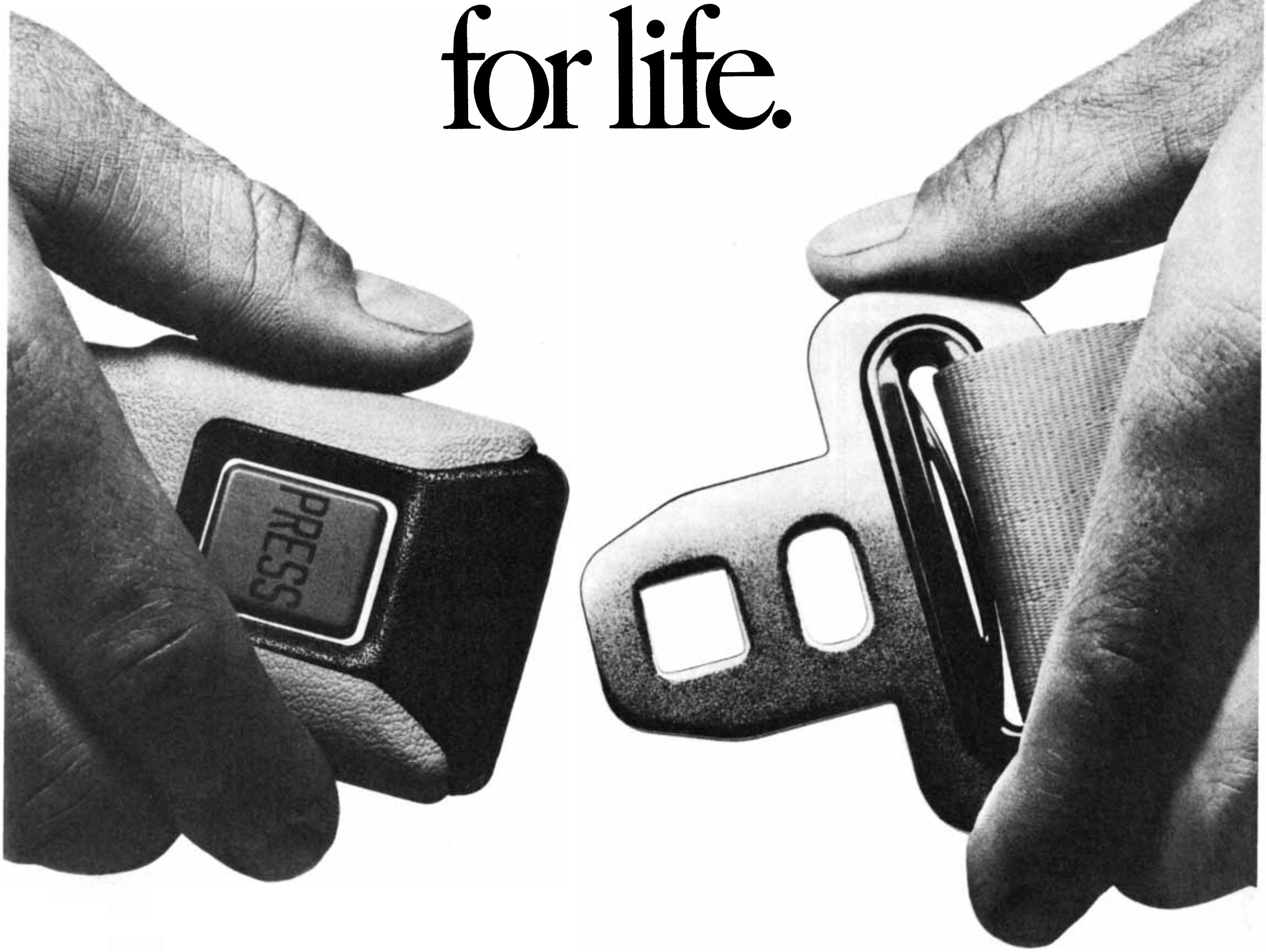
"At the Royal Institution recently, Professor James Dewar exhibited the method he employs for the production of solid oxygen. The successful device depends upon allowing liquid oxygen to expand into a partial vacuum; the enormous absorption of heat which accompanies the expansion results in the production of the solid substance. Oxygen in this condition resembles snow in appearance and has a temperature about 200 degrees Centigrade below the freezing point of water. A supply of this material will enable chemists to approach the absolute zero of temperature and so to investigate many interesting changes in the physical properties of bodies under the primordial condition of the temperature of space."

"In these days of remarkable exhibitions of skill in playing baseball by professional exemplars of the game, one cannot look back to the early period without being struck by the great contrast between the work done on the diamond fields at Hoboken, in the 'fifties,' and that which marks the play of the leading professional teams of the present era. The game has been wonderfully improved, and in nothing so much as in the great degree of skill now shown in the pitching department. Modern pitching excels the old method in one special feature, and that is in *the horizontal curve of the ball through the air*. If the ball be made to rotate on its own axis from right to left or left to right, to the extent of this increased lateral friction is the ball retarded on the side on which the increased friction bears. The result is naturally a curve in the direction of the side on which the ball's progress has been retarded."



How a pitcher throws a curveball

Seat belts. Guaranteed for life.



Seat belts help save lives. Seat belts help reduce injuries. Seat belts work. Guaranteed.

But sometimes even seat belts don't work properly. For some people that's the perfect excuse not to wear one.

Well, when it comes to seat belts, Honda does not believe in excuses. That's why if any seat belt in any Honda ever fails to function properly, Honda will repair or replace it.* Free. You see, Honda wants you to wear your seat belt whenever you're in a car.

All you have to do is bring your Honda to an authorized Honda dealer. It doesn't matter if you drive a 1970 Honda or a 1986 Honda. It doesn't matter if you bought it third hand or brand new. The only thing that is important is that your seat belts work. If they don't, we fix them, free.* Regardless of automobile's age or mileage.

After that, it's up to you to wear one. Just remember, seat belts are a simple fact of life.

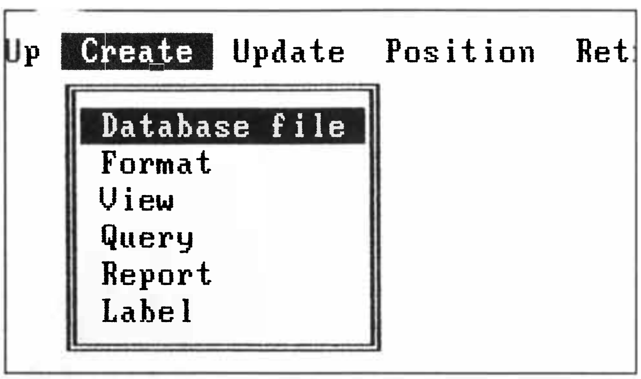
*Limited written warranty. Some restrictions apply. See dealer for details. Applies to model years 1970-1986.

HONDA

The database used now be used

Introducing dBASE III® PLUS.

The PLUS stands for all the improvements we've made to the world's number one selling database management software.



The Assistant helps beginning users accomplish day-to-day data management tasks without programming.

Mind you, dBASE III PLUS still has the powerful dBASE programming language, dot prompt, and all the features that have made dBASE III the standard of the industry.

We've simply raised the standard.

And just as dBASE III introduced more power to the people, our new dBASE III PLUS introduces more people to the power.

People who aren't all that crazy about programming, for example.

The Assistant feature in dBASE III PLUS now provides them with new easy-to-use pull-down menus for creating, using and modifying multiple databases.

So now anyone who can manage a simple cursor can manage day-to-day data management tasks. Without programming.

And by using our new Screen Painter, anyone can create custom screens. Without programming.

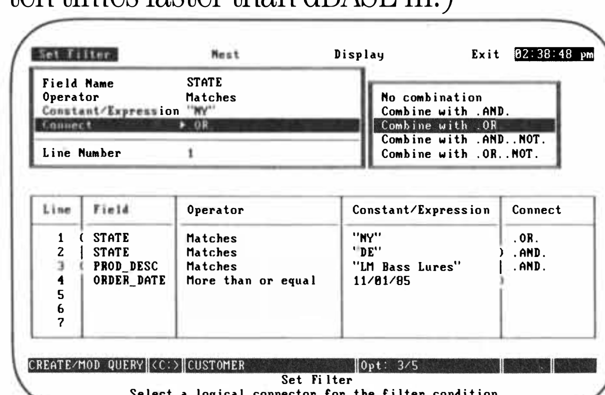
Or using View, access related information in several databases at one time. Without programming.

With Advanced Query System, another new non-programming feature, any user can build complex queries just by selecting from the dBASE III PLUS pull-down menus.

For rapidly creating entire programs, there's even a new Applications Generator.

And for all those who wish to learn to program, the Assistant can be of further assistance. By teaching you programming commands as you go along. Without disrupting your work flow.

These are only a few of the dBASE III PLUS features that can help new users quickly get up to speed. And experienced users quickly increase their speed. (Sorting, for example, is up to two times faster and indexing up to ten times faster than dBASE III.)



Advanced Query System lets you set up and answer complex queries without programming.

ASHTON·TATE
dBASE III PLUS
The data management standard.

THE AUTHORS

C. WILEY HINMAN ("Potential New Crops") is adjunct professor at the University of Arizona's College of Agriculture. He is also president of Hinman Associates and vice-president of Vega Biotechnologies, Inc. He was graduated in 1951 from Grinnell College with a bachelor's degree in chemistry and earned a Ph.D. in organic chemistry and biochemistry from the University of Illinois at Urbana-Champaign in 1954. Hinman has worked for the Dow Chemical Company and the Diamond Shamrock Corporation and has held several Government advisory positions.

JACK O. BURNS ("Very Large Structures in the Universe") is director of the Institute for Astrophysics and associate professor of physics and astronomy at the University of New Mexico. He received his B.S. in astrophysics at the University of Massachusetts in 1974; his M.S. (1976) and Ph.D. (1978) are from Indiana University. From 1978 to 1980 he was a postdoctoral research associate in the National Radio Astronomy Observatory's Very Large Array program and then he joined the faculty at New Mexico. The subject of his article, cosmology, represents a relatively new area of research for Burns, his first research interest having been extragalactic radio astronomy.

RONALD C. KENNEDY, JOSEPH L. MELNICK and GORDON R. DREESMAN ("Anti-idiotypes and Immunity") share an interest in viral diseases and immunity. Kennedy is assistant scientist at the Southwest Foundation for Biomedical Research in San Antonio and adjunct assistant professor of microbiology at the University of Texas Health Science Center at San Antonio. He got his M.S. and Ph.D. from the University of Hawaii at Manoa before going on to postdoctoral work under Melnick and Dreesman at the Baylor College of Medicine, where he became assistant professor in 1983. Kennedy joined the Southwest Foundation in 1984. Melnick is distinguished service professor of virology and epidemiology and dean of graduate sciences at Baylor. He earned his Ph.D. in biochemistry at Yale University, where he taught for 18 years before moving to Baylor in 1958. He began investigating viral vaccines during World War II and has been a member of the World Health Organization's expert advisory panel on virus diseases for 28 years. Dreesman is chairman of the department of virology at the Southwest Foundation and an adjunct full professor of virology at the University of Texas Health Science Center and at Baylor. He received his Ph.D. in microbiology from the University of Hawaii at Manoa (1965) and taught at the St. Louis University School of Medicine before moving to Baylor in 1969. Dreesman became a full professor at Baylor in 1979 and left in 1984 to join the Southwest Foundation.

PETER MOLNAR ("The Structure of Mountain Ranges") is associate professor of earth, atmospheric and planetary sciences at the Massachusetts Institute of Technology, a position he has held since 1978. He is a graduate of Oberlin College and got his Ph.D. in geophysics from Columbia University in 1970. Molnar writes: "I try to spend at least three months each year in remote mountainous environments doing fieldwork and savoring the virtues of the particular indigenous culture. Free time at home is often spent struggling to learn a little of the language spoken in the next field area or reading about the geography, anthropology, archaeology and architecture of some of the extraordinary cultures of the world."

J. H. HAMILTON and J. A. MARUHN ("Exotic Atomic Nuclei") are respectively Landon C. Garland professor of physics at Vanderbilt University and associate professor of theoretical physics at the University of Frankfurt. Hamilton earned his B.S. (1954) at Mississippi College and his M.S. (1956) and Ph.D. (1958) from Indiana University. After finishing his graduate studies he joined the faculty at Vanderbilt, where he has been since, except for time spent on various visiting research appointments and fellowships in Europe. Maruhn studied at the University of Frankfurt, receiving a Ph.D. in 1973 for a thesis on mass distributions in nuclear fission. After working as a research associate at the Oak Ridge National Laboratory in Tennessee from 1974 to 1977, he returned to Frankfurt to take up his current position at the university's Institute for Theoretical Physics.

FRANK A. GELDARD and CARL E. SHERRICK ("Space, Time and Touch") collaborated for many years in studies of the stimulus-processing capabilities of the skin. Geldard was Stuart Professor of psychology emeritus as well as senior research psycholo-

COME FACE TO FACE WITH SPACE.

Climb inside a real space capsule. Encounter simulated zero-gravity. Fire rocket engines and lasers. And even pilot a spacecraft by computer. At The Space and Rocket Center in Huntsville, Alabama, over 60 hands-on exhibits put the space program at your fingertips.

You'll also experience our spectacular Spacedome theater. Here, engulfed by a wraparound screen and 48-speaker sound system, you'll thrill to our new film, "The Dream is Alive".

Next door at Marshall Space Flight Center, you'll witness the development of NASA's Space Station and see where our astronauts prepare for missions.

During your visit, you can catch your breath at the Huntsville Marriott, located on the grounds.

Without a doubt, it's the best place to really get in touch with the space program. So, take a day or two and come face to face with 30 years of American adventure. At The Space and Rocket Center.

Get the right stuff for planning a visit, plus information on the incredible U.S. Space Camp. Call toll-free for a copy of our free brochure.

NASA

THE SPACE AND ROCKET CENTER.
Huntsville, Alabama
800/633-7280 or 205/837-3400

gist at Princeton University until his death in 1984. His undergraduate and graduate degrees were all from Clark University. After getting his Ph.D. in 1928, he joined the faculty at the University of Virginia. During World War II he was chief of the psychology section in the Office of the Surgeon General of the Army Air Forces Training Command. After the war he remained at Virginia until 1962, when he moved to Princeton. He seredipitously discovered the saltation phenomenon described in the article while giving his last laboratory course in perception before retiring from teaching in 1972. Sherrick is senior research psychologist and lecturer in psychology at Princeton University. He did his graduate work in experimental psychology under Geldard at the University of Virginia, receiving his Ph.D. in 1952. After a year as a research associate at Virginia, he moved to St. Louis, where he taught and did research at Washington University and at the Central Institute for the Deaf. In 1962 he returned briefly to Virginia and then went with Geldard to Princeton. Sherrick is particularly interested in the potential of touch as a substitute for impaired senses.

RALPH NORMAN HABER ("Flight Simulation") is professor of psychology at the University of Illinois at Chicago Circle. He holds a B.A. in philosophy from the University of Michigan (1953), an M.A. in psychology from Wesleyan University (1954) and a Ph.D. in psychology from Stanford University (1957). He taught at Yale University for six years and at the University of Rochester for 15 years before going to Illinois. He wrote the current article while on leave working as a visiting scientist in the Air Force Human Resources Laboratory at Williams Air Force Base in Arizona. Haber's basic interest is in how humans perceive the layout of the visual space surrounding them.

NEIL F. HADLEY ("The Arthropod Cuticle") is professor of zoology at Arizona State University. He got his B.A. in 1963 at Eastern Michigan University. After receiving his Ph.D. from the University of Colorado in 1966, he joined the faculty at Arizona State. Hadley edited *Environmental Physiology of Desert Organisms* (Academic Press, 1975) and is the author of *The Adaptive Role of Lipids in Biological Systems* (John Wiley & Sons, 1985). His research interests encompass comparative and environmental physiology, with emphasis in recent years on the retention and loss of water by terrestrial arthropods.

COMPUTER RECREATIONS

A sublime flight of fancy over a deserted data base

by A. K. Dewdney

Nervously I studied the instrument panel of my Cessna 182. Was I foolish to try a solo flight without instruction? Should I find some experienced hand to take me aloft? Outside on the bleak landscape of Meigs Field in Chicago not a soul could be seen, not even a mechanic out for a smoke in the sun. My only companion on this mission of madness was the flight manual in my lap. I had spent 10 minutes reading it.

I suppose I would have been more confident with a joystick or a wheel in my hand. As a child I had seen *Hellcats of the Navy* at least 10 times; I already knew the basic moves. But this airplane was evidently far more advanced. It was controlled entirely from a keyboard in front of me.

With a sigh I turned to page 46 of the manual. The F2 key would nudge the throttle; the F4 key would push it wide open. Not entirely reckless, I tentatively pressed F2. In the background the soft drone of the engine rose a semitone in pitch. The runway started rolling toward me, and suddenly I had a moment of panic. The plane almost immediately began to veer off the concrete strip. In my ignorance I had neglected to read anything about the rudder or the nosewheel. My only salvation was the ailerons, the keys labeled 4 (left bank) and 6 (right bank). To my amazement one of them brought me slowly back to the runway.

Only 100 yards ahead loomed the deep blue waters of Lake Michigan. More throttle. F2 F2 F2 F2. Still more. The engine sounded fuller but the plane refused to lift. The runway rolled ever faster below me and I had another wave of panic. What magic control would lift me away? My life flashed before my eyes. Scenes from my childhood: there I was flying a model airplane, controlled only by a pair of flaps on the tail. The elevator. Feverishly I looked up "elevator" in the manual. The elevator controls an aircraft's angle of attack; raising it would lift the nose of the plane. I pressed the key labeled 2 just in time to avoid a watery grave. The severe geometry of land and water receded smoothly. I was airborne! The plane seemed to climb of its own accord. My perspective on Lake Michigan's western shoreline improved steadily, and I flew north for a time, hypnotized by my achievement.

As many readers will have guessed by now, the cockpit was the study in my house and the controls were on the keyboard of my personal computer, which was running a program for simulating flight. My familiar surroundings, however, in no way diminished my sense of adventure; the fact that both the airplane and the landscape were mere simulations only served to expand my imagination. Here one can try things one would never attempt in a real aircraft. The bills for damage would be imaginary too. I banked the plane to the southwest, heading for Chicago. My intention was to swoop down Michigan Avenue and impress the traffic with my derring-do. According to a clock on the instrument panel, it was 4:30, the start of evening rush hour.

As I approached Chicago something seemed terribly amiss. Only a few buildings were visible along the lakefront, and the streets were empty of traffic. There was absolutely no one about. Had there been a nuclear war? I noticed I had mistaken the two-hand altimeter for a clock, but even the discovery that it was much earlier than 4:30 could not explain the depopulation of the great metropolis. With reluctance I concluded the flight-simulation program had no people in it. Its data base was already crammed full of North American landscape, not to mention the geographic details for simulating four metropolitan areas and more than 20 airports.

On more recent flights I have become less and less concerned about the lack of company. It is just too much fun to fly. Many other people apparently feel the same way. The program is called FLIGHT SIMULATOR, and it has been at or near the top of the recreational software best-seller charts for many months. There are many other flight-simulation programs available both commercially and free through computer bulletin boards accessible by telephone. There is a single presence, however, behind FLIGHT SIMULATOR and a number of other panic inducers: Bruce A. Artwick, a software engineer from Champaign, Ill. Perhaps more than anyone else, Artwick is responsible for the takeoff of flight-simulation programs that can be run on home computers. Flight simulation has long been the almost exclusive province of pilots in training [see "Flight Simulation," by Ralph Norman Haber, page 96], but thanks to Artwick and other programmers it is now virtually common property.

Consider what a flight-simulation program must do in order to sustain the user's willing suspension of disbelief. First it must create and re-create a landscape from a specific vantage point several times a second. As the simulated aircraft plies its simulated skies, the vantage point changes and the appearance of the landscape must be recomputed. Furthermore, for each vantage point there must be an appropriate viewing angle, which is derived by simulating the physics of an aircraft in flight. For example, as the aircraft banks and turns, the viewing angle as well as the vantage point must be shifted accordingly. The basic computational cycle exploits a number of tricks and trade-offs that are fascinating in their own right.

The geographic data base is fundamental to FLIGHT SIMULATOR. It is a vast list of the objects needed to make a picture of the landscapes one might want to fly over (or avoid crashing into). You gain access to the scene of your flight by tracing a top-down hierarchy of information. A block of memory labeled U.S.A. holds the names of regions such as Chicago, New York and Seattle. The position of the aircraft is then given by an ordered triplet of windscreen-oriented coordinates: x, its latitude, y, its height above sea level in meters, and z, its longitude. FLIGHT SIMULATOR then compares the coordinates with the limits defined for each region in the U.S.A. block. For example, if x and z specify a point in the Chicago area, the program executes an operation called a data-base load: the list that describes the Chicago area is loaded from disk into the memory of the computer.


DODGE 600
5/50 PROTECTION, STANDARD.

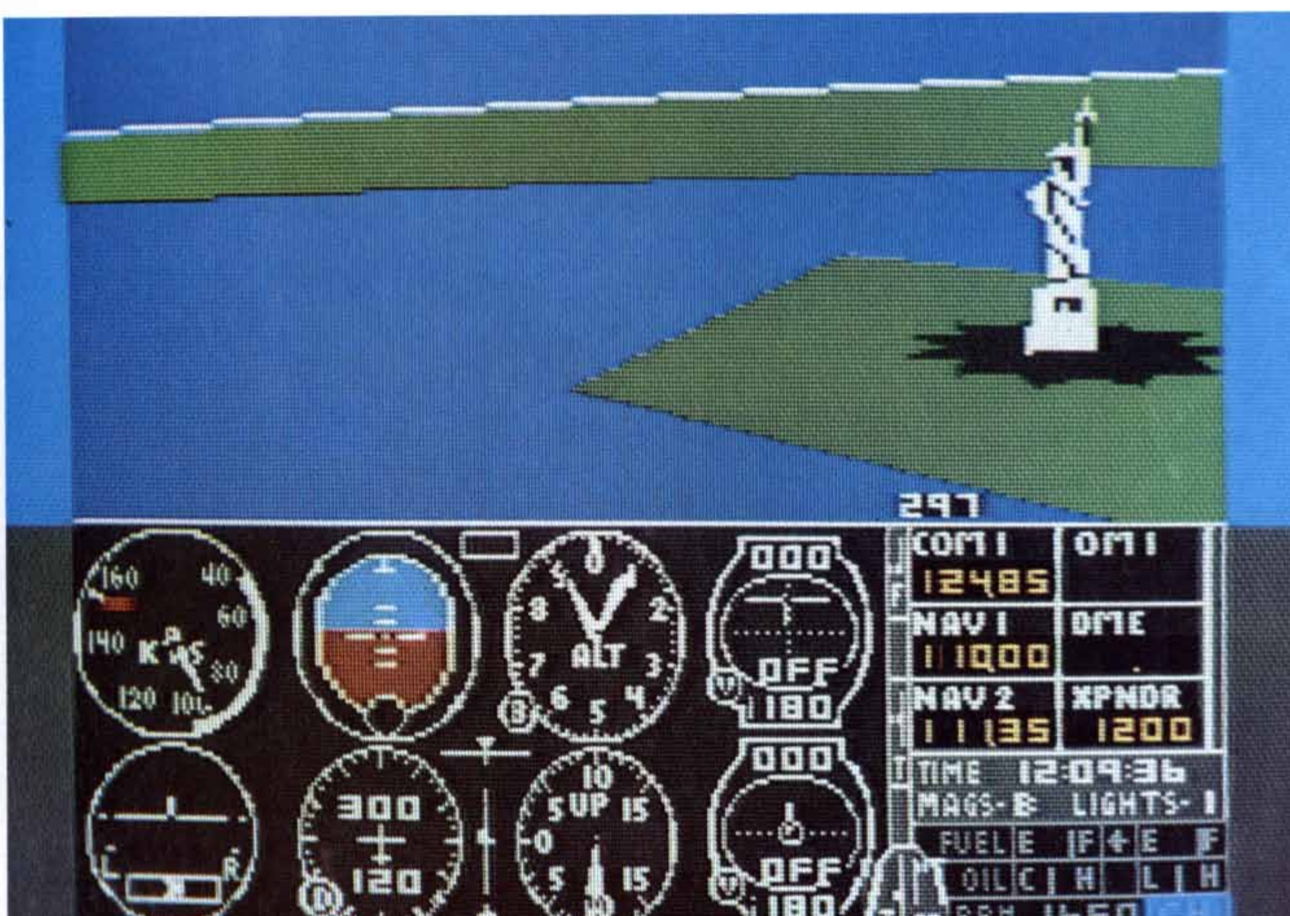
When I first banked the plane toward Chicago, no buildings were visible. After a while, however, the Sears Tower popped into view. Later it took on detail. Such is the strategy of FLIGHT SIMULATOR for staying within the speed and memory limits of a microcomputer: a feature of the landscape is not displayed until it is really needed. There are two special test routines in the program that determine the appearance or disappearance of a feature. The first of them, *ifin2d,* tests only two coordinates, the longitude and latitude, of the plane. I have already described how it associates a major geographic region with the two coordinates.

The second test routine, *ifin3d,* determines whether a given feature of the surrounding landscape is close enough to warrant a display. As I approached Chicago *ifin3d* continually recomputed my three-dimensional distance from certain downtown landmarks. When I approached within 12,800 meters of the Sears Tower, a miniature replica of the building suddenly sprouted from the landscape as a wriggling stack of pixels in the distance. Closer, at a distance of 7,680 meters, more detail suddenly became visible. For each new set of coordinates FLIGHT SIMULATOR zips down the display list through all the features of the landscape in memory and tests them according to the display rules in *ifin3d.* If an object passes the test, the program draws it according to its appearance from the current vantage point. If an object fails the test, it is not drawn and the program branches to the next object on the list.

In general every object or feature of the simulated scenery is defined by a list of points, lines and surfaces. Because the frames of visual experience are computed five times per second, fast enough to exploit the persistence of vision, lines and surfaces must be drawn in the shortest time possible. Hence even the elementary graphic operations in the program were designed with great care. Artwick has described them in his fascinating book *Microcomputer Displays, Graphics, and Animation.*

For example, the basic method for displaying a line eschews multiplication and division because they take between 10 and 100 times as long to accomplish on a microcomputer as addition and subtraction do. Instead the method is based on a careful analysis of the computational overhead of each instruction. The clever line-drawing algorithm that results is essentially made up of only four instructions: an addition, a test, a branch and a plot. Suppose the point (a,b) has just been plotted on a line of (positive) slope 1/4. The program enters a loop in which it repeatedly adds 1 to the a coordinate, adds the slope 1/4 to a special sum called the tracking error and then tests the tracking error. When the tracking error is greater than or equal to a positive integer, the program exits the loop and plots the next point, in this case $(a + 4, b + 1)$.

FLIGHT SIMULATOR draws surfaces in a way that seems cumbersome to describe but is extremely fast to execute. Briefly, the program draws the polygon bounding the surface and then fills in the interior. When the shape of the polygon has been computed, the graphic routine first identifies the scan lines on the monitor that cross the polygon. For each such scan line the edges of the polygon that are crossed are also identified and then sorted according to the order in which the scan line crosses



A view of the Statue of Liberty from the cockpit in FLIGHT SIMULATOR

Pleasure Trip

Don't let radar spoil it: Remember your Passport

Imagine setting out on a long day's drive and not having to worry about radar. Seeing America might be fun again if you didn't have to watch it unfold in the rearview mirror.

PASSPORT means low-profile protection

PASSPORT has exactly what the vacationing driver needs: advanced radar protection, yet none of the nuisances you've come to expect of ordinary radar detectors.

The experts at *Car and Driver* magazine gave us a rave review: "In a word, the Passport is a winner."

Our customers agree, even though we allow them ample opportunity to do otherwise. Every Passport is sold with the promise that it will live up to the customer's highest hopes, not just our claims. If it doesn't satisfy within the first 30 days, it can be sent back for a full refund.

PASSPORT has been on the market just over a year and already our money-back offer has been declined a quarter of a million times. So we'll stand on our record.

PASSPORT means the size of a cassette tape

PASSPORT provides superheterodyne performance in a package the size of a cassette tape. This miniaturization is possible only with SMDs (Surface Mounted Devices), microelectronics common in satellites but unprecedented in radar detectors.

All you do is clip PASSPORT to your dashtop, windshield or visor, plug it in, and adjust the volume. Upon radar contact, the alert lamp glows and the variable-pulse audio begins a slow warning: "beep" for X band radar, "brap" for K band. Simultaneously a bar graph of Hewlett-Packard LEDs shows radar proximity.

PASSPORT comes complete with all accessories.

As you get closer, the pulse quickens and the bar graph lengthens. PASSPORT watches for radar so you can watch the road.

PASSPORT comes with a leather travel case and everything necessary to get you on the road in seconds, no matter what car you drive. And one more convenience—we deliver. We make PASSPORT in our own factory in Cincinnati and we sell direct to you.

PASSPORT means an easy fit in your travel case

Call toll free. We'll answer your questions. If you decide to buy, we'll ship by UPS the next business day at our expense. For $6.00 extra, Federal Express will deliver to you within two business days of shipment.

And once again, we make this promise: if you're not satisfied within 30 days, return PASSPORT. We'll refund your purchase and your shipping costs. There are no hidden charges. Moreover, PASSPORT is backed by a one-year limited warranty.

Next time you set out on a pleasure trip, don't forget your PASSPORT.

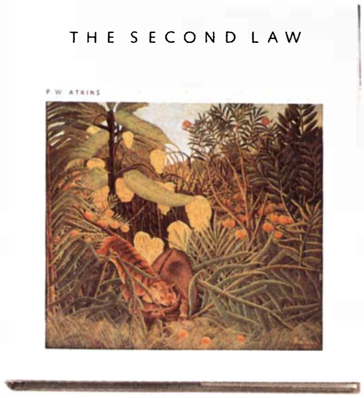
$295 (OH res. add $16.23 tax)
Slightly higher in Canada

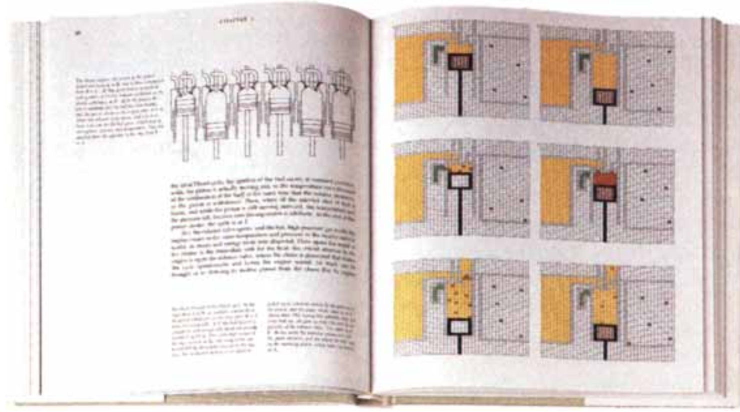
VISA MasterCard

PASSPORT®
RADAR • RECEIVER

Call Toll Free 800-543-1608

Cincinnati Microwave
Department 9167
One Microwave Plaza
Cincinnati, Ohio 45296-0100





YOUR PREMIER VOLUME

THE SECOND LAW

P.W. Atkins

In his famous essay on the two cultures, C.P. Snow made the Second Law of thermodynamics the litmus test of scientific literacy: Not to know the Second Law is the same as not having read a work of Shakespeare.

This is the law, of course, that explains why hot objects grow cool whereas cool objects do not spontaneously become hot; why a bouncing ball must come to rest and a resting ball cannot, of its own, begin to bounce. To some people, the Second Law conjures up visions of clanking steam engines, intricate mathematics and incomprehensible physical theory. It is better known to many in its restatement as Murphy's second law: "Things, if left alone, will gradually go from bad to worse!"

That restatement captures, in its way, the universality and power of the law that inspired Arthur Eddington to call it "the arrow of time."

In *The Second Law* P. W. Atkins, of Oxford University, breaks through the mathematical barriers to understanding of the law. By vivid example he shows it at work in heat engines, refrigerators, and heat pumps; the cosmos, in the ecosphere and in the living cell; in the physical, chemical, and life processes that drive the changing world in all its richness and variety.

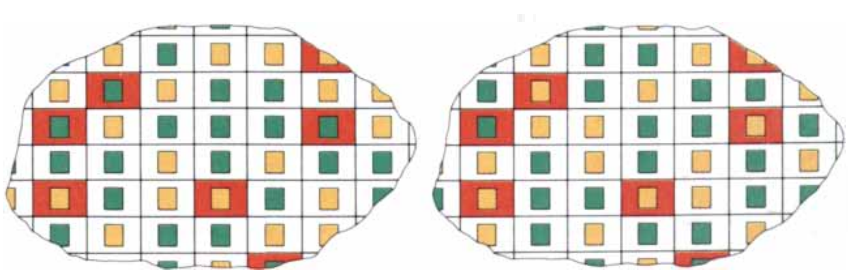
A special appendix equips readers who have computers with a kit of tools to put the Second Law to work.

230 pages, 104 illustrations

P. W. ATKINS is a lecturer in physical chemistry at the University of Oxford and a fellow of Lincoln College. He is the author of numerous distinguished works in his field, including *Physical Chemistry,* Second Edition, *The Creation,* and *Molecular Quantum Mechanics,* Second Edition.

"The interested person can do no better than to peruse this fascinating new volume by Peter Atkins . . . . Even those who feel they already know all about the Second Law can hardly fail to find new angles and fresh insights into what is going on in the world around them. I hope that this remarkable book will be widely read by scientists as well as by the non-scientists for whom it is primarily intended."
NEW SCIENTIST

"A lovely book, beautifully illustrated and presented, and clearly commensurate with its companion volumes in the Scientific American Library. *It should engender affection as a sophisticated latterday Hogben, complete with social allusions implicit in the call to conserve, not energy, but entropy, and to use the virtuous entropy-conserving heat pump rather than crude and profligate combustion...."*
NATURE



The Mark II universe model of a paramagnetic material at some arbitrary temperature. In the absence of a magnetic field, the opportunity for electrons to be UP (yellow) or DOWN (green) is an additional contribution to the entropy of the sample. Atoms may still be ON (red) and OFF (white), but the spins they carry may be UP or DOWN at random. The two illustrations show the same distributions of ON and OFF, but different distributions of UP and DOWN.

them, from left to right. The routine then considers each pixel, or picture element, of the scan line in left-to-right order. When the line encounters the first edge of the polygon, it begins to display points colored appropriately for the surface being represented. When the line encounters the next edge, it stops displaying points. At the edge after that one it begins again.

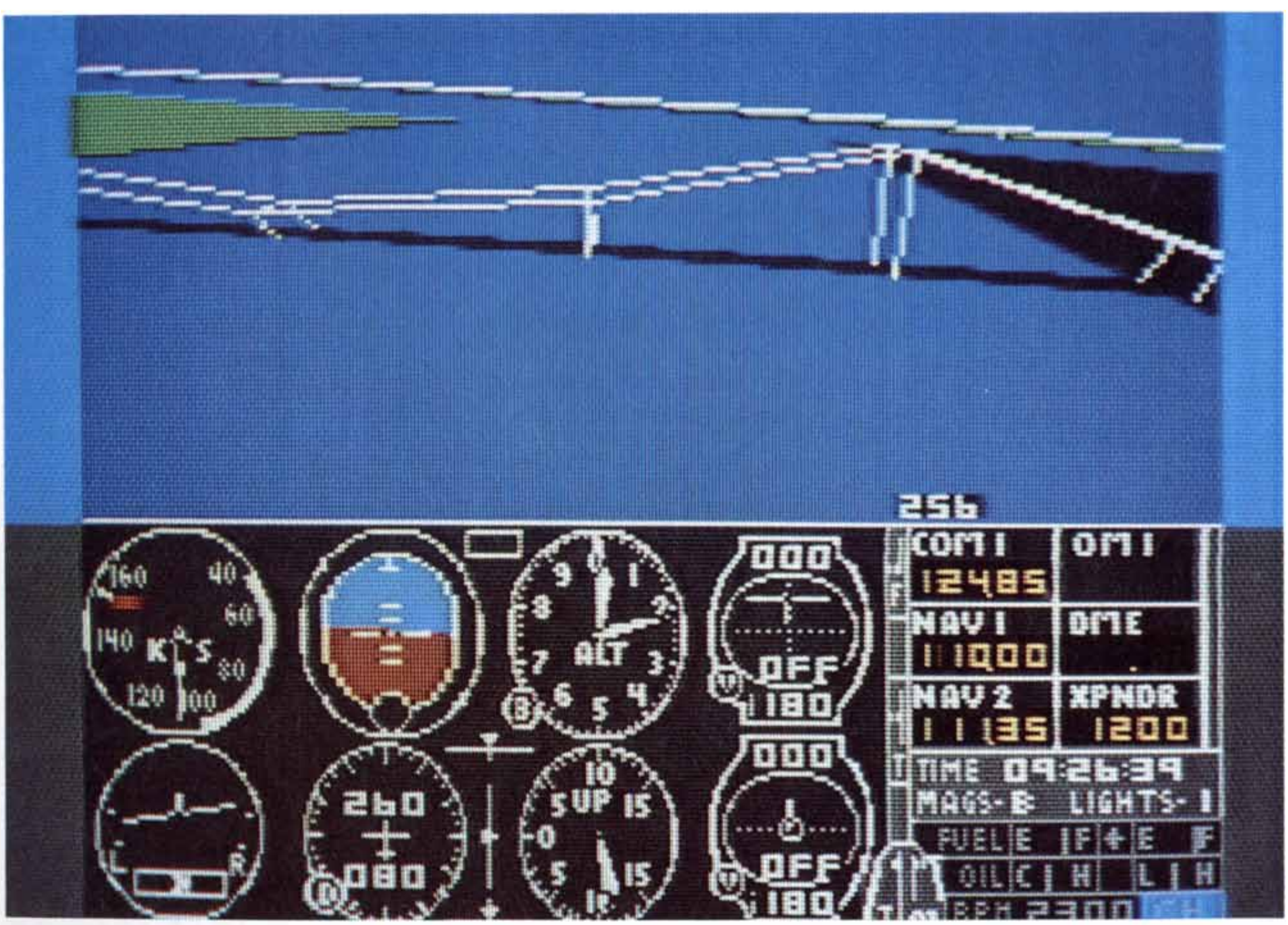
The on-again, off-again scan-oriented drawing technique is based on a fundamental result of plane topology: any polygon divides the plane into two parts, an inside and an outside. Accordingly any line through an edge of the polygon passes from the inside (the surface) to the outside (nonsurface) or vice versa. Most surfaces drawn by FLIGHT SIMULATOR are bounded by fairly simple polygons. For example, soon after I left Meigs Field the shoreline of Lake Michigan extended from below the airplane to a distant horizon. On the way it followed a polygonal embayment partially enclosed by a north-pointing spit of land. The scan lines of blue water in the bay were interrupted by the spit and then continued to fill in the main body of the lake.

The writers of graphic-display programs have often pondered the so-called hidden-line problem: How is a program to draw an opaque, three-dimensional object in such a way that only the side nearest the observer is visible? Some methods require the solution of time-consuming equations, a computational load that FLIGHT SIMULATOR can ill afford. Artwick's solution, largely applicable only to simple landscapes, is to draw the entire object in such a way that the side nearest the observer is done last. Parts of the object not visible from the observer's point of view are simply erased by the parts that are. Representing the twin towers of the World Trade Center in New York affords a pretty example of this crude but effective technique. From every angle the nearest sides of each building obscure the farthest ones and the nearer tower eclipses the farther one.

After my discovery of Chicago's depopulation, I decided to head back to Meigs Field. The plane's real owner might be getting anxious about its whereabouts. Still, there was much I wanted to learn. For one thing, I felt embarrassed to know only a few controls. For another, I had yet to try a truly dramatic stunt. I flew out over Lake Michigan for the attempt.

The ailerons were my favorite controls. Pressing the key for the left aileron made the plane bank to the left; ahead of me the horizon tilted to the right, and the scenery appeared to slide gently down the slope. In accordance with aerodynamic laws the nose of the Cessna turned steadily to the left. Flying straight and level once again, I typed a few 8's and almost physically felt the nose drop. Outside my window the horizon and scenery rose to the top of my view. The plane was diving. The airspeed indicator slid past 150 miles per hour. I tried the rudder controls. The addition key moved the rudder slightly to the right and the plane began to slue badly. I centered the rudder and contemplated the landscape below climbing toward me with deceptive slowness. It was time for my stunt.

Pilots attempt loops far less often than programmers. I knew instinctively, however, that I needed a good deal of airspeed for the maneuver. First down, then up and over. Pressing keys was not my idea of pulling back steadily on the stick, but I dutifully pressed key 2. The plane came roaring out of



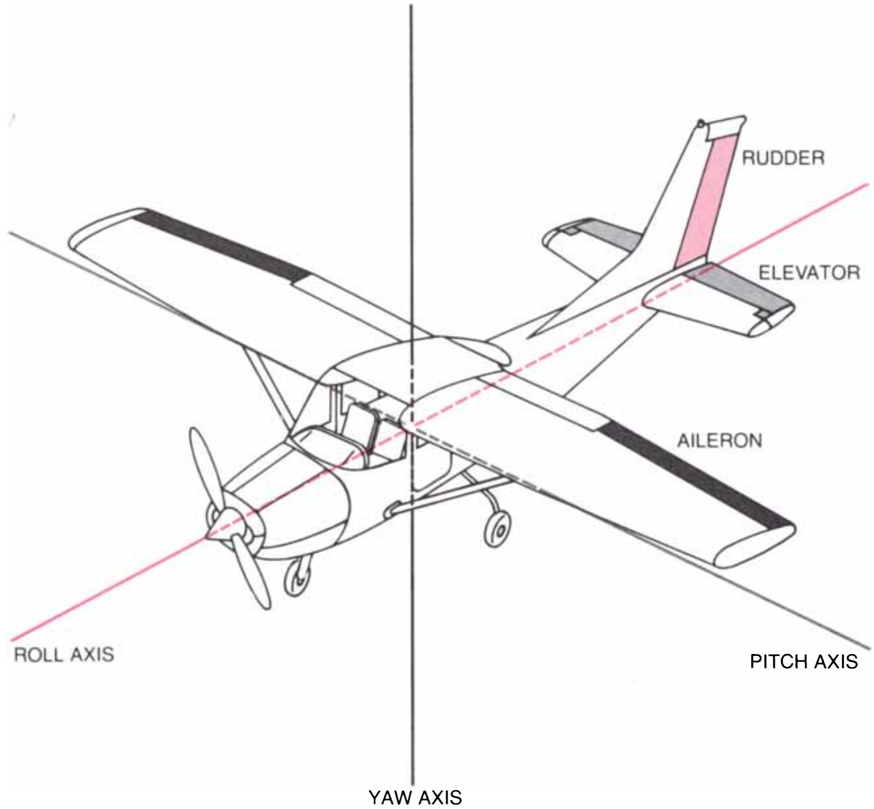
On a death-defying course under a simulation of the Manhattan Bridge its dive into level flight and then up into the empyrean of solid blue rasters. I watched the horizon plunge below my windscreen and then waited. I had not the faintest idea whether I would make it or not. All was silence. Would I stall and plummet to the earth? My instruments were no help; I had only a vague idea what half of them were for.

Suddenly an upside-down landscape swung down into my window, green ground above and blue sky below. Had I made it? I was too confused to answer my own question. Grimly, I held on. Now all was ground ahead of me. Then the horizon rose to its proper level. I tried to drop the nose to prevent another loop. The horizon wavered and then stayed. I had made it! It was becoming clear that I was one of those gifted people who can fly expertly with almost no instruction. Ahead of me Meigs Field popped up into view. I would land and take a shower.

The main runway stretched from south to north. From my position southeast of the field I planned to fly west and then turn north when lined up with the runway. Somehow the plan went awry. I never quite got the runway in line, and I approached the field from too great an altitude. It was like a nightmare that gets steadily worse. I dropped the nose of the plane to lose altitude quickly, while still fighting for a few more yards of easterly movement. But the runway left the screen. I lost my horizon. A warning buzzer sounded as the ground rushed up. Damn! I crashed. I knew I had crashed because a network of cracks covered the windscreen and "Crash" was printed across it. Was I dead? It would be a great pity if the Chicago data base should lose its only living inhabitant.

In order to simulate these adventures FLIGHT SIMULATOR must constantly recompute the appearance of landscape from the Cessna's cockpit. The controls of a plane can be grouped according to the three kinds of rotation they induce [*see illustration on this page*]. The ailerons rotate the plane about a line that extends from nose to tail, the roll axis. The elevators cause a rotation about a line parallel to the wings, the pitch axis. The rudder tends to rotate the plane about its yaw axis, a line that is vertical with respect to the pilot. If the current position of the plane is given by coordinates x, y and z, how does the program determine where it will be a split second later?

The direction of the Cessna's nose and the current airspeed constitute a vector, and the vector gives rise to a new position. The program calculates a new direction for the plane—and consequently a new viewing angle—from the equations of flight. For each display cycle the inputs to the equations include the current values of the controls and environmental conditions such as air pressure and density, wind velocity and temperature. The flight equations incorporate the plane's flight characteristics and apply the laws of physics to the inputs. The forces of thrust, gravity, lift and drag are added, and the inertial effects of the plane's mass and geometry are taken into account. The output of the equations is the new flight direction.



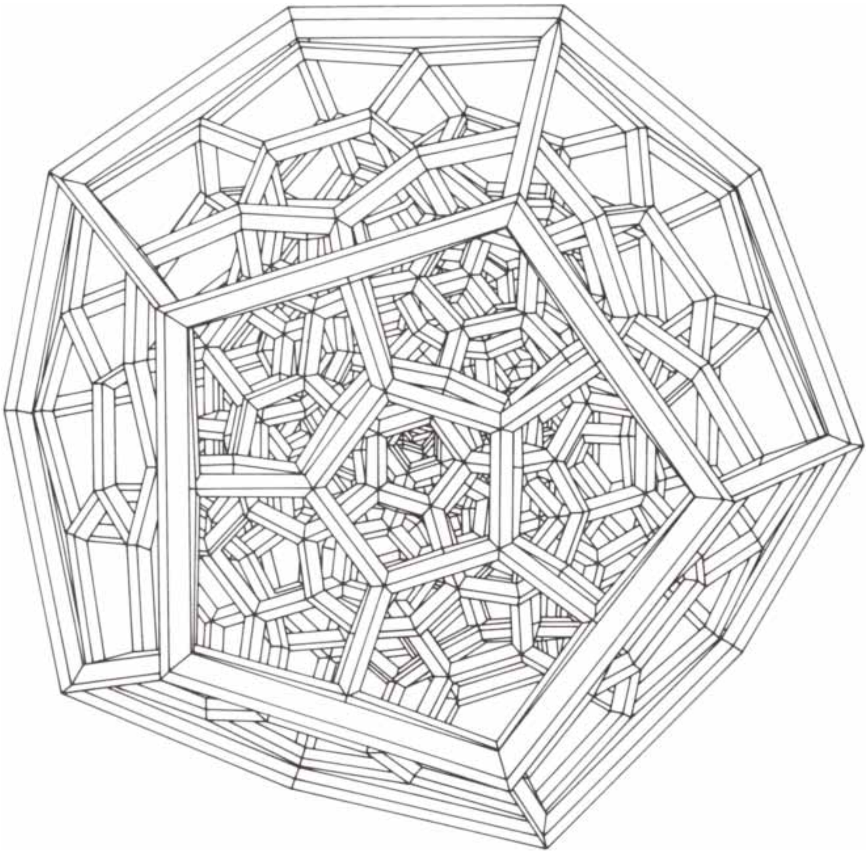
Three kinds of control surface for the rotation of an aircraft

The equations are too complicated to solve quickly, and so Artwick has replaced them with tables of values. For example, one table lists the lift of the plane corresponding to each angle of attack; the angles are given for each increment of a tenth of a degree. There is an additional advantage of tables: flight characteristics can be altered simply by changing the tables; new formulas need not be derived.

Readers thinking of a pilot project in the vertical direction can buy a flight-simulation program for any of the more popular makes of home computer. Artwick has had a guiding hand in the best-known program, sold by the Microsoft Corporation of Bellevue, Wash. He has also played a major role in developing a program called FLIGHT SIMULATOR II, marketed by the Sublogic Corporation of Champaign, Ill. Both programs are available for Apple II, IBM, Commodore 64 and Atari personal computers. A third Artwick production was recently announced for the Macintosh computer.

After a few maiden flights the reader may want to try some of the hair-raising adventures suggested in a new book by Charles Gulick, modestly titled *40 Great Flight Simulator Adventures*. I am grateful to David Wehlau of Waltham, Mass., for calling the book to my attention. Each adventure begins when the user starts up any of the flight-simulation programs mentioned above and types in the necessary data from the book. The data include latitude, longitude, altitude, control positions, time of day and so on.

Instantly you may find yourself a few hundred yards from the Manhattan Bridge, on a course that takes you directly under it. Alternatively, in an adventure titled "Dead Stick off San Clemente," your engine has just quit 5,000 feet above San Clemente Island off the California coast. Can you land on the island in a dead glide? A third adventure is titled "Twilight Zone." You are parked on a grass landing strip. The clock on the instrument panel says three minutes past midnight, but it is light outside. The aircraft, to the extent that it is visible from the side and rear windows, is all black. The takeoff is slow and, once aloft, neither windows nor radar show any land-



Edward B. Becker's projection of the four-dimensional hyperdodecahedron

scape at all. What is to be done? Personally, I prefer a deserted data base.

Core War, the subject of my inaugural column in this magazine (May, 1984), will be fought on the U.S. eastern seaboard later this summer. The Computer Museum in Boston plans to host the first international Core War on the last weekend in August. The cleverer programs will render their weaker colleagues inoperable by destroying vital instructions. The (human) winner will be Core War champion for a year.

For interested competitors a document that defines the standards for the entries and an entry form are available from Mark Clarkson, 8619 Wassall Street, Wichita, Kans. 67210. Clarkson directs the International Core War Society, an organization of some 150 core warriors. Many of them are currently preparing their entries, but it is not too late to order the standards document and write a program of your own. I shall have more to say about the tournament in next month's column.

High-I.Q. programs, the subject of this department in March, led many readers to try a hand at the minitest. A few others even wrote HI Q programs of their own. Paul A. Cook of Sudbury, Mass., among others, wondered how Marcel Feenstra, the originator of HI Q, arrived at a measured I.Q. of 160 for his program. As Feenstra explains the matter, the idea came from an Australian friend, Chris Harding, who holds the Guinness record for the highest identified I.Q. Harding suggested a test of the program based on all the sequence-completion questions appearing in the book *Know Your Own I.Q.*, by Hans J. Eysenck. Feenstra adapted the method of measurement suggested in Eysenck's book.

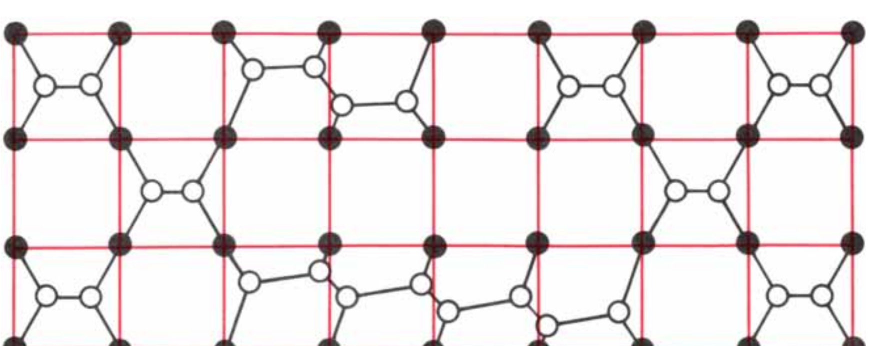
The April Fools' edition of "Computer Recreations" apparently fooled no one. Everyone who wrote a version of HYPERCUBE retained both sanity and corporeality. In a startling case of precognition David Ecklein of Rumney, N.H., developed the program several months before I described it. Ecklein's program develops all the effects that HYPERCUBE does, but it uses only four planes of rotation instead of six. Eric Halsey of Seattle, Wash., was not satisfied with twiddling HYPERCUBE. He went on to other higher-dimensional objects including the four-dimensional cross polytope, the four-dimensional simplex and a special polytope called the 24-cell.

Finally, I must mention Edward B. Becker of Riverhead, N.Y., who has had a close encounter with the hyperdodecahedron [*see top illustration on this page*]. An ordinary dodecahedron is the classical, 12-faceted platonic solid. Becker's hyperdodecahedron is made up of 120 ordinary dodecahedrons constructed from 1,200 lines. The image shown was obtained by projecting the monstrous figure onto three-dimensional space, then shifting each component dodecahedron by a small amount in a specified direction. It happens that each line belongs to three of the component dodecahedrons. Becker writes: "The planes are bounded by the edges in their shifted versus unshifted positions. In this manner I replaced each line...with a kind of Y beam, which better suggests some of the overall structure, at least to my 3D perception."

Last month Martin Gardner challenged readers to find a Steiner tree shorter than 32.095 units that would span a 4-by-9 array. No one yet has improved on the result obtained by Ronald L. Graham and Fan R. K. Chung of AT&T Bell Laboratories [*see illustration at left*]. Its length is

$$7[1 + \sqrt{3}] + [(([3(2 + \sqrt{3})] - 2)^2 + 1]^{1/2}) / 2] + [(([5(2 + \sqrt{3})] - 2)^2 + 1]^{1/2}) / 2] ,$$

or approximately 32.094656....

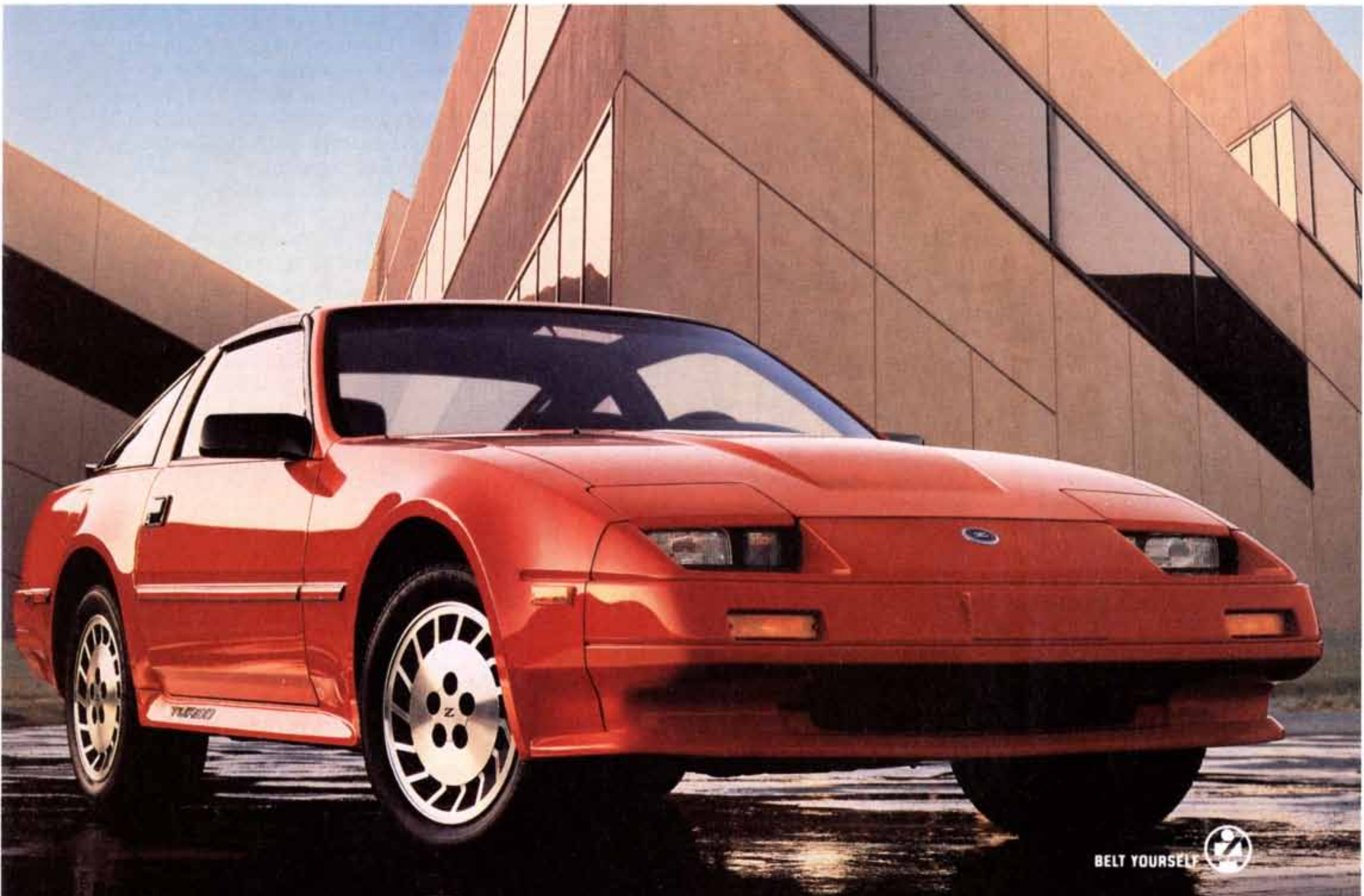


Shortest Steiner tree known for the 4-by-9 rectangle

THE NEW NISSAN 300 ZX

PACE CAR FOR THE PERFORMANCE GENERATION.

NISSAN



Very few automobiles have generated the excitement and emotional involvement of the Z-car. It was not surprising then, to read that Motor Trend Magazine called the 300 ZX, "the best all-around Z-car ever built."

And now, Nissan has taken one more step in the thoughtful evolution of this classic.

At the heart of this Z is one of the most technologically advanced, 3-liter, V-6 engines in the world. Track proven, this sophisticated design includes computerized electronic multi-port fuel injection and high-power, hemi-shaped combustion chambers. The turbo model puts out a rousing 200 horsepower. When you combine the rigid, lightweight powerplant of the 300 ZX with electronically adjustable shocks, the result is startling.

Outside, fender flares, housing wider tires, are integrated into the body. The air dam is extended and rocker panel extensions reduce air turbulence. All this, plus a wider track results in better handling than ever.

Inside, a choice of electronic or analog instrumentation along with every conceivable luxury, including a resounding 80-watt, 6-speaker stereo system.

The 300 ZX, turbo or fuel injected. Once you get inside a Z, a Z will get inside of you.

THE NAME IS NISSAN

BOOKS

Antarctic antiresource, molecular therapeutics, Dr. Dee's Shew-Stone, Sculptor Moore's words

by Philip Morrison

THE SEVENTH CONTINENT: ANTARCTICA IN A RESOURCE AGE, by Deborah Shapley. Resources for the Future, Inc. Distributed by the Johns Hopkins University Press ($35). Fabian von Bellingshausen, a veteran captain in the imperial Russian navy, seated among his officers stiff in their regalia, listened gravely. The two lofty frigates of his expedition rode at anchor among the mists of an icy South Shetlands summer. Between them rode a little sloop, not much bigger than the longboats of the warships. It was the Yankee captain of that sloop who spoke, beardless Nathaniel Palmer, aged 21, master over his mate and four men.

Von Bellingshausen, so far from Kronstad, had been charged that year of 1820 to search out new southern lands for the maritime glory of Czar Alexander and all the Russians. Palmer too was an explorer dispatched in quest. His modest aim was not high discovery; he sought unvisited seal beaches, whereon the fleet of sealers for whom he sailed out of Stonington, Conn., could club more little southern fur seals, a most lucrative prey while they lasted.

Dressed in sealskins, Captain Nat told his hosts how earlier that season he had caught sight of and then charted the long continental shore only 50 miles south of their island anchorage. His logbook confirmed his words. "What shall I say to my master?" cried the forestalled Russian commander.

The tale (the three anchored ships are shown in an old print) holds more than such stirring folklore. Since Captain Cook's long ocean circuit of the 1770's, geographers had come to accept the inference he first drew: south beyond the ubiquitous floes of ice in rough seas there lay an icy island continent of little worth. Beginning in about 1820 American, British and Norwegian sealers had pressed south along the coastline of South America. Then they sailed to the offshore Antarctic islands, also thronged with seal; we simply do not know who was the first among many hardy captains to make out the peninsular mainland nearby, past its guardians of ice and fog. There are British claims; one logbook out of Nantucket records an actual Antarctic landing in 1821; there are even those who maintain that von Bellingshausen himself was not behind Palmer but ahead of him.

In this single tale of Antarctic discovery are hinted the chief themes of almost two centuries of history. They are explored in a penetrating work by Deborah Shapley, an intelligent investigative journalist familiar with contemporary science and its politics. Hers is an intricate political tapestry woven from national pride and rivalry, glittering but transient sources of wealth, the ways of military and civil authority, and a pervasive sense of daring, adventure and discovery. Error and incomplete knowledge often hamper the principal figures. Underlying the pattern are both the impersonal findings of science and the personal relations existing within the community of science.

That lonely seventh continent is distinct. It accounts for a fifteenth of the land surface of the earth (conceded, the ground there is mostly buried under thick ice). Yet today's Antarctic population reaches a peak of only 2,000 under the summer sun and dwindles to a far smaller number during winter night. At most only a couple of human beings have been born there, and a major fraction of those who have died on the continent were passengers on a single tourist flight that crashed into Mount Erebus in 1979.

This singular land is governed in a singular way. Since 1961 the inhabitants of Antarctica have lived under a brief document drawn up by diplomats from a dozen countries, who met biweekly in a board room of the National Academy of Sciences in Washington for a year or two before the treaty went into effect.

The treaty binds its signatories to a rational regime that has about it the openness of the Enlightenment, before the nation-states had seen the power they could derive from science-based technologies. It holds that the Antarctic can be used for peaceable purposes only: no military bases, forts or weapons tests are allowed; military personnel and equipment engaged in peaceful work are acceptable. Nuclear explosions and nuclear-waste disposal are forbidden. Research is to be open and cooperative and its results made available to all. Advance notice is to be given of all stations and expeditions, and of the introduction of military personnel and equipment; the right of inspection from the air at any time and place, along with free access on the ground for designated observers, is assured. Any disputes are to be settled by agreement if possible, otherwise the International Court is to intervene.

In 1991, by its own terms, that treaty is expected to undergo review by the states that are parties to it. Those dozen nations were the original ratifiers. By now four more countries have anted up the requisite investment for participation: "substantial scientific research activity" in the Antarctic. The total includes four of the Big Five, all explorers long active in the south for one reason or another. (The fifth, the People's Republic of China, has no record of past interest, but it has indicated its intention of joining soon, and a few Chinese scientists are now at work as visitors to stations on the ice.) The five southernmost countries, closest to the ice, are motivated, as one might expect, by practical and even territorial interests. The other countries include five more explorers from the heroic past, Belgium, Japan, Norway, Poland and West Germany, and two big nations whose scientists have worked on the ice only during the past few years, India and Brazil.

The treaty contains a key provision that has defused the disputes of a long history. All territorial claims are simply finessed; the signatories agree that no claim is to be diminished or furthered by any action under the treaty. Certainly problems might arise here: treaty breakdown, mineral rights.... But so nice a political balance has been struck that the meager benefits of sovereignty over so frozen a land have not yet moved any state to act against the treaty. Even brief war in the Falklands between two treaty states did not disturb their correct treaty relations.

Interest is the nub. Brevity, the innocent good will of the scientists, even textual ambiguity, although they have their uses, are not the main points. Nations calculate what they stand to gain or lose. Only today the game appears


GLENLIVET
WELL, THE GLENLIVET Scotch whisky does cost around $20.00. Which some say is a small price to pay for a Scotch which has been made in the same unique way since 1747. A 12-year-old single malt Scotch with a smoothness and unique character that is unsurpassed to this day. All of which could explain why people are so strangely possessive about The Glenlivet. Which is a pity. You might just have to buy a bottle of your own.
THE GLENLIVET. JUST SLIGHTLY OUT OF REACH.

to be gaining newer and richer stakes.

Two potentially large resources are hinted in the Antarctic. The million whales taken in its waters during the 20th century drew whalemen by the tens of thousands who worked there offshore and on the islands. Only a remnant of the whales remain. But their primary fodder, the krill, thrive. The take of those shrimplike creatures, whose red-brown swarms staining the cold green waters have long fed whale, seal, squid and seabird, has risen to half a million tons a year, a haul equivalent to the biomass of 10,000 whales. There is an old dream that protein from krill might one day even become a major foodstuff for a hungry world; that would imply growth yet to come in the krill fishery of more than an order of magnitude.

The treaty structure has managed to meet the intensifying challenge of economic interest by indirection. Its research-centered provisions were not relevant to such a scale of activity; instead the parties to the treaty devised a new agreement, the Convention on the Conservation of Antarctic Marine Living Resources; it took effect in 1982. A legal "umbilical cord" ties the convention to the original treaty. The new agreement is open wide to states not previously involved. (The EEC has signed on, along with both Germanys.) It provides for what appears to be a very farseeing management of all marine resources by a new commission sitting in Hobart, Tasmania, and by its scientific arm, charged with the task of coming to understand the unique ecosystem of the circumpolar ocean.

Minerals around and under the ice are still only of potential value, not likely to be realized on a large scale in this century. Yet the first deep-sea drillship did find traces of hydrocarbons beneath the Ross Sea, and several treaty states are focusing geophysical operations there. Would a bonanza revive old claims to newly oily territory?

Hard-rock mining is further off, although the geologists who match plate boundaries mutter about a vast new Rand hidden somewhere under the ice. Economics will in the end determine, but the calculus will be of tomorrow's return on investment, not our own.

It may well turn out that the most important geophysical value is a negative one, a kind of antiresource; it is the effect the ice can have on weather and on sea level during climatic change. The solution of that equation may mean more to humankind than krill, petroleum or diamonds. Close monitoring of changes and, even more important, new insight into their cause and course are likely to be the highest prizes won on the Antarctic ice.

There were no Americans on the southern ice during the heroic age of Robert F. Scott and Ernest Shackleton before World War I. Our time came when in 1929 Richard E. Byrd flew in his trimotor over the South Pole. His energy and charm moved the distant Antarctic into the spotlight back home. Who did not know his base, called Little America? Who did not become aware of eager Boy Scout explorers, ham radio operators, comical penguins, wealthy sponsors? Behind those engaging scenes there were diplomatic and strategic plans, covert before World War II, more visible thereafter, to stake U.S. claim to much of the area, or at the least to deny such claims made by any adversary.

The Antarctic Treaty that now rules was no outcome of years of diplomatic calculation. It sprang from a dinner party for scientific guests that James Van Allen—he of the famous Van Allen belt—held in 1950 at Silver Spring, Md. There the idea grew of steering the Third Polar Year of 1957 toward the Antarctic, where new scientific methods promised new geophysical results. The scheme soon blossomed into worldwide enthusiasm for an International Geophysical Year, deliberately setting hopeful cooperation within science against the ominous divisiveness of the Cold War.

The I.G.Y. was a hit around the world. Although many of its organizers did not suspect it, the I.G.Y. plans fitted well enough into a classified policy of the National Security Council to play out a rather modest American hand in that southern continent. The improvising scientists seemed to President Eisenhower an open, hopeful and much cheaper alternative to the expensive stratagems of the diplomats and the armed services. The I.G.Y. got what it needed from the U.S., including generous logistic support, and remained a bargain by bureaucratic standards.

As the I.G.Y. began, Navy Seabees were building red structures around a barber pole marking the South Pole itself. We are there still at the symbolic heart of the ice cap, the permanent station to this day supplied exclusively by military air transport. But our policy has shifted with that of all 16 nations on the ice (and 15 more that are less than full partners in the treaty) into "an amalgam of nationalistic elements subordinated to international cooperation in the name of science." It is likely that another shift toward applied science, a wider environmental ecology at sea and onshore, with geophysical exploration inland, is at hand, still under the improvised regime of the philosophers.

The objections to the treaty regime by the some 100 nations that are in no position to occupy a station on the ice seem to have been met in part by the evolution of the treaty group from a self-chosen club of the powerful into a majority of humankind. The participants should become a little more generous to outsiders. The I.G.Y. spirit, never quite as lofty or disinterested in reality as it seemed in the abstract, has proved a hopeful basis for working out one solution, if a local one, to the daunting riddle of national rivalry. We can look to the year 1991 with modest satisfaction.

When the U.S. began construction of its South Pole station in 1957, no one was certain that the snow would bear the load of the aircraft ski. A 45-year-old photograph posed by Scott's doomed party fixed the hardness of the snow at the pole itself by showing just how far their boots had sunk in. The future often puts to good use what the past did not quite intend.

MOLECULAR BIOLOGY AND HUMAN DISEASE, edited by Alexander MacLeod and Karol Sikora. Blackwell Scientific Publications (paperbound, $17). A dozen molecular biologists and physicians around the laboratories and clinics of the Medical Research Council in Cambridge, England, with help from some of their expert friends in London, Edinburgh and Baltimore, join here to survey the state of the molecular approach to disease. Their story is brief but full, aimed at a readership of students and of physicians in practice. It is the firsthand expertise of these redoubtable authors that distinguishes the fine text, although it cannot compete with the pages of this magazine for visual support. There are some valuable illustrations, but they are few.

The first half-dozen articles lay out the new technologies that enable us to examine and then to edit the genetic message, down to the level of errors in spelling. The first task is to publish a gene—a molecular rarity, parts per billion by weight—in an edition large enough for study. That is cloning. The methods are of course biological: enzymatic selection from the DNA text, followed by cellular multiplication and then recognition on fusing with the complementary strand fragment, itself prepared from natural templates. The text must next be read. This is done through letter-by-letter sequencing, still difficult; reading the full sequence of the Epstein-Barr virus, some 250 kilobases long, is the greatest achievement of this technique today. The blotting techniques, which combine spatial separation of pieces of the

DNA by electrostatically driven diffusion with recognition by specific partial markers, facilitate the task.

One chapter summarizes the remarkable processes leading from gene to protein, and another is a revealing essay on the cytoskeleton, that complex dynamic framework, based on three distinct kinds of protein microfilaments, that supports the cell in life and in motion. The cytoskeleton's most classical display occurs during the chromosome ballet in mitosis. (This chapter has, surprisingly, not one illustration.)

The powerful technique of recognition by monoclonal antibodies is well described; particular *B* lymphocytes, each producing a single antibody, can now be propagated indefinitely. The lymphocyte is usually fused with a cell from a malignant line; the pathological propensity for indefinite self-replication copiously amplifies production of the particular antibody.

The physicists have added to this biological armory. The technique here is flow cytometry. Down a fine capillary tube cells flow swiftly one by one, a few thousand per second. For simple counting, laser light can be scattered off the cell stream. Deeper characterizations at similar speeds can be made by first marking each cell in some specific way with a fluorescent dyestuff. Optical pulse-height sorting and computer display are by now familiar in many laboratories. A million cells can be sorted in a few minutes by their content of as few as 50 to 100 specific molecules per cell. Cells have been sorted according to the ratios of the base pairs their chromosome DNA contains, double-dyeing them with two colors of base-specific dyes and exciting the two pulses of color with appropriate lasers. That scheme already works for half of the human chromosomes.

What can these techniques do for medicine? Somewhat more than half of the book outlines six or eight groups of molecular diagnoses and therapies—essentially all for genetic diseases—now in clinical service at some level. The most mature is our remarkable understanding of the wide variety of molecular lesions found in the red protein hemoglobin. A widespread class of genetic disorders, the thalassemias, spans the clinical spectrum from lifelong carrier states without any symptoms to fetal death. A few different mutations in a cluster of genes can account for this phenotypic diversity. One form, which lacks one of the globin chains that join to make the working tetrameric molecule, is a common cause of fetal death in Southeast Asia. Only 20 percent of the infant's hemoglobin is functional; "it is surprising that these infants survive to term." Prenatal diagnosis, for this condition as well as more generally, is in wide use; sampled cells can be probed for faulty DNA sequences.

An account of oncogenes, first steps to cancer, is presented, along with a discussion of the opportunities for potential therapy that this knowledge implies. The forecast is for a full-blown course of treatment in 20 years or so, not before.

There are known molecular disorders of muscular tissue, some in the mitochondria, some in specific neuroreceptors in the muscle-nerve interface. Hope is offered for eventual control of DNA repair processes; it may even be possible to restore genes damaged by ionizing radiation.

Two brief papers treat genetic fixes to nonhuman DNA that could have medical importance. One paper outlines efforts to generate a strain of wheat safe for those who cannot tolerate one of the gluten proteins. The other discusses the production of more than a dozen classes of complex, therapeutically interesting molecules by engineering the genetic implantation and expression of new biochemical capabilities in one or another microorganism. This chapter reads a little like the business columns of a newspaper; the market has long admired the pharmaceutical potential envisioned here.

Insulin synthesis is a case worth citing. Pig and cow have been the only practical sources of the hormone. The distressing immune response of some patients to porcine insulin has promised a fine reception for a synthesized molecule with exactly the amino acid sequence of human hormone. "Unfortunately for the genetic engineer" pig and human insulin differ by only one amino acid. It may be cheaper and simpler to snip off the 30th link in the *B* chain than to genetically engineer a bacterial product.

Genetic disease is fascinating; the side-by-side discussion of many approaches in this book makes a strong impression on the reader, a kind of review of molecular biology with therapy as exemplar. It is clear that much of the impact must be made before birth. A couple of percent of human beings carry some kind of "potentially lethal inherited blood diseases." Therapy is then in fact the termination of pregnancy, with all its social and ethical problems. Manipulating genes in embryos is possible in vitro, but it too is surrounded by profound problems in ethics. Adult gene therapy is surely the most difficult technically; how can the involved tissues be replaced? The only disorders that show promise are some of those of the blood. The bone marrow is accessible for manipulable samples; what remains can be eradicated easily. It is not too much to imagine laundering the bone marrow of all stem cells, say, with a given defect. For success the genetic controls must accompany the pure protein information. That may demand "the manipulation of some very sophisticated genetic switches."

A reader comes away from this rewarding and difficult book with deep admiration for the evident ingenuity of the workers, but one also feels that the major achievements of biomolecular engineering may already be among us in the form of those symbiotic organisms we call domesticated plants and animals, the foundation of civil society for 10,000 years.

THE MIRROR AND MAN, by Benjamin Goldberg. University Press of Virginia ($20). What an intelligence scandal it must have been! First, the dozen foreign experts who had been cleverly cajoled and then smuggled into France found themselves confronted by pleading letters from their wives back home, urgent letters that turned out to be forgeries. The king himself was induced to pay the anxious little group a visit, bringing presents and praise. (In the same manner, Jack Kennedy would one day come to Huntsville.) Then some of the lonely wives were brought over as well. The stakes grew. The expatriates' factory was attacked by armed men. That pitched battle ended only on the intervention of a company of soldiers. Later a couple of the best workers were even poisoned.

Thus was the long-secret process first winkled out of its home in Venice to be set up for the Sun King by an enterprising subject, Jean-Baptiste Colbert, in the Faubourg St.-Antoine. The secret was not the design of some weapons system but the technique for the manufacture of fine glass mirrors. These were mirrors not far from modern quality; they could be made in any size for which there was crystal-clear glass sheet, a key material developed two centuries earlier in Venice. The reflecting surface was a later Venetian invention, a mercury-tin amalgam applied to suitable glass at room temperature to form a coating that is bright, adhering and durable. Commonplace mirrors of good quality are still made in that way, although vacuum-evaporated aluminum has ended almost four centuries of tin-amalgam monopoly.

The French proved able journeymen; within five years the much-needed Venetian experts went home with gifts and a full pardon from their Republic. By then the new shop in Paris could already meet the mushrooming national needs, easily grasped on a visit to the endless mirrored galleries of Versailles. Within 20 years the French artisans had perfected the manufacture of superb mirrors from plate glass, which they cast into flat sheets while the glass was fluid instead of shaping it from large heat-softened blown cylinders. The Compagnie de Saint-Gobain, a world leader today in the glass industry, is the direct descendant of Jean-Baptiste Colbert's secret enterprise, moved from Paris to the castle town for which it is now named.

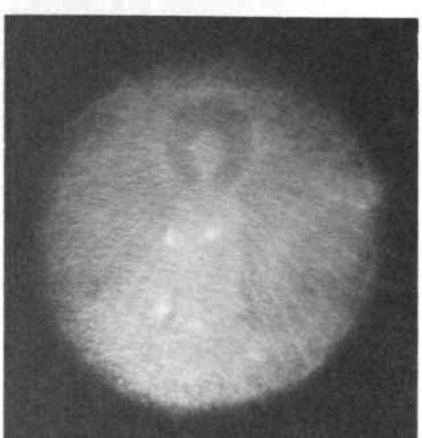
That tale out of the history of technology is taken from one delightful chapter that divides into two parts this monograph by a veteran optical physicist. The latter third of his book briefly surveys with diagrams and data (but not mathematics) the nature and development of the outward-looking mirrors of modern science and technology, those of the multiple-mirror telescope, the 60-inch shopping-mall searchlight, the lighthouse, the space probe, the laser and the 20 acres of computer-steered solar-power mirrors in the desert pilot plant at Barstow.

Benjamin Goldberg has understood how ancient and important is the device he treats; human beings have been fascinated by the image in the mirror for a very long time. He documents with sympathy and erudition the long biography of the inward-looking mirror, that in which we see our own double, the visible hint of the soul that must one day depart the body. He follows this theme from Narcissus through Egypt, India, China, Japan and pre-Columbian America. He expands on European mirror thought and practice from early times to the growth of Renaissance perspective theory under Alberti, Brunelleschi and Leonardo. There are good discussions of the mirrors that the Dutch masters employed as tools and represented in their paintings, and of the mirror-decorated rooms of the well-to-do clients of Thomas Chippendale and Robert Adams.

We possess 2,000 bronze Etruscan hand mirrors from as long as 2,500 years ago, as old as any mirrors in Europe. The Greeks used mirrors freely too, although we have fewer from them. The engraved backs of Etruscan mirrors mostly depict Greek myths; Hercules was a favorite. Those mirrors were used mainly by women, and they are found associated solely with female burials. The oldest glass mirror in Europe is Roman, plainly made on an Egyptian model, dark and distorting when compared with the fine white-bronze mirrors of later Rome; glass mirrors do not reappear until the 13th century.

Chinese cast-bronze mirrors were in vigorous development for 20 centuries, from the Shang to the Tang. The backs of those mirrors are themselves a kind of mirror of both the cosmological and the domestic concerns that occupied Chinese culture over the dynasties. A poignant poem by Li Po is cited. In it a woman polishes again and again with her red sleeve "the bright moon, / ...its splendor lighting up everything. / In its center is my reflection...." That moon is the mirror that her long-absent husband, on indefinite duty at some border, had given her. There is a brief explanation of the images cast by certain "magic" concave mirrors that reflect onto a wall the relief pattern seen cast in the mirror back. A photograph shows the effect as it is now exhibited in the Shanghai Museum.

In Japan the mirror was revered; the oldest ones are preserved in Ise as sacred attributes of the ancestral Shinto sun-goddess herself. The three artisans who create bronze mirrors—designer, molder and polisher—were until the past century considered an official part of the imperial court. An ingenious composite magic mirror that reflects a hidden design of the Buddha is shown producing its visible effect. Once glassmaking finally arrived in Japan the three bronze mirror makers had to seek other trades.



Projected reflection of Buddha (left) *and X ray of mirror, showing concealed relief* (right)

Mosaic mirrors made up of polished pieces of iron oxides and iron pyrites have been known from Olmec ruins in Mexico for almost 50 years. They are concave and well polished still. Too small to ignite fire from sunlight, they could in the tropical sun certainly raise smoke from dry tinder. They may be the source of the theme of smoking mirrors that is so widespread in Mesoamerica.

In the British Museum there can now be seen the very black mirror, or Shew-Stone, in which its owner, the scholar-sorcerer Dr. John Dee, did often scry for hints and glimpses of the spirits of the dead. In 1575 he noted that his patron Elizabeth the Queen had "willed me to fetch my glass so famous...." Dee's stone is a polished mirror of obsidian from Mexico, no doubt brought from the New World to Europe during the generation of the conquest of Mexico; the Aztecs there were reputed to be great scryers. The iron minerals are not hard to work and to polish with sand and rouge; obsidian makes sterner demands on lapidary skill, but its hardness and luster are prized. The skillful Aztec craftsmen needed patience as well.

The oldest-known manmade mirror is of selenite, crystalline gypsum, with a few traces of what was once a wood frame. It was found in the predynastic graves of El-Badari in middle Egypt, datable before 4500 B.C. Earlier than that reflections of self were surely seen in still water held in hand, cup or pool. Goldberg's final chapter returns to the issue of self-image. There is an account of contemporary studies of the subject that presents both psychological experiments and the observations of artists and writers from Lewis Carroll to Robert Coles.

No summary of the behavior of animals toward reflected images is given, but not much else that is relevant seems missing from this graceful and yet compact study. The lists of sources given chapter by chapter are dangerously inviting.

HENRY MOORE: MY IDEAS, INSPIRATION AND LIFE AS AN ARTIST, by Henry Moore and John Hedgecoe. Chronicle Books ($22.50). Intimate, unpretentious and beautiful, this book offers Henry Moore's direct reflection on the springs and flow of his art over nearly 60 years. He is 88 now, still drawing, although his arthritic joints can no longer support his lifelong campaign to extract from matter "the full three-dimensional world with air around it."

The recorder and editor of his narrative is his longtime friend and teaching colleague, whose apt and brilliant photographs document here much of what was around the artist, from the heaps of mine waste "like pyramids, like mountains" where the sculptor played as a boy, through the mean streets of his youth, to his bone-strewn studios and his models, on to the triumphant sculpture. Besides these sympathetic illustrations, the volume presents a systematic selection of smaller photographs of about 50 of Moore's works arranged to bring out his development of a few recurrent themes. There are also a number of photographs made by Moore himself, landscapes, textures, masks and sculpture of the past, and his own drawings.

"I have always been excited about natural strata and the actual forms of stone.... I like ruins, the Parthenon, for instance. Now that light passes through it, it is far more sculptural than if it were all filled in.... A sculptor wants to know what a thing is like from on top and from beneath—a bird's eye view and a worm's eye view. It's infinite.

"Brancusi's dictum of truth to material...had a big influence on me... a thing in stone should look different from a thing in bronze.... Later, though, this became a restriction....it is, of course, the vision behind a work that matters most, not the material.

"I liked carving stone.... But, after a time, I began to realize that it was preventing me from including the full three-dimensional world with air around it.... I was trying to make the sculpture as fully realized as nature, not just two reliefs. For me, this was a revelation....

"I used to collect pebbles and flints from the beach—the back of the car used to sag with the weight of them all.... I regarded them as nature's sculpture. Nobody is sure how flintstones came about. I think some were formed by a natural casting process.... Nature produces the most amazing variety of shapes, patterns, and rhythms. What we see with our eyes has been added to by the use of the camera, the telescope and the microscope...enlarg[ing] the sculptor's vision. But merely to copy nature is no better than copying anything else."

If today the sculptors dare to explore motion, stress and change, the work of Moore seems a celebration of the substantial, of matter itself, seat alike of art and of science.

JOJOBA SHRUBS grow on land owned by Agrifuture, Inc., in California. Jojoba (*Simmondsia chinensis*) is a desert plant valued for its seeds, which yield an oil that can replace sperm-whale oil (now banned by the U.S.) in a number of commercial applications.

Established 1845 SCIENTIFIC AMERICAN July 1986 Volume 255 Number 1

Potential New Crops

Buffalo gourd, crambe, jojoba and kenaf are approaching commercial production. A number of other plants show promise as sources of food and as industrial materials

by C. Wiley Hinman

Wheat, corn, soybeans and other crops pour out of the U.S. agricultural system at such a prodigious rate that one tends to overlook the economic potential of several plants not now grown on farms. Yet there are good reasons to look for such plants, and over the past decade a few workers in government, university and industrial laboratories have been doing so. They have come up with some surprises: plants that yield not only food and fiber but also an amazing variety of commodities with applications in industry. The commodities include a long-chain fatty acid useful as a lubricant in rolling steel; medium-chain fatty acids for making soaps, detergents, surfactants and lubricants; an oil that can substitute for sperm-whale oil in cosmetics and that is a good lubricant as well, and organic compounds that can serve as feedstock in making hydrocarbon fuels and rubber.

Some of the main reasons for seeking new crop varieties were cited in a report to the National Science Foundation in 1978 by Soil and Land Use Technology, Inc., of Columbia, Md. One reason, according to the report, is that the genetic pool of existing agricultural crops might become too shallow, leaving the plants vulnerable to disease or environmental shock, unless it is enriched and diversified by new genetic material from wild plants. Another is that "present crops are either very demanding environmentally," meaning they require more water or nutrients than nature provides, "or narrowly adapted ecologically," meaning they have limited tolerance to temperature or soil conditions beyond a narrow range. Third, they require "high energy inputs in the form of fuel, fertilizers, pesticides, processing and irrigation." Finally, farmers locked into the production of a crop such as wheat would be less vulnerable to price instabilities if they could turn to alternative crops.

So far the search for new crops has turned up four that are within reach of commercialization: buffalo gourd, crambe, jojoba and kenaf. Several other plants that are under study offer exciting possibilities for eventual commercialization and the solution of agricultural and economic problems.

The buffalo gourd (*Cucurbita foetidissima*) is a wild member of the genus that includes cultivated squash and pumpkins. It is indigenous to the arid and semiarid regions of North America. The species reproduces asexually as a perennial vine yielding a large crop of seeds valuable in various ways and exceedingly large roots rich in high-grade starch.

Although the vine is killed by temperatures below freezing, the gourd's roots survive even when the temperature is as low as −25 degrees Celsius (−13 degrees Fahrenheit), particularly when an insulating layer of snow covers the ground. These properties indicate that the buffalo gourd could be grown successfully in the entire region overlying the Ogallala aquifer, an underground reservoir extending from South Dakota and Missouri to Texas and New Mexico.

All parts of the plant can be put to use. The vines, which have a high content of protein and are readily digested, can serve as a forage. The roots, which reach a weight of nearly 40 kilograms in only three or four growing seasons, contain a starch similar to the starch from the cassava root. It can be transformed into the sweetening agents dextrin, maltose and glucose. The seeds are rich in oil and protein. Moreover, the fruit that contains the seeds lends itself to mechanical harvesting. The fatty acid composition of the oil (65.3 percent lineolic acid, 23 percent oleic, 6.13 percent palmitic and 2.2 percent steric) makes it excellent for human consumption. (A fatty acid contains tne COOH group of the elements carbon, oxygen and hydrogen, usually in straight chains with even numbers of carbon atoms.) The meal remaining after the oil has been extracted contains about equal amounts of protein and fiber and is suitable as feedstuff for animals.

Scientific examination of the buffalo gourd began in the U.S. Department of Agriculture in the 1940's and is continuing today in the U.S., Mexico, Australia and Lebanon. Studies of the plant and its domestication have been part of a research project at the University of Arizona since 1963. On the basis of these studies it appears that seed yields of two tons per hectare are possible. (A hectare is 2.47 acres.) In comparison, the yield of soybeans is about 3.5 tons per hectare and of sunflower seeds about 1.5 tons. Less information is available on the yield of buffalo-gourd roots, but current data point to the likelihood of an annual harvest of eight tons or more per hectare. Little information is available on the yield of vines.

Buffalo gourd may also turn out to be a source of liquid fuels. Preliminary investigations indicate that it is capable of good annual yields of oil that can serve as diesel fuel and as a source of ethanol. Studies are continuing at the Energy Institute of New Mexico State University, focusing in particular on whether the plant yields more oil as an annual or as a perennial and on how well the gourd grows when the water available to it varies in salinity.

More research, both systematic and well financed, will be needed to establish the conditions that will make the gourd a commercial crop. The best results would probably be achieved if a single company coordinated efforts to grow buffalo gourd as a commercial crop and to manage the development of its various products.

Crambe (*Crambe abyssinica*) holds promise as a source of erucic acid, a long-chain fatty acid that serves commercially in the production of lubricants and plasticizers. The traditional source of this major commodity was imported rapeseed oil. In recent years this source has become increasingly scarce as rapeseed growers have switched to varieties that give higher yields but contain less erucic acid. The Agriculture Department, in its new-crop program, has stressed the need to find alternate crops that could serve as a domestic source of erucic acid and other long-chain fatty acids.

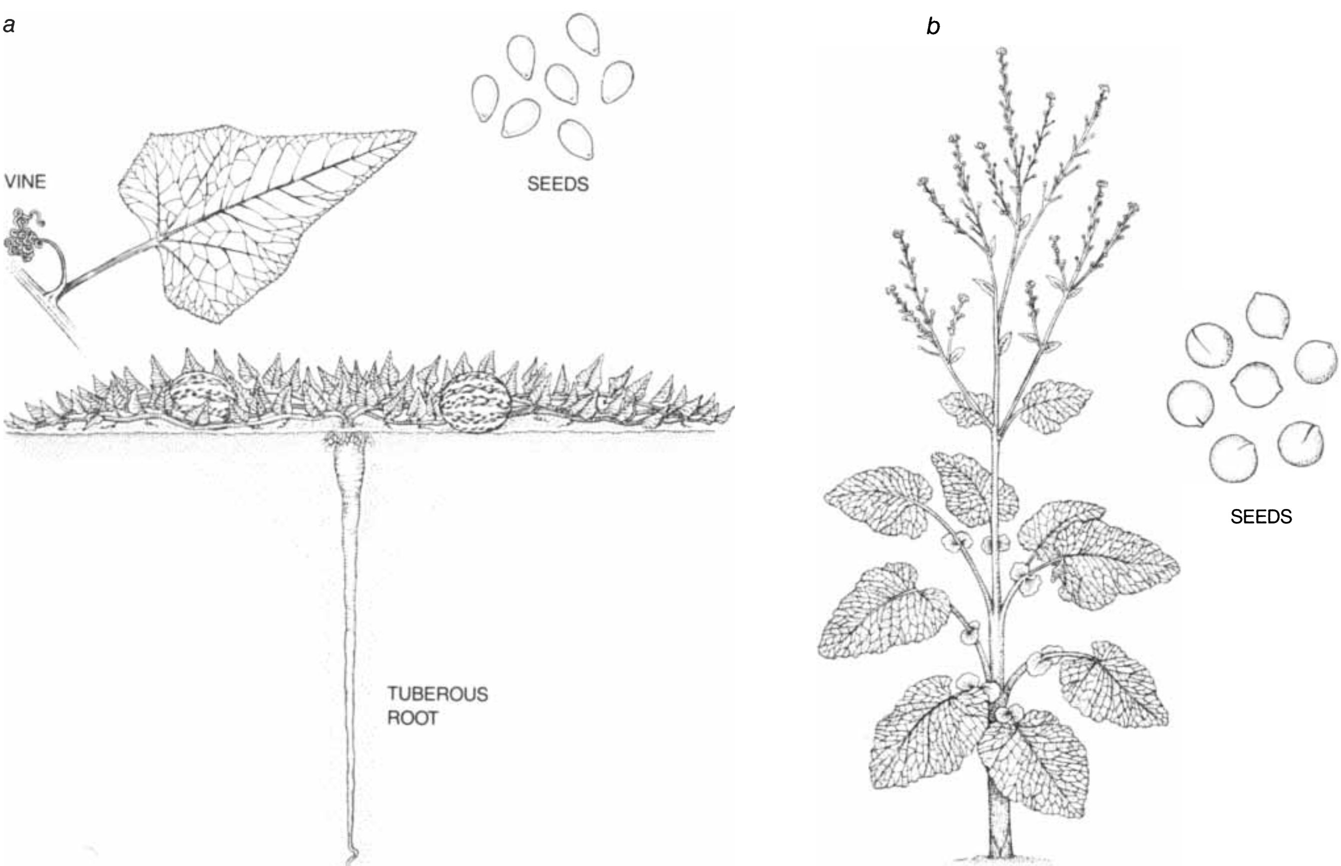
Crambe is a member of the Cruciferae family (also called Brassicaceae), which includes the mustards, cabbages and rapes. *C. abyssinica,* grown originally in countries around the Mediterranean Sea, is a leading candidate to replace rapeseed. Indeed, whereas rapeseed oils contain from 40 to 50 percent erucic acid, crambe oil contains as much as 60 percent.

The plant offers the advantage of a short growing season (90 to 100 days). Preliminary trials have shown that it grows from North Dakota to Texas and from Connecticut to California. In each area it yields economically attractive quantities of oil. Collections of the plant's germ plasm, or hereditary material, were made at Purdue University, and extended study of the material led to the development of improved seed lines. Three cultivars, or strains, have been registered with the Agriculture Department.

Early field trials yielded enough seeds to enable workers at the Agricultural Research Service's Northern Regional Research Center (NRRC) in Peoria, Ill., to do chemical studies. The workers also explored methods of processing aimed at obtaining high-quality oil and nutritious meal from the seeds, evaluated the oil as a substitute for rapeseed oil and even developed new products.

One problem with crambe as a livestock feed is that, like other members of the Cruciferae family, the plants contain glucosinolates, often concentrated in the seeds. These compounds are toxic to humans and animals. Before the effort to develop crambe it was rare to include meal from seeds containing glucosinolates in feeds for animals. Crambe meal could not be fed to animals without the approval of the Food and Drug Administration.

To gain that approval it was necessary to show that feeding a crambe-meal ration to beef cattle for up to six months was as beneficial as feeding them soybean meal in terms of weight gain and of edible products containing no more than one part per million of glucosinolates or compounds known to result from the chemical breakdown of glucosinolates. Tests over a period of several years showed that crambe-meal rations constituting up to 15 percent of the total diet caused no serious problems in beef cattle. The FDA thereupon granted clearance for the use of the meal at a maximum of 4.2 percent of the diet.



FOUR PROMISING PLANTS for commercial development are the buffalo gourd (*a*), crambe (*b*), jojoba (*c*) and kenaf (*d*). The buffalo gourd (*Cucurbita foetidissima*), in the squash and pumpkin family, yields seeds rich in oil and protein, vines with a high protein content and roots with starch similar to the starch of the cassava root. Crambe (*Crambe abyssinica*) is a source of erucic acid, which

The NRRC study compared crambe oil with rapeseed oil in commercial applications. Crambe oil was found to be satisfactory in every case and superior to rapeseed oil in several applications. There appears to be no technical obstacle to the direct substitution of crambe oil for high-erucic rapeseed oil in commerce.

The new-product studies by the Peoria investigators showed that hydrogenation of crambe oil yields a solid harder than beeswax. It could be used in making candles and several cosmetics. Another finding was that the erucic acid from the oil can be converted chemically into brassylic acid and pelargonic acid. Brassylic acid can be transformed into liquid waxes useful in high-pressure lubrication, and even into a form of nylon that has excellent properties as a plastic to be made into such industrial products as electrical insulation and gear wheels. Pelargonic acid produces salts that in aqueous solutions help to dissolve substances that do not dissolve well in plain water, and it also serves in the manufacture of lacquers and plastics.

The current outlook for crambe as a new agricultural product is favorable. The plant can be grown in virtually all the 48 contiguous states. Economic evaluations of the plant suggest that it could be a profitable item for both farmers and chemical processors.

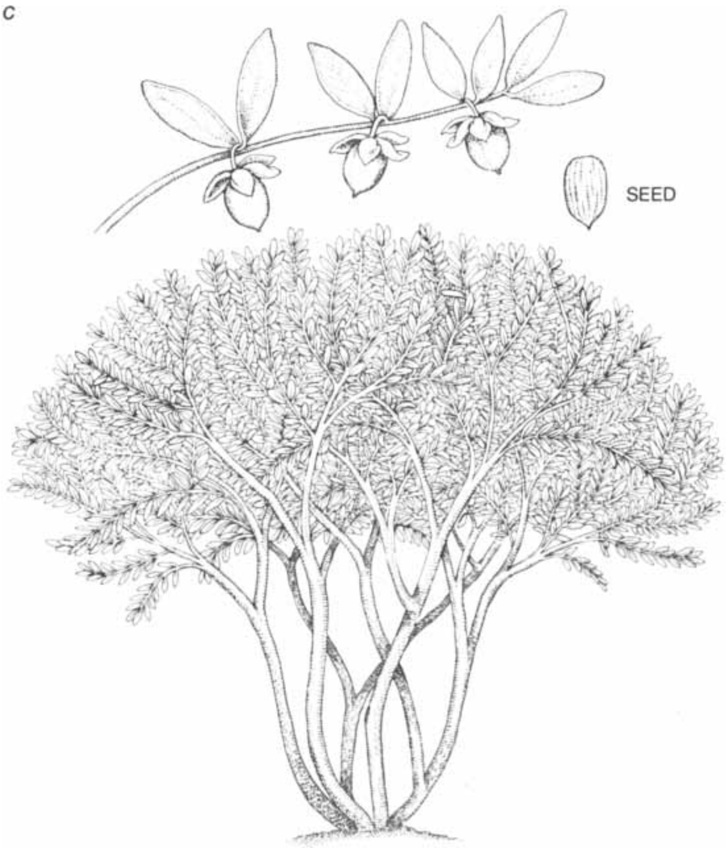
Jojoba (*Simmondsia chinensis*) is a shrub native to the Sonora desert of Mexico and the southwestern U.S. It became of agricultural interest somewhat more than a decade ago when oil from the sperm whale was banned from the U.S. in an effort to protect the endangered animals. Sperm-whale oil had for many years been the only commercial source of liquid-wax esters for cosmetic products (such as lotions, shampoos and conditioners), high-performance lubricants and even transmission fluid.

Studies done in several laboratories showed that oil from the jojoba seed could serve as a replacement for sperm-whale oil. Data from these chemical studies, together with interest on the part of environmentalists and financial backers, stimulated commercial plantings of the species. By last year roughly 16,000 hectares had been planted commercially in the Southwest. The oil currently commands a high price (more than $30 per gallon), but through increased production the price can be expected to reach about $15 within a few years. At that price the oil should be able to gain a larger market.

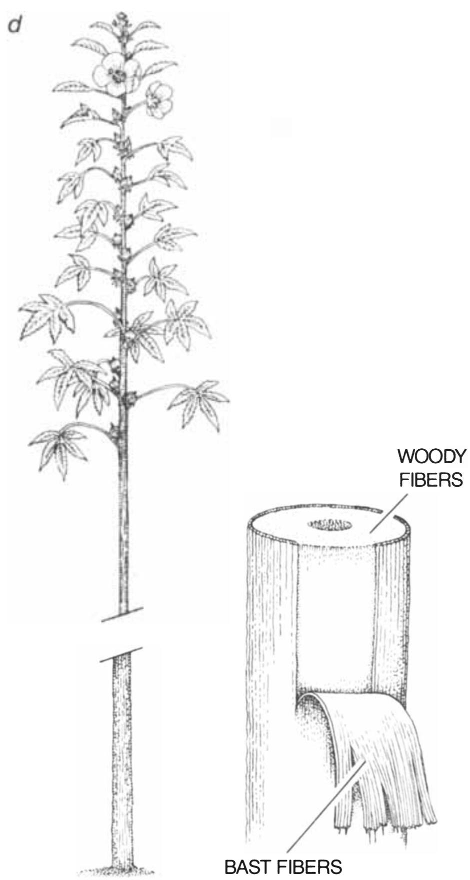
The jojoba plant must be three or four years old before it bears seeds and up to 10 years old to achieve maximum yields. The maximum yield has not yet been determined from either wild or cultivated shrubs, but wild plants have been known to produce between 14 and 18 kilograms of clean, dry seeds per plant per year. Several studies are under way to determine such matters as the optimal spacing of rows, the best ratio of male to female plants, the optimal amounts of irrigation and fertilization, efficient methods of harvesting the seeds and effective ways of controlling pests.

For every pound of oil obtained by pressing jojoba seeds one obtains a bit more than a pound of meal. This meal contains residual oil and also proteins, fiber, tannins and other substances. At present the meal has no commercial use. It has possibilities as animal feed, but before it could be sold for that purpose it would require extensive processing to remove the residual oil, a toxic component called simmondsin and materials (probably tannins) that make the meal unpalatable to animals. The cost of this processing equals or surpasses the value of the meal as a

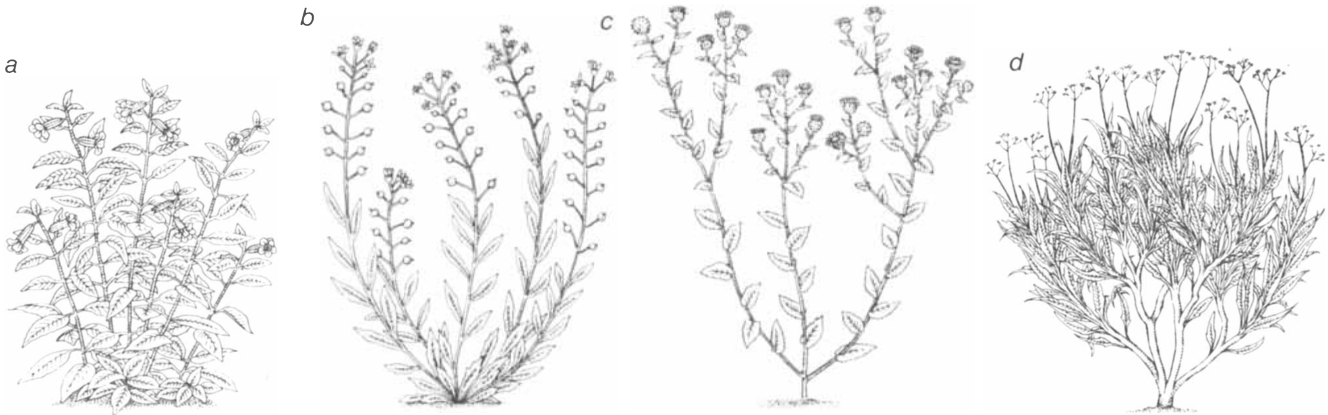
c



d



is important in the production of lubricants and plasticizers. The jojoba shrub is valued not only for its oil-producing seeds but also for its ability to thrive in arid regions where few other agricultural crops will grow without extensive irrigation. Kenaf (*Hibiscus cannabinus*) is a potential source of fiber for making paper and paperboard. It yields a pulp as good as or better than pulp from wood.



OTHER POTENTIAL CROPS include *Cuphea,* bladderpod, gumweed and guayule (*a–d*). *Cuphea* can be a source of fatty acids. Bladderpod is a desert genus (*Lesquerella*) producing seeds that are a source of an acid similar to the ricinoleic acid from imported castor oil. Gumweed (*Grindelia camporum*) could be a substitute for the wood rosin employed industrially in making adhesives and paper sizings. Guayule (*Parthenium argentatum*) produces rubber in its stems and roots; the shrub could also yield hydrocarbon fuels.

feed. Because of a low content of nitrogen, the meal is worth even less as a fertilizer. Perhaps its greatest value will be as a heating fuel, valued at an estimated $50 to $75 per ton at the site of origin. Transportation costs suggest that the likeliest markets will be close to such sites.

About 30 percent of the meal is protein. A significant part of the protein could be extracted for sale as a supplement to animal feed. It also has possibilities as a component of cosmetic products and as a sizing material in the paper industry.

The tannins in the meal, which are at an unusually high concentration of 8 percent, may also have a commercial future. All the tannin consumed in the U.S. today (in tanning, dyeing, inkmaking and various medical applications) is imported. If jojoba tannins turn out to be effective in such applications, the plant may become the first domestic source of the material.

The possibility of isolating other profitable products from jojoba meal is good. The meal contains the starting material for a potential new drug for appetite control; sugars capable of serving as substrates in fermentation; lignins for conversion into adhesives, plastics and chemicals, and cellulose fibers useful in many ways.

Perhaps the most valuable attribute of jojoba is that it will grow well in arid regions where most other crops require extensive irrigation. Many of the world's warm and arid regions have fertile soil and a highly developed agriculture, but the irrigation needs of conventional crops have caused alarming declines in groundwater levels. Both jojoba and buffalo gourd could help to alleviate the problem if further research and development efforts are successful.

Kenaf (*Hibiscus cannabinus*) is a herbaceous annual from eastern central Africa. It grows as thick stands of straight, slender stems six meters or more in length and about five centimeters in diameter at the base. The plant is important as a potential source of fiber for making paper and paperboard.

In 1983 the worldwide output of paper and paperboard by the major producing countries exceeded 150 million metric tons. Nearly all the fiber required for the manufacture of those products was wood pulp obtained from forests. The production and consumption of these materials generally increase at a rate of about 10 percent per year. Shortages of fiber are appearing in many regions. The pulp industries in Finland and Sweden, for example, now need to import wood chips, and as of 1982 at least two pulp mills in the U.S. were resorting to recycled newsprint to fill out an inadequate supply of wood. One authority has predicted a worldwide wood shortage of some 200 million cubic meters by the year 2000.

The greatest potential for increasing the fiber supply lies in nonwoody fibrous plants. Forests, which have a 30-year regeneration cycle, cannot produce as much fiber per year as suitable annual plants do. An annual crop also provides many other advantages, such as a yearly return on investment, improved planning through short-term forecasting and greater flexibility in the use of land.

With these considerations in mind the Agriculture Department began in the 1950's a screening program to identify annual fibrous plants that could be cultivated specifically for papermaking. Data were accumulated for 506 species and reported in 1960. On the basis of preliminary appraisals 92 species were tested by pulping on a laboratory scale. Kenaf made the best showing and was chosen for detailed development work.

Field trials designed to evaluate kenaf for its potential as a pulp crop were undertaken in 1957. By 1965 it had been established that the species would grow without irrigation in the lowlands of Alabama, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina and eastern Texas. Good yields were reported for areas as far north as Indiana, Iowa, Kansas and Nebraska.

Kenaf can yield from 25 to 45 metric tons (dry weight) of stems per hectare per year. The stems contain two

PRODUCT	USES
ERUCIC ACID	LUBRICANTS, PLASTICIZERS, FOAM SUPPRESSANTS
OIL	COSMETICS, LUBRICATION, FOAM SUPPRESSANTS
LAURIC ACID	SOAP
RICINOLEIC ACID	TEXTILE FINISHING, PLASTICS, SOAPS
VERNOLIC ACID	PLASTICS, COATINGS, ADHESIVES

ACIDS AND OILS important in industrial processes could be obtained from several of the potential new crops. The uses and sources of some of the items are shown. In the diagram of chemical structure black dots represent carbon atoms and colored dots oxygen atoms. Hydrogen is not shown.

distinct types of fiber: the outer-bark fiber, which can be a substitute for jute, and a thick inner core of short, woody fibers. Both types serve for making pulp. It is also possible to combine wood and kenaf fibers.

Processed on an industrial scale, kenaf yields pulp with properties usually as good as or better than wood pulp. In recent years the American Newspaper Publishers Association has expressed interest in paper made from kenaf pulp, causing manufacturers of newsprint to explore the possibility. Their interest is buttressed by the fact that in suitable conditions kenaf can be produced in greater quantities per hectare than pulpwood and at about half the cost. In the South particularly the crop now appears to be competitive with the standard commercial crops.

Buffalo gourd, crambe, jojoba and kenaf are at or close to commercial development. Currently many other plants that offer interesting possibilities for commercialization are being studied.

One of the plants is the group that makes up the genus *Cuphea*. A screening program at Peoria identified the genus as a potential source of short- and medium-chain fatty acids. The current source of lauric acid and acids with shorter chains is coconut oil from the Philippines; the chemical industry processes much of this imported oil to make surfactants and lubricants. The discovery of *Cuphea* stimulated research at a number of institutions, which demonstrated that oil extracted from the seed of various *Cuphea* species is of satisfactory quality for industrial applications. Developing one of the species as a crop will depend mainly on harvestable yields and the price of oils of this type.

Another possibility is *Lesquerella*, a desert genus of plants often known as bladderpod. The plants are a good source of hydroxy fatty acids, chiefly lesquerolic acid, which is quite similar to ricinoleic acid from castor oil. U.S. imports of castor oil amount to some 60,000 tons per year. The acids serve in making many products, including plastics and coatings. In addition lesquerella oil can be polymerized to yield a new class of tough plastics. Bladderpod can be grown with less water and on poorer soil than the castor-oil plant (*Ricinus communis*) and does not have the toxic and allergenic properties of castor beans.

Several plants screened at Peoria show potential as a source of epoxy fatty acids, which are consumed by U.S. industry at a rate of 70,000 tons per year in a variety of products. Currently the acids are prepared from soybean or linseed oil in a process that is expensive in both money and energy. The leading alternative species are *Vernonia anthelmintica, V. galamensis* and *Stokesia laevis.*

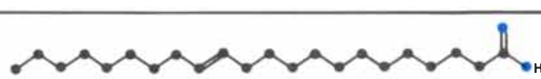
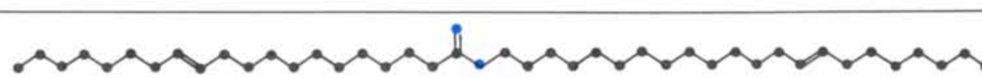
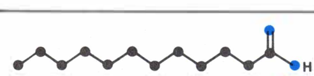
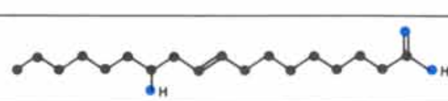
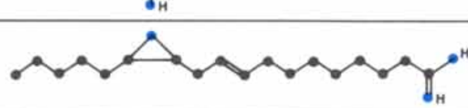
Gumweed (*Grindelia camporum*) is a possible substitute for wood rosin, which industry consumes in large quantities to prepare adhesives, paper sizings and many other products. The traditional source, old southern-pine stumps, is approaching exhaustion. *G. camporum* is a desert plant that could thrive with much less water than the conventional crops now grown in arid regions require.

Other plants have been given research attention in recent years because they produce organic chemicals that could serve as sources of hydrocarbon products from which rubber and fuel could be made. Guayule (*Parthenium argentatum*) is probably the leading candidate for rubber. Interest in plant sources for fuel hydrocarbons is temporarily at a low point because of the large worldwide supply of fossil crude oil. Nevertheless, promising candidates have been identified and doubtless will be developed when it is economically feasible to do so.

A final possibility is *Sesamum indicum,* the source of sesame seeds and oil. It was grown commercially in the U.S. more than a century ago but is not a commercial crop now in spite of the fact that some 40,000 metric tons of sesame products were imported in 1983. The plant's decline as a cash crop in the U.S. is attributed to uneven ripening of the seed and to the lack of mechanical harvesting equipment. Encouraging efforts to overcome these problems are under way in Arizona and South Carolina.

When a plant with crop potential is identified, much remains to be done to make it a commercial success. The first step is a research and development effort encompassing workers in many disciplines: botanists, geneticists, agronomists, chemists, engineers, nutritionists, economists and others.

In addition there is the problem of gaining the interest of organizations that have the imagination and the capital to exploit a new crop to the fullest extent. Many such crops yield not only a primary product but also secondary products that may be outside the developer's field of interest. Accordingly it is hard to find organizations in the private sector that are willing to provide all the money and effort needed to make a crop a commercial success. Often such an organization wants financial backing or an assured market to see the undertaking through the crucial early years. It is therefore likely that new crops will become commercial only when government and private organizations cooperate to develop them.

OLD SOURCE	NEW SOURCE	STRUCTURE
RAPESEED	CRAMBE	
SPERM WHALE	JOJOBA	
COCONUT	*CUPHEA*	
CASTOR OIL	*LESQUERELLA*	
LINSEED OIL OR SOYBEAN OIL	*VERNONIA ANTHELMINTICA*	

Very Large Structures in the Universe

Clusters of galaxies make up enormous superclusters, which alternate in space with immense voids. Such structures were born in the high-energy interactions of the early universe

by Jack O. Burns

A decade ago most astronomers believed that on large scales the universe is dull and featureless. Such aggregations of matter as clusters of galaxies, it was thought, are scattered uniformly throughout space. Recent observations have shown this is not the case. Even in the largest observable dimensions the luminous matter in the universe is gathered into formations that seem to have a few characteristic shapes. Enormous superclusters—that is, clusters of clusters of galaxies—share the universe with equally enormous voids, regions containing little or no luminous matter. What is the origin of such structures? How did they come to exist and how can they be explained?

During the past few years astronomers, particle physicists and cosmologists have joined forces to answer these questions. Their joint effort has led to a dramatic new comprehension of the structure observed in the universe today. The astronomers' role is to find and map the very large structures, determining the characteristics that are common to all; the physicists' role is to discover the nature of the physical laws that governed the early universe and influenced the formation of the earliest structures; the cosmologists' role is to describe precisely how the earliest structures formed, under the influence of these laws, and how they evolved into the structures contemporary astronomers have located and mapped.

The largest structures found and mapped so far, superclusters, are long filaments or shells that are themselves made up of many clusters of galaxies. Clusters of galaxies tend to be approximately spherical in shape and may contain hundreds or thousands of galaxies each; superclusters may consist of tens of clusters linked like beads on a string. The largest supercluster is more than a billion light-years long. Voids, which are hundreds of times less dense than the superclusters, separate superclusters from one another.

The discovery of superclusters and voids is one of the major advances that has occurred in observational astronomy in the past decade. Equally major advances in the field of particle physics are now making it possible to understand the structures' creation and evolution. The very large structures we observe today are actually the results of very small perturbations that arose in the early universe. As the universe expanded and cooled after the big bang, the single unified force of nature separated into the four aspects we know today: gravity, electromagnetism and the strong and weak nuclear forces. The separation of the forces, and the quantum nature of interactions in the early universe, caused tiny fluctuations in the density of an otherwise smooth background. The fluctuations grew as the universe expanded, eventually reaching the proportions of superclusters. In a sense, then, these structures are the fossilized remains of high-energy processes that occurred soon after the creation of the universe.

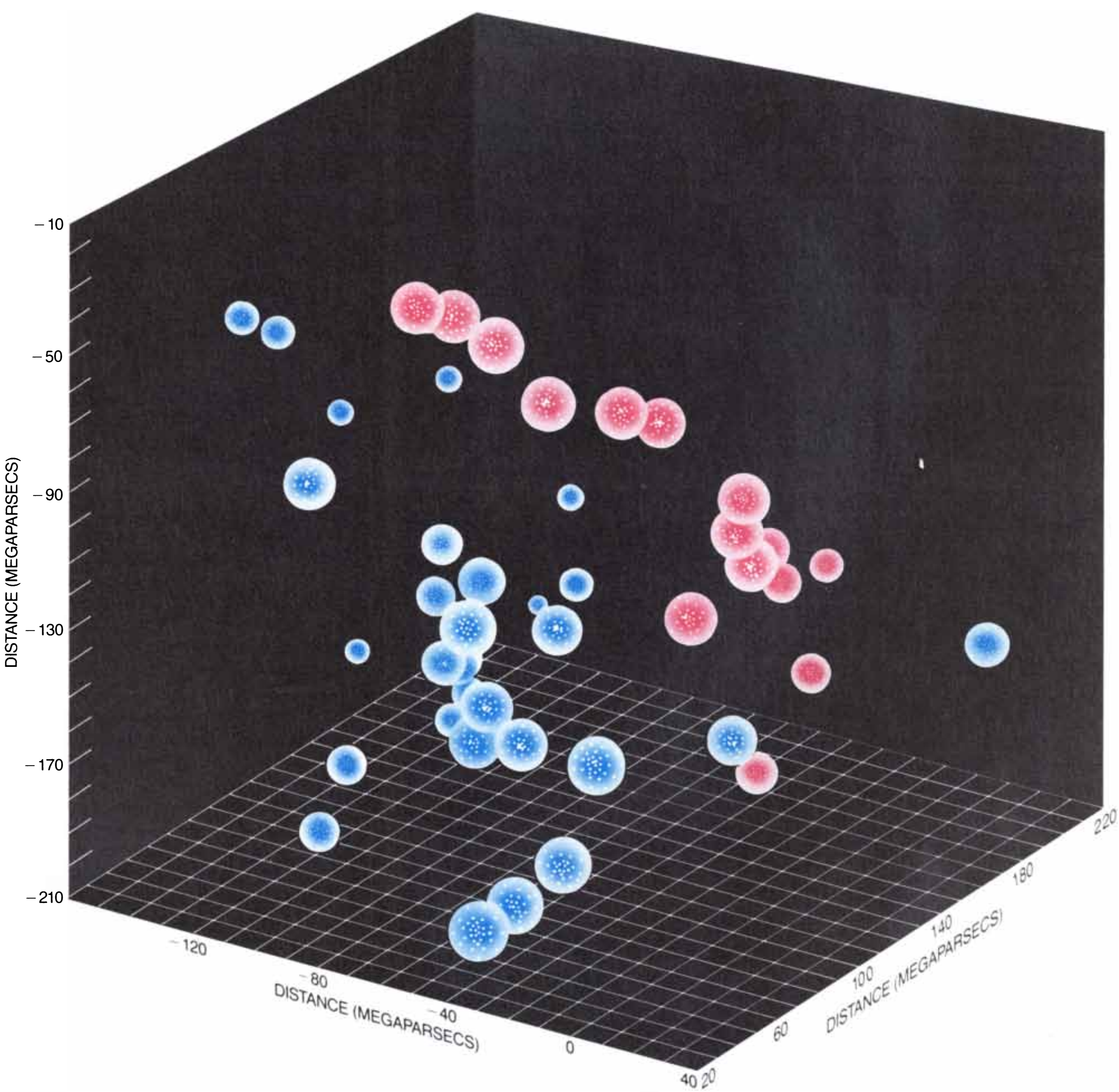
The starting point in understanding the very large structures is to map their shapes and distribution in space. Making such maps requires in turn that investigators ascertain the three-dimensional positions of galaxies and clusters. Obtaining these data is a two-step process.

The first step is to form a two-dimensional image of the positions of galaxies and clusters as they are seen from the earth. In the 1950's the Palomar Observatory, in conjunction with the National Geographic Society, photographed the entire sky of the Northern Hemisphere, forming images of features with brightnesses a million times fainter than those visible to the unaided eye. In 1958 the late George O. Abell of the California Institute of Technology examined glass copies of the Palomar sky survey to find clusters of galaxies. He defined a cluster as any agglomeration of galaxies in which there can be found at least 50 bright galaxies within a sphere whose radius is 6.5 million light-years. Abell found 2,712 clusters, some of which, when larger radii are considered, contain thousands of galaxies.

The second step in mapping large-scale structures is to determine each object's position in the third dimension of space: to find how far each object is from the earth. This can be done by examining the spectrum of light emitted by the object. Because of atomic processes in the atmospheres of their constituent stars, galaxies emit light that is particularly intense at certain frequencies. Since all other galaxies are moving away from the Milky Way, the spectrum observed from each galaxy will appear to be shifted toward the red, or longer-wavelength, end of the spectrum (just as the horn of an automobile that is moving away seems to emit a lower tone than the horn of a car that is standing still). The amount by which a galaxy's spectrum is red-shifted is a measure of how fast the galaxy is receding. In 1929 Edwin P. Hubble showed that galaxies are receding from the earth because of the expansion of the universe as a whole, and that more distant galaxies are receding at greater velocities and hence have greater red shifts. The measured red shift of a galaxy therefore gives a good indication of its distance from the earth.

Until recently surveys of galaxy red shifts have been limited in scope because they require large amounts of time and telescopes with particularly large apertures. New detectors, which are faster and gather more of the available light, have made extensive surveys much easier by drastically reducing the amount of time needed at the telescope. The most ambitious survey so far, which required five years of observation and data analysis, was recently completed by John P. Huchra, Marc Davis and their colleagues at the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory. They measured the red shifts of all the galaxies in the Northern Hemisphere that are brighter than magnitude 14.5 (a brightness about 2,500 times dimmer than that visible to the unaided eye). The farthest galaxies of this magnitude are about 300 million light-years distant.

These data, when taken together with additional red shifts determined by R. Brent Tully of the University of Hawaii at Honolulu, have provided a complete map of the local supercluster (in which our galaxy lies). It is composed of a core of at least 11 separate galaxy clusters, joined by low-density bridges that consist mainly of single galaxies. Surrounding the core is a halo of up to 50 additional groups of galaxies and thousands of individual galaxies, distributed within a sphere whose diameter is roughly 100 million light-years. There are strong indications that this local supercluster has



LARGEST KNOWN STRUCTURE in the universe is a filament-shaped supercluster (*red spheres*), or cluster of clusters of galaxies, that lies in the region of the constellations Perseus and Pegasus and is more than a billion light-years long. Nearly all the luminous matter in the universe is gathered in superclusters, which are separated from one another by enormous spherical or elliptical voids. In this illustration, which is based on a model built by David J. Batuski of the University of New Mexico, spheres represent clusters of galaxies. The earth is at point (0,0,0) in this coordinate system, which is in megaparsecs (one megaparsec is 3.26 million light-years).

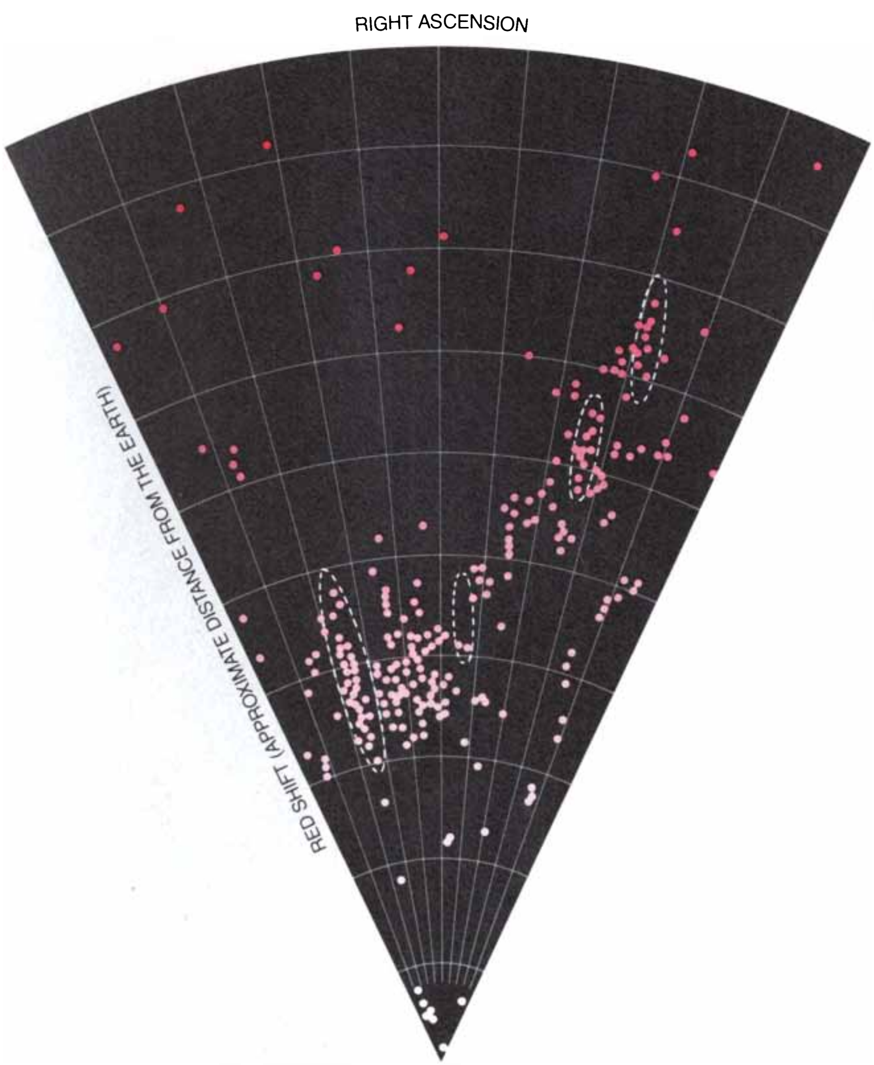
branches reaching outward to link it with several distant superclusters.

Other red-shift surveys, which have covered smaller angular scales but greater distances, have also yielded critically important information on superclusters. For example, the finding that there are significant structures hundreds of millions of light-years in scale was made in 1978 by Stephen A. Gregory, Laird A. Thompson and William G. Tifft, then all at the University of Arizona, who had observed a strip of sky only 15 degrees wide. They found that a thin chain of galaxies connects two rich clusters, known as the Coma cluster and Abell 1367. They also found a roughly spherical void, which lies in the foreground of the supercluster and is nearly as large as the supercluster itself.

Since the work of Gregory, Thompson and Tifft more than a dozen other superclusters have been discovered and explored, including several that lie in the directions of the constellations Perseus, Hercules, Ursa Majoris and Pisces. The Hercules supercluster is more than 260 million light-years long and consists of six Abell clusters and the bridges connecting them.

One exceptionally large void, in the constellation Boötes, was discovered in 1981 by Robert P. Kirshner of the University of Michigan and several colleagues. This region is about 250 million light-years in diameter and is ringed by "walls" made up of galaxies. The central region of the void is remarkably free of normal elliptical and spiral galaxies, although recent work by Kirshner and J. Ward Moody of the University of New Mexico indicates that a few dim dwarf galaxies may lie in the void's core.

In more recent work Valérie de Lapparent, Margaret J. Geller and Huchra at the Harvard-Smithsonian Center for Astrophysics have extended a 117-degree-by-six-degree slice of the Center for Astrophysics survey by measuring the red shifts of galaxies with magnitudes as low as 15.5. They have found a series of filamentary extensions emanating from the rich Coma galaxy cluster. Between the filaments are voids whose diameters measure about 100 million light-years. These observations have led the investigators to propose a "bubble" theory of the universe, according to which galaxies and clusters actually form the surfaces of giant bubblelike structures. The largest bubble would have a diameter comparable to that of the void found in 1981 by Kirshner.

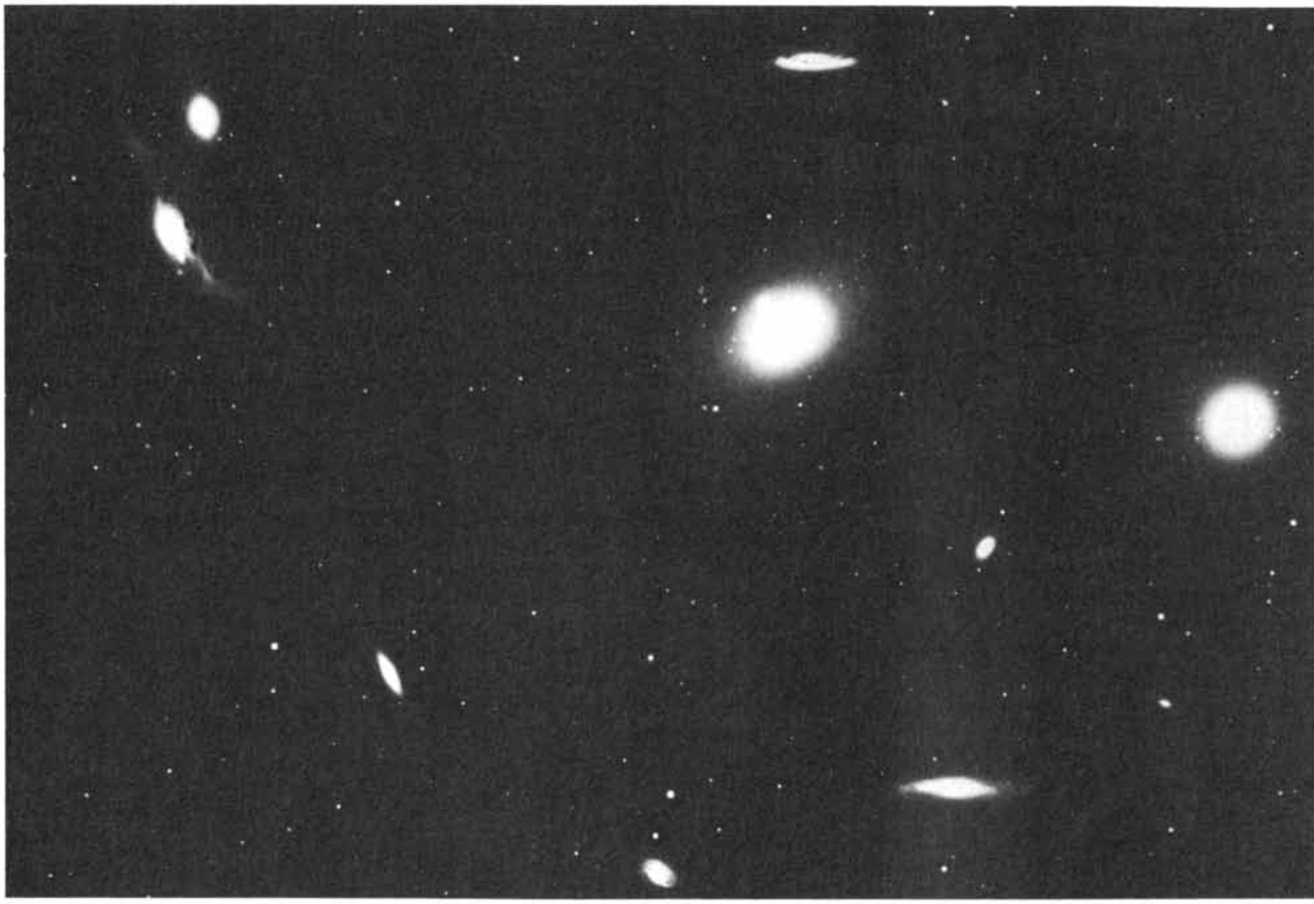


WEDGE DIAGRAM depicts a section of the Perseus-Pegasus supercluster (*filament-shaped formation at right*). Points represent galaxies, whose positions are plotted according to their right ascension (a variable partially describing an object's position as it can be observed from the earth) and their approximate distance from the earth (a variable that cannot be observed directly). A galaxy's distance from the earth is gauged by measuring the amount by which the spectrum it emits is shifted toward the red, or long-wavelength, end of the spectrum. A galaxy's red shift is directly proportional to the speed with which it is receding from the earth. The recession of the galaxies is caused by the overall expansion of the universe, and so more distant galaxies generally have greater red shifts. Abell clusters (the best-catalogued and brightest clusters of galaxies) are circled by broken white lines. Many of the red shifts here were provided by John P. Huchra of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory.

How large might the largest supercluster structures be? The question has remained unanswered in part because less than 1 percent of the volume of the universe lying within a few billion light-years of the earth has been surveyed adequately. There may be more than a million galaxies within this volume, and the amount of telescope time needed to make red-shift surveys out to these distances with current technology makes such surveys impossible.

Perhaps a starting point would be to try to get some rough indication of where in space candidate supercluster structures might lie and then to examine those regions. If some such tracer of supercluster structure, that is, some indication of regions that might contain superclusters, could be found, astronomers would have a powerful method of quickly mapping out some of the large structures in the universe.

For about four years David J. Batuski (a graduate student of mine at the University of New Mexico) and I have been trying this approach. We started from the observation that all the superclusters found to date have consisted in part of rich clusters (clusters containing a large number of galaxies)



CLUSTER OF GALAXIES in the constellation Virgo is roughly spherical and includes nearly every known type of galaxy. Superclusters consist of tens of such clusters (which may in turn contain hundreds or thousands of galaxies), linked like beads on a string.

linked by low-density bridges. Closely spaced groups of rich clusters might therefore be a reasonable place to look for large-scale structures. In particular, we reasoned, the Abell clusters, which are generally quite rich and whose positions are well catalogued, might provide the tracer we needed to find and map the positions of at least the richest superclusters.

Abell had recognized in the early 1960's that the clusters of galaxies he had found might themselves be clustered together, but he was not able to pursue the speculation further because he lacked the necessary red-shift measurements. Today the red shifts of more than 560 of the Abell clusters have been measured. (The red shift of a cluster is determined by measuring the red shifts of the three to five brightest galaxies in that cluster; this process requires only modest amounts of time and has been shown empirically to give accurate results.)

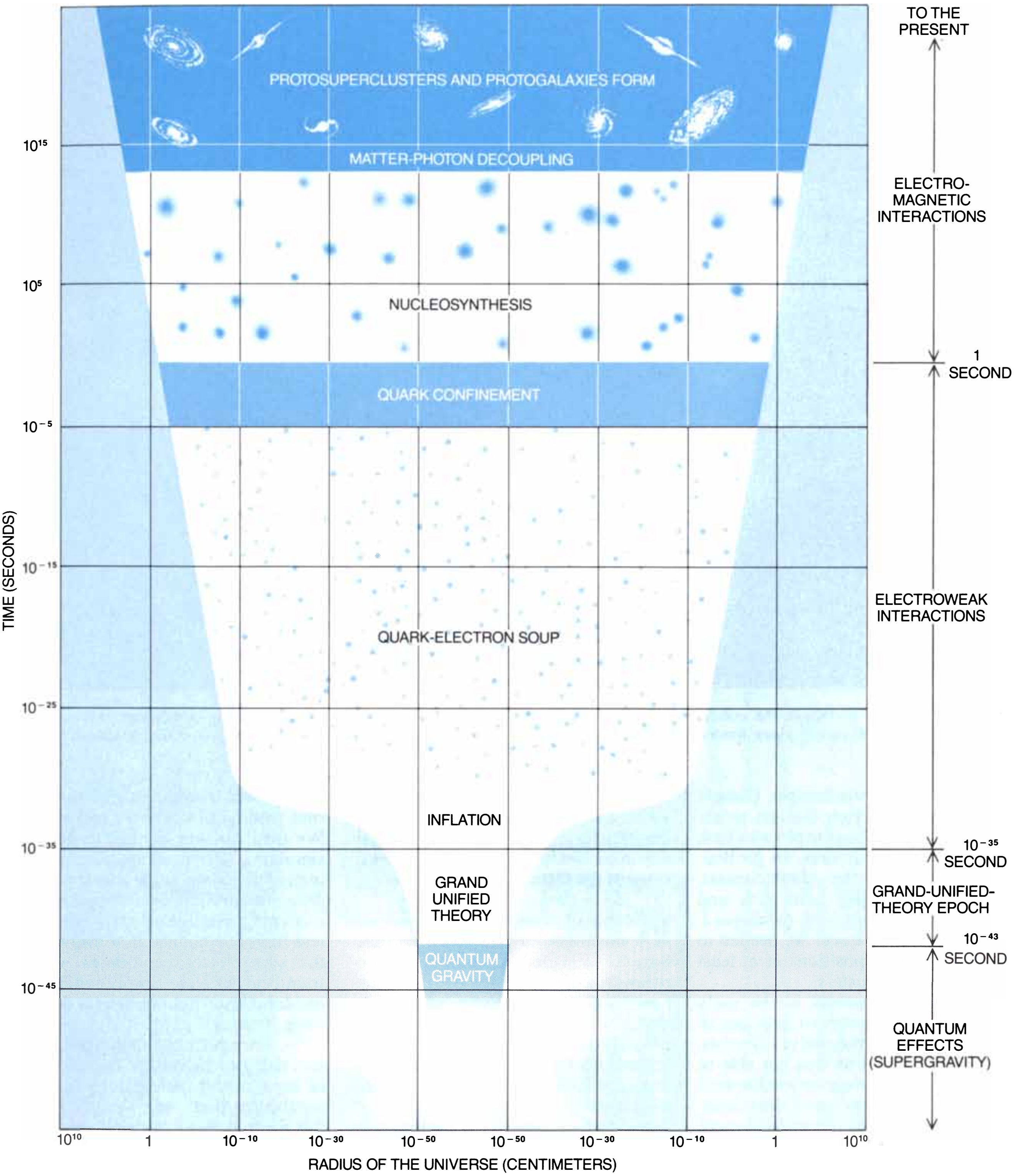
Batuski and I first assembled, from the literature and from unpublished sources, a collection of all the measured red shifts of Abell clusters within a distance of 1.4 billion light-years. We estimated the distances of Abell clusters whose red shifts had not been measured by applying the relatively well-established rule of thumb that the magnitude of the tenth-brightest galaxy in a cluster is usually a good indicator of the distance to that cluster.

Having obtained three-dimensional positions for 652 Abell clusters, Batuski designed a computer algorithm that searched for overall structure by a method known as percolation. The computer examined a spherical volume with a radius of 130 million light-years centered on a randomly selected Abell cluster. If any neighboring Abell cluster was found within that volume, the computer considered the two clusters to be part of the same formation and examined a sphere centered on the second one as well. The process was repeated over the entire volume in which Abell clusters are known to lie. Any regions in which Abell clusters were found to have near neighbors were identified as candidate superclusters.

With this algorithm we found a total of 102 candidate superclusters. To check the legitimacy of our algorithm (and so the significance of our findings) we ran a series of simulations in which 652 clusters were scattered randomly over a volume equal to the actual volume of space we had studied. We then ran our algorithm on these simulated distributions to see how many candidate superclusters a random distribution would give rise to. Our comparison showed there is much less than one chance in a million that the supercluster candidates we had found earlier were simply random coincidences and not indicative of underlying structure.

We were further encouraged by the fact that our algorithm had identified as supercluster candidates all the superclusters that were already known; this made it less likely that there were superclusters we had failed to identify. The major difference between these superclusters as they were previously known and as our algorithm had identified them was that our algorithm suggested they are much more extensive than had previously been supposed. For example, the Hercules supercluster had been thought to contain six Abell clusters; our algorithm indicated it might contain as many as 42.

In addition to searching for candidate superclusters we searched for candidate voids; using a separate algorithm, we identified volumes of space



EVENTS IN THE EARLY UNIVERSE determined the nature of the large-scale structures observed today. In the 10^{-43} second immediately following the big bang, the four fundamental forces that are known to act today were unified into a single force. No adequate theory has yet been constructed to describe this period; such a theory would have to combine the theoretical structures of general relativity and quantum mechanics. Quantum fluctuations during this phase of the universe may have created the seeds for the very large structures that were to form later. After 10^{-43} second had passed, gravity "froze out" from the other three forces, acquiring different characteristics and mediating different interactions. The other three forces remained unified. When 10^{-35} second had passed, the strong nuclear force froze out from the electroweak force (the combination of the weak nuclear force and electromagnetism) in a way that caused the universe to expand rapidly. From that time until 10^{-6} second later the universe was a hot soup of quarks and electrons. When the universe was one second old, the electromagnetic and weak forces dissociated. By this time the universe had cooled enough for quarks to be able to bind together to form protons. For the next 100,000 years matter and radiation were strongly coupled: regions that were dense with matter were also dense with radiation. During this period protons and electrons merged to form neutrons, and neutrons and protons bonded to one another to form nuclei of helium, deuterium and lithium atoms in the process known as nucleosynthesis. Matter then decoupled from radiation and the universe became transparent: radiation could stream freely. During this epoch fluctuations in the density of matter began to form into protogalaxies and protosuperclusters, structures that were later to evolve into galaxies and superclusters.

with diameters of at least 160 million light-years that contain no Abell clusters. We found 29 candidate voids, all of which are spherical or elliptical.

Having identified the candidate superclusters, our next step was to try to determine whether the Abell clusters that made up each candidate were actually connected. This is a long-term project that will take years to complete. We began by measuring the red shifts of the rich clusters whose positions and estimated red shifts indicated they were part of three of the largest candidate superclusters; we also examined a few poorer clusters and individual galaxies that seemed to lie between the Abell clusters in candidate superclusters. To ensure that our search pattern did not preferentially find only filament-shaped superclusters, we also measured the red shifts of clusters and galaxies lying in fields on each side of our putative bridges.

Our examination produced one particularly interesting result: we found a linear string of galaxies and clusters, in the region of the constellations Perseus and Pegasus, that is more than one billion light-years long. This is the largest known structure in the universe. It begins with the Perseus supercluster, which had already been discovered, and it curves gently southward and away from the earth along our line of sight. The filament is composed of 16 remarkably well-aligned Abell clusters; no gap between the clusters exceeds 160 million light-years. Those gaps that do exist are partially filled by a combination of poor clusters and individual galaxies.

The filament is particularly striking because of the three voids that surround it. Each void is roughly 300 million light-years in diameter, is free of rich clusters and is relatively free of bright galaxies. We are confident that these are real voids and not just regions that appear dark because of dust clouds in the Milky Way; most of the filament and the great majority of the voids lie well away from the plane of our galaxy.

The filament may be even longer than we envision; it may not end at the Perseus supercluster. Several years ago Riccardo Giovanelli and Martha P. Haynes of Cornell University suggested that the Perseus supercluster may in fact extend to the northeast, where it might be connected to the Ursa Majoris supercluster. This conjecture will be difficult to confirm, because any bridge between the superclusters would lie behind the highly obscured plane of the Milky Way. Giovanelli and Haynes are making observations in the radio frequencies with wavelengths of about 21 centimeters, which are generally not obscured by foreground dust. If the two superclusters do indeed connect, the entire filament would be more than 1.3 billion light-years long.

Is this filament just an accidental alignment of clusters and galaxies? As a test Batuski and I did another series of simulations, in which the computer randomly scattered clusters throughout a volume equivalent to the one actually occupied by the Abell clusters; the pseudoclusters were constrained to follow the correlation statistics actually found in real Abell clusters. We were not able to reproduce randomly a filament resembling the Perseus-Pegasus supercluster, and we determined that an accidental alignment of so many clusters over such a distance has a probability of less than .3 percent. I should add, however, that the Perseus-Pegasus filament seems to be a rarity among the distributions of clusters in the Abell catalogue; we know of no other filament of such length and degree of alignment. It will be known for certain whether other such filaments exist only when a volume of space much larger than the one with which Batuski and I were concerned has been explored.

Our analysis of the Abell catalogue indicates that superclusters and voids are very common structures in the universe. This leads us to ask whether there is an overall pattern to these structures. Are the superclusters, filaments and voids independent of one another, or do they interrelate in some way to form a single cosmic tapestry? In 1980 Jaan Einasto, M. Joeveer and Enn Saar of the Struve Astrophysical Observatory in Estonia suggested that superclusters and voids might form a cellular structure. The universe would then resemble a giant sponge or Swiss cheese, since most of the luminous matter would be concentrated into cell walls that are separated by large empty regions. The elongated, linear filaments would occur at intersections of cell walls. It is still too soon to tell whether the universe does indeed have such regularity or whether it is composed of an irregular distribution of supercluster structures, some of which interconnect.

What accounts for the existence of such large structures? The answer is to be found by looking back in time to the energetic processes that occurred not long after the big bang. Recent progress in the physics of high-energy particles has revealed much about the conditions then prevailing. Attempts to devise a grand unified theory, a theory that describes all the forces of nature as aspects of a single force, have led physicists to examine regimes of very high temperature and energy. Such conditions could have existed only within the tiny fraction of a second immediately after the creation.

There is not yet an adequate theory to describe the behavior of the universe in the 10^{-43} second following the big bang. At that time the gravitational fields in the universe would have been exceedingly strong and would have acted over very short distances. To describe such conditions adequately would require a theory that combined the general theory of relativity, which is the currently accepted theory of gravity, and quantum theory. No complete theory of quantum gravity has yet been formulated [see "Quantum Gravity," by Bryce S. DeWitt; SCIENTIFIC AMERICAN, December, 1983]. Many of the fluctuations in the density of matter that developed in the later universe into superclusters and galaxies may have originated in quantum fluctuations (randomly occurring nonuniformities in the density of matter and energy) that arose during this period, when the universe was less than 10^{-50} centimeter across.

After about 10^{-43} second gravity began to "freeze out," that is, it began to acquire unique characteristics distinct from those of the remaining unified force and to mediate different interactions. Since that time gravity has gone its own way; it is now the dominant force at long ranges. The forces that remained unified were the electromagnetic force, which in the present-day universe operates on both macroscopic and microscopic scales and is the dominant force on the scale characteristic of atomic distances; the weak nuclear force, which operates only within the atomic nucleus and is responsible for radioactive decay and certain other radioactive processes, and the strong force, which also operates only within the nucleus and binds quarks together to form such particles as protons.

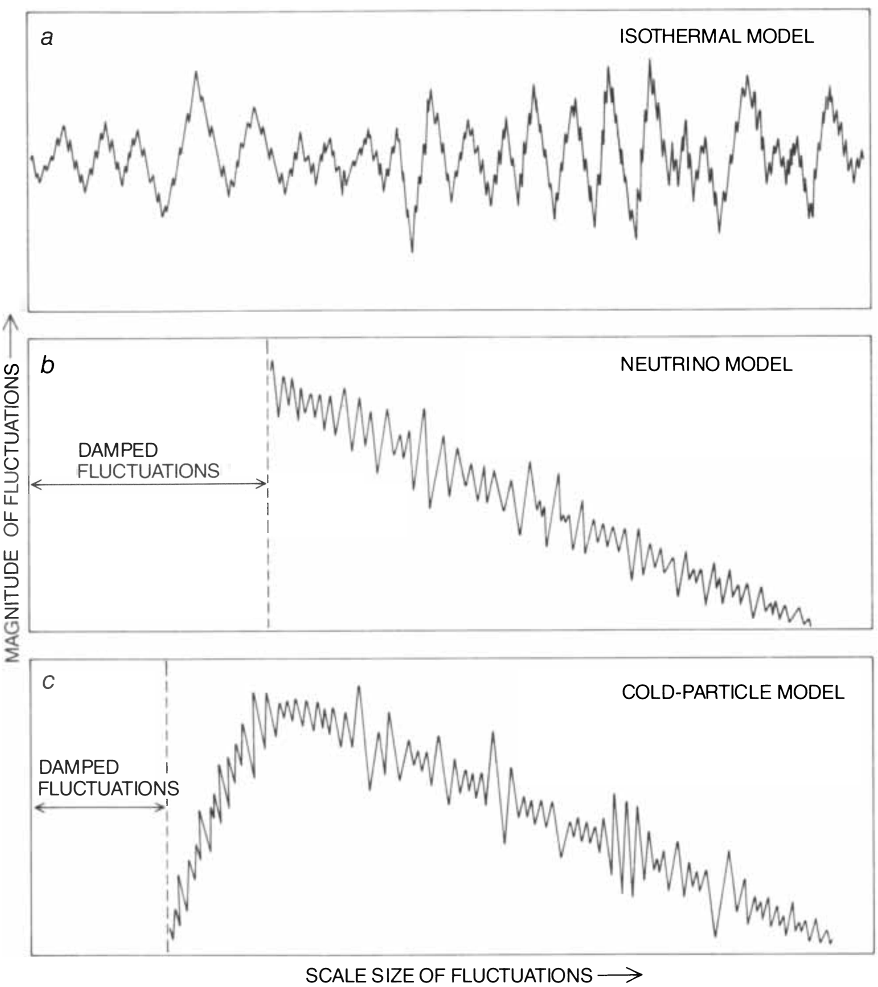
About 10^{-35} second after the big bang the strong force separated from the electroweak force (the combination of the electromagnetic and weak forces). Alan H. Guth of the Massachusetts Institute of Technology has suggested that the separation of the weak force precipitated a phase of rapid expansion, or inflation, during which the universe grew by 50 orders of magnitude in 10^{-32} second [see "The Inflationary Universe," by Alan H. Guth and Paul J. Steinhardt; SCIENTIFIC AMERICAN, May, 1984].

Between 10^{-35} and 10^{-6} second the universe consisted of a very hot soup of electrons and free quarks (quarks that have not been bound together to form larger particles). The temperature of the universe was still so high that no quarks could bind together to form such particles as protons: any quarks that had been bound together would have been shaken apart by thermal energy. By 10^{-6} second the universe had cooled sufficiently for protons to be able to form. Strong experimental evidence suggests there are no longer any unbound quarks.

When the universe was about one second old, the electromagnetic and weak forces dissociated. A great deal is now known about this epoch because, unlike the energies involved in the grand unification of the electroweak and strong forces, the energies necessary to observe unification of the electromagnetic and weak forces can be obtained in high-energy particle accelerators. In the mid-1970's Steven Weinberg, Abdus Salam and Sheldon Lee Glashow first described the way in which the electromagnetic and weak forces are unified at high energies, and in 1983 Carlo Rubbia and his colleagues were able to observe the particles, called *W* and *Z* bosons, that mediate electroweak interactions.

For roughly 100,000 years after the first second of its birth the universe was dominated by radiation remaining from the primeval fireball. Matter and radiation were strongly coupled; that is, those regions in which matter was dense were also dense with radiation. During this period some protons merged with electrons to create neutrons. The process of nucleosynthesis, in which protons, which are hydrogen nuclei, merge to form the heavier nuclei of helium, deuterium and lithium, also began in this period.

At the end of the period, about 100,000 years after the creation, matter began to decouple from radiation and the universe became transparent for the first time: radiation could flow through the universe without much interference from matter, and fluctuations in the density of matter were no longer matched with fluctuations in the density of radiation. This epoch also marked the beginning of the period in which superclusters and galaxies grew rapidly from the seeds formed in earlier density fluctuations. This matter-dominated era has extended to the present time, some 15 billion years after the big bang.

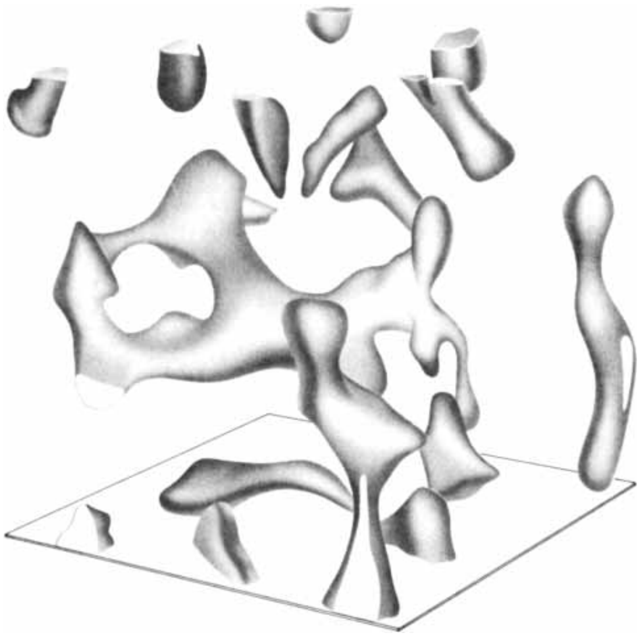


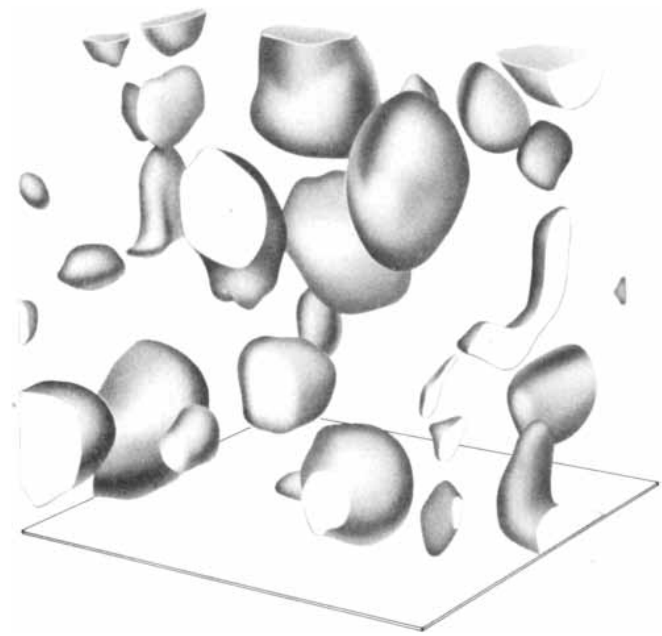
FLUCTUATIONS in the density of matter in the early universe could have arisen on several scale sizes. The size of the earliest fluctuations to develop would have determined the size of the earliest structures to grow and so would have played an important role in determining the nature of large-scale structures in the later universe. In an isothermal model (*a*), in which the background thermal radiation in the early universe is roughly the same throughout space (even though matter is distributed in clumps), fluctuations may arise on all scale sizes. The first structures to grow would then be the smallest, simplest ones, such as protogalaxies, which would later have aggregated through mutual gravitational attraction to form protoclusters and protosuperclusters. The model is called the bottom-up model, because the smallest structures form first. In another model (*b*) the early universe is dominated by fast-moving neutrinos, which break up any density fluctuations much as a fast-moving cannonball might scatter a stone wall. Once the neutrinos have slowed down, fluctuations can form. The smallest-scale fluctuations have all been damped out, however, and so the first fluctuations are on the same scale as the horizon of the universe (the distance that could have been traversed by a photon since the big bang). Fluctuations on other, even larger scales grow at later times as the universe expands, but they are not nearly as dominant in the structure of the universe. This is called the top-down model, because the early, large-scale fluctuations condense first to form the largest objects (protosuperclusters), which then fragment into smaller structures. In a variant model (*c*) the early particles are not neutrinos but certain very massive particles predicted by the new theory of supersymmetry. These would be slower (colder) than neutrinos, and so fluctuations could form earlier, and on smaller scales, than in the neutrino model. These small-scale fluctuations are soon followed by large-scale ones. This is called the hybrid bottom-up model.

The particles produced in various phases of the early universe strongly influenced the formation of structure. Small fluctuations in the density of these particles in space eventually grew into superclusters.

Even during the earliest epochs the universe was never perfectly smooth. Quantum-mechanical fluctuations, always present, were particularly important in the era of quantum gravity. Moreover, the freezing out of the fundamental forces was a turbulent, disorganizing process. The fluctuations in density that evolved into galaxies, clusters and superclusters could have originated in any of a number of ways. Tracing such mechanisms is the task of the cosmologist.

As physics and astronomy have advanced, the cosmologist has had better information to work with and has been better able to refine his theories. In recent years three general kinds of model have been proposed to describe the formation of the early structures.





SUPERCLUSTERS AND VOIDS that would form in a neutrino-dominated universe have been modeled on supercomputers by Joan M. Centrella of Drexel University and Adrian L. Melott of the University of Chicago. These models predict that matter would aggregate into shell-like structures and filaments (*left*) alternating in a cellular structure with spherical or elliptical voids (*right*).

One, called the bottom-up model, derives from work done in the late 1960's and early 1970's by P. J. E. Peebles and his colleagues at Princeton University. They were studying fluctuations in the density of matter in a hypothetical early universe that was dominated by baryons (ordinary matter, such as protons and neutrons). For the sake of simplicity they assumed a universe in which fluctuations in density were isothermal, that is, one in which the background radiation (and hence the temperature) was smoothly distributed, although the matter in the universe was strongly clumped. Under this set of assumptions the fluctuations in the distribution of mass first form protogalaxies that have masses from 100,000 to one million times the mass of the sun. The protogalaxies become galaxies, which are then grouped into clusters by mutual gravitational attraction. This scenario is called the bottom-up model because smaller structures (namely galaxies) form first and then grow into larger structures (namely superclusters).

The bottom-up model was faced with a severe problem in the late 1970's, when superclusters hundreds of millions of light-years long were discovered. There simply has not been enough time since the big bang for such large structures to have formed by the gravitational aggregation of galaxies.

A second kind of model considers a universe of adiabatic density fluctuations (fluctuations in which the amount of thermal radiation, or heat, in a given amount of matter remains constant even as the spatial density of the matter fluctuates over time). In such a universe the ratio between the energy density of matter and that of radiation must be the same in all regions of space, and so the distribution of radiation is clumped in a way that matches the distribution of matter. The size and distribution of such density fluctuations was studied independently in the early 1970's by Edward R. Harrison of the University of Massachusetts at Amherst and Yakov B. Zel'dovich of the University of Moscow. Guth has shown that a similar distribution would emerge naturally in the inflationary model he proposed.

In the earlier, isothermal model, variations in density occurred on all scales of length. In the adiabatic model, on the other hand, only fluctuations on very large scales are allowed because the dynamics of the interactions among particles and between particles and radiation suppress the smaller perturbations. In this model, then, protosuperclusters, diffuse objects that precede superclusters, form first, later collapsing to become denser and then fragmenting to form galaxies. This is called the top-down model.

Zel'dovich has demonstrated that in the environment of the early universe the protoclusters would tend to collapse rapidly in one dimension, taking on the shape of a flattened pancake. Linear filaments would be found wherever two pancakes intersected [see "The Large-Scale Structure of the Universe," by Joseph Silk, Alexander S. Szalay and Yakov B. Zel'dovich; SCIENTIFIC AMERICAN, October, 1983].

Early versions of the top-down model had one major flaw. In his initial formulation of the model, Zel'dovich assumed that the universe today is composed mainly of baryons. Because this model of the universe is adiabatic, however, the large fluctuations in the density of matter necessary to produce the observed level of structure would have to correspond to equally large fluctuations in background radiation. This is inconsistent with the well-established observation that the universe is filled with background radiation that is virtually isotropic (the same in all directions).

The top-down model was salvaged by discarding the assumption that the universe is made primarily of baryons. Although the earth consists almost exclusively of such "normal" matter, the universe as a whole could be dominated by some other kind of matter. In particular, it could be made of matter that, unlike baryons, does not tend to undergo interactions in which photons are emitted. Such matter would be dark and therefore hard to observe directly. There is already good observational reason to believe some kind of dark matter is present in great abundance in certain systems. For example, certain galaxy clusters must be made up of tens or hundreds of times more mass than can be directly inferred from the amount of luminous matter they contain; otherwise they would not be able to hold together all their constituent galaxies by the force of gravitation alone.

What might the dark matter consist of? One candidate is neutrinos. The neutrino is a weakly interacting particle (one that tends not to interact with other particles) that had been assumed for many years to be massless. Theoretical considerations and several unconfirmed experiments have recently suggested, however, that the neutrino might indeed have mass and could therefore be the constituent particle of dark matter.

If the neutrino does have mass, it could play an extremely important role in the pancake model. Neutrinos decoupled from radiation before ordinary matter did. When they first decoupled, they were very hot and had velocities close to the speed of light. Because of their high speed and the low probability that they would interact with one another, the neutrinos "free-streamed" through the early universe: they flowed through the universe in every direction at high velocities and their motion was not appreciably perturbed by the distribution of energy and matter. Such high-velocity free-streaming particles tend to eliminate any fluctuations in the density of matter, much as a cannonball moving at high speed might scatter a loosely built wall of stones without being appreciably slowed by the collision.

As the universe expanded, the neutrinos cooled; once their velocities were less than a tenth the speed of light, they no longer damped out fluctuations in density, and so density perturbations began to grow. These first perturbations were on the scale size of the horizon (the distance that could have been traversed by a photon in the time since the big bang), because all smaller perturbations had been erased by the free-streaming neutrinos. As in the original pancake model, then, the first structures to condense would have been the largest: the protosuperclusters. Protosuperclusters would have begun as fluctuations in the density of neutrinos; a slow-moving neutrino might be stopped and gathered in by a density fluctuation, much as a cannonball moving at low speed might bounce into a stone wall and stop moving, adding its own mass to the wall. Later, baryons fell into the neutrino superclusters, eventually producing the structures observed today.

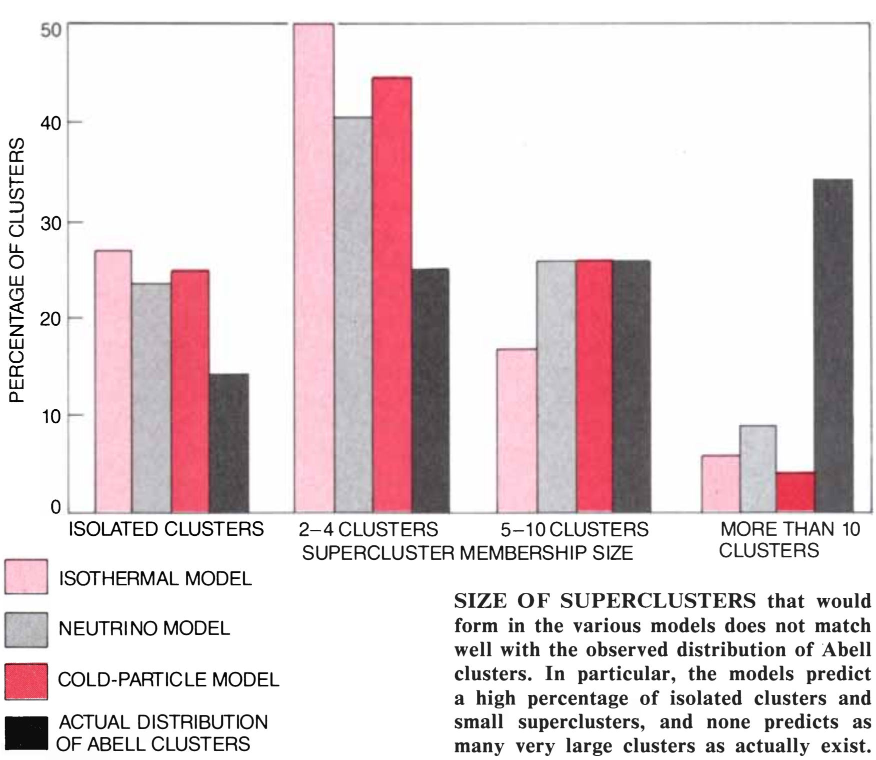
This model of a universe dominated by neutrinos leads to three predictions that can be tested by new observations. First, it predicts that there is little or no matter, either luminous or dark, in the voids: any aggregations of matter must have descended from the earliest, supercluster-forming perturbations, which formed around the voids. New observations to test this prediction are being made by Kirshner and his colleagues, who are examining the Boötes void, and Gregory and his colleagues, who are examining the Coma void. There are already preliminary indications that some dwarf galaxies may lie in the Boötes void. If these indications bear up under detailed scrutiny, the model is in serious observational trouble.

The second observationally testable prediction of the neutrino-dominated pancake model is that the major axes of galaxies should tend to be aligned with the short axis of the supercluster in which they lie (the axis along which the pancake was first flattened). Such an alignment, which would represent a preserved memory of the one-dimensional collapse that initially formed the supercluster and its constituent galaxies, was suggested by Stephen Gregory and Laird Thompson. Bryan E. Laubscher, Gregory and I are now checking this prediction by making a detailed study of the orientations of galaxies in the Perseus supercluster.

The third way to test the neutrino model is derived from calculations done on supercomputers by Joan M. Centrella of Drexel University and Adrian L. Melott of the University of Chicago [*see illustration on preceding page*]. They predict that in a neutrino-pancake universe relatively simple structures will form: large pancakes and the filaments that form at the intersections of pancakes should alternate regularly with spherical voids. Observationally, it is still too soon to tell whether such simplicity and regularity actually exists in the large-scale structures.

Regardless of how these observational tests turn out, the neutrino model is faced with some potentially devastating problems, according to computer simulations done by Simon D. M. White of the University of Arizona and Marc Davis of the University of California at Berkeley. First, because the neutrinos had such a high velocity after decoupling, and because neutrinos must have a very low mass if they have any mass at all, neutrinos could be expected to free-stream for quite a long time before cooling down enough to allow appreciable fluctuations in density to grow. Perturbations would not grow until the universe was relatively large and old, and the first galaxies would not form until quite late in the evolution of the universe—so late, in fact, that their ages would be inconsistent with the currently accepted ages of some of the oldest stars, which lie in the so-called globular clusters found near many galaxies.

White and Davis also argue that the neutrino model produces a much greater amount of clustering than has been observed, at least in the local region of the universe. In addition, because of the neutrinos' low masses and high velocities, they would not bind easily to the halos of galaxies and to clusters of galaxies, and so one would



SIZE OF SUPERCLUSTERS that would form in the various models does not match well with the observed distribution of Abell clusters. In particular, the models predict a high percentage of isolated clusters and small superclusters, and none predicts as many very large clusters as actually exist.

have to invent a second particle (or some other mechanism) to explain the anomalies in galaxy dynamics that are now explained by the hypothesis of dark matter. This is not aesthetically pleasing.

Recent advances in theoretical physics may provide a way out of the dilemma in the form of a new model of the formation of large structures, a model that combines features of the isothermal model with those of the neutrino model. A relatively new concept called supersymmetry [see "Is Nature Supersymmetric?" by Howard E. Haber and Gordon L. Kane; SCIENTIFIC AMERICAN, June] predicts the existence of certain particles called photinos, gravitinos and axions, which might have been created very early in the history of the universe. These particles would be much more massive than neutrinos, and so they would not free-stream after decoupling from the background radiation. They would be slower—colder. Fluctuations in density could then grow at much earlier times and on much smaller scales than in the neutrino-dominated model. In the new model, galaxies would form first and larger-scale structures would evolve from fluctuations that form during a later period in the early history of the universe. Because large-scale structures derive from early fluctuations, the slow process of gravitational aggregation is not the only source of supercluster formation, as it was in the isothermal model. This model, which relies on slow, or cold, particles, and in which galaxies form before superclusters but both form from early fluctuations, is called the cold-particle model or the hybrid-bottom-up model.

Three observationally testable predictions arise from the cold-particle model. First, galaxies should be older than superclusters. The interrelations among the alignments of superclusters and their constituent galaxies should therefore not be as strong as they are in the neutrino model. Second, because the spectrum of density fluctuations in this model prefers smaller structures over larger ones, such enormous structures as the Perseus-Pegasus filament should be rare in the universe. Third, although the voids should still be relatively barren, they might be expected to contain some matter, which would have descended from matter that was not included in the relatively late fluctuations that brought about the superclusters.

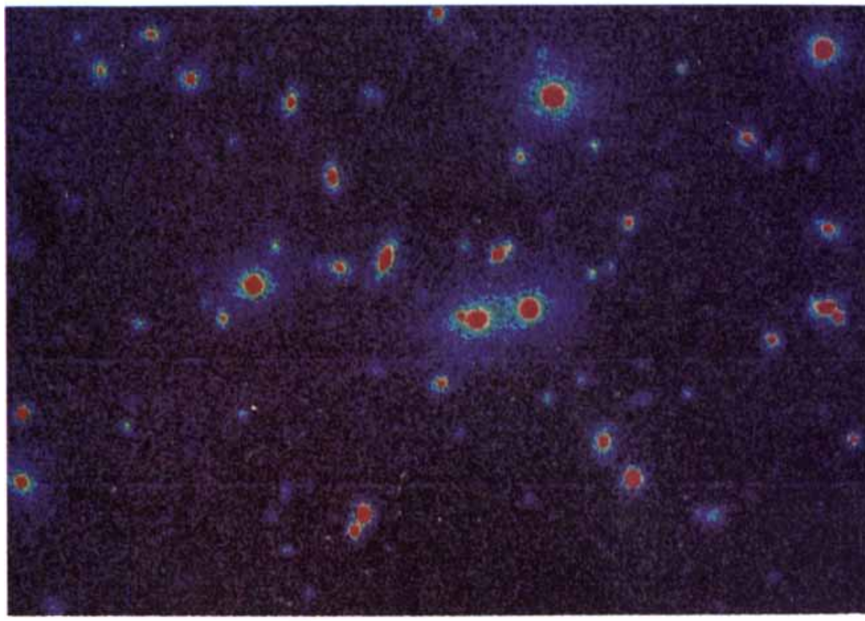
Batuski, Melott and I have undertaken to test the relative merits of the isothermal, neutrino and cold-particle models by simulating on a supercomputer the distribution of rich clusters that might be expected in each model. Preliminary comparisons of the clustering predicted by each model with that actually found in the Abell clusters are not encouraging. None of the models is able to reproduce the apparent abundance of supercluster filamentation on large scales. In particular, the neutrino model produces too many small, tight superclusters and too many galaxy clusters overall; the isothermal and cold-particle models, on the other hand, provide much less structure than is seen in the Abell clusters. We are beginning more extensive modeling in order to confirm and enlarge these tentative conclusions.

The simulations we have done so far leave us with an interesting theoretical quandary. What if, after we have completed our simulations, none of the proposed models is able to reproduce the very large structures that have recently been discovered?

In such a case I can foresee two alternatives. First, the distribution of luminous matter, such as galaxies and superclusters, might not match very well the distribution of dark matter. The dark matter is really what we are mapping in our simulations, and the distribution of the dark matter may be much smoother than the relatively clumpy distribution of the luminous matter, which can be observed and mapped directly.

The second alternative is that density perturbations may originate from a totally different mechanism than we have been considering. Perhaps the formation of galaxies and clusters was stimulated by cosmic strings: long, one-dimensional defects in the structure of spacetime that produce abnormally strong gravity in regions where they kink. Alternatively, the new "superstring" theories of fundamental physics, in which particles are viewed not as zero-dimensional points but as tiny one-dimensional vibrating strings, may have astrophysical implications that have not yet been fully explored. In particular, it might be that superstring theories call for some kind of "shadow matter" to have been created during the era of quantum gravity. Shadow matter would interact with ordinary matter only through the force of gravitation.

Challenging puzzles and problems remain for both the theoretician and the observer. We are living in a particularly exciting time in cosmology and in high-energy particle physics, because the conjoining of the two disciplines promises to answer questions about the largest and the smallest objects in the universe.



RICH CLUSTER of galaxies in the Corona Borealis is captured in an image made by a new charge-coupled device at the University of New Mexico's Capilla Peak Observatory. In this false-color image red represents the most intense optical light and blue the least intense. Charge-coupled devices require less light and shorter observing times than photographic plates, and they also provide more precise information about the galaxies' sizes, colors and orientations; they are making possible much more extensive surveys of the structure of deep space than have been done in the past. Better knowledge of the shapes and distribution of very large structures will provide the data necessary to confirm theories about their origin. The image was made with equipment provided by grants from the National Science Foundation, the Research Corporation and the University of New Mexico.

Anti-idiotypes and Immunity

Anti-idiotype antibodies recognize the molecular individuality of other antibodies. They modulate the immune response and open the way to experimental and therapeutic manipulations of the immune system

by Ronald C. Kennedy, Joseph L. Melnick and Gordon R. Dreesman

The immune system of an individual can make millions of different kinds of antibodies: the complex molecules that recognize and bind to alien substances that invade the body, marking them for destruction. Each antibody can in turn be the target of other antibodies that recognize its unique molecular characteristics. By means of such antibody-antibody reactions the immune system interacts with itself. Networks of reactions seem to modulate the normal immune response, and disorders of the networks may underlie autoimmune diseases, such as rheumatoid arthritis, systemic lupus erythematosus and myasthenia gravis, which stem from an immunological attack on the body itself.

Work done in our laboratory and elsewhere has shown that these interactions, known as idiotype-anti-idiotype reactions, can be exploited to manipulate the immune system. The possibility could have profound implications for the therapy of a spectrum of illnesses ranging from infections through many kinds of cancer to autoimmune diseases.

Idiotype-anti-idiotype reactions are a product of the diversity and specificity of antibodies themselves. Immunoglobulins, the class of proteins to which antibodies belong, are made by white blood cells called *B* lymphocytes. The population of *B* lymphocytes is divided into many millions of different clones, each clone consisting of one or a few cells that make a unique antibody. The multiplicity of clones means that almost any large molecule can act as an antigen and elicit a specific immune response.

Ordinarily the antibody molecules are bound to the membrane of *B* cells as surface receptors. When one of the membrane-bound antibodies encounters an appropriate antigenic determinant (the part of an antigen to which the antibody binds directly), the *B*-cell clone expands into a large population of plasma cells. The plasma cells secrete free antibody molecules, all of which are specific for the same antigenic determinant. The antibody binds to the antigen, setting in motion a complex process that leads to its destruction or neutralization. Most antigens on the surface of a foreign cell or substance bear numerous antigenic determinants and hence stimulate the secretion of many different antibodies.

The basis of each antibody's specificity resides at one end of its constituent chains of amino acids. There are four such chains in an antibody molecule: two identical heavy, or long, chains flanked by two identical light chains, or shorter segments. Much of each chain is taken up by a constant region whose makeup is similar for every chain within a broad class; there are five broad classes of heavy chains and two of light chains. Each chain also includes a variable region that has a unique chemistry and conformation for each *B*-cell clone. At the end of each of the antibody molecule's two arms the variable regions of a heavy chain and a light chain form a combining site, where the antibody interacts with its specific determinant.

In the 1950's it was demonstrated that molecules of immunoglobulin themselves bear antigenic determinants. Myeloma proteins, which are immunoglobulins that are spontaneously made in large amounts by *B*-cell cancers, were injected into experimental animals. The animals responded by producing antibodies that reacted with the myeloma proteins.

In 1963 Henry G. Kunkel and his colleagues at the Rockefeller Institute for Medical Research and Jacques Oudin of the Pasteur Institute independently made an analogous observation for antibodies made normally, by cells stimulated by antigen. An experimental animal was exposed to antigen, eliciting the production of antibody. The initial antibody (Ab-1) was isolated from the animal's serum and injected into another animal, where it induced a second antibody (Ab-2). The Ab-2 bound only to the Ab-1 and not to other antibodies against which it was tested. It could therefore be said to recognize the primary antibody's individuality. Oudin referred to the unique antigenic determinants on the Ab-1 as its idiotype (from the Greek roots for "individual" and "form"). For the Ab-2 produced in response to the idiotype he coined the term anti-idiotype.

Other work showed the immune system can make various kinds of antibodies in response to an Ab-1. Some of them bind not only to the Ab-1 that induced them but also to other antibodies with the same class of heavy or light chains. Presumably those second antibodies, which are known as anti-isotypes, recognize antigenic determinants in the constant regions of the heavy and light chains. Another set of antibodies, called anti-allotypes, react with every antibody made by the individual from which the Ab-1 was taken, although not necessarily with antibodies from other individuals. The allotypes to which such second antibodies respond are analogous to the blood-group antigens of red blood cells: all the antibodies in an individual share the same allotype, but the allotype varies within a species. The determinants that make up an allotype lie primarily in the constant regions as well.

An anti-idiotype, in contrast, was defined as an antibody that is able to react only with the Ab-1 elicited by a single antigen. The idiotype therefore must lie in the variable region of the Ab-1. During the late 1960's and early 1970's studies done by Alfred Nisonoff, now at Brandeis University, Klaus Eichmann of the Max Planck Institute for Immunobiology in Freiburg and a series of independent investigators yielded a more precise picture of idiotypes and anti-idiotypes.

A simple organic molecule or a bacterial sugar served as the antigen in some of the demonstrations, inducing the Ab-1, to which an Ab-2 (the anti-idiotype) was then generated. By incubating the Ab-1 with the antigen, the workers were often able to prevent the anti-idiotype from reacting with the idiotype. Presumably the antigen blocked the binding of the anti-idiotype by filling the combining site on the Ab-1. Hence the idiotype must lie at or close to the combining site.

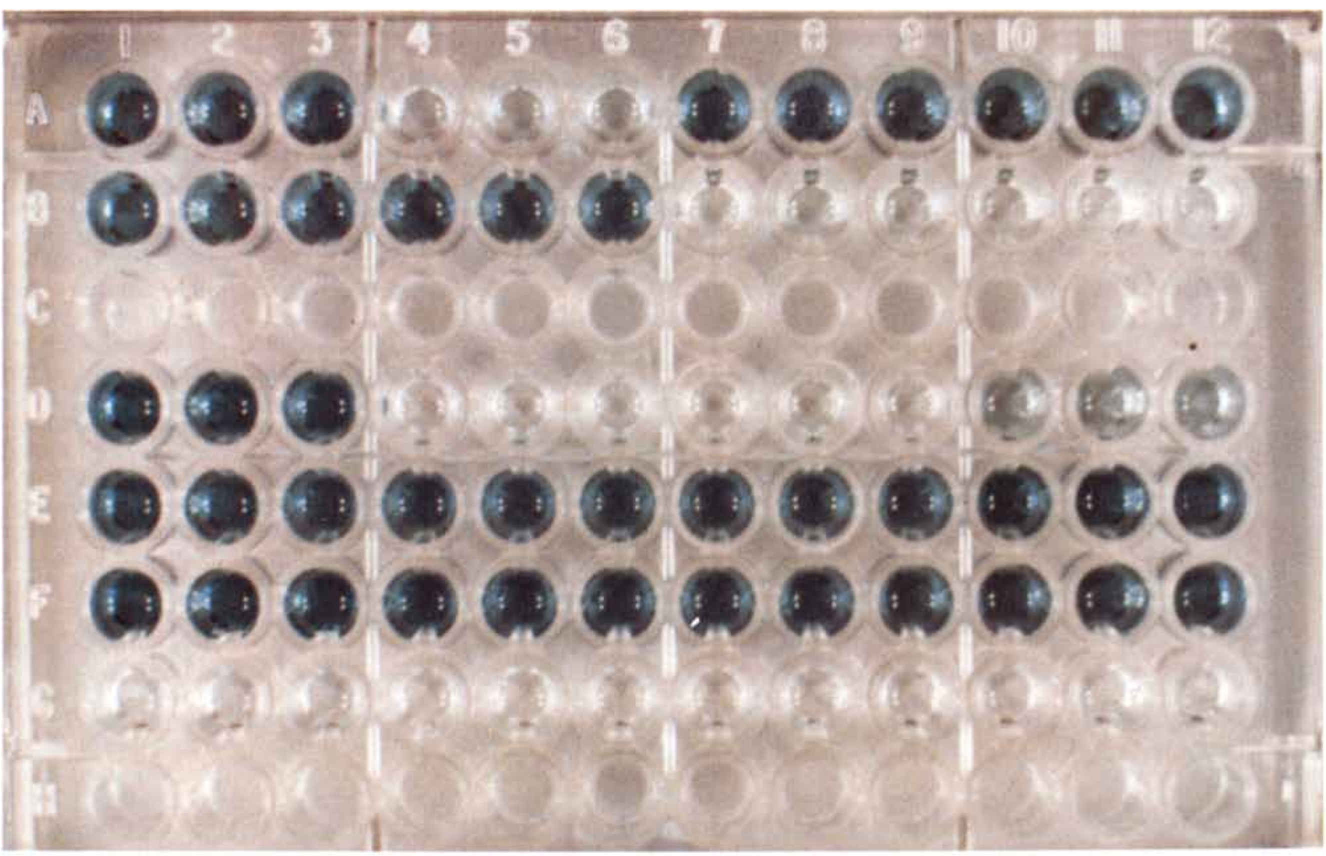
Other studies showed that the association between the idiotype and the antigen-binding site can imply a similarity between the anti-idiotype and the antigen. When a hormone such as insulin served as the antigen, the resulting anti-idiotype was able to compete with the hormone itself for binding to its cell-surface receptor, which suggested the Ab-2 resembled the antigen. Because it arises within the immune system and yet mimics a molecule that originates elsewhere, such an Ab-2 is said to bear the "internal image" of the antigen. It was found, however, that these internal-image anti-idiotypes are a minority; most anti-idiotypes do not mimic the effects of the original antigen.

The idiotype, moreover, does not always coincide with the combining site; not all anti-idiotypes are prevented from binding when the Ab-1 is incubated with antigen beforehand. The combining site of an antibody consists of amino acid sequences that vary considerably among antibody molecules; sandwiched among these segments are "framework" regions that vary much less. An Ab-2 that recognizes antigenic determinants on a framework region still qualifies as an anti-idiotype, since it will bind only to a specific Ab-1. But because it does not interact with the combining site, it is not inhibited by the presence of bound antigen. Such an anti-idiotype need not bear any resemblance to the antigen.

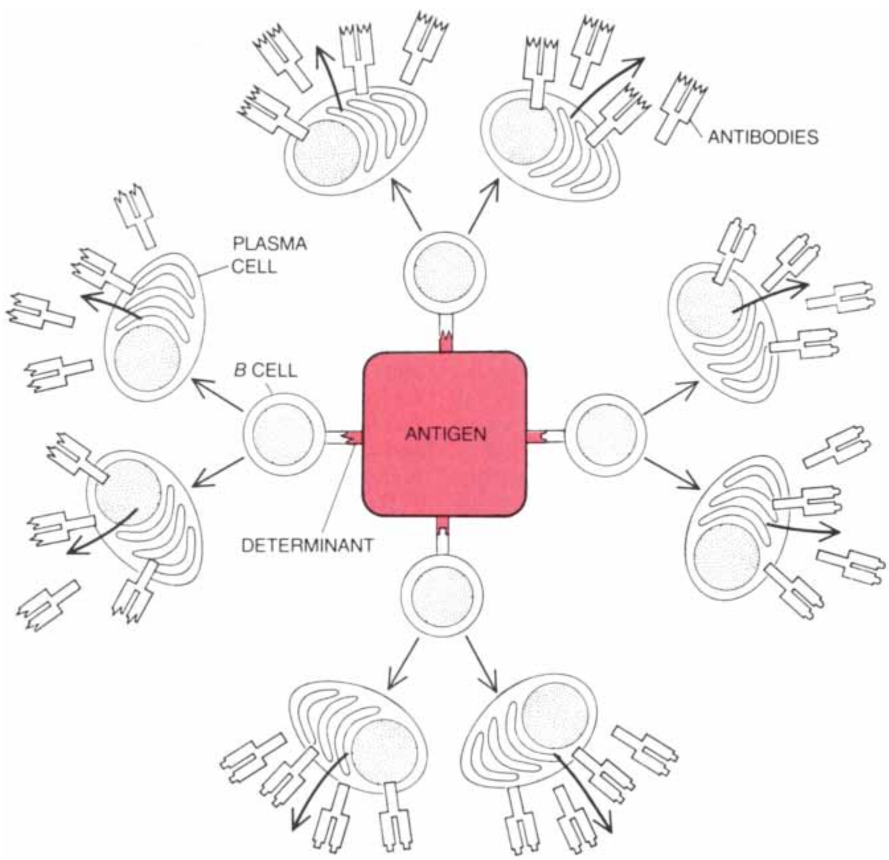
In the studies that established the concept of idiotype and anti-idiotype experimental animals were challenged with antibody generated in other experimental animals. Do similar interactions occur naturally, within a single immune system? In the early 1970's Niels Kaj Jerne of the Basel Institute for Immunology proposed that they do—that a network of idiotypes and anti-idiotypes regulates the immune response. Jerne suggested that idiotypes are displayed on antibodies and the surface receptors of *B* cells and *T* cells. (*T* cells are the other class of lymphocytes, which kill foreign cells directly and have regulatory roles [see "The *T* Cell and Its Receptor," by Philippa Marrack and John Kappler; SCIENTIFIC AMERICAN, February].) When, in response to an antigen, *B* cells and *T* cells proliferate and antibody levels rise, the concentration of idiotypes increases.

The body is normally "tolerant" of molecules of its own manufacture: it does not mount an immune response against them. Jerne proposed that individual idiotypes ordinarily exist at a level too low for tolerance to develop. Hence the rising concentration of idiotypes stimulates the proliferation of other sets of lymphocytes bearing anti-idiotype receptors.

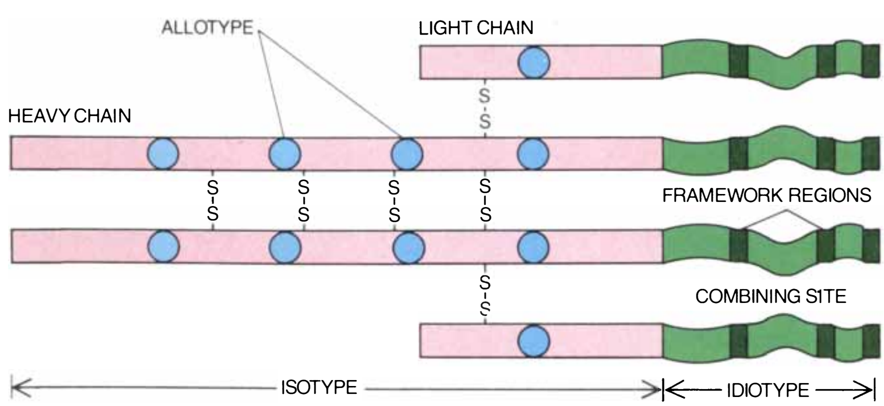
The *B* cells in the second population



BINDING OF ANTI-IDIOTYPE to its target antibody is signaled by green color in an enzyme-linked immunosorbent assay (ELISA). The plastic wells were first coated with various samples of human or mouse antibodies. Rabbit serum containing an anti-idiotype made against human antibody to hepatitis-*B* antigen was incubated in the wells and then washed off. Next the wells were exposed to a goat antirabbit antibody to which the enzyme horseradish peroxidase had been coupled; last they were filled with a colorless solution that changes to green in the presence of horseradish peroxidase. The color reveals bound goat antibody, which in turn indicates the presence of bound anti-idiotype. Hence the samples in the colored wells contain the anti-idiotype's target, antibody to hepatitis-*B* antigen.



ANTIBODY RESPONSE to an antigen, or foreign molecule, ensues when *B*-cell clones, each consisting of one or a few identical lymphocytes, or white blood cells, recognize surface patterns known as antigenic determinants. The membrane receptors of each *B*-cell clone bind to a specific determinant. The clone then multiplies and develops into a population of plasma cells, which secrete free antibody molecules. Each Y-shaped antibody molecule has two combining sites, one on each arm, which have the same specificity as the receptors of the parent clone. The combining sites enable the antibodies to bind to other molecules of antigen and effect their destruction or neutralization. Because every antigen bears many determinants, many *B*-cell clones collaborate in the antibody response. The diagram simplifies the actual process; in fact *B* cells require aid from other cells, including the kind of white blood cells known as *T* cells, if they are to mount an antibody response.



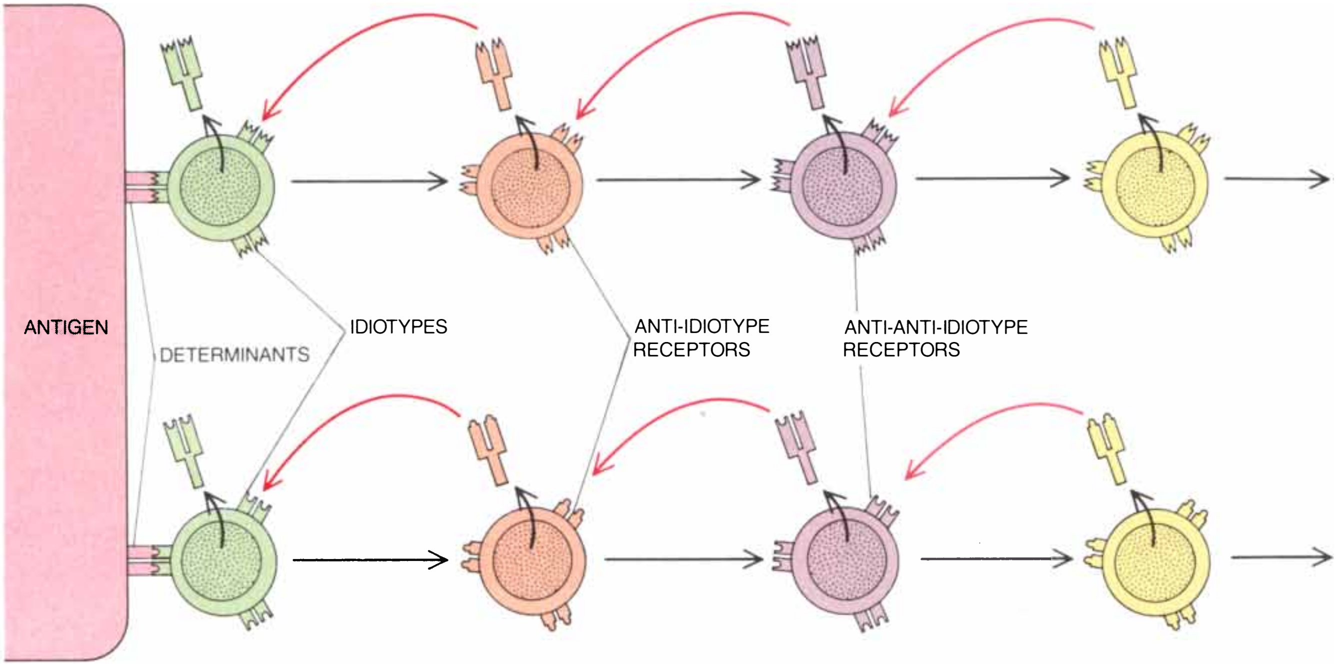
MAP OF AN ANTIBODY MOLECULE reveals two heavy, or long, chains of amino acids flanked by two light, or shorter, chains. Each chain is divided into a constant region (*red*) and a variable region (*green*). The constant regions are similar in all antibodies belonging to the same broad class, but antibodies from a given *B*-cell clone are distinctive at their variable regions. The segments that make up the combining site vary the most among clones; the intervening framework regions vary less. Each region of an antibody bears antigenic determinants that can stimulate the production of other antibodies. The constant regions carry determinants that are identical for antibodies belonging to the same class; they make up an antibody's isotype. Determinants that are identical for all the antibodies from one individual but that vary among individuals also occur in the constant regions (and occasionally in the variable regions as well). They constitute an antibody's allotype. Determinants that lie only in the variable regions, either at the combining site or in framework regions, are known as the idiotype. Antibody from each *B*-cell clone has a distinct idiotype.

of lymphocytes secrete anti-idiotype antibodies, which regulate the initial immune response. Jerne suggested that they do so either directly, for example by binding to the idiotype-bearing antibodies and inactivating them, or indirectly, by binding to idiotypes on the surface of *T* cells that regulate other lymphocytes. The effects of anti-idiotypes on the immune response may be either stimulatory or suppressive.

The interactions triggered by an antigen, Jerne proposed, do not end with the anti-idiotype (Ab-2), since anti-idiotypes themselves bear idiotypes. The huge diversity of antibodies has the result that every determinant in an idiotype is recognized by the combining site of some other antibody. Thus the anti-idiotype antibodies evoke anti-anti-idiotype antibodies, or Ab-3, which stimulate the production of anti-anti-anti-idiotypes, or Ab-4, and so on. The initial antigen thereby triggers a far-reaching perturbation in an intricate network [see "The Immune System," by Niels Kaj Jerne; SCIENTIFIC AMERICAN, July, 1973].

The first experimental evidence that such an idiotype network may operate within an individual immune system came in 1974. Leo S. Rodkey, who is now at the University of Texas Health Science Center in Houston, exposed rabbits to an antigen and purified the antibody they made in response (Ab-1). He allowed the animals' immune system to rest for 14 months and then reinjected Ab-1 into the individual animals from which it had come. He tested for an Ab-2 response in their serum by means of a radioimmunoassay. Ab-1 was labeled with radioactive iodine and treated with an enzyme that removed the effector region of the immunoglobulin chains, a crucial part of their constant regions, but left their idiotype-bearing variable regions intact. The Ab-1 was mixed with serum and exposed to a goat immunoglobulin capable of binding to the effector region on any rabbit antibody and causing a precipitate to form. Because the effector region had been removed from the Ab-1, it could not react on its own with the goat immunoglobulin. An unaltered antibody from the rabbit serum had to be bound to the Ab-1.

The formation of a radioactive precipitate indicated that a second, intact antibody was bound to the Ab-1. To confirm that the intact antibody was indeed an anti-idiotype, Rodkey incubated radioactively labeled Ab-1 with the original antigen and repeated the assay. This time much less Ab-1 was precipitated by the goat immunoglobulin, which indicated that the second antibody's target was the variable re-



IDIOTYPE NETWORK regulates the immune response to an antigen, according to a scheme proposed by Niels Kaj Jerne of the Basel Institute for Immunology and supported by recent findings. An antigen stimulates the proliferation of specific lymphocytes that recognize determinants on its surface. The surface receptors of the cells and the combining sites of the antibodies they secrete bear idiotypes, which cause a second population of lymphocytes, having anti-idiotype receptors that interact with the idiotypes, to expand. Those cells also bear idiotypes, which stimulate a third population of lymphocytes. Jerne proposed that a disturbance propagates through the idiotype network as such interactions are repeated indefinitely, but it is now believed they may be limited to a finite number. The interactions also take place in reverse: each population of cells secretes anti-idiotype antibodies that bind to idiotypes on the preceding population of cells and antibodies (*colored arrows*). Those reactions can shut down the immune response or bolster it.

gion of the Ab-1. Rodkey called the Ab-2 auto-anti-idiotype. Since then other investigators have demonstrated that auto-anti-idiotypes are secreted during a normal immune response.

Constantin A. Bona of the Mount Sinai School of Medicine made a more extensive demonstration of an idiotype network in mice. A monoclonal antibody (an antibody made in culture by a single clone of antibody-secreting cells) that was specific for a bacterial sugar served as the Ab-1; Bona injected it into the mice to produce an Ab-2, which he isolated and reinjected into the animals. In this way he elicited an Ab-3 and then, by repeating the cycle, an Ab-4. Just as others have shown that an Ab-2 can mimic the antigen, so Bona found patterns of resemblance in his network. Since the Ab-4 and the Ab-2 both react with the Ab-3, Bona reasoned, their combining sites should be similar. Therefore the Ab-4, like the Ab-2, should bind to the Ab-1. Indeed, 60 percent of Bona's Ab-4 did bind to the Ab-1, the initial idiotype.

Other investigators have found evidence that, as Jerne postulated, the idiotype network embraces *T* cells as well as *B* cells. In one study *T* cells proliferating in response to an antigen were taken from an experimental animal and injected into another animal. Antibodies to idiotypes on the *T*-cell receptors were isolated from the serum of the second animal and injected into the first one. Their effect was to inhibit the immune response. By binding to surface receptors on *T* cells of the class whose function is to suppress the immune response, the anti-idiotypes may have triggered suppression.

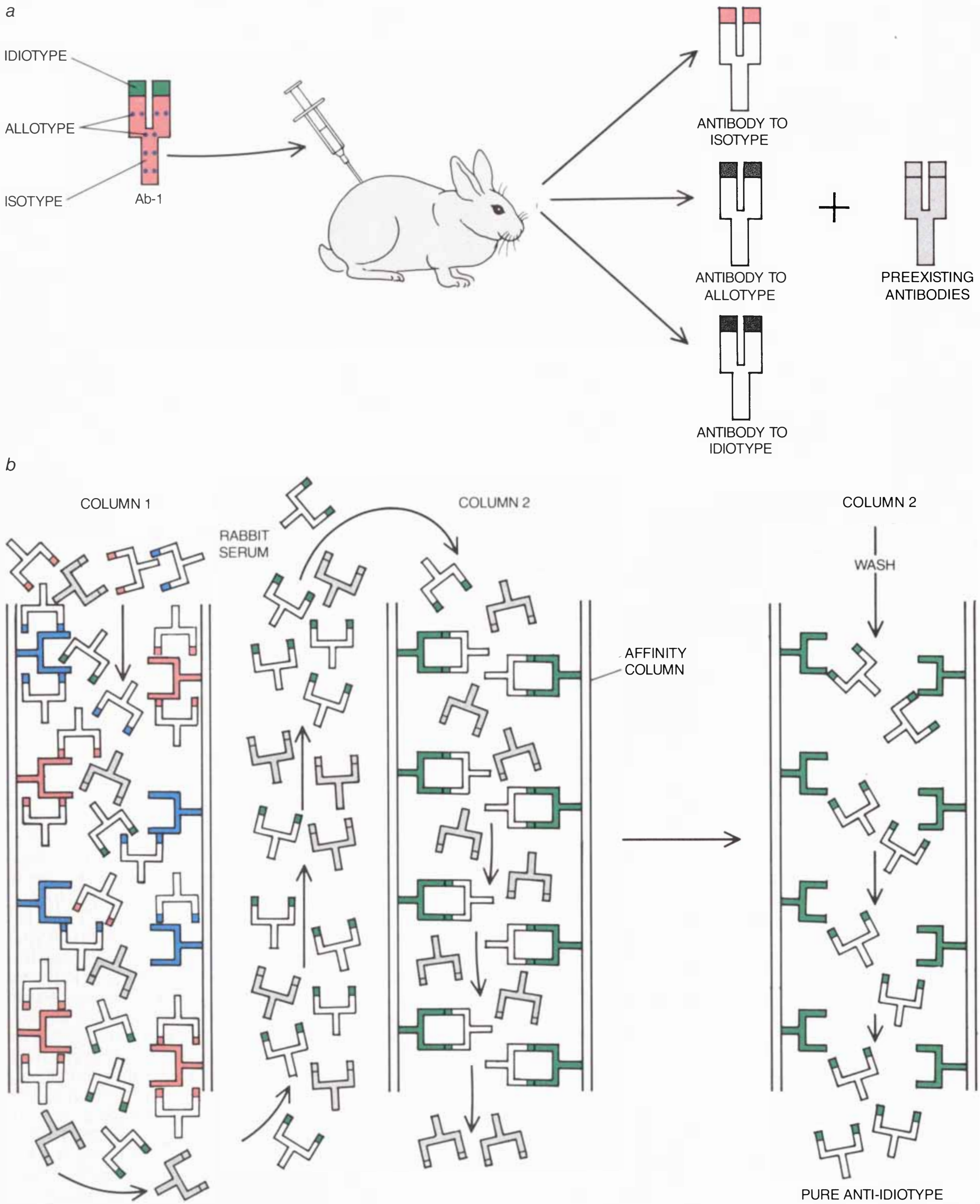
Our group has investigated the idiotype networks that develop in human beings infected with hepatitis-*B* virus (HBV). The motivation for the work is practical as well as scientific: HBV causes a serious primary infection, and chronic infection often leads to the development of other liver diseases and primary liver cancer. An understanding of the immune networks stimulated by the virus, we thought, might suggest immunological means of preventing the disease.

Liver cells infected with HBV release large quantities of viral envelope protein. The envelope material is antigenic and is known as hepatitis-*B* surface antigen (HBsAg). HBsAg contains three distinct antigens, designated by letters. The *a* antigen is present in every strain of HBV. The other two antigens vary in a consistent way: in every strain the *d* or the *y* antigen is paired with the *w* or the *r* antigen. Antibody to the *a* antigen provides protection against reinfection with HBV of any strain [see "Viral Hepatitis," by Joseph L. Melnick, Gordon R. Dreesman and F. Blaine Hollinger; SCIENTIFIC AMERICAN, July, 1977].

We asked whether human antibodies produced in response to HBsAg display a variety of idiotypes, corresponding to combining sites recognizing the various antigens. If they do, our task of unraveling the idiotype-anti-idiotype reactions initiated by infection would be complicated considerably and our chance of finding a way to manipulate them would be reduced.

We demonstrated, however, that human antibody molecules made after exposure to HBsAg through hepatitis-*B* infection or vaccination all express a common idiotype. Our strategy was to inject antibody from a single individual into rabbits, from whose serum we then isolated an anti-idiotype. Antibodies from several different people were equally capable of reacting with the anti-idiotype, suggesting the antibodies bore the same idiotype. That idiotype, we found, is induced by the *a* antigen. We tested HBsAg from each of several viral strains individually for its ability to block the reaction of anti-idiotype with human antibody. All the strains were equally effective at inhibiting the reaction, which indicated that the idiotype was binding to an antigenic determinant shared by all the strains. That determinant must reside on the *a* antigen.

Does the rabbit anti-idiotype bear



MANUFACTURE OF ANTI-IDIOTYPE to a human antibody against hepatitis-*B* virus begins with the injection of the purified antibody (Ab-1) into a rabbit (*a*). Serum extracted from the rabbit contains preexisting antibodies, antibodies to the isotype and allotype of the human antibody, and anti-idiotype. The unwanted antibodies are removed in affinity columns, which contain a matrix to which other antibodies have already been allowed to bind (*b*). One column has been treated with human antibodies that have the same isotype and allotype (*red and blue*) as the Ab-1 but a different idiotype; the second column has been treated with Ab-1, the original idiotype (*green*). The rabbit serum is passed through the first column, which removes the anti-isotype and anti-allotype antibodies, leaving the anti-idiotype and the preexisting antibodies. Passage through the second column removes the anti-idiotype, leaving only the extraneous antibodies. The second column is then flushed with a compound that releases bound antibody, yielding pure anti-idiotype.

an internal image of the *a* antigen? The fact that HBsAg can prevent the anti-idiotype from binding to the human Ab-1 suggests it does. But HBsAg is a large molecule; it might block an anti-idiotype reaction sterically—simply by virtue of its bulk—even if the anti-idiotype interacted with a framework region of the Ab-1 rather than with the combining site itself. We therefore tested a much smaller molecule—a synthetic peptide (a short sequence of amino acids) mimicking one of the determinants in the *a* antigen—and found that it too could inhibit the binding of the anti-idiotype.

Other observations also suggested that the rabbit anti-idiotype mimics the shape of the hepatitis-*B* antigen. Antibodies that all recognize the antigen but are made in different species may differ in their biochemistry, even when the shape of their combining site is the same. Hence an anti-idiotype might react with Ab-1 from only one species, responding to the chemical makeup of the variable region rather than the shape of the combining site. We found, however, that the rabbit preparation was able to bind to Ab-1 produced in mice, chimpanzees and five other species of mammals besides human beings. Evidently the anti-idiotype, like the hepatitis-*B* antigen itself, interacts with the combining site of the Ab-1 on the basis of shape.

Could the rabbit anti-idiotype simulate the protective effect of conventional hepatitis-*B* vaccine, which consists of HBsAg? The notion that an idiotype bearing an internal image could serve as a vaccine against an infectious agent had been proposed by Nisonoff and Edmundo Lamoyi, now at the National Institute of Allergy and Infectious Diseases, and independently by Ivan M. Roitt and his colleagues at the Middlesex Hospital Medical School in London. Early results that seemed to confirm the possibility included work on skin-graft rejection by Jeffrey Bluestone and David H. Sachs at the National Cancer Institute. In rejecting foreign tissue the immune system responds to proteins known as histocompatibility antigens on the membrane of the foreign cells. Bluestone and Sachs isolated antibodies to histocompatibility antigens from the serum of mice that had received skin grafts and then generated an anti-idiotype in other species.

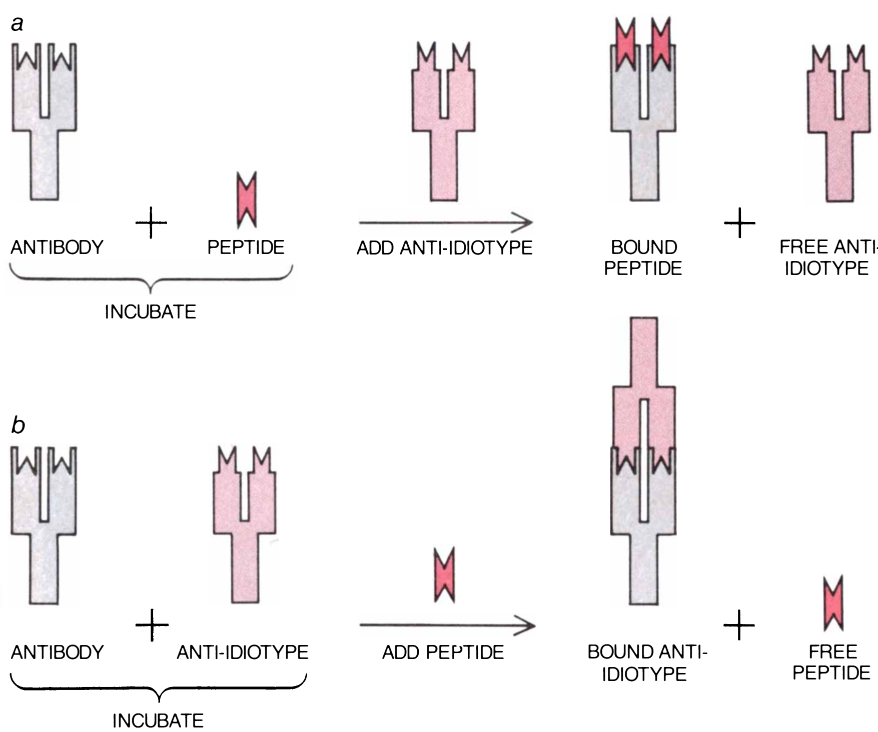
When mice that had not been exposed to foreign tissue were inoculated with the anti-idiotype, they secreted antibodies that could bind to histocompatibility antigens. Apparently the anti-idiotype preparation contained internal images of the histocompatibility antigens. The internal-image Ab-2 must have stimulated the immune system to make an Ab-3 with a specificity similar to that of an Ab-1.

In a second study bearing on the possibility of anti-idiotype vaccines, David L. Sacks of the National Institute of Allergy and Infectious Diseases injected mice with a monoclonal Ab-1 that recognized a surface antigen on *Trypanosoma,* a protozoan that in humans causes African sleeping sickness. Other mice given the resulting anti-idiotype were protected against a dose of *Trypanosoma* that would ordinarily be lethal. Later studies showed, however, that the anti-idiotype protected mice only if they were closely related to the mouse that was the source of the monoclonal Ab-1.

Such genetic restriction suggested that the anti-idiotype did not bear an internal image of the antigen; otherwise it should have been able to stimulate the immune system of any mouse to make protective antibodies. Presumably the idiotype that induced the Ab-2 was associated with parts of the variable region whose makeup differed in genetically diverse mice, even on equivalent antibodies. Thus the anti-idiotype could induce a protective Ab-3, identical with the monoclonal Ab-1 in its antigen specificity, only in the same inbred mouse strain that provided the Ab-1. (Recently Sacks and his colleagues succeeded in making an anti-idiotype that does appear to bear an internal image of a *Trypanosoma* antigen and should not be subject to such genetic restriction.)

As a first step in determining whether our anti-idiotype could serve as a vaccine against hepatitis *B* we investigated the effect of pretreatment with the anti-idiotype on a subsequent immune response to HBsAg. We injected groups of mice with either our rabbit anti-idiotype or, as a control, a nonspecific rabbit immunoglobulin preparation. We then exposed the animals to HBsAg and analyzed their antibody response by means of a hemolytic-plaque assay. *B* cells from the spleens of the mice were placed in layers of sheep red cells precoated with HBsAg. When complement, a system of blood proteins that ruptures cells to which antibody is bound, was added, the appearance of clear zones called plaques indicated the destruction of red cells near antibody-producing *B* cells.

To our pleasant surprise significant-



EQUIVALENCE of hepatitis-*B* surface antigen (HBsAg) and an anti-idiotype made by injecting human antibody to hepatitis *B* into rabbits was demonstrated by the authors. When another batch of the human antibody is incubated with a synthetic peptide (a short chain of amino acids) that mimics a crucial antigenic determinant on HBsAg, the anti-idiotype can no longer react with the antibody (*a*), suggesting that the idiotype lies at the occupied antigen-binding site of the antibody. Conversely, incubation with the anti-idiotype prevents the human antibody from subsequently reacting with the peptide (*b*). By showing that the anti-idiotype and the antigen interact with the same site on the human anti-hepatitis antibody, the results suggest that at its combining site the anti-idiotype may be structurally similar to the antigen: the anti-idiotype may bear an "internal image" of the antigen.

ly more plaques appeared around spleen cells from mice that had been pretreated with the anti-idiotype than around cells from control mice. The pretreatment, it seemed, had bolstered the later immune response. We also detected antibody secretion in cells from a group of mice that had received the anti-idiotype alone, without a later injection of HBsAg. A separate assay, in which the serum rather than the spleen cells of the mice was tested, confirmed that the anti-idiotype by itself was able to elicit antibody capable of reacting with the viral antigen.

Are the Ab-3 antibodies induced by the anti-idiotype (Ab-2) equivalent to those made during a natural infection with HBV? Those antibodies recognize the *a* antigen, which is common to all strains of the virus. To find out whether the antibodies induced by the anti-idiotype also react with the *a* antigen we isolated the Ab-3 by passing serum from treated mice through a column lined with HBsAg and extracting the bound Ab-3 from the column. We then tested the ability of HBsAg from various viral strains to inhibit the binding of the Ab-3 to the Ab-2. In every case binding was inhibited, which indicated that the combining site of the Ab-3, like that of the human Ab-1, recognizes the *a* antigen. We also compared the mouse Ab-3 and the human Ab-1 directly, by allowing them to compete for binding to the rabbit Ab-2. The human antibody, which is protective against HBV, and our experimental mouse Ab-3 bound to the anti-idiotype with comparable reactivity.

These findings suggested to us that the Ab-3 induced by the anti-idiotype can confer protective immunity to hepatitis *B*. In order to test the potential of rabbit anti-idiotype as a hepatitis vaccine for humans, however, we needed an appropriate animal model. HBV does not actually cause disease in mice, even though the animals do mount an antibody response when they are injected with HBsAg. Chimpanzees do develop hepatitis *B* and so are the appropriate animal in which to test a vaccine against the pathogen.

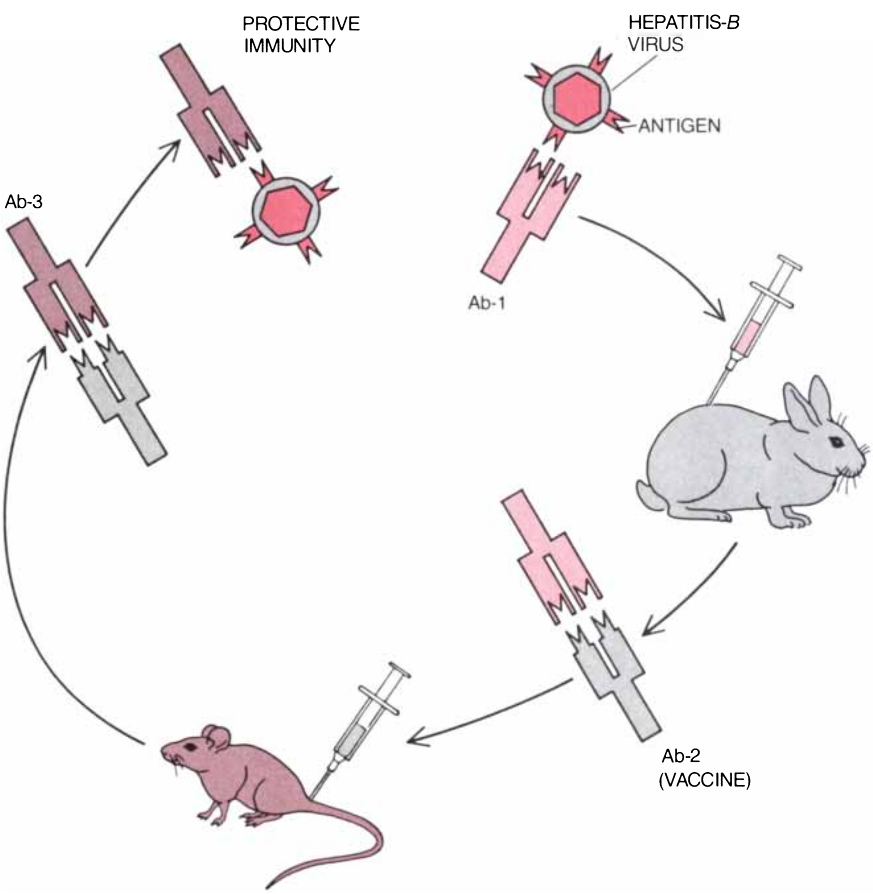
In collaboration with Jorg W. Eichberg and Robert E. Lanford of the Southwest Foundation for Biomedical Research, we injected two chimpanzees with the rabbit anti-idiotype and two with a control preparation made of nonspecific immunoglobulin. We challenged all four animals with infectious HBV. The serum of control chimpanzees showed elevated levels of liver enzymes and other markers of infection, but the animals that had received the anti-idiotype failed to develop any signs of disease.

Our work with chimpanzees is the most direct demonstration up to now of the vaccine potential of an anti-idiotype. There are also preliminary indications that anti-idiotypes could serve as vaccines against other diseases. Pretreatment with anti-idiotypes has been shown to protect mice partially or completely against a range of infectious agents: Sendai virus and reovirus (which cause mild brain infections), rabies virus, herpes simplex virus, poliovirus and a number of disease-causing bacteria and protozoa.

If anti-idiotype vaccines prove feasible in humans, they would have a number of advantages. Current vaccines consist either of the infectious agent itself, killed or attenuated, or of purified antigen. Many antigens, such as HBsAg, are difficult to purify, although protein antigens can now be made through the techniques of recombinant DNA. The antigenic parts of many bacteria and protozoa consist of carbohydrates or fats, however, and cannot be made by recombinant methods or synthetically. The alternative of administering the pathogen itself is not always attractive; attenuated organisms, for example, sometimes regain their virulence. Anti-idiotype vaccines would be easy to produce through monoclonal-antibody technology and could not cause the disease they are designed to prevent.

Their specificity is also appealing. An anti-idiotype triggers an antibody response only to the one antigenic determinant whose internal image it bears. An attenuated infectious agent, in contrast, may carry a number of antigenic determinants besides the one that induces protective immunity. Some may resemble determinants on tissue in the body itself and hence may trigger an autoimmune response.

Anti-idiotype vaccines might be of particular value in newborn infants, whose immature immune system cannot make antibodies to the envelope material of certain pathogenic bacteria, which consists of a sugar. Conventional vaccines against the bacteria are



ANTI-IDIOTYPE VACCINE against hepatitis *B* uses an internal image of the viral antigen. Human antibody (*Ab-1*) to the antigen is injected into a rabbit, which makes an anti-idiotype (*Ab-2*). The anti-idiotype is injected into a mouse or a chimpanzee, which makes antibody (*Ab-3*) to the anti-idiotype. Because of the similarity in shape of the hepatitis antigen and the anti-idiotype, the third antibody structurally resembles the first antibody, even though they differ biochemically. The similarity makes the Ab-3 able to react with the viral antigen and protect a vulnerable animal, such as a chimpanzee, against infection.

ineffective in newborns. An anti-idiotype could mimic the structure of the envelope material, but because the anti-idiotype is a protein, it would elicit an immune response. The effectiveness of such anti-idiotype vaccines has already been shown in newborn experimental animals.

Several factors could limit the use of anti-idiotypes as human vaccines. The injection of an anti-idiotype serum made in another animal, such as a rabbit, could cause fever and allergic reactions in some people. Moreover, because of the complexity of the idiotype network, the effect of an anti-idiotype may not always be salutary. Many anti-idiotypes appear to elicit a response equivalent to the one evoked by the antigen itself; others, however, may exert the suppressive effects on the immune system that were postulated by Jerne. We observed that the administration of an anti-idiotype made against mouse monoclonal antibody that was specific for herpes simplex virus significantly shortened the survival of mice later infected with the virus.

Anti-idiotypes may prove to be effective against cancer, perhaps as both a vaccine and a treatment. Cancers of the immune system can arise in a single clone of *B* or *T* cells; in such cancers all the cells express the same idiotype on their surface. A suitable anti-idiotype could in theory eliminate all the cancerous cells while leaving normal tissue unharmed. Ronald Levy and his co-workers at the Stanford University School of Medicine used specific anti-idiotypes made in mice to treat many patients suffering from *B*-cell lymphomas. One patient has been in complete remission for five years, and about half of them have shown clinically significant responses. Recent studies by Levy's group and by Jeffrey Cossman and his colleagues at the National Cancer Institute indicate, however, that other cancer cells expressing a different idiotype can emerge during anti-idiotype treatment. Coping with the population of variant cells might require the administration of multiple anti-idiotypes.

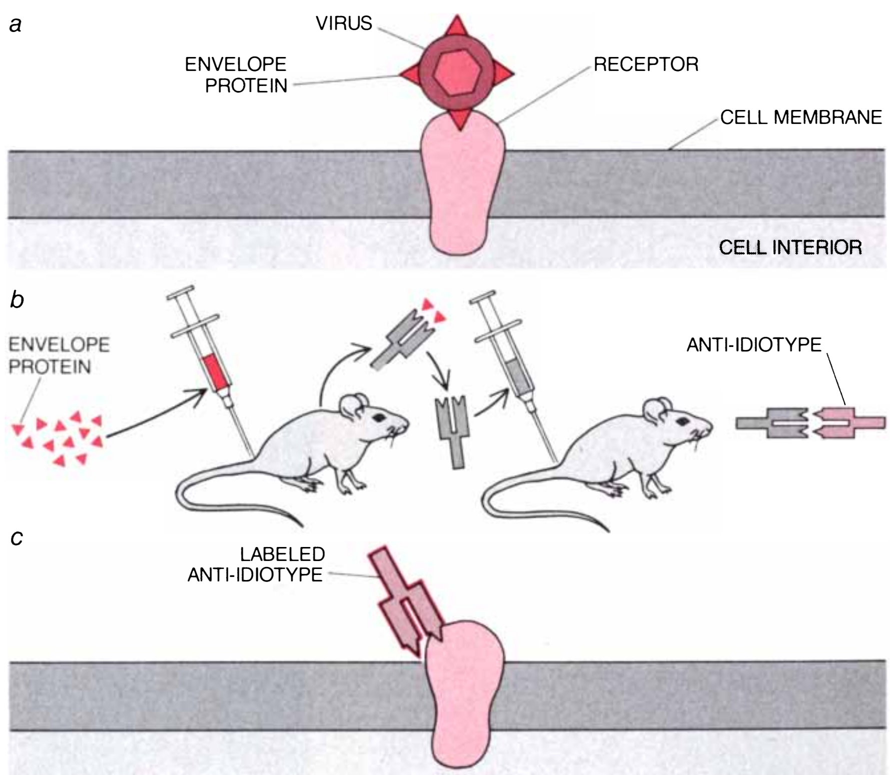
Keith A. Krolick of the University of Texas Health Science Center at San Antonio and Ellen S. Vitetta and Jonathan W. Uhr of the University of Texas Health Science Center at Dallas took a different approach to treating cancer with anti-idiotypes. They coupled ricin, a plant toxin, with an anti-idiotype and found that the preparation destroyed mouse leukemic *B* cells in culture. The anti-idiotype bound to the surface receptors of the *B* cells and thereby positioned the toxin molecules to destroy the cancer cells specifically. The workers also tested their preparation in living mice afflicted with *B*-cell leukemia and observed that it slowed the growth of the cancer.

It may even be possible to mobilize the body's own anti-idiotypes in fighting cancer. In a study done by Dorothee Herlyn and Hilary Koprowski of the Wistar Institute of Anatomy and Biology, patients with cancer of the colon and cancer of the rectum improved when they were treated with a mouse monoclonal antibody to an antigenic determinant on the surface of the cancer cells. Many of the patients made anti-idiotype to the mouse antibody. Koprowski proposed that the anti-idiotype carried an internal image of the cancer antigen and thereby stimulated the patients' immune system to react against the cancer itself.

Recently the group reported producing such an anti-idiotype in laboratory animals. The anti-idiotype bore an internal image of the major colon-cancer tumor antigen. When the anti-idiotype was injected into other animals, it elicited antibody (Ab-3) that could react with the tumor antigen and that was identical with the mouse monoclonal antibody (Ab-1), which had induced the anti-idiotype.

The unusual manipulations to which anti-idiotypes lend themselves also make them valuable as a tool for research into disease mechanisms. Anti-idiotypes have served, for example, to pinpoint the site on the cell surface through which a virus infects a cell. Anti-idiotype made against antibody recognizing the part of the virus's protein envelope that adheres to the cell membrane can mimic the crucial envelope molecule. Such an anti-idiotype can seek and bind to the matching receptor on the cell surface. Mark I. Greene and Bernard N. Fields, who are now respectively at the University of Pennsylvania School of Medicine and at the Harvard Medical School, adopted that strategy to identify the receptor for reovirus on the surface of nerve cells and cells of the immune system. The same method has enabled us to identify what is probably the receptor for HBV on liver cells.

The results of anti-idiotype studies have suggested the possibility that idiotype-anti-idiotype reactions may



SITE OF VIRAL ATTACK on a cell surface can be pinpointed with an anti-idiotype. For a virus to infect a cell, part of its protein envelope must bind to a receptor: a complementary protein on the cell surface (*a*). The viral protein can be injected into an experimental animal to make antibody, which in turn can induce an anti-idiotype when it is injected into another animal (*b*). If the anti-idiotype bears an internal image of the original antigen (the viral protein), it will be able to bind to the cell-surface receptor (*c*). The anti-idiotype can be labeled with a fluorescent or electron-dense substance so that its target on the cell surface will be identifiable under the light or the electron microscope. The strategy is valuable for identifying a cell-surface receptor when the virus itself is hard to isolate and use as a probe.

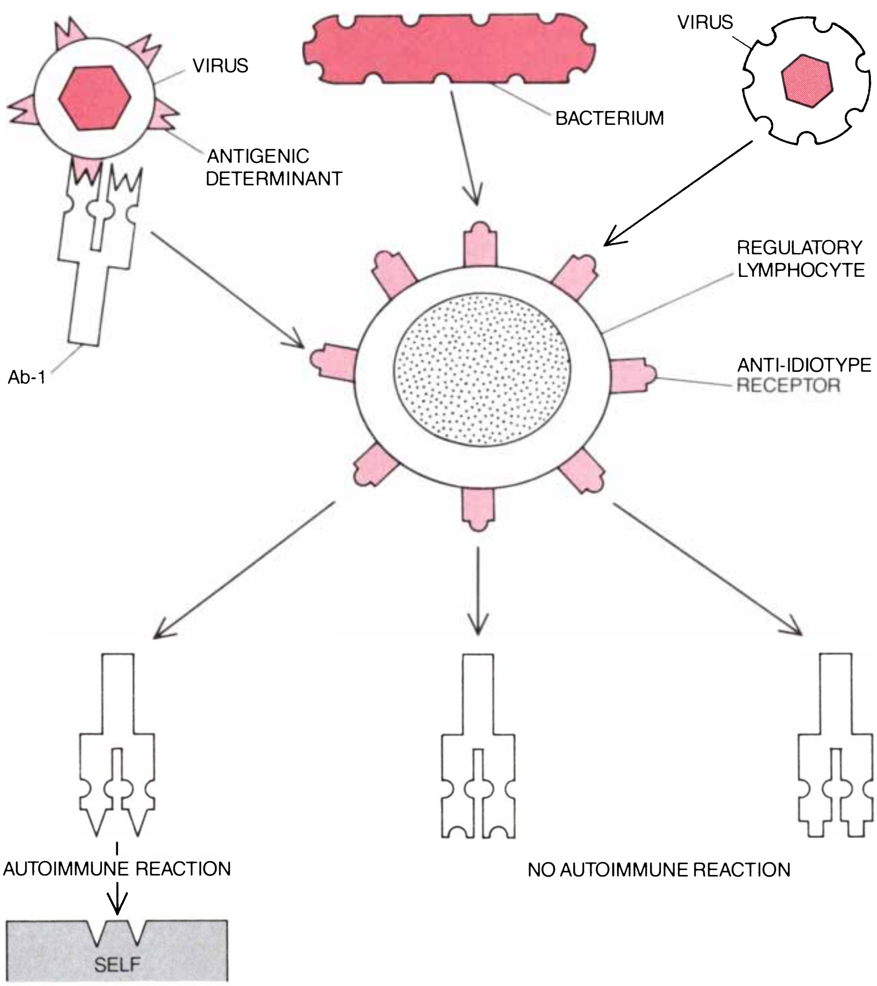
themselves precipitate some autoimmune diseases. One such disease is myasthenia gravis, a disorder in which muscles are afflicted with extreme weakness and fatigue. The symptoms reflect an immunological attack on the acetylcholine receptor, a cell-surface molecule through which muscles receive chemical signals from nerves.

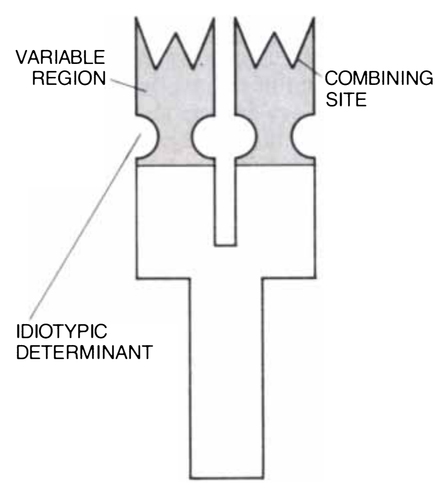
Bernard F. Erlanger, Norbert H. Wassermann and their colleagues at Columbia University sought to characterize the attack. The workers raised rabbit antibodies to a substance that binds to the acetylcholine receptor and then injected the antibodies into other rabbits to make an anti-idiotype. The anti-idiotype seemed to carry the image of a crucial part of the binding substance: it was able to bind to the rabbit acetylcholine receptor. In doing so it produced symptoms of myasthenia gravis in the animals. Recently W. L. Cleveland of Columbia University, along with Erlanger and Wassermann, induced the disease in mice by injecting them with cells secreting a similar monoclonal anti-idiotype.

In myasthenia gravis, then, an anti-idiotype somehow induced in the body may react with the acetylcholine receptor, blocking or destroying it. In Graves' disease, an autoimmune disorder marked by abnormal growth of the thyroid gland, an anti-idiotype may have the opposite effect. Workers have been able to induce the disease in experimental animals by administering an anti-idiotype made against antibody to thyrotropin, a hormone that stimulates thyroid growth. The anti-idiotype presumably mimicked the hormone and so was able to bind to its cell-surface receptor, stimulating thyroid growth. The same interactions, occurring spontaneously in the body, could bring about the human disease.

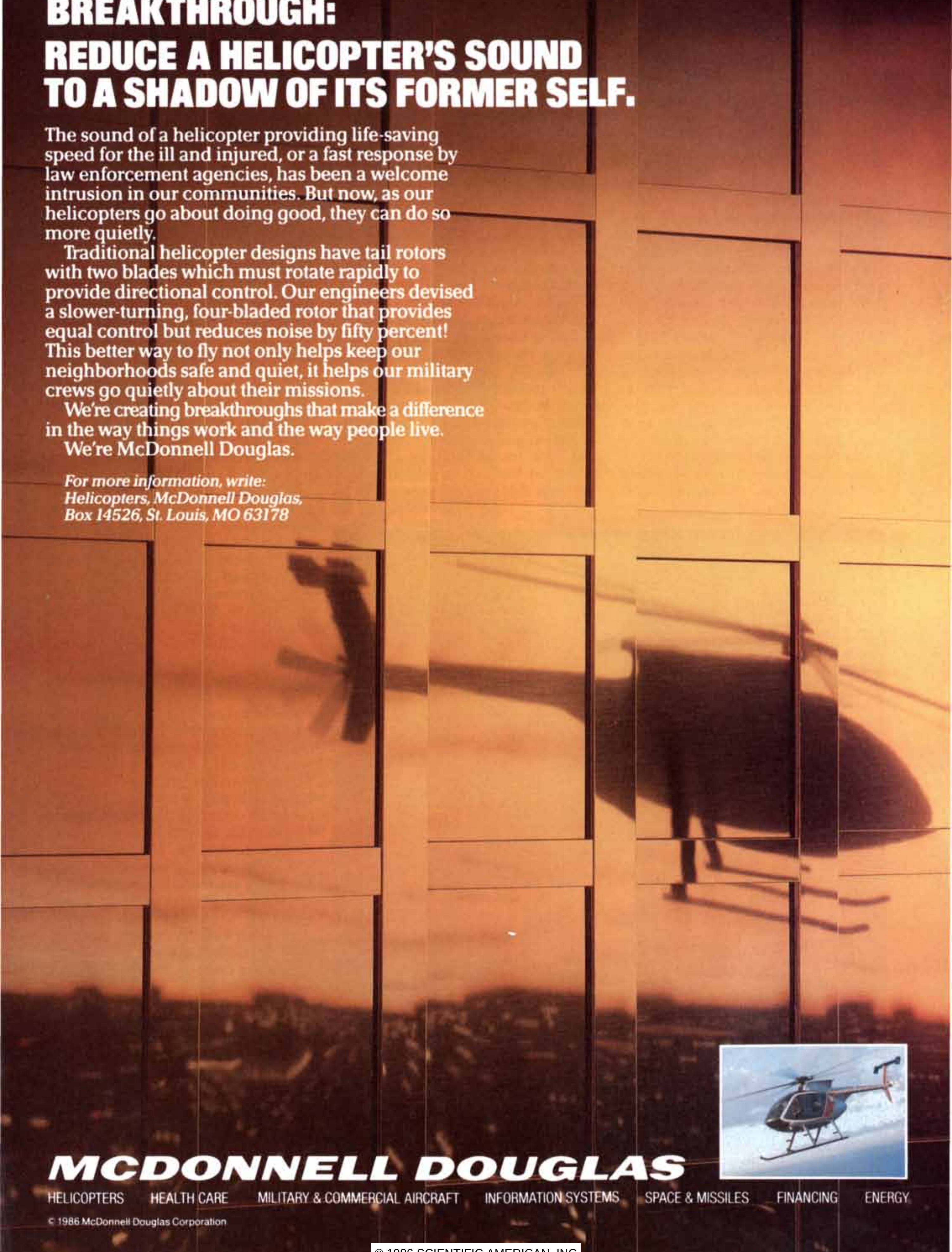
In two other autoimmune diseases, rheumatoid arthritis and systemic lupus erythematosus, a disorder of connective tissue, there are also some indications that anti-idiotypes participate. A sign of both diseases, and perhaps a fundamental aspect of their mechanisms, is the presence in the serum of particular antibodies: rheumatoid factors (antibodies to a segment of the immunoglobulin molecule) in rheumatoid arthritis and anti-DNA antibodies in systemic lupus erythematosus. Such antibodies have been shown to express a common idiotype in each victim, which suggests that they interact with a single anti-idiotype. It is conceivable that the anti-idiotype helps to initiate the disease by stimulating the production of the serum antibodies.

Thus the intricacy of idiotype-anti-idiotype reactions opens the way to the development of complex and unlikely disorders. It also makes subtle preventive and therapeutic interventions in the immune system conceivable. Manipulating the immune network by means of anti-idiotypes in order to prevent certain infections and to treat some kinds of cancers now seems within reach. A deeper understanding of idiotype-anti-idiotype reactions could make it possible to cure diseases in which the defect lies in the immune network itself. It is not unreasonable also to speculate that an anti-idiotype may in the future serve as a vaccine or a therapeutic agent for a very different kind of disease affecting immunity: AIDS.





AUTOIMMUNE DISEASE might arise from idiotype-anti-idiotype reactions. One proposed mechanism depends on the fact that, as shown at the right, an antibody may bear idiotypic determinants that do not lie precisely at its combining site. Such an antibody, produced in response to a bacterium or a virus (*top left*), may stimulate a population of regulatory lymphocytes bearing anti-idiotype receptors that do not mimic the original antigen. Alternatively, the regulatory lymphocytes might be stimulated not by the antibody to the pathogen but by the pathogen itself, that is, by a surface determinant resembling a host idiotype (*top center and right*). The regulatory cells would then induce the production of a new set of antibodies (*bottom*). Because the idiotypic determinant is unrelated to the combining sites of antibodies on which it is found, the new antibodies need not bear any functional similarity to the original antibody: they may have quite different specificities. If the target of the antibodies is host tissue, an autoimmune reaction ensues. Preliminary results suggest the possibility that this mechanism might precipitate rheumatoid arthritis and systemic lupus erythematosus.

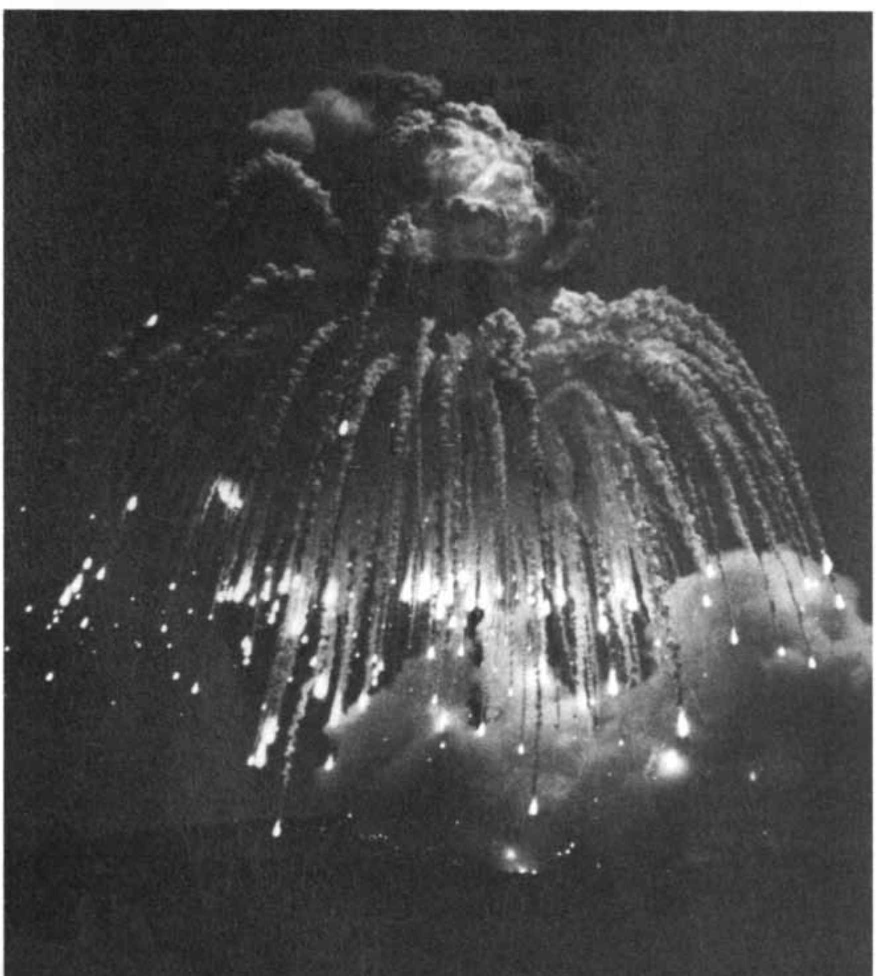

BREAKTHROUGH:
REDUCE A HELICOPTER'S SOUND TO A SHADOW OF ITS FORMER SELF.
The sound of a helicopter providing life-saving speed for the ill and injured, or a fast response by law enforcement agencies, has been a welcome intrusion in our communities. But now, as our helicopters go about doing good, they can do so more quietly.
Traditional helicopter designs have tail rotors with two blades which must rotate rapidly to provide directional control. Our engineers devised a slower-turning, four-bladed rotor that provides equal control but reduces noise by fifty percent! This better way to fly not only helps keep our neighborhoods safe and quiet, it helps our military crews go quietly about their missions.
We're creating breakthroughs that make a difference in the way things work and the way people live.
We're McDonnell Douglas.
For more information, write: Helicopters, McDonnell Douglas, Box 14526, St. Louis, MO 63178
MCDONNELL DOUGLAS
HELICOPTERS HEALTH CARE MILITARY & COMMERCIAL AIRCRAFT INFORMATION SYSTEMS SPACE & MISSILES FINANCING ENERGY
© 1986 McDonnell Douglas Corporation

SCIENCE AND THE CITIZEN

Out of Space

It is no secret that the U.S. space program is in trouble. Four accidents—the explosion of a Titan rocket carrying a Department of Defense satellite last fall, the loss of the space shuttle *Challenger* and its crew of seven in January, the explosion in April of another Titan (and the resulting damage to that rocket's launch pad) and the failure early in May of a Delta launcher carrying a weather satellite—have severely limited, at least temporarily, the U.S.'s ability to put satellites in orbit. The current crisis is leading to a full-scale examination of the nation's strategy in space. This examination has turned up deficiencies that may in the long run have even more serious consequences than the current shortage of launchers has.

How bad is the current situation? According to Albert Wheelon, an executive vice-president of the Hughes Aircraft Company who is a member of the commission examining the *Challenger* accident and of the President's Foreign Intelligence Advisory Board, "we have an unqualified disaster on our hands.... We are essentially out of business." All shuttle missions are on hold for at least a year and a half to two years; Delta and Titan missions are delayed until the causes of the recent accidents can be determined. Moreover, the Titan accident in April caused significant damage to the only operating Titan launch pad at Vandenburg Air Force Base, which is the only base in the continental U.S. from which satellites can easily be put in polar orbits. Since most spy satellites and military weather satellites require polar orbits, and since the Titan is the only U.S. rocket that can carry the largest satellites, those programs have had to be delayed until the pad can be repaired. That might take as long as a year.



EXPLOSION in April of a Titan rocket at Vandenburg Air Force Base caused heavy damage to the only launch pad able to put large satellites in polar orbits, the ones occupied by most spy satellites and military weather satellites. Photograph from Wide World.

Among other payloads affected by the recent series of accidents are communications satellites that were due to be launched for India, Indonesia, the British Ministry of Defence, Hughes and RCA; several scientific missions, including the *Galileo* and *Ulysses* planetary probes (which can be launched only in limited "windows" of time) and the Space Telescope; diverse payloads for the U.S. Department of Defense, and several weather satellites. The weather satellites are crucial because the failure in 1984 of the *GOES-5* weather satellite and the destruction of the *GOES-7* that was on board the failed Delta have left the National Oceanic and Atmospheric Administration with only one satellite to cover the entire country, in place of the two (one on each coast) that are needed for full coverage. (The remaining satellite is now shuttled back and forth from coast to coast to cover major meteorologic events.) According to Chester M. Lee, the head of customer services at the National Aeronautics and Space Administration, "some of those satellites will be delayed, and some may never get up."

Many analysts point out that the current situation is not quite as bad as has been suggested; it is not completely impossible for the U.S. to launch critical payloads. Two Atlas launchers, one carrying a weather satellite (but not a replacement for the *GOES-5*) and the other a naval fleet-communications satellite, are scheduled to be launched in mid- to late June. (The Atlas is not an acceptable replacement for the Titan because it can carry only about a third of the Titan's payload.) More significant, a number of older launchers are available that could be reconfigured, in an emergency, to carry certain smaller payloads. According to one industry source, it would even be possible to make available an emergency launcher for a large polar-orbit satellite.

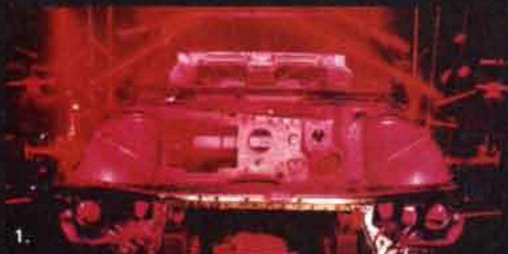
The real problem is cost. Modifying a launch pad to make it compatible with an emergency, jury-rigged launcher would require a minimum expenditure of \$40 to \$50 million. Modifying the vehicles themselves would cost as much as \$60 to \$70 million per vehicle, even if the costs of tooling up to reconfigure the rockets could be spread over 10 to 12 vehicles. The emergency polar-orbit launcher, according to the industry source, "would be a hand-built system. You

DESIGNING GM BYTE BY BYTE.

Designing General Motors to become the first 21st century corporation means going back to the drawing board and looking at ourselves in the light of a new age: the Computer Age.

It means thinking in a new mode, accessing the future in a daring, creative new way.

Our goal? A sleeker, more streamlined, computer-driven GM—an organization powered by technology, fueled by brainpower, and outclassing all competition. Our inventive use of computer technology in design, engineering, manufacturing, and safety is producing a GM programmed for quicker response to our customers, better efficiency and outstanding performance. A GM designed to bring you into the future. Byte by byte.



1.

2.

THE GM ODYSSEY: SCIENCE NOT FICTION

GM

3.

4.

5.

1. AUTOMOTIVE DIMENSIONAL CHECKER. Probing car bodies to assure solid, tight-fitting assemblies are one hundred and twenty-two lasers and cameras.
2. ADVANCED CONCEPTS CENTER. Where creative people work together studying lifestyles of today's consumer to help provide for transportation needs of tomorrow.
3. INFORMATION PROCESSING CENTER. Electronic Data Systems control center helping to streamline data processing and telecommunications functions, enabling General Motors to become more responsive to its customers' needs.
4. INSTRUMENT PANEL. Touch-sensitive cathode-ray tube with multiple functions that include diagnosing service problems in seconds.
5. MULTIMATCH. Machine vision system used in various quality-control procedures that increase manufacturing quality and productivity.
6. NUMERICALLY CONTROLLED ROBOT PAINTERS. These robotic spray painters provide consistent, high-quality paint finishes on GM automobiles.

6.

The fruits of Hitachi's ongoing research (from left to right): computed tomography, lasers, robots, a computer-scanned human eye, and optical memory disks.

would be launching a Rolls-Royce."

Modifying older launchers is unsatisfactory for another reason as well: there are not many of them left. They are being phased out of operation in order to let the shuttle take on the role of primary launch vehicle. (Only two Delta's remain in NASA's stockpile.) The only replacements in sight are a limited number of "complementary expendable launch vehicles" (CELV's) that were bought by the Air Force two years ago ("over NASA's vigorous opposition," according to Wheelon) and are due to be delivered in 1988. The CELV's would probably not be used to launch commercial telecommunications satellites.

The phaseout of older launchers is at the root of what may be an even more serious crisis. Enough telecommunications satellites are now in orbit to meet the needs of most users. By 1989 or 1990, however, a majority of those satellites will have to be serviced or replaced; there may be no launcher available for these missions. When the satellites begin going out of service, the resulting shortage will have effects throughout the economy. A spokesperson for the Department of Transportation (the agency responsible for private-sector commercial use of space) asks: "What about all those chains that run their businesses on computers linked by satellite and that use no hard copy? The dimensions of the problem loom much larger than is currently being reported."

Within the department there is a good deal of enthusiasm for "privatising" commercial space operations: encouraging private-sector companies to buy their own launch vehicles (and perhaps even to prepare their own launch sites) and to offer space transportation as an ordinary commercial service. To encourage such activity, the Department of Transportation suggests, NASA (and the shuttle in particular) should be barred from carrying commerical payloads. Private firms would thus be assured of a market and would not have to compete with Government-subsidized prices.

What broader changes need to be made in U.S. space strategy to ensure that this crisis does not develop again? Almost any new national space strategy would include building a larger, more diverse fleet of launch vehicles to complement the shuttle. Analysts within and outside the Government are now saying with increasing frequency that too much emphasis has been put on the shuttle as a single system capable of carrying out most U.S. space missions. Among many observers the consensus is that not enough attention has yet been paid to developing a coherent overall plan for gaining access to space. The emphasis may now turn toward saving the shuttle for the tasks only people can perform and depending on small unmanned vehicles for the others. It is not yet clear whether the same philosophy will be applied to NASA's proposal to build a large, inhabited space station to act as the center of in-space manufacturing and development.

Dirge

Warnings to Congress and the Administration about deterioration in U.S. science and technology have reached a crescendo in recent weeks. A panel of the White House Science Council declared that "our present leadership in science and technology is challenged" by shortages of faculty and students in important fields and by obsolete equipment. The National Science Board pointed to "serious deficiencies" in undergraduate education in science, mathematics and engineering. A committee of the National Research Council warned that the nation may soon be short of physicists at a time when "major advances are to be found in every field."

The White House panel pointed out that Federal support of research in universities has declined for the past six years as a percentage of the gross national product. It should be increased, the panel said. Other recommendations included the award of merit scholarships to encourage talented students to pursue mathematics, engineering and the natural sciences, the establishment by the Government of university centers for broad-based research and the creation of a fund the National Science Foundation could employ to help universities rebuild their infrastructures by refurbishing aging buildings and replacing obsolete equipment.

The National Science Board, which is the policymaking body of the National Science Foundation, said the major problems in undergraduate education in science, mathematics and engineering are that laboratory instruction is often dull and done with obsolete equipment, that faculty members have difficulty keeping up with their field and that courses and curricula are often out of date. The board urged the National Science Foundation to deploy "$100 million in new funds" by fiscal 1989 in an effort to remedy the problems.

The committee of the National Research Council, which is an arm of the National Academy of Sciences and the National Academy of Engineering, said research in physics is jeopardized because many physics teachers are due to retire and the number of people able to replace them is declining. According to the committee, the number of students receiving a doctorate in physics has dropped from 1,500 in 1970 to about 900 this year. Moreover, the proportion of Americans in the group is declining. In addition, increasing numbers of the students are going on corporate payrolls rather than into academic research. These trends, the committee said, could have serious consequences both for the economy and for national defense, which rely heavily on the development of new technologies.

Pocketbook Voting

The concern of political parties for the "image" of their presidential candidate is largely misplaced, according to a study by Gregory B. Markus of the University of Michigan. Voters, he said in a paper presented at a conference on economics and politics at the California Institute of Technology, make up their minds primarily on the basis of personal and national economic conditions. Hence the "expensive and exhausting" presidential campaigns are "ultimately a wash in terms of affecting election outcomes."

Markus, who is associated with the Center for Political Studies of the university's Institute for Social Research, examined national economic statistics in conjunction with data from the institute's studies of national elections from 1956 through 1984. He found that he could have predicted the outcome of the presidential elections in those years by means of econometric models alone, disregarding the personalities of the candidates, policy issues and campaign themes. The economic data in the models include the percentage change in real disposable personal income per capita and the percentage of people who feel they are better or worse off financially than before.

Markus estimates that each increase of 1 percent in disposable income increases the vote for an incumbent by 2.2 percentage points, other things being equal. He points out that a decline in real disposable personal income has occurred in only one presidential-election year since 1952, namely in 1980, when President Jimmy Carter lost to Ronald Reagan. A close analysis of that election, he says, "provides no evidence for the contention that Reagan's victory was the result of his policy or ideological positions."

Markus intends to apply his econometric-model approach to the 1988 presidential election to see whether he can predict the outcome in advance.

INCREDIBLE VACATIONS FOR INQUISITIVE MINDS.

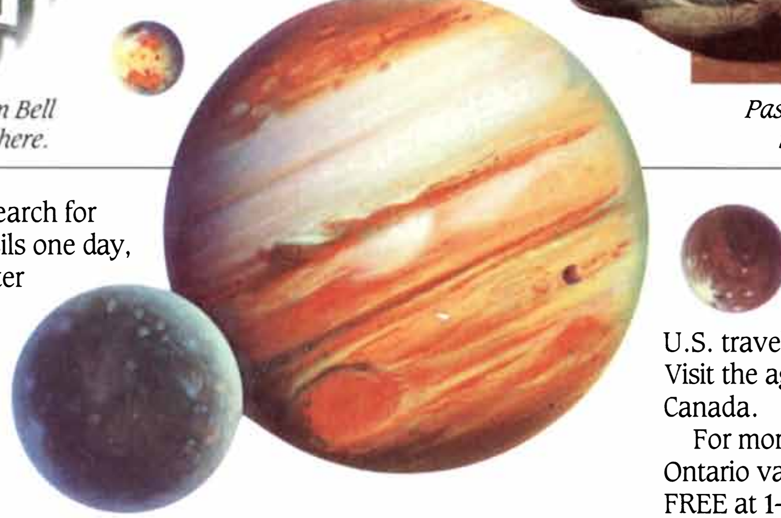
Close encounters with sunken wrecks.

There's more of Moore in Ontario.

Alexander Graham Bell lived and worked here.

Say hello to a fossil.

Past, present and future mingle in Ontario.



Ontario is timeless. Search for 135 million year old fossils one day, take a round trip to Jupiter in 2019 A.D. the next.

Or, enjoy a superb meal in 1986. The ages mingle in Ontario. The sciences, the arts and history live side by side. Ontario is the vacation spot made for time travellers.

With the great exchange rate, your U.S. travel dollar goes a lot farther. Visit the ages in incredible Ontario, Canada.

For more information about Ontario vacations, call us TOLL FREE at 1-800-268-3735.

ONTARIO *Incredible!*

Even though there will be no incumbent running, he notes, it can be argued that the effect of the economic situation on the Republican nominee will be the same as it would be if President Reagan were running again.

PHYSICAL SCIENCES

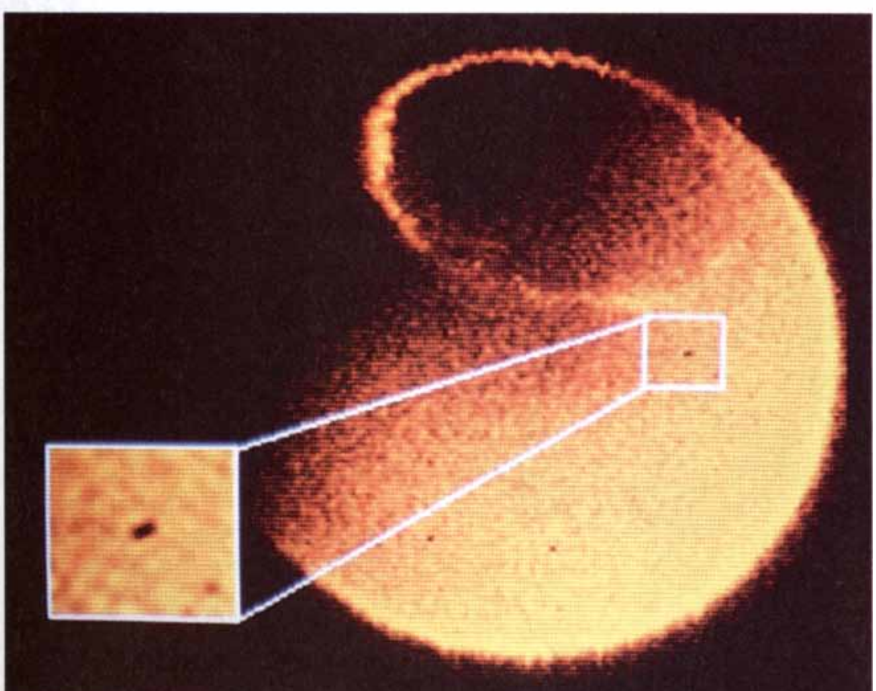
Holes in the Atmosphere

Where did the earth get its water? The conventional answer is that water, along with the main atmospheric gases, was "cooked" out of molten rock and discharged from volcanoes soon after the planet formed. A new hypothesis challenges that view. According to Louis A. Frank of the University of Iowa and his colleagues, the earth has acquired water gradually over billions of years—from small, icy comets that continue to plunge into the upper atmosphere at a rate of about 20 per minute.

The evidence for this startling proposal, presented by Frank, John B. Sigwarth and John D. Craven in *Geophysical Research Letters,* comes from satellite observations of the ultraviolet dayglow emanating from the upper atmosphere. Ultraviolet sunlight is scattered in all directions by atomic oxygen, which is the dominant atmospheric constituent at altitudes of more than about 100 kilometers. In ultraviolet images made from much higher altitudes by the polar-orbiting *Dynamics Explorer 1* the day side of the earth is suffused in a seemingly uniform glow.

The appearance is deceptive: after closely inspecting thousands of images over the past five years, Frank and his colleagues have found that the dayglow is not so uniform after all. It is interrupted by transient dark spots, called holes, in which the intensity of the ultraviolet emissions falls to between one-fifth and one-twentieth of its normal value. Over the entire day side of the earth the holes appear approximately 10 times a minute and persist for several minutes. Although they are sometimes as much as 150 kilometers in diameter, the typical diameter is about 50 kilometers. That means a typical hole covers just one pixel, or picture element, on a computer-processed satellite image. The Iowa workers contend that the holes must nonetheless be real, in part because they occur much too frequently to be attributed to an instrumental artifact or to noise in the data-transmission system.

After considering several possible explanations for the holes, the workers concluded that the most plausible one is an influx of comets. Comets consist mainly of water ice or snow, and water vapor is an effective absorber of ultraviolet radiation. If an incoming comet were to vaporize at an altitude between the glowing oxygen layer and *Dynamics Explorer 1,* the resulting water cloud would block the ultraviolet radiation and produce a dark hole on the satellite image.



ATMOSPHERIC HOLES show up as dark spots on an image of the earth's dayglow made at ultraviolet wavelengths by the satellite *Dynamics Explorer 1*. The hole in the inset is 150 kilometers in diameter and is unusually large. About 10 holes appear every minute. They may be clouds of ultraviolet-absorbing water—the vaporized remains of small, icy comets.

To produce a hole 50 kilometers in diameter, according to Frank and his colleagues, the comet would have to be a fluffy 100-ton snowball about 12 meters across. As the comet approached the earth it would be disrupted by tidal or electrostatic forces, the snow would vaporize and the water vapor would spread out, ultimately plunging into the atmosphere as a "piston of gas." Since 10 holes per minute are observed in the dayglow, and since an equal number of comets must be falling into the night side of the atmosphere unobserved, the Iowa group estimates that the earth is currently bombarded by 20 100-ton comets every minute. An influx of water at that rate—about one ten-thousandth of an inch of rainfall worldwide per year—would be more than enough to form the oceans over the 4.6-billion-year history of the planet.

Frank and his co-workers suggest a broad range of other implications for their hypothesis. Ice ages, for instance, might be caused by dramatic increases in the comet flux: if the flux were from 100 to 1,000 times greater than its present value, a global, sunlight-blocking ice cloud might form at high altitudes. Climate changes induced by peaks in the small-comet flux might also account for the periodic mass extinctions that other investigators have attributed to catastrophic impacts of large comets or asteroids. Moreover, the effects of the small comets would not be limited to the earth: presumably the other planets, at least those farther from the sun, would also be subjected to a persistent hail of icy bodies from outside the solar system.

Indeed, some critics of the small-comet hypothesis have focused on Mars. It should have received about as many comets as the earth, yet it has no ocean. Where has all the water gone? According to Thomas M. Donahue of the University of Michigan, there is no way much of the water could have escaped from Mars, where the average temperature is -53 degrees Celsius; if the water were still on the planet, the surface layer of permafrost would have to be hundreds of kilometers deep, which Donahue considers physically unthinkable.

More fundamentally, other skeptics doubt that small comets could have produced holes in the earth's dayglow. To make it all the way to the earth a fluffy comet would have to be protected from the sun's heat by a mantle of dust. According to Ian Stewart of the University of Colorado at Boulder, the gravitational field of a comet 12 meters in diameter would not be strong enough to keep such a mantle intact. Furthermore, Stewart argues, to produce a water-vapor cloud 50 kilometers wide at the level of the dayglow, the comet would have to begin disintegrating and vaporizing at a distance of about one million kilometers from the earth. In the process a lot of ionized hydrogen and oxygen would be injected into interplanetary space—much more, Stewart says, than is compatible with the measured concentrations of those elements.

Stewart is one observer who is not yet convinced that the atmospheric holes are real, but neither he nor anyone else has identified a source of error in the Iowa group's data. Other investigators apparently do believe in the reality of the phenomenon. And so far no one has come up with a good alternative to the surprising explanation proposed by the Iowa workers themselves.

Missing Neutrinos

A tankful of cleaning fluid under the South Dakota prairie has posed a major conundrum. The cleaning fluid is the sensing element in a solar-neutrino detector. Since it began operating in 1968 the device, emplaced in the Homestake gold mine, has detected only a third of the number of these virtually massless particles that theory predicts should be streaming toward the earth from the sun.

It seems unlikely that the neutrinos are getting lost along the way. Equally unlikely is the possibility that the sun is in fact producing the elementary particles at an unexpectedly low rate, since the prevailing theory successfully accounts for other important characteristics of the sun, such as the relation of its mass and composition to its luminosity and lifetime. Last year at a meeting in Finland two Soviet physicists suggested a possible explanation for the discrepancy. Hans A. Bethe, who won the Nobel prize in physics in 1967 for his work on energy production in stars, comments: "I think this is the first explanation that could be right." Bethe offers some modifications and extends the Soviet calculations in *Physical Review Letters.*

The apparent resolution of the problem hinges on the fact that there are different "flavors," or types, of neutrinos: an electron neutrino, a muon neutrino and a tau neutrino. Electron neutrinos are produced during thermonuclear fusion processes within stars. S. P. Mikheyev and A. Yu. Smirnov of the Institute for Nuclear Research in Moscow suggest electron neutrinos emitted by the core of the sun might interact with matter in the sun in such a way that they change into muon neutrinos on their way out. The fraction of neutrinos that undergo the metamorphosis would be quite large owing to the high density (130 grams per cubic centimeter) of matter at the center of the sun. (The density of water, in comparison, is only one gram per cubic centimeter.)

In contrast with electron neutrinos, muon neutrinos would not interact with the chlorine atoms in the Homestake cleaning fluid. Plans to build detectors that would be sensitive to muon neutrinos are under way in both the U.S.S.R. and Western Europe. The Soviet experiment will use 60 tons of gallium and the European experiment 6,500 tons of chilled argon.

Seeing Double

A team of astronomers led by Edwin L. Turner of Princeton University has collected evidence showing that what were thought to be a pair of distinct and separate quasars, or starlike objects, may in fact be two images of the same quasar. Although instances of such astronomical image splitting have been reported before, the observed angular separation between the two putative split images is at least 20 times wider than any previously reported separation.

The only known mechanism that can account for such image splitting is one first predicted by Albert Einstein in 1936: a gravitational lens. Such a lens is formed when spacetime is distorted by the gravitational field of a massive object. Because the path of light conforms to the local geometry of spacetime, in the vicinity of a massive object light rays are forced to curve. Hence diverging light rays from a distant source that pass by the object can be "bent" around and made to converge. If the earth intercepts some of these rays, each approaching from a slightly different direction, multiple images of the light source can be observed spread across a small area of the sky.

Until now the widest angular separation between a pair of gravitational-lens images had been about .002 degree. When two close quasars were discovered in 1979, the possibility that they were in fact two images of one quasar was therefore dismissed: the celestial objects (respectively designated 1146+111 B and 1146+111 C) were separated by .04 degree, which seemed to be implausibly large for a gravitational-lens image separation. Yet there was some suggestive evidence for a possible connection between the two quasars.

Quasars are characterized by a substantial displacement of their light spectra toward the long-wavelength (red) end of the spectrum, the effect called red shift. Such a displacement of the spectrum can be the result of a quasar's high recessional velocity (which is thought to be directly correlated with its distance from the earth) or perhaps of its own strong gravitational field. Curiously, the spectra of 1146+111 B and C appeared to be red-shifted by the same amount, implying they were equally far from the earth or of equal mass or both.

In a letter to *Nature* in February, Bohdan Paczyński, also of Princeton, pointed out that the 1146+111 pair, among other pairs and triplets of quasars, might in fact be the kind of illusion that would be expected if a gravitational lens were in front of a single quasar. Paczyński suggested that accurate red-shift measurements and profiles of the emission spectra of these quasars could serve as "fingerprints" by which to determine whether any of the quasar groups are images of one and the same quasar.

It was Turner's team of astronomers that finally matched the spectral fingerprints of 1146+111 B and C and found them to be remarkably similar. The investigators collected the spectrographic data at the Kitt Peak National Observatory; they present their findings in *Nature.*

Preliminary searches using optical telescopes have not revealed any superclusters of galaxies that are large enough to account for the separation between the images of 1146+111 B and C. As a result astronomers are speculating that more exotic manifestations of huge masses may be involved: black holes or cosmic strings. Cosmic strings are extremely fine but enormously dense threads of mass and can extend perhaps the entire length of the universe; they are predicted by certain theories of the universe's origin. The possibility that such a cosmic string is at the heart of the gravitational lens affecting the observation of 1146+111 B and C has aroused considerable interest in astrophysical and cosmological circles because no cosmic string has yet been detected.

BIOLOGY

Tumor-Cell Armor

Many cancers respond at first to a cytotoxic drug but eventually become resistant to it—and, disastrously, to a host of other, unrelated drugs to which they have not been exposed. Why? Different lines of investigation are converging to produce a picture of a single versatile defense mechanism that is activated under the pressure of drug treatment.

Two groups, one led by Ira Pastan and Michael M. Gottesman of the National Cancer Institute and the other by Igor B. Roninson of the University of Illinois College of Medicine, have concentrated on the genetic basis of multiple-drug resistance in human cells. Working with cancer-cell lines in culture, the investigators have studied a specific segment of DNA called *mdr*1. Earlier work with a line of cells from the Chinese hamster had implied that the *mdr*1 sequence might play a crucial role in the development of multiple-drug resistance.

Roninson, then at the Massachusetts Institute of Technology, in collaboration with an M.I.T. group led by David E. Housman, had compared the genetic material of drug-resistant hamster lung cells with that of the parent drug-sensitive cells. The experimental procedures were designed to detect DNA sequences that were present in more copies in the resistant cells than in the sensitive ones. Such amplified sequences, the investigators surmised, would bear some relation to the cells' capacity to cope with drugs. Two resistant cell lines contained a similar amplified sequence.

Roninson and his colleagues isolated and cloned the amplified sequence and from it constructed a DNA probe that could recognize and bind to similar sequences in the total DNA from lines of human tumor cells. The probe bound to two different segments of DNA, both of which were amplified in drug-resistant cells. Those sequences in turn were isolated; they served as probes for the matching messenger RNA (mRNA) in the human cells. The presence of mRNA corresponding to a DNA sequence indicates the sequence is expressed and may be directing the production of a protein. Only one of the DNA sequences, designated *mdr*1, seemed to be expressed.

Pastan, Gottesman and their colleagues, working with Roninson, have now determined that the level of *mdr*1 mRNA in cultured cells representing several kinds of human cancer corresponds to the degree of multiple-drug resistance. They report in *Science* that a thousandfold increase in resistance to colchicine (a cytotoxic drug that is ordinarily given as a remedy for gout) was correlated with an 820-fold increase in *mdr*1 mRNA. In cell lines that were highly resistant to colchicine and conventional anticancer drugs the DNA sequence itself was amplified; in less resistant cells the *mdr*1 gene was no more prevalent than it was in sensitive cells, but it was expressed more abundantly.

These observations have led the authors to propose that selection for multiple-drug resistance takes place in two stages: first the *mdr*1 gene is activated by a mutation or a nongenetic change in the DNA, leading to increased expression, and then the gene is amplified, leading to higher levels of resistance.

How might *mdr*1 enable a cell to survive a varied onslaught of cytotoxic drugs? Studies of the surface proteins characteristic of multiple-drug-resistant cells, carried out over several years by Victor Ling of the Ontario Cancer Institute and his colleagues, suggest an answer. When such cells are exposed to cytotoxic drugs in culture, they tend to accumulate much smaller amounts of the drugs than sensitive cells do. The difference seemed likely to reflect a change in the properties of the cell membrane.

Ling and other workers found that the membrane of resistant cells bears unusually large amounts of a protein he called P-glycoprotein. Monoclonal antibodies to the protein showed an affinity for cell membranes that was correlated with the cells' level of drug resistance. The antibodies also detected an overabundance of P-glycoprotein on tumor cells taken from patients with drug-resistant ovarian cancers. The workers suggest P-glycoprotein modulates the permeability of the cell membrane; in unusual quantities it might change the properties of the membrane and thereby limit the intracellular accumulation of foreign compounds—including cytotoxic drugs.

The gene for P-glycoprotein appears to be closely related to or identical with *mdr*1 itself. Ling and his colleagues have identified and cloned the DNA encoding P-glycoprotein, as have groups led by Peter W. Melera and June L. Biedler of the Memorial Sloan-Kettering Cancer Center and by Piet Borst of the Netherlands Cancer Institute. The workers have found that, like *mdr*1, the DNA sequence is amplified in multiple-drug-resistant cell lines. Other factors also suggest that the two DNA segments are related: they are found on the same chromosome, and they encode mRNA's of the same size.

The close association of the glycoprotein and its gene with multiple-drug resistance points to some new approaches to cancer therapy. Monoclonal antibodies to P-glycoprotein or genetic probes for *mdr*1 mRNA could detect the activity of the gene in cells biopsied from a tumor, yielding a measure of the cancer's sensitivity to cytotoxic drugs. Such antibodies might also have a direct therapeutic role if P-glycoprotein is crucial to multiple-drug resistance. By incapacitating the protein, they might make resistant cells sensitive again.

Setting the Clock

Everyone who has flown across more than one or two time zones is familiar with the chronological disorientation that accompanies east-west travel. Innumerable travelers, lying wide awake in the middle of the night in a new city, have wished for a pill that would make possible a quick, painless adjustment to local time. Recently that weary traveler's dream moved toward fulfillment when Fred W. Turek and Susan Losee-Olson of Northwestern University found that a commonly prescribed medication can reset the internal clock of a golden hamster.

Turek and Losee-Olson worked with a drug called triazolam, which is used in treating insomnia. Triazolam is a member of the family called benzodiazepines; the most frequently prescribed member of that family is valium. The benzodiazepines are thought to work by heightening the effect of gamma-aminobutyric acid (GABA), a neurotransmitter that is widely distributed in the central nervous system. Nerve cells containing GABA are found in the suprachiasmatic nuclei, paired structures in the hypothalamus that appear to form the basis of the central biological pacemaker in mammals.

The golden hamster is an excellent species for tests of circadian (daily) rhythms because its behavior in captivity is remarkably regular. Caged golden hamsters begin their workout on the treadmill at exactly the same time every day, and so changes in the time they begin to run can indicate alterations of internal rhythms. Turek and Losee-Olson recorded daily activity in a group of hamsters for two weeks to provide a baseline and then injected each hamster with triazolam at one of eight times spread equally over the animal's circadian cycle.

Intriguingly, the results showed the effect of the drug varied considerably with the timing of the injection. When triazolam was injected six hours before running would otherwise have begun, the circadian cycle was advanced by about 90 minutes. When the drug was injected from six to nine hours after the beginning of running, however, the circadian cycle was retarded by as much as an hour. Injections given at other intervals yielded smaller phase shifts.

Turek and Losee-Olson's findings were published in *Nature* along with a commentary by Arthur T. Winfree of the University of California at San Diego. Winfree writes: "The potential practical interest of [Turek and Losee-Olson's] results lies in the hope that the same or a related fast-acting, quickly eliminated chemical without conspicuous side effects may also abruptly reset the circadian clock in humans. If so, then we may soon find ourselves carrying pills and a schedule whenever we cross time zones at a pace much in excess of one per day."

Chief of Staph

The bacterial species *Staphylococcus aureus* gives rise to an array of human diseases that includes boils, toxic-shock syndrome and skin, blood, heart and lung infections. It also causes the antibiotic-resistant staphylococcus infections that can plague hospitals. The bacteria are thought to cause disease mainly by releasing enzymes and toxic proteins into cells and tissues. Now investigators seem to have found a trigger for the production of these substances; it is a "master gene"—called the accessory gene regulator (*agr*)—that activates the genes encoding several of these exoproteins. Richard P. Novick and his colleagues at the Public Health Research Institute in New York and at the University of Minnesota Medical School report their results in *Molecular and General Genetics*.

The investigators knew that mutations of *S. aureus* affecting exoprotein production often alter the levels of many proteins simultaneously. This suggested that there might be a single gene coordinating multiple exoprotein genes. The team's demonstration of the gene's existence was nonetheless somewhat serendipitous. In the course of attempting to isolate the gene encoding the toxin associated with toxic-shock syndrome (the toxin had been identified earlier), the group was testing a recombinant strain of *S. aureus* to see how well it produced the toxin. The workers knew the strain was not able to produce alpha hemolysin, a toxin that damages red and white blood cells, presumably because the alpha hemolysin gene was inactivated by a transposon, or "jumping gene," inserted into it. They did, however, expect that the gene for the toxic-shock toxin (which was definitely present) would be expressed.

The investigators were therefore surprised to find that when the strain was grown in culture, it produced depressed levels not only of alpha hemolysin but also of the toxic-shock toxin and of two additional toxins that damage blood cells. This suggested that the transposon had been inserted at a site outside the alpha hemolysin gene—into a master gene controlling the activity of not one but several exoprotein genes. Indeed, when the investigators examined the alpha hemolysin locus of the *S. aureus* strain, they found no transposon there.

On further examination of the culture, the team also found a reduced level of messenger RNA (mRNA) for both the toxic-shock toxin and alpha hemolysin. Since mRNA acts as a template for the production of substances encoded by DNA, such a reduction suggested that the master gene, *agr*, acts at the level of mRNA, perhaps by enhancing the rate of its transcription from DNA.

Novick says he and his colleagues are about to complete the isolation and sequencing of *agr*. Having done that, they expect to be able to identify the product encoded by the master gene and to trace its activities. The product quite likely is a protein and certainly is a mediator of *agr*'s activity. In the more distant future the group also hopes to discover the trigger that activates the activator.

TECHNOLOGY

Chernobyl, U.S.A.?

The explosion at the Chernobyl nuclear reactor in the U.S.S.R. and the accompanying spread of radioactive fallout have raised the question of whether a similar series of events could happen in the U.S. Determining the answer is not easy, largely because the details of the cause of the Soviet disaster are so far not known. Yet the consensus among nuclear physicists and engineers suggests that the U.S. does not have a shield of inherently superior technology.

The cause of the Chernobyl accident was initially linked to a reactor component called the moderator. It reduces the speed at which a neutron travels, thereby enhancing the likelihood that the particle wiil be captured by a nucleus and induce it to fission. The moderator in the Chernobyl reactor is graphite. The moderator in all but two U.S. reactors is water. (The reactors at the Hanford site near Richland, Wash., and the Fort St. Vrain commercial reactor in Platteville, Colo., do have graphite moderators. Both reactors are believed to have more safety features than the one at Chernobyl, but only the Fort St. Vrain reactor meets all the U.S. safety codes for commercial reactors.)

Western experts initially thought the Chernobyl accident was the consequence of annealing, a process that restores the neutron-controlling properties of a graphite moderator. Graphite consists of a lattice of carbon atoms. When neutrons pass through it, they knock carbon atoms out of their resting positions, raising the internal energy of the lattice. The phenomenon is called the Wigner effect because it was postulated by Eugene P. Wigner of Princeton University. Over a period of many months the internal energy can reach dangerously high levels. To purge the lattice of its energy, that is, to allow the carbon atoms to return to their normal resting positions, the graphite is typically annealed, or heated, once a year. Although heating the graphite is a standard procedure (the flow of coolant water is simply reduced), overly rapid heating can cause an explosion. Such an explosion did occur in 1957 at the Windscale reactor in England.

As more information has become available, however, it seems increasingly unlikely that the Wigner effect was responsible for the Chernobyl accident. Richard K. Lester, associate professor of nuclear engineering at the Massachusetts Institute of Technology, notes that at the time of the explosion the reactor was operating at full power, which is not usually the case during the annealing procedure. Moreover, reports indicate the normal operating temperature of the Chernobyl reactor appears to have been from 250 to 300 degrees Celsius—high enough so that carbon atoms can return to their resting positions without annealing.

The currently favored scenario to explain the start of the Chernobyl accident, which seems to have culminated in a fire in the moderator and at least a partial meltdown, has little to do with the choice of moderator—be it graphite or water. Most experts, including Harold R. Denton, director of regulation for the U.S. Nuclear Regulatory Commission (NRC), believe the difficulties began when cooling water stopped flowing to the reactor core, perhaps because of a leak in the cooling system or a turbine failure. As the core heated up, cooling water inside fuel rods was converted into steam. The steam may in turn have interacted with zirconium cladding on the rods to produce free hydrogen. Enough hydrogen may eventually have been generated so that explosive reactions between hydrogen and oxygen took place. A brief explanation given by Soviet leader Mikhail S. Gorbachev during a televised address appeared to confirm this course of events.

Could there be a hydrogen-oxygen explosion in a U.S. reactor? Thomas B. Cochran, a senior scientist at the

Natural Resources Defense Council and a former adviser to both the NRC and the U.S. Department of Energy, points out that hydrogen was released during the 1979 accident at the Three Mile Island-2 reactor near Harrisburg, Pa. There was no chemical explosion because little or no free oxygen was present. Nevertheless, both Cochran and Robert D. Pollard, a nuclear safety engineer with the Union of Concerned Scientists and a former member of the staff of the NRC, caution that there is little reason to believe U.S. reactors are immune to hydrogen-oxygen explosions. Brian W. Sheron, deputy director of safety review and oversight for the NRC, takes issue with that point. He comments: "Zirconium can react with steam to produce [free] hydrogen. But I must qualify my statement. U.S. plants are designed so that you don't get that situation."

Proponents of nuclear power and industry officials have argued that even if an accident did occur, the containment buildings surrounding U.S. reactors would prevent the release of significant amounts of radiation. One version of a containment building is a domed concrete structure with an airtight steel liner. The containment building at Three Mile Island is credited with having prevented the release of large amounts of radiation into the atmosphere even though a partial fuel meltdown took place.

Initially word was spread by the Atomic Industrial Forum, Inc., and other organizations in the West that the Chernobyl reactor had no containment building. Subsequent information gathered by the Central Intelligence Agency and American nuclear experts who have visited Chernobyl, as well as information from technical literature, indicates otherwise. Western experts now agree that a large structure of heavy steel and concrete did in fact enclose the Chernobyl reactor. Moreover, it appears to have been designed to withstand pressures that are comparable to those in many American reactors.

It remains unclear how effective U.S. models are compared with the Soviet structure. Cochran thinks the containment buildings in the U.S. cannot confine full-core meltdowns: "Large power reactors of very different designs can undergo a core meltdown.... I think it would be a mistake to seek assurances in the technical differences among various confinement structures." Cochran has testified before a House subcommittee on energy conservation and power that "the most important lesson to be learned from Chernobyl [is that] a core melt can happen. It can happen anywhere, anytime, in any large operating reactor in the U.S. or abroad."

Sheron, on the other hand, remarks that "the Soviet enclosures seem to be philosophically very different from those in Western nations." He continues: "Based on everything we know about Chernobyl, there is no reason for us to change our regulation. There is nothing jumping out saying what we should fix." He argues that although "we can possibly postulate accidents that would fail confinement..., the resulting risk is still acceptably low. We can't live in a risk-free society."

Sunspots

The economical conversion of sunlight into electricity suitable for the utility grid has been a goal long pursued by solar-energy advocates. Up to now attempts to generate electricity directly from sunlight have been frustrated by the expensive production techniques and low efficiencies associated with photovoltaic, or solar, cells. A new, record-setting level of efficiency in solar-energy conversion attained by a research team at Stanford University therefore represents a significant step toward cost-effective generation of electricity by means of solar cells.

Solar cells are made of semiconductor materials, such as crystalline silicon, which release electrons from their bonding sites when they are exposed to light of certain energies. An electron thus released can travel through the silicon, leaving behind it a "hole" (the vacated bonding site). By interspersing atoms of other elements in silicon's crystalline structure (a process called doping), silicon can be made preferentially receptive either to the negatively charged electrons (in which case it is called *n*-type silicon) or to the positively charged holes (*p*-type silicon). Charges generated by photons at the interface between layers of *n*-type and *p*-type silicon therefore segregate themselves, the electrons migrating to the *n*-type layer and the holes migrating to the *p*-type layer.

The electric current generated by the moving charges can be tapped simply by placing electrical conductors on the outside surfaces of the layers. Unfortunately, because the conductors cover the surfaces of a conventional solar cell, they unavoidably block out some of the impinging sunlight. Moreover, the presence of the doping impurities in the regions under the current collectors tends to cause freed electrons to recombine with holes before they have had a chance to generate a current, further reducing the efficiency of the cell.

The Stanford investigators managed to reduce both types of losses by doping tiny spots on the back surface of a layer of silicon dioxide. Fine "fingers" of deposited aluminum serve as current collectors at each *p*-doped and *n*-doped spot. Such a novel design eliminates the blocking of sunlight at the front surface and minimizes the doped regions. As a result the Stanford solar cell attains an efficiency of 27.5 percent (in concentrated sunlight). Although the nearest competitor has an efficiency that is only about 1.5 percent less, it is made of gallium arsenide, which costs considerably more and is difficult to process.

The Stanford researchers are confident that they can further improve their silicon solar cell to reach a maximum efficiency of 29 percent. At the same time studies are being done to determine whether the cell can be mass-produced at a cost low enough for practical large-scale power generation.

Mussel Power

In order to attach itself to a surface the mussel secretes a sticky glue that is as powerful as epoxy; it offers the additional advantage of working even under water. After 10 years of effort J. Herbert Waite of the University of Connecticut has succeeded in identifying the main ingredient of the substance and has synthesized the material. He and the National Institute of Dental Research, which supported his work, foresee applications for the adhesive in dentistry, medicine and industry.

Over a period of three years Waite and his associates laboriously extracted tiny amounts of glue from more than 3,000 mussels. They found that, like epoxy, it consists mostly of a resin called polyphenolic protein. Chemical analysis showed that a repeating sequence of 10 amino acids makes the glue work. Apparently the surface tension of the molecule displaces water, so that the molecule can form three types of bond (hydrogen, ionic or valence) with the substrate. By means of solid-phase peptide synthesis Waite's group was able to reproduce the adhesive in the laboratory.

Because the adhesive binds in a wet, saline environment, it is expected to be effective for tooth repair and perhaps for mending broken bones or injured tendons. Moreover, it has the remarkable property of not accepting another adhesive once it sets. Hence the U.S. Navy, in what a mussel might regard as the ultimate irony, hopes to employ the adhesive as a coating for ship hulls to prevent fouling by marine organisms.

The Structure of Mountain Ranges

What holds mountains up? Some stand on plates of strong rock; others are buoyed by crustal roots reaching deep into the mantle. The latter may collapse when their flanks are not pushed together

by Peter Molnar

In looking at mountains one is generally first struck by the topography: the extraordinary scale, the shapes carved by glaciers and streams, the contours smoothed and decorated by vegetation. Many people are inspired by the same awe they feel before certain manmade structures, such as the soaring arches and stained glass of a Gothic cathedral. Yet as the eye steps from detail to detail across a landscape it is easy to forget that enormous forces are required not only to build but also to support a mountain range. Each range, like each cathedral, stands on a foundation, without which it would collapse. If one wants to feel more than inarticulate wonder before mountains or buildings, it helps to understand the invisible mechanisms that support the visible beauty. Hence the purpose of this article: to describe the underlying structure—the tectonics, if not the architecture—of mountain ranges.

Two Kinds of Support

The analogy to architecture is not merely rhetorical; the different solutions architects have found to the problem of supporting buildings have parallels in the structure of mountain ranges. One solution is to build on a basement of strong, inflexible rock. Some of the world's tallest buildings, for example, stand on the Manhattan schist, a rock formation that has not been significantly heated or deformed (and thereby weakened) since the Precambrian era ended some 600 million years ago. The world's highest mountains, the Himalayas, are like the Manhattan skyscrapers: they stand on a thick shield of strong Precambrian rock, the northern edge of the Indian subcontinent.

A basement of strong rock, however, is not necessary for the support of a large structure. I work in a 20-story building in Cambridge, Mass., that rests on pilings pounded 40 meters into artificial fill and glacial moraine in what was once a tidal basin along the Charles River. To some extent the building floats on water-saturated deposits, and in that respect it is not unlike a large ship. Mountains too can be supported by the buoyancy of light material floating on heavier material. An example is the Tibetan plateau north of the Himalayas, nearly all of which lies above 4,500 meters. Unlike the substrata of the Himalayas themselves, the substrata of the plateau appear to be weak and easily deformed—like the landfill under my office building or the water under a ship.

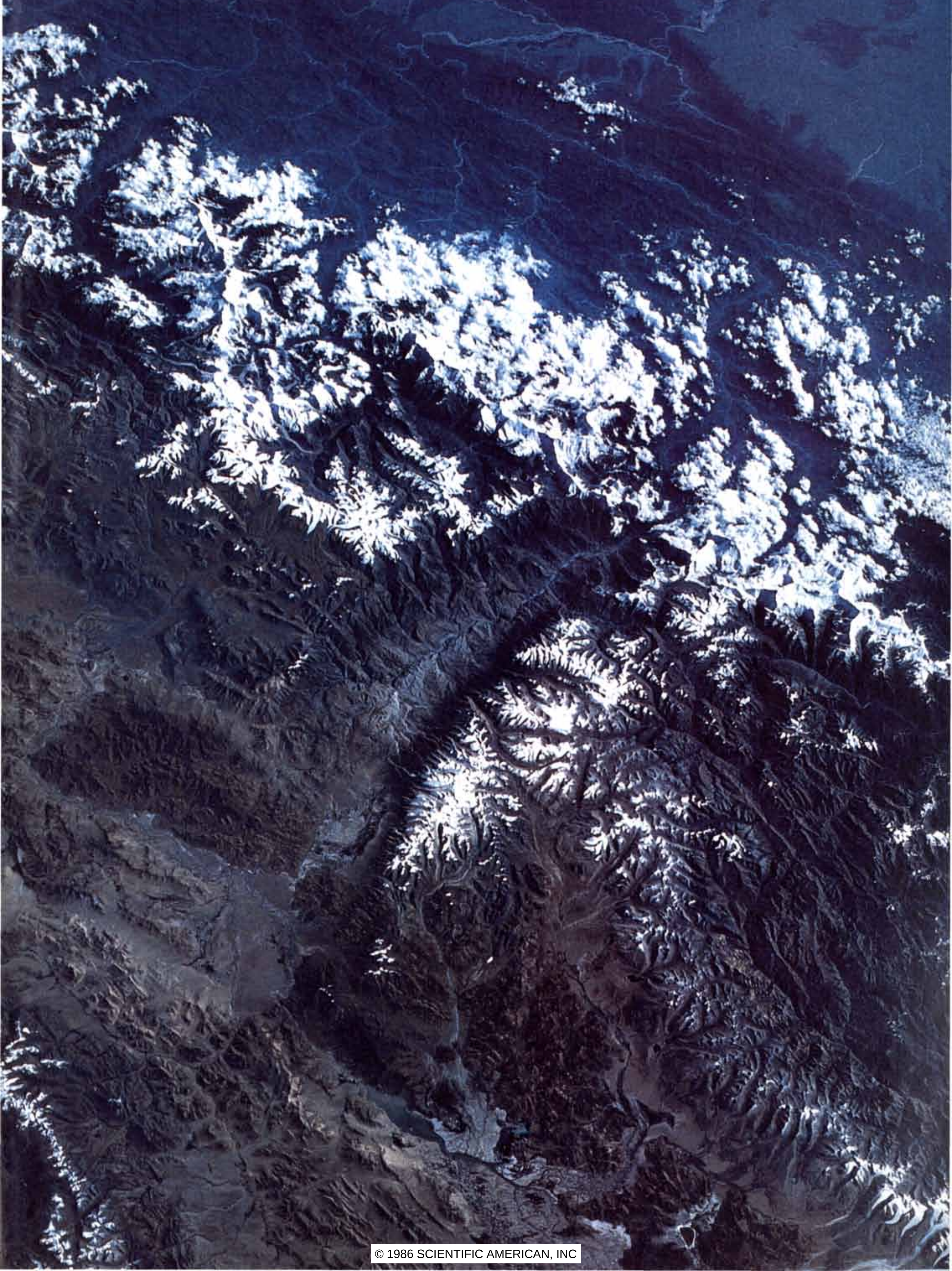
The Survey of India

The Himalayas and the neighboring Tibetan plateau thus exemplify two quite distinct mechanisms for supporting mountain ranges (which is not to say that both mechanisms cannot operate in the same range). In fact, it was the study of this area 140 years ago that led to the first advances in understanding the structure of mountains. The pioneers were the surveyor George Everest, J. H. Pratt, the scientifically inclined archdeacon of Calcutta, and George B. Airy, the eminent mathematical physicist and Astronomer Royal of Britain. The story of their work is in itself a fascinating bit of intellectual history.

In the 1840's Everest was directing the first topographical survey of the Indian subcontinent. His crews had two methods of measuring distances. First, they could measure short distances by the conventional surveying technique of triangulation, arriving at longer distances step by step. Second, they could determine the relative positions of two widely separated points directly by observing the position of a reference star from both points at the same time of day. In principle the two methods should have yielded similar results, but in practice there were large discrepancies. The most celebrated of these concerned the distance between the towns of Kaliana and Kalianpur, respectively some 100 and 700 kilometers south of the Himalayan front. The astronomical survey placed the two towns about 150 meters closer to each other than triangulation did.

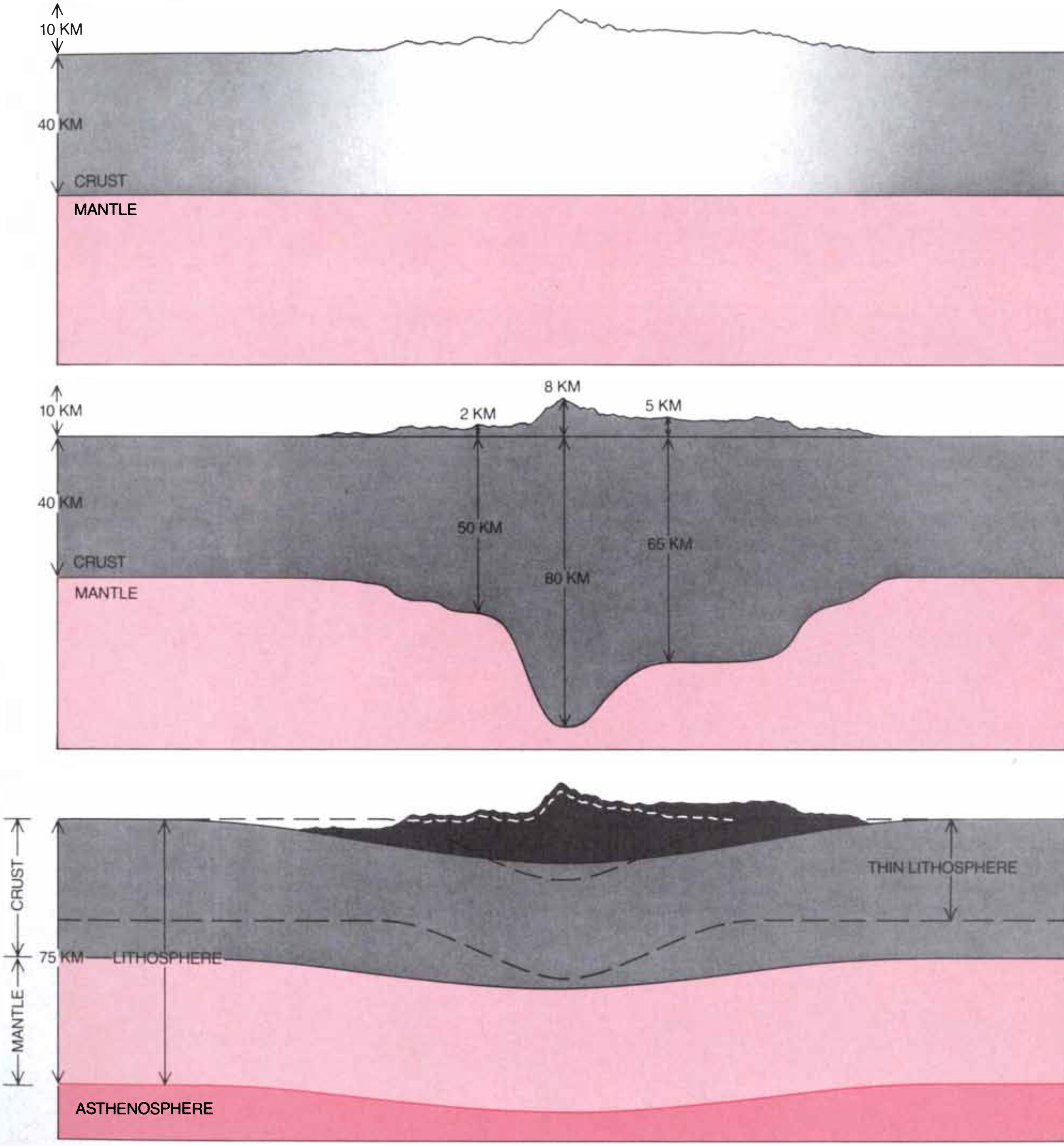
Everest assumed that cumulative errors in triangulation accounted for the discrepancy, but in 1854 Pratt showed that the error lay instead with the astronomical measurements. To determine the position of a star on the celestial sphere the surveyors had to know precisely the direction of the zenith (the vertical direction), which was defined by a plumb line. Pratt suggested

HIMALAYAS AND TIBET exemplify the two mechanisms by which mountain ranges and high plateaus can be supported. The view in this photograph, taken from the space shuttle *Challenger,* is to the south over southern Tibet, Nepal and northern India; the area covered is about 175 kilometers wide. The snowcapped Himalayan peaks are partially masked by clouds. They stand on the northern edge of the strong Indian plate, which bends down to distribute their weight; the Ganges plain, just visible beyond the mountains, owes its existence to the bending. In contrast, the high Tibetan plateau is supported by a deep root of buoyant crust. Thick crust tends to be weak, and if it is not held together by horizontal forces, it can spread. Tibet is slowly spreading and collapsing. The large valley in the center of the photograph is a graben that has formed where a block of crust has dropped down along normal faults as the adjacent crust has spread apart. The graben cuts through east-west-trending folds left by the collision between India and Eurasia that raised the Himalayas. To the south it continues into the Kali Gandaki valley, which is flanked by Annapurna to the east and Dhaulagiri to the west. Both peaks are more than 8,000 meters high.



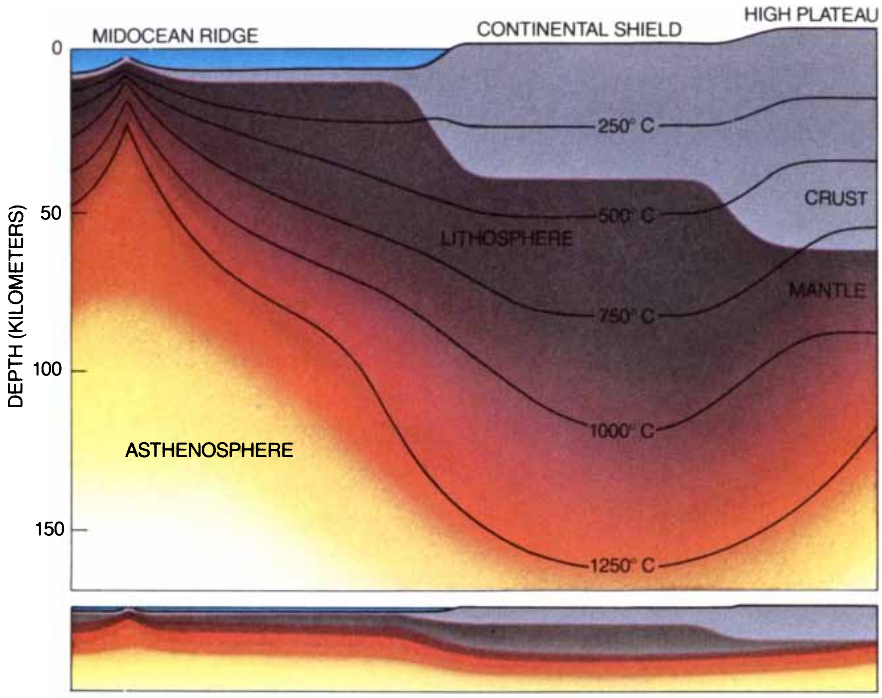
that the gravitational attraction exerted by the large mass of the Himalayas and the Tibetan plateau would deflect the plumb bob to the north, and the deflection would be greater at Kaliana than at Kalianpur because Kaliana is closer to the mountains. The resulting difference in the measured directions of the zenith would introduce an error into the calculation of the relative positions of the two towns.

When Pratt tried to determine the size of the error by estimating the mass of the Himalayas and the Tibetan plateau, he made a puzzling discovery. His results indicated that a plumb bob should be deflected by 28 seconds of arc at Kaliana and by 12 arc seconds at Kalianpur. The 16-arc-second difference may seem small, but actually it corresponded to an error in the astro-



THREE MECHANISMS of isostatic compensation embody different ideas of how mountain ranges are supported. In the 19th century J. H. Pratt and George B. Airy suggested that mountains are supported solely by the buoyancy of the crust floating on the denser mantle; in this view the excess mass in a mountain range is compensated by a mass deficit directly under the range. Pratt (*top*) proposed that the density of the crust varies laterally and that mountains stand high where the density is low (*light gray*). Airy (*middle*) assumed that the crust is of uniform density but that it is thicker under mountain ranges: like icebergs, mountains are supported by deep roots of buoyant material. Felix A. Vening-Meinesz, however, proposed that a mountain range is compensated regionally rather than locally (*bottom*). The weight of the mountains bends a strong layer, now called the lithosphere, that usually includes the crust and part of the mantle and that overlies a weak, fluidlike layer called the asthenosphere. The bending of the lithosphere distributes the load over a broad region. Mountains stand higher on thick lithosphere because it bends less than thin lithosphere (*broken lines*).

nomical distance measurement that was three times larger than the observed discrepancy of 150 meters. Apparently, Pratt concluded, the real difference in the gravitational deflection of the plumb bob was only about five arc seconds, which implied that he had overestimated the mass of the mountains; there was a lot less mass under the Himalayas and Tibet than his analysis of their topography had suggested. Indeed, had Pratt had access to accurate topographic maps (his maps put most of Tibet at just over half its true altitude) he would have inferred an even greater "missing" mass.



LITHOSPHERE AND ASTHENOSPHERE are shown in vertically exaggerated (*top*) and unexaggerated (*bottom*) cross sections. The two layers differ in temperature: the lithosphere is cooler and therefore stronger than the asthenosphere. The boundary between them is not sharp; it can be taken as lying roughly at 1,300 degrees Celsius. In contrast, the chemical boundary between crust and mantle is sharp. The thickness of the lithosphere varies considerably. Under midocean ridges, where it is created by the upwelling of hot material from the asthenosphere, it is very thin; under continental shields, which are made of crust that has not been heated for 600 million years or more, its thickness may increase to more than 150 kilometers. Under some high plateaus, however, the lithosphere is not thick. Indeed, under Tibet the thermal boundary between lithosphere and asthenosphere appears to fall within the crust instead of in the upper mantle. The crust under Tibet is warm and pliable.

Crust and Mantle

Airy read Pratt's paper at his desk in London. At first the idea of missing mass surprised him, but then he quickly realized that the surface of the earth is probably not strong enough to support a huge mass of mountains without deforming in some way. The deformation leads to a mass deficit under the mountains that compensates the excess mass on the surface. Compensation of this type is familiar to most people as the phenomenon Archimedes discovered when he got into a full bathtub and it overflowed; to geologists it is now known as isostasy.

In Airy's conception of isostasy the earth's light crust floated on a heavier but weak, fluidlike substratum: the mantle. The chemistry of the crust is well known today, and it is indeed lighter than the mantle. Although both layers consist primarily of oxygen and silicon, the relatively heavy elements iron and magnesium are much more abundant in the mantle. In contrast, larger fractions of the relatively light elements, including sodium, calcium, aluminum and potassium, are concentrated in the crust. As a result the crust is less dense than the mantle, and so it is not unreasonable to think, as Airy did, that the crust floats on the mantle as cream floats on milk. Like the boundary between cream and milk, the boundary between crust and mantle is quite sharp.

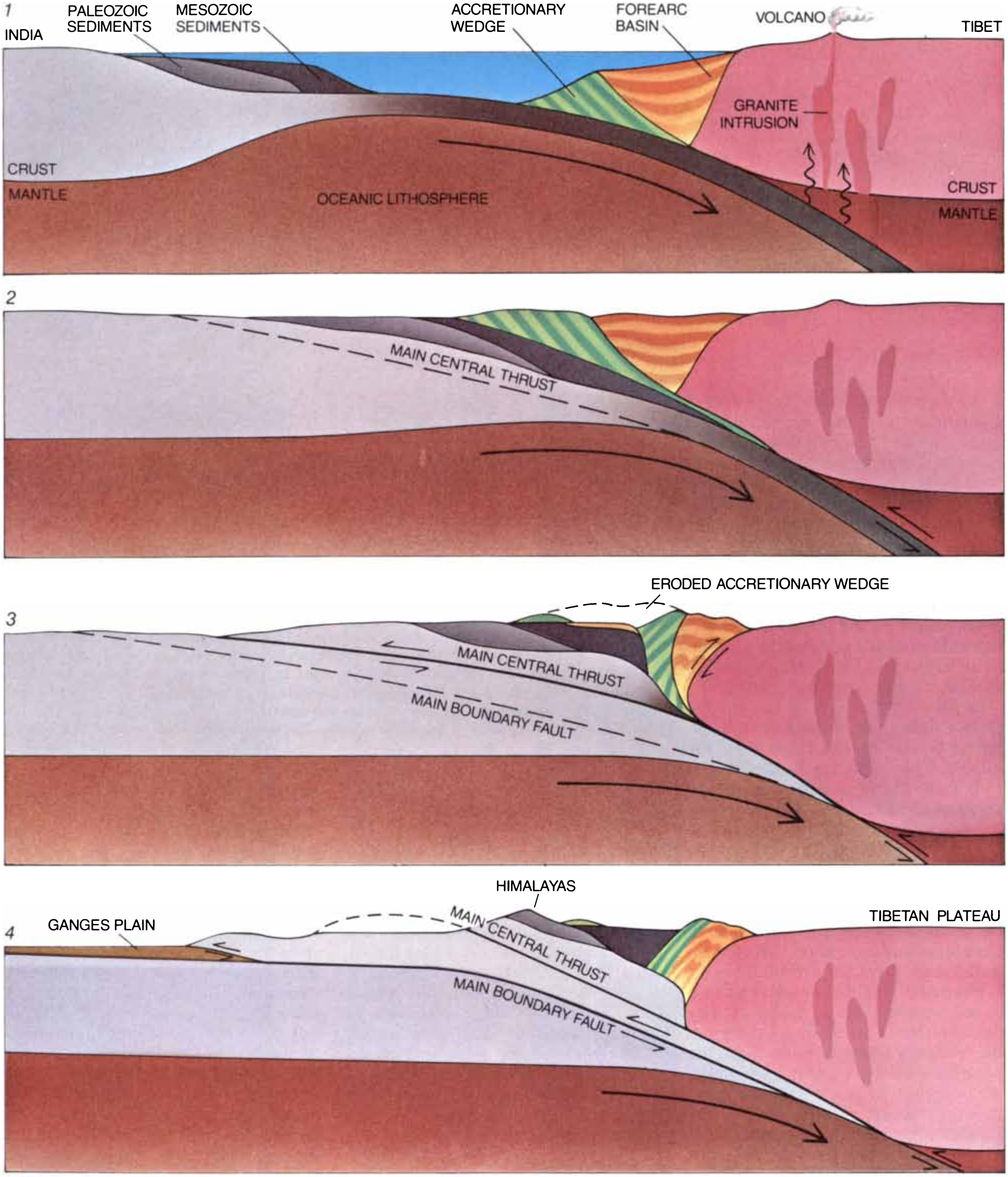
Pratt shared Airy's conception of a floating crust, but the two men disagreed on the mechanism underlying isostatic compensation. Pratt thought the temperature and hence the density of the crust vary from place to place. Where the crust is hotter and lighter than average it rises to form mountains; where it is cold and dense it subsides to form vast lowlands. Airy, on the other hand, thought the density of the crust is fairly uniform but that its thickness varies. The crust is thicker under mountains, he argued, than under lowlands; the visible mountains are like the tips of icebergs, and like icebergs they are supported by deep, invisible roots.

Seismological studies over the past several decades have confirmed Airy's prediction that the thickness of the crust varies substantially. Continental crust is on the average between 30 and 40 kilometers thick, but under mountains the thickness may increase to as much as 75 kilometers. The crustal roots compensate the excess mass of the mountains by displacing denser mantle rock. Conversely, the crust under the deep oceans compensates the low density of water by being only about six kilometers thick. Gravity-induced rock movements keep the earth in approximate isostatic equilibrium, such that the mass of an imaginary column through the earth is roughly the same whether its surface is a mountain range or part of an ocean.

Lithosphere and Asthenosphere

In spite of such impressive confirmations Airy's version of isostasy is only approximately correct. Early in this century, even before seismologists had confirmed that the thickness of the crust varies, they had found that the mantle, like the crust, is solid rather than liquid. Hence the image of the crust floating on the mantle is an oversimplification, and the same must be true of Airy's hypothesis. In the 1930's the Dutch geophysicist Felix A. Vening-Meinesz suggested that the isostatic compensation of a topographic load should take place on a regional rather than a local scale and should involve more than just the formation of crustal roots. (In the limiting case this is obvious; the crust does not poke hundreds of meters into the mantle under the Empire State Building.)

Specifically, Vening-Meinesz proposed that a large load such as a mountain range deflects the earth's strong outer layer, now called the lithosphere. The lithosphere usually includes not only the crust but also the uppermost part of the mantle. It overlies a weak, fluidlike layer called the asthenosphere. To a first approximation the lithosphere can be treated as elastic, and under a mountain range it bends downward, thereby distributing the weight of the range over a broad region. The bending of the lithosphere creates a trough parallel to the range. As a result the excess mass of the mountains is compensated in part by a



HIMALAYAS WERE FORMED when the Indian lithospheric plate drifted northward and collided with the Eurasian plate. The collision is shown here in simplified, vertically exaggerated diagrams. Some 60 million years ago the oceanic lithosphere at the leading edge of the Indian plate was being subducted under southern Tibet (*1*). Magma rising above the Indian plate erupted from volcanoes and formed granite intrusions. Sediments and oceanic crust scraped off the descending plate piled up in an accretionary wedge, which created a forearc basin that trapped sediments eroded from Tibet. Sometime between 55 and 40 million years ago the two landmasses collided (*2*). Presumably the Indian crust was too buoyant to plunge far under Tibet; as a result a new fault, the Main Central Thrust, broke through the Indian crust. Subsequently motion continued along the fault (*3*). A slice of Indian crust, topped by Paleozoic and Mesozoic sediments that had been deposited on the continental shelf, was thrust up onto the oncoming subcontinent. The accretionary wedge and the forearc sediments were thrust northward onto Tibet. (Much of this material has since been eroded away.) About 20 to 10 million years ago the Main Central Thrust became inactive. Since then India has slid northward along a second fault, the Main Boundary Fault (*4*). A second slice of crust has been thrust up onto the subcontinent, lifting up the first slice. The two uplifted slices make up the bulk of the Himalayas; many of the peaks are capped by Paleozoic sediments. The Indian plate bends slightly under the weight of the mountains, and the resulting trough, now filled with sediments, can be detected under the Ganges plain.

mass deficit in the trough and not just by a deficit directly under the range.

The lithosphere, it is now known, is not a continuous layer but instead consists of 20 or so separate plates. The movements of the plates over the asthenosphere account for the formation of ocean basins and mountain ranges and for the other phenomena known collectively as plate tectonics. Although descriptions of such horizontal plate movements generally treat the plates as being rigid rather than elastic bodies, this does not contradict Vening-Meinesz' hypothesis that the lithosphere bends elastically under a topographic load. A lithospheric plate is somewhat like a wood table: the table moves rigidly when it is pushed across the floor, but it may sag in the middle when a heavy load is placed on it.

The asthenosphere offers buoyant resistance to the sagging of the lithosphere, but the lithosphere does not float on the asthenosphere. Unlike the boundary between crust and mantle, the boundary between lithosphere and asthenosphere is not a chemical one, and the density contrast across the boundary is therefore negligible. The buoyant restoring force exerted by the asthenosphere arises from its being much denser than the layer of air or water overlying the flexed lithosphere.

The lithosphere and asthenosphere differ in temperature rather than in composition: the lithosphere is cooler, which explains why it is stronger. Within the lithosphere temperature increases rapidly with depth, reaching a level of roughly 1,300 degrees at the boundary of the asthenosphere. The boundary is not sharp like the chemical boundary between crust and mantle, and there is no general agreement among investigators on how to define it. What is clear is that like the crust the lithosphere varies widely in thickness, from as little as 10 kilometers to more than 150.

The thicker a wood table is, the greater the load it can support and the less it sags. The same is true of the lithosphere. A thick plate is stronger than a thin one and bends less under the weight of a mountain range. Consequently a range should stand higher on a thick plate, other things being equal, than it would on a thin one. High mountains can nonetheless exist on a thin plate if they are supported in the way Airy envisioned, by deep crustal roots. The isostatic mechanisms put forward by Airy and Vening-Meinesz are not mutually exclusive. On the contrary, it has been found that a mountain range can be supported by a strong foundation of thick lithosphere (like the New York skyscrapers), by deep roots of light crust (like a ship) or by a combination of both mechanisms. The relative importance of the mechanisms varies from range to range.

The Himalayas and Tibet

To determine which mechanism is more important in the case of the Himalayas and the Tibetan plateau one must first consider how those mountains were formed. Some 70 million years ago India and the rocks that now make up the Himalayas were about 8,000 kilometers south of their present position, drifting northward from Antarctica toward Asia on a large plate consisting primarily of oceanic lithosphere. Southern Tibet was at that time on the south coast of Asia and lay about 2,000 kilometers south of where it is now. As the Indian and Eurasian plates collided the oceanic lithosphere north of the Indian landmass was bent down and thrust under Tibet, much as the plates under the Pacific Ocean are now being thrust under Japan, the Aleutians and South America. (Geologists call this process subduction.) It was as if the Indian plate were a conveyor belt trundling around a spool under southern Tibet.

Sometime between 55 and 40 million years ago the Indian landmass itself struck the south coast of Asia, and at that point the conveyor belt began to jam; the speed of the Indian plate was reduced from between 10 and 20 centimeters per year to about five centimeters per year (the rate at which India still plows into Eurasia today). As India plunged under Tibet with tremendous force, a northward-dipping fault tore through the northern edge of the subcontinent. The crust under the fault plane continued to move northward and downward, but a slice of continental shelf and deep crust above the fault plane was in effect shaved off the oncoming subcontinent and thrust backward on top of it. Between 20 and 10 million years ago the process was repeated: the first fault became inactive and a second fault formed at a deeper level. A second slice of Indian crust was thrust onto the subcontinent, lifting up the first slice. The eroded remnants of these two slices of ancient Indian crust are exposed today in the Himalayas, and they constitute the bulk of the range.

The heavy weight of the Himalayas bends the Indian plate downward south of the range. Over millions of years sediments eroded from the mountains have filled the resulting trough, forming the broad plains of the Ganges and Indus rivers. Seismological and drilling results obtained by the Oil and Natural Gas Commission of India have documented the presence of the trough in the Precambrian bedrock under the sediments. The bedrock dips smoothly down toward the mountains, reaching a depth of about five kilometers at the front. Some 200 to 300 kilometers south of the front, at the edge of the trough, the bedrock is exposed at the surface.

Considering the great weight of the Himalayas the trough is not very deep. The Indian plate does not bend much because it is particularly thick and strong. Just how thick it or any plate is cannot be determined precisely, but by making some simplifying assumptions one can calculate the comparative thicknesses of different plates. Hélène Lyon-Caen of the University of Paris has shown, for example, that the Indian plate is more than twice as thick as the Pacific plate under Hawaii.

The strength and thickness of the Indian lithosphere are a major reason the Himalayan peaks are so high. They are definitely not supported by deep crustal roots in the manner proposed by Airy. The thickness of the crust under the Himalayas is only about 55 kilometers, which is more than the 35 to 40 kilometers observed under the rest of India but much less than the 80 or so kilometers that would be needed to support the mountains solely by crustal buoyancy. The Himalayas are a clear case in which Airy's notion of local isostatic compensation fails and Vening-Meinesz' notion of regional compensation through bending of the lithosphere is required.

The Tibetan plateau, on the other hand, does fit Airy's conception. The plateau extends north of the mountains for hundreds of kilometers, and only in a few valleys near the edges does its altitude drop below 4,500 meters. Seismological work by Wang-Ping Chen of the University of Illinois at Urbana-Champaign and Barbara Romanowicz of the University of Paris suggests the Tibetan crust is generally between about 65 and 70 kilometers thick—thicker than the crust under the Himalayan peaks. The weight of the high plateau is compensated primarily by the buoyancy of its deep crustal root, as Airy proposed 130 years ago.

Other Ranges

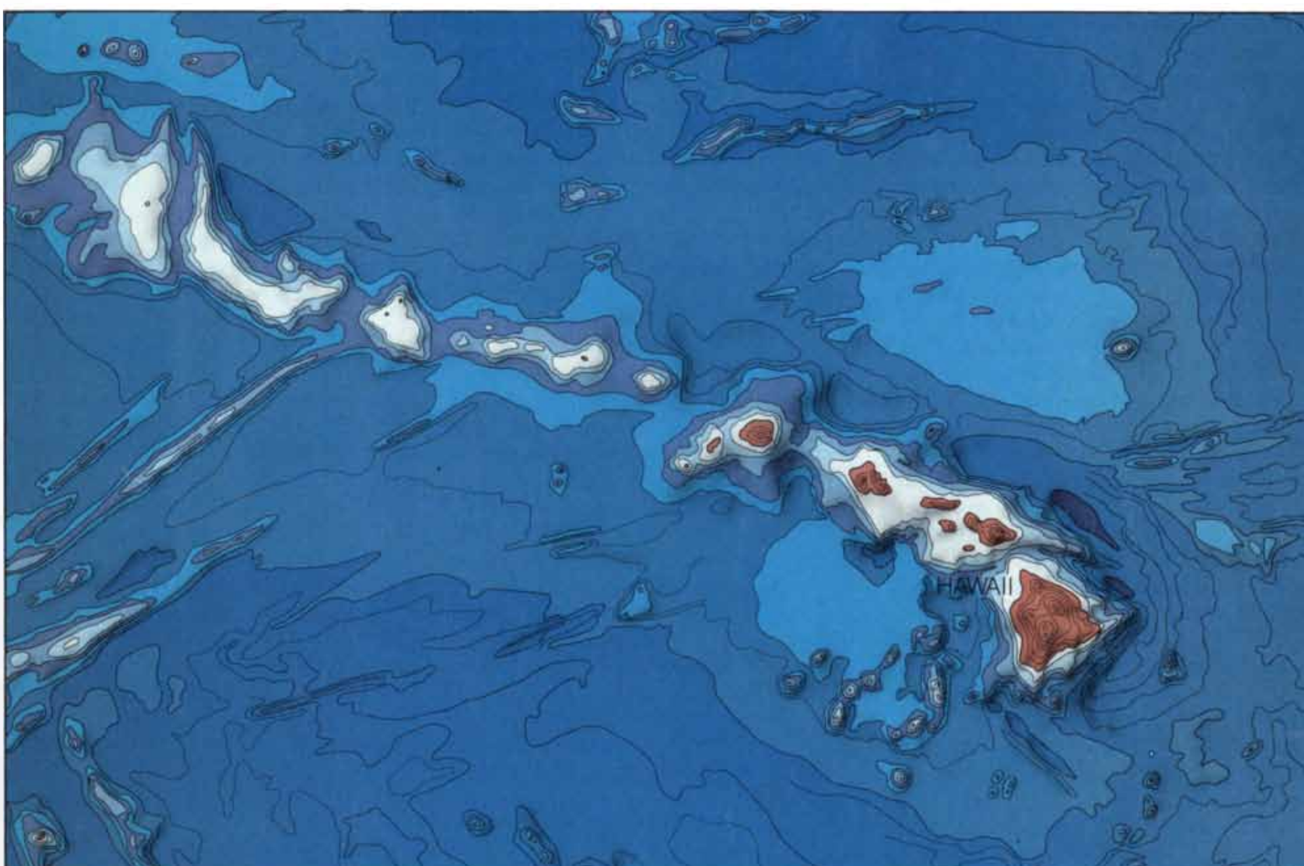
Other mountain ranges and high plateaus bear out the structural dichotomy exemplified by the Himalayas and Tibet. The abundant folding of rock layers in the Alps is evidence that they were formed in much the same way as the Himalayas, by crustal material that was sheared off the southern edge of Europe and overthrust northward onto the European plate when it collided with the Italian prong of the African plate. The Molasse basin in northwestern Switzerland is analogous to the Ganges and Indus plains: it is filled with debris eroded from the Alps, and its existence is due at least in part to the downward flexing of the European plate under the weight of the mountains. Garry D. Karner, now at the University of Durham, and Anthony B. Watts of the Lamont-Doherty Geological Observatory of Columbia University have shown that the European plate is less than half as thick as the Indian plate. The difference probably helps to explain why the Himalayas are nearly twice as high as the Alps. The Himalayas stand on a stronger foundation.

The Rocky Mountains in Canada also rest on a downward-flexed lithospheric plate. The exact manner in which the Rockies formed is still a matter of debate. It is clear, however, that the Canadian Rockies consist of slices of sedimentary rock that were successively detached from the underlying basement rock and thrust eastward on top of one another. A detailed exploration of this area has shown that the basement rock, part of the Precambrian Canadian shield, dips gently toward the west under the mountains. The dip in the lithosphere indicates that the weight of the mountains is compensated regionally, as Vening-Meinesz predicted.

Although they were formed in an altogether different way, the Hawaiian Islands provide another example of regional compensation. The islands were built up volcanically, by molten rock rising from the asthenosphere through the Pacific lithosphere and pouring out onto the ocean floor. The resulting peaks are gigantic: Mauna Kea on Hawaii stands more than 4,200 meters above sea level and some 9,000 meters above the surrounding ocean floor. The weight of the islands bends the Pacific plate downward by a few hundred meters, creating a "moat" around the islands. Just outside the moat the plate is warped slightly upward. The upwarping arises because the asthenosphere resists the bending of the lithosphere.

Whereas the Alps, the Canadian Rockies and the Hawaiian Islands have foundations similar to those of the Himalayas, the Andes, the highest mountains in the Western Hemisphere, are more akin to Tibet. The weight of the range seems to be supported by a buoyant crustal root as much as 70 kilometers deep. Indeed, the Andean crust has been the focus of debate on one of the main unresolved questions about mountain building: the question of how (as opposed to why, which Airy explained) the crust is thickened under many ranges.

There are two conceivable answers. First, the crust can be thickened by volcanic magma that wells up from the mantle and cools in the crust, forming intrusions of granite and other igneous rocks. Second, a block of

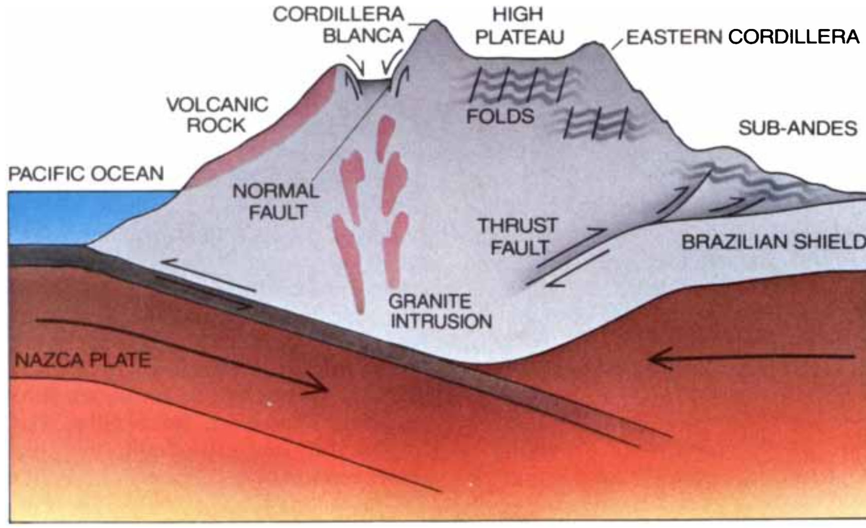


BENDING OF THE LITHOSPHERE is evident in this bathymetric chart of the region around the Hawaiian Islands. The islands are volcanic structures built of lava erupted onto the sea floor. Their weight bends the lithosphere downward by several hundred meters, forming a "moat" that is deepest around the island of Hawaii. A broad region of the sea floor around the islands is probably swelled by an upwelling of hot rock in the asthenosphere; to the north and south of the Hawaiian region the ocean gets progressively deeper.

crust becomes thicker if its edges are pushed together by horizontal forces and it is thereby shortened. In the Andes intrusive volcanism and crustal shortening occur side by side; the question is which process contributes more to the thickening of the crust.

The western cordillera of the Andes is a volcanic arc of the kind typically found above a subduction zone, where one lithospheric plate dives under another. As the Pacific-crust-bearing Nazca plate dives into the asthenosphere it is heated, and molten rock—either from the plate itself or from the asthenosphere above it—rises into the crust of the overriding South American plate, forming volcanoes and granitic intrusions. Consequently rocks in the western Andes and on the coastal plains of Peru and Chile are predominantly volcanic. Most of the rocks in the high central plateau and in the eastern cordillera, however, are not volcanic. Instead they are primarily sedimentary rocks folded and thrust on top of one another. The folding and overthrusting is evidence that the crust in those regions has been shortened in a direction perpendicular to the range.

Crustal shortening continues today on the eastern flank of the Andes. Seismograms analyzed by Douglas S. Chinn and Bryan Isacks of Cornell University, Gerardo Suárez of the National Autonomous University of Mexico and William Stauder of St. Louis University indicate that earthquakes on the eastern flank occur along faults where the Brazilian continental shield is thrust westward under the mountains. The rate of underthrusting is apparently only a few millimeters per year, but it may have been higher in the past. I have argued, along with Suárez, Lyon-Caen and B. Clark Burchfiel of the Massachusetts Institute of Technology, that this crustal shortening, and not volcanism, is largely responsible for the thick crust under the eastern cordillera. As the Andes are squeezed by the eastward thrust of the Nazca plate and the westward thrust of the Brazilian shield, the crust is thickened.



PERUVIAN ANDES are shown in a simplified cross section. The mountains are supported by a deep crustal root. Under the western Cordillera Blanca the crust has been thickened by intrusions of volcanic material rising above the Nazca plate as it plunges under South America. The convergence of the two plates also thickens the crust by pushing it together, or shortening it. Folded rock formations in the eastern sub-Andes prove that the crust there is being shortened and lifted up as the Brazilian shield is thrust under the mountains. Folded sedimentary rock in the high plateau suggests it was formed earlier by the same process. Hence many workers think crustal shortening rather than volcanism is primarily responsible for the height of the Andes and the thickness of the crust. Although the sides of the range are still being pushed together, the crust in the high Andes is stretching: on the western side of the Cordillera Blanca great blocks of crust have dropped down along normal faults. The Andes may be collapsing as the horizontal forces that support them diminish.

The Mountains Are Falling

The buoyancy of the crustal root supports the weight of the mountains, but the horizontal forces that created the root also seem to have a more direct effect. It appears that they help to buttress the Andes and prevent the range from spreading and collapsing. Ironically, evidence for this view comes in part from the observation that the buttresses are beginning to fail. While the crust on the sides of the range is being pushed together, the crust in some regions of the high Andes is being pulled apart. The Cordillera Blanca, a western chain that includes Peru's highest peak, Huascarán, is a good example. The chain is bounded on its western side by a steep fault running parallel to the range; to the west of the fault the crust has dropped and moved away from the mountains. This type of fault, along which one block of crust drops in relation to another, is called a normal fault. It is a clear indication that the crust is extending.

My introductory analogy between a mountain range and a Gothic cathedral may help to clarify the significance of the normal faults in the Andes. The Andean peaks and high plateaus, the roof of the Western Hemisphere, are like the vaulted ceiling of a cathedral. The vaults exert outward thrusts on the walls that tend to push the walls apart. (In the case of the cathedral the thrust is due not only to gravity but also to wind loading.) To prevent the ceiling from collapsing, Gothic architects built huge flying buttresses that countered the outward forces on the walls. Another way of solving the problem, which a present-day architect might favor, would be simply to stretch steel cables between the opposing walls; the cables would have enough tensile strength to hold the walls together.

In a sense the Nazca plate and the Brazilian shield are the flying buttresses of the Andes. Their inward horizontal thrust on the flanks of the range helps to support the high peaks and plateaus. What the normal faults in the high Andes suggest is that the horizontal buttressing forces are no longer strong enough to discharge their function satisfactorily; nor are the rocks that make up the mountains strong enough to serve the function of steel cables and hold the mountains together. Although the crust is still being pushed together on the eastern flank of the mountains, in the high Andes it is being stretched apart: the roof is falling. The range as a whole may be entering a stage of decline that will ultimately lead it to collapse completely under its own weight.

If the Andes do collapse, they will probably not have been the first mountains to suffer such a fate. Many workers, myself included, consider the Andes a modern analogue of a mountain range that dominated the western U.S. between 80 and 30 million years ago, when the Rockies were just being built farther east. At that time a lithospheric plate under the eastern Pacific Ocean was converging with the North American plate, and the oceanic lithosphere was being subducted, just as the Nazca plate is now being subducted under South America. Sometime between 30 and 10 million years ago the subduction under North America stopped.

With the cessation of convergence the horizontal forces that had shortened and thickened the crust under the mountain range would have diminished or even disappeared.

When that happened, the crust would have begun to spread. Signs of crustal extension are plentiful in the Basin and Range province west of the Rockies, between central Utah and the Sierra Nevada: the alternating basins and tilted ranges are bounded by normal faults like the ones west of the Cordillera Blanca. As the crust has spread apart, blocks of crust have dropped along the normal faults to form the basins. (One of the basins is Death Valley, which is now below sea level but may once have been at an altitude of several kilometers.) According to this view, the Basin and Range province is the remnant of a broad belt of mountains and high plateaus that collapsed after the horizontal forces supporting the belt were removed. One day the Andes may look like the Basin and Range.

Tibet too may be collapsing. Although the pressure applied by India's northward motion to the rest of Asia seems to be enough to prevent Tibet from extending in a north-south direction, the high plateau has no similar buttress on its eastern flank. Accordingly the plateau is laced with northerly trending normal faults, along which one side moves down and away from the other. Tibet is spreading to the east, and in the process it is pushing southeastern China eastward with respect to the rest of Asia.

Why are the Andes and Tibet in particular susceptible to collapse? It is precisely because they are supported mainly by deep crustal roots. The strength of crustal rock decreases rapidly with increasing temperature and hence with increasing depth, probably more rapidly than the strength of mantle rock. Thick crust therefore tends to be weak. Moreover, for reasons that are not fully understood, the crust and upper mantle under Tibet and the Andes seem to be comparatively warm; indeed, the boundary between the cool lithosphere and the warm asthenosphere may actually lie within the crust rather than well below the crust, as it does in most areas. As a result the crustal roots of Tibet and the Andes appear to be weak and fluidlike, and to the extent that they are not restrained by horizontal buttressing forces they tend to spread out. The horizontal forces are what keep the roots deep and the plateaus high. Paul Tapponnier of the University of Paris and I have suggested that the plateaus can be thought of as pressure gauges: the more they are subjected to horizontal pressure, the greater their elevation is. In Tibet and the Andes the pressure may have begun to drop.

The Himalayas, the Alps and the Rockies, on the other hand, are supported primarily by strong, thick lithosphere consisting of relatively cold crust and mantle. (The crust under the Himalayas, for example, is much colder than the Tibetan crust because it is underthrust by the cold Indian plate.) Although these ranges were formed by horizontal forces, they have no need of horizontal buttressing to remain standing. Judging from the general absence of normal faults in them, they do not seem to be collapsing.

Dynamics

Some mountain ranges are like pressure gauges, others are like loads on elastic plates—the analogies are correct as far as they go, but I must stress that they are simplifications. At a finer level of detail they reveal their limitations. Karner and Watts have shown, for instance, that the weight of the Alps is not great enough to bend the European plate down as much as it is under the Molasse Basin; some additional force must pull the plate down. In contrast, Lyon-Caen and I have found that the strength of the Indian plate, great as it is, does not fully ac-

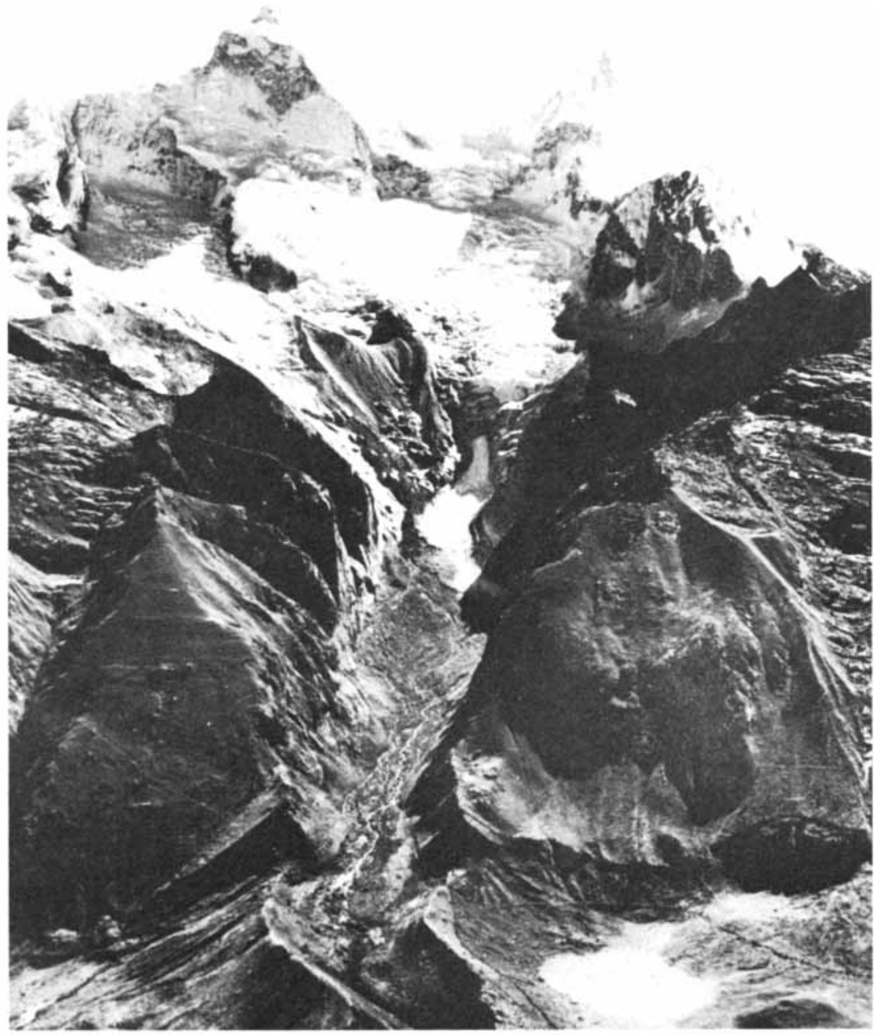
TILTED AND FOLDED sedimentary rock in the high plateau of the Andes indicates the plateau was formed by shortening of the crust. The view is to the southeast from east of the Cordillera Blanca. The folded limestone, sandstone and shale layers were pushed together from the east and west. The rocks in the distant snow-capped mountains, part of the western cordillera, are also folded.

count for the high altitudes in the Himalayas; an additional force apparently flexes the northern edge of the plate up. Hawaii also is pushed up. The depth of the ocean floor in a large area around the islands outside the Hawaiian moat is only about 4,500 meters, whereas 1,000 kilometers or so to the north and south the ocean depth is about 5,500 meters.

These deviations indicate that the simple model of a plate bending under the weight of the mountains is incomplete. What is missing is a consideration of plate dynamics, of the forces that drive continents together, shorten the crust and cause huge terrains to be thrust onto the edges of strong plates. Plate motions are widely thought to be the surface manifestations of a convective circulation that extends deep into the mantle, but the overall pattern of the circulation is not well known.

Nevertheless, some conclusions can be drawn. It seems clear, for example, that Hawaii lies above a region of the asthenosphere where hot material wells upward. Some of the material erupts at the volcanoes on the islands, but the upwelling column is much broader than the islands themselves. The upward thrust of the hot material accounts for the broad swell in the sea floor around Hawaii.

Under other mountain ranges one might expect to find a downwelling of relatively cold material. Under the Himalayas the Indian plate, stripped of the crustal slices that make up the mountains, may be plunging into the asthenosphere. The material in the upper part of the plate is significantly colder and therefore denser than the asthenosphere, and so it should sink. The weight of the sinking material may help to pull the plate down. At the same time, as Lyon-Caen and I have contended, the part of the plate just behind the leading edge would be flexed upward, which would help to push the mountains up. (To visualize the phenomenon take a plastic or metal ruler and bend one end over the edge of a table.) The sinking material might also drive a circulation in the mantle that helps to push the Indian and Eurasian plates together.



NORMAL FAULTS are responsible for the steep western face of the Cordillera Blanca. The peak on the left, Huandoy, is 6,356 meters high. The escarpment running along the base of the mountains and across moraines left by a receded glacier is an active normal fault. In relation to the mountains the valley in the foreground has dropped several kilometers along this and parallel faults. Apparently the range is collapsing as the crust spreads.

Measuring Gravity

How can one study the dynamics of the mantle and determine in particular if dense sinking material is present under mountain ranges? One method is to measure variations in the earth's gravity field; the field should be slightly stronger above regions of the earth that are underlain by dense material. Unfortunately the differences in gravity caused by density variations in the mantle are small, probably less than about .01 percent of the average value of 9.8 meters per second squared. In a mountainous region they are masked by the much larger differences caused by variations in topography. To correct for topographic effects one must have extremely accurate maps, which in areas such as the Himalayas are simply not available.

The solution is probably to measure gravity with satellites. A satellite travels far above the gravitational influence of ridges and valleys, but its orbit is slightly perturbed by gravitational anomalies resulting from density variations in the mantle. By tracking the perturbations one can map the gravity field and the density variations. So far only large-scale gravity anomalies, thousands of kilometers wide and unrelated to mountain ranges, have been mapped. With improved tracking or, alternatively, with new satellite-borne instruments that directly measure lateral variations in gravity, it should eventually be possible to detect the smaller anomalies caused by density variations under mountain ranges.

When such measurements become available, a major step will have been taken toward understanding mountains not as static features but as features that grow and decay, as elements of an evolving earth. A deeper appreciation of the dynamics of mountain ranges will undoubtedly force a modification of some of the simple concepts I have presented here. Until then geophysicists fascinated by mountain architecture will remain in a situation not unlike that of the Gothic architects, who found they could support giant cathedrals with flying buttresses but who never really understood the underlying physical principles.

Exotic Atomic Nuclei

They have abundances of neutrons and protons quite different from those in the nuclei found in nature. Their unusual properties are yielding new views of motions and structure within nuclear matter

by J. H. Hamilton and J. A. Maruhn

When an object is less than a millionth of a millionth of a centimeter in size, even its simplest properties cannot be measured directly but have to be deduced. This makes it all the more remarkable that the properties of the atomic nucleus, the cluster of protons and neutrons at the center of an atom, are becoming rather well known. It was long supposed, for example, that all nuclei are spherical. It then emerged that some nuclei are prolate (shaped like an American football). It now turns out that atomic nuclei seldom are spherical; research made possible by a new generation of particle accelerators and detection systems shows that nuclei can be not only spherical or prolate but also oblate (like a discus), triaxial (like a partially deflated football) or octupole (like a pear). Indeed, an interplay of deformations can render a nucleus peanut-shaped or make it look like a football with a bulge around the middle. Certain newly discovered nuclei are found to have deformations far greater than it was ever thought a stable nucleus could have. Moreover, in contradiction of ideas of only a decade ago, certain nuclei turn out to have more than one fixed shape.

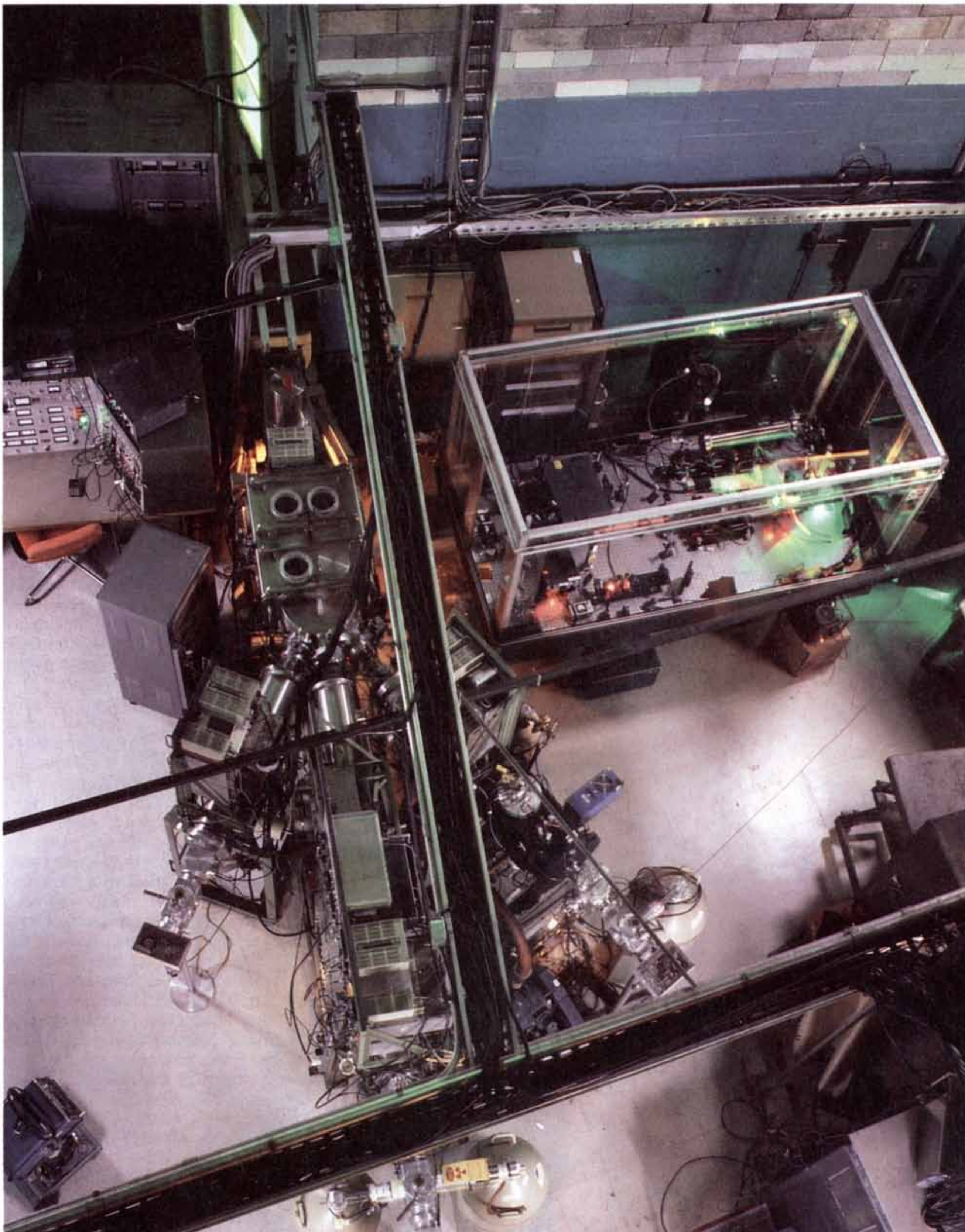
Many of these new insights come from studies of atomic nuclei under exotic conditions far removed from those of the stable nuclei found in nature. By exotic we mean that a nucleus has a total number of neutrons or protons markedly different from those of the nuclei found naturally in the crust of the earth. For example, in the course of our experiments at the Oak Ridge National Laboratory the isotope thallium 184 was discovered. Like all other thallium atoms its nucleus includes 81 protons. (The number of protons is the same for all the isotopes of a given chemical element.) On the other hand, thallium 184 has only 103 neutrons, or 19 fewer than the lightest thallium isotope found naturally in the earth. (The number of neutrons varies from one isotope of an element to another.)

How has the study of exotic atomic nuclei modified the understanding of what might be called the nuclear landscape? For one thing, new "magic numbers" have been identified. The established magic numbers represent quantities of protons or neutrons that make a spherical nucleus resistant to changes in properties such as shape; the new numbers significantly augment the list of the old ones by bestowing a similar stability on deformed nuclei. The existence of the new numbers sometimes leads to conflicts between one magic number (40, for example) that can stabilize a spherical nuclear shape and a nearby number (38) that can stabilize a deformed shape. Moreover, in certain exotic nuclei the abundances of protons and neutrons are both near deformed magic numbers. A strong reinforcement then occurs. The result is superdeformation: the most deformed ground states, or states of least energy, ever observed in atomic nuclei.

The study of exotic atomic nuclei has also led to the discovery of exotic modes of radioactive decay. In the familiar modes, called alpha, beta and gamma decay, a radioactive nucleus emits an alpha particle (two neutrons and two protons, which amounts to a helium-4 nucleus), a beta particle (an electron or a positron) or a gamma ray (a high-energy photon, or quantum of electromagnetic radiation). The newly discovered, exotic modes begin with a beta decay. After that the nucleus emits one or two protons, one, two or three neutrons, or even a hydrogen-3 nucleus, which consists of one proton and two neutrons. In other and still more exotic modes the decaying nucleus emits a cluster of eight neutrons and six protons (a carbon-14 nucleus) or even 14 neutrons and 10 protons (a neon-24 nucleus). The cluster decays offer clues to the ways nucleons (protons and neutrons) are organized inside a nucleus.

The insights provided by the study of exotic atomic nuclei are best understood as an outgrowth of explorations of the nuclei found in nature. These explorations span less than 80 years; before the end of the 19th century the nucleus was unknown, and the atom was thought to be the smallest unit of matter. Then in 1896 radioactivity was discovered. In effect the alpha, beta and gamma rays emitted by certain atoms were ambassadors bringing the first news of the existence of an unknown world, the nucleus of the atom. It was another 14 years before Ernest Rutherford established that all the positive electric charge in an atom, and nearly all the mass, are contained in the nucleus, a minute central body 100,000 times smaller than the atom itself.

The ensuing 30 years brought little new knowledge. The nucleus was conceived of as resembling a spherical liquid drop that could vibrate or be broken apart into smaller drops. Then came two insights. The first of these, the spherical-shell model of the atomic nucleus, was proposed in 1949 by Maria Goeppert Mayer of the Argonne National Laboratory and the University of Chicago and J. Hans D. Jensen of the University of Heidelberg. It imposes on the nucleons in a nucleus the same principles of quantum mechanics that had been applied earlier to the electrons orbiting the nucleus. In particular, quantum mechanics establishes that the neutrons and protons confined in a nucleus have available to them a set of discrete states, each corresponding to a particular amount of energy and a particular pattern of motion. On the other hand, the law of quantum mechanics called the Pauli exclusion principle dictates that in a nucleus no neutron can occupy the

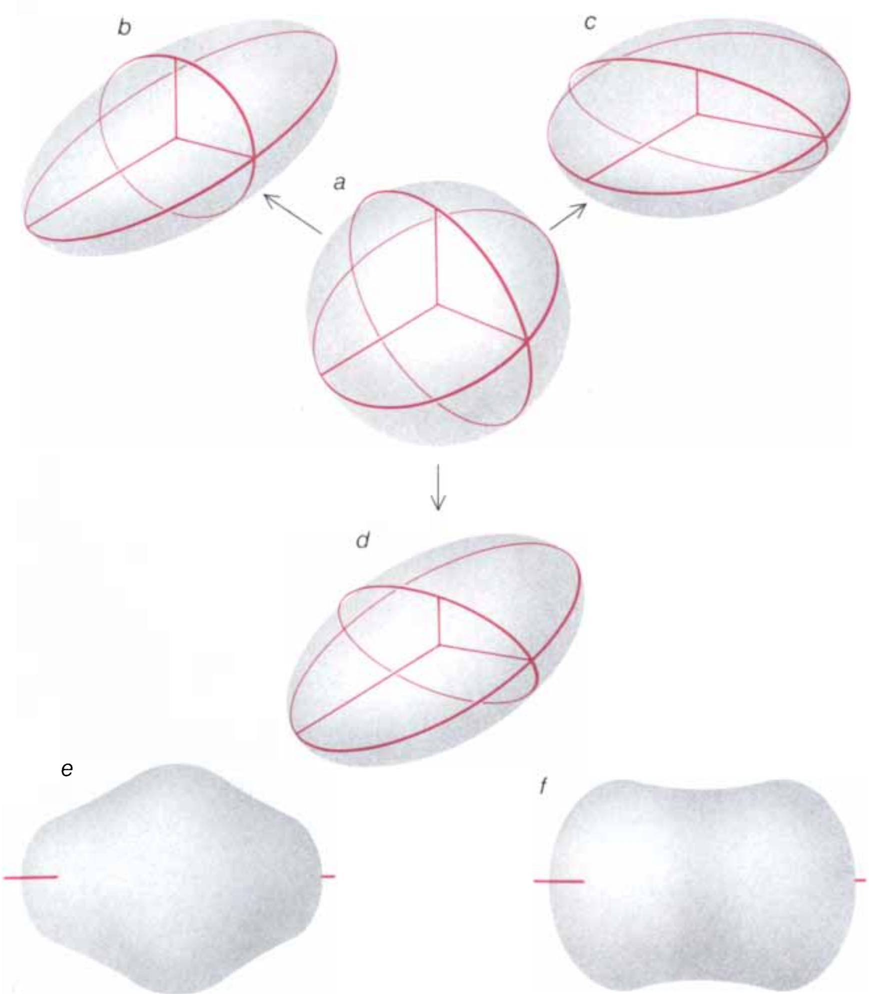


ON-LINE ISOTOPE SEPARATOR at the Holifield Heavy Ion Research Facility of the Oak Ridge National Laboratory in Tennessee serves investigations of exotic atomic nuclei. The exotic nuclei emerge, behind concrete shielding, from collisions of heavy ions and the nuclei of atoms in a target, typically a metal foil. A set of magnets then divides the exotic nuclei into streams, each stream consisting of nuclei of a particular mass. A given stream can be shunted into any one of three channels, which diverge toward the bottom of the photograph. The channel at the right also receives a beam of laser light, which excites the electrons orbiting each exotic nucleus. In turn the electrons interact with the nucleus, providing investigators with a means of determining the average radius of the nucleus, a clue to the nuclear shape. The laser light is produced by a sequence of two lasers installed in the enclosure at the center right.

state filled by another neutron and no proton can occupy the state filled by another proton. As a result a nucleus is "constructed" by filling a succession of states with neutrons and a succession of states with protons, beginning in each case with the state that has the least energy.

The states tend to cluster in sets; hence large energy gaps separate the most energetic state in one set, or shell, and the least energetic state in the next-higher shell. Here the similarity between the quantum mechanics of nucleons in the nucleus and that of electrons orbiting the nucleus becomes particularly striking. The filling of an atom's electron shells yields an inert chemical element: helium, neon, argon and so on. In nuclei the case is much the same: according to the spherical-shell model, the filling of nuclear shells yields a nucleus whose sphericity is notably "hard," or unchanging. Thus the model becomes pervaded by the magic numbers 2, 8, 20, 28, 40, 50, 82, 126 and 184. In the jargon of nuclear physics they identify spherical nuclei that have closed shells of neutrons or protons.

The spherical-shell model therefore establishes some nuclear categories. Nuclei that match two spherical magic numbers (for example oxygen 16, with eight protons and eight neutrons, or lead 208, with 82 protons and 126 neutrons) have particular stability in their properties, including a spherical shape: they are spherical-double-magic. Nuclei that match one magic number and thus have one closed shell, either of neutrons or of protons, gain in their stability and are thought to be nearly spherical. The same applies to nuclei whose endowments of neutrons and protons place them only a few nucleons away from a magic number. Conversely, among deformed nuclei the "hardest" ones, namely the ones with the greatest stability in their properties, are the ones that are farthest from any spherical magic number: their shells are half-filled in regard both to neutrons and to protons. "Soft," slightly deformed nuclei are found in the periodic table between spherical nuclei and deformed ones.



RANGE OF SHAPES of atomic nuclei is much broader than was once thought. For spheroids *(top)* each shape is described by the lengths of three mutually perpendicular axes *(colored lines)*. In a spherical nucleus *(a)* the axes are all equal. In a prolate spheroid, or football-shaped nucleus *(b)*, one axis is longer than the other two, which are equal. In an oblate spheroid, or discus-shaped nucleus *(c)*, one axis is shorter than the other two, which are equal. In a triaxial nucleus *(d)* the axes are all unequal. Some nuclei manifest additional, higher-order deformations, which make their shape more complex. Examples include the ground state, or state of least energy, of the nuclei of uranium 234 *(e)* and hafnium 180 *(f)*.

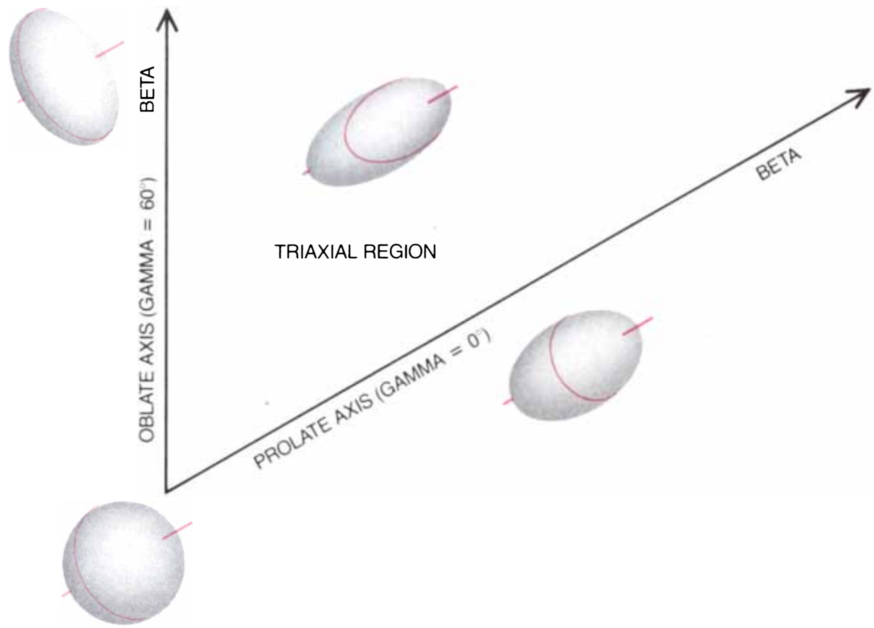
The second insight into the shape and structure of the atomic nucleus was the collective model of nuclear deformation, proposed in 1952 by Aage Bohr and Ben Mottelson of the Niels Bohr Institute in Copenhagen. In essence it imposes quantum mechanics on the earlier understanding of the nucleus as a liquidlike droplet of nuclear matter. In the collective model one imagines that the droplet could take any shape; one then calculates, in a framework of quantum mechanics, the energy of each shape and of departures from that shape. Some nuclei turn out to have a pattern in which the energy has its minimum at a particular degree of deformation and rises steeply for any departures; such nuclei are hard and deformed. Others show little change in energy over a range of deformations; such nuclei are soft.

The theoretical program of devising models of atomic nuclei has of course been complemented by experimental investigations. In one technique investigators record the energies of the gamma rays and other particles emitted by nuclei in excited states. The atomic nucleus absorbs and emits energy only in quanta, or discrete units. Each absorption marks its transition to a state of higher energy and each emission marks its transition to a state of lower energy. In each state the nucleus has a particular amount of energy and angular momentum. Moreover, it has a particular shape and supports particular rotations and vibrations, or motions among its nucleons. From the recorded data the energy levels of the nucleus are determined. The resulting energy-level diagram is in essence a shorthand representation of what nature allows the nucleus to do.

This is not to say the experiments are easy. Indeed, by the late 1960's more than 5,000 nuclei had been predicted on theoretical grounds but fewer than 1,600 had been identified and fewer than 400 of those had been examined well enough so that at least some of their low energy levels were known. The problem is that unstable atomic nuclei tend to transform themselves into nuclei that are more stable, as manifested by their resistance to radioactive decay. For nuclei near stability the energy available for such transformation, in the form of a beta decay, is less than one million electron volts (MeV); the half-life of such a nucleus is typically measured in months or years. For nuclei far from stability the available energy is 5 to 10 MeV and the half-life is seconds or less.

By the late 1960's, therefore, the surveying of the nuclear landscape had begun to demand the study of exotic atomic nuclei. It would be a difficult project. In each experiment highly complex data (the energies of hundreds or even thousands of gamma rays) would have to be analyzed in searches for the few nuclear transitions revealing nuggets of new information. Moreover, in each experiment the data would have to be gathered fast, sometimes no more than hundredths or even thousandths of a second after a nucleus had been created. Investigators were eager to solve such problems. New regions of deformed nuclei were expected, and new double-magic nuclei, such as tin 100 (50 protons and 50 neutrons) or tin 132 (50 protons and 82 neutrons). Would such nuclei be spherical and "hard"? Or would the spherical magic numbers that yield closed-shell nuclear structures lose their magic character in regions of the periodic table far from stability? Are the magic numbers different there? Do the nuclei in such regions show new types of deformation?

The resolution of the experimental difficulties associated with producing and probing exotic atomic nuclei came in the form of an on-line isotope-separation system (or ISOL: isotope separator on-line). A target, typically a metallic foil, is bombarded on one side by a beam of particles emerging from a particle accelerator or a nuclear reactor. The particles fuse with the nuclei inside the target. On the other side the resulting exotic nuclei, propelled from the target, come under the influence of a strong magnetic field. Since nuclei have electric charge, they are deflected by the field; the amount of the deflection depends on their mass. As a result a continuous stream of exotic, short-lived nuclei can be divided into beams, each beam consisting of nuclei of a particular mass. Each beam still contains a number of different types of nuclei, incorporating combinations of the neutron number N and the proton number Z that sum to the same mass number A. Nevertheless, the characteristic X rays that the nuclei of each chemical element emit in coincidence with other modes of decay serve to identify each nucleus.

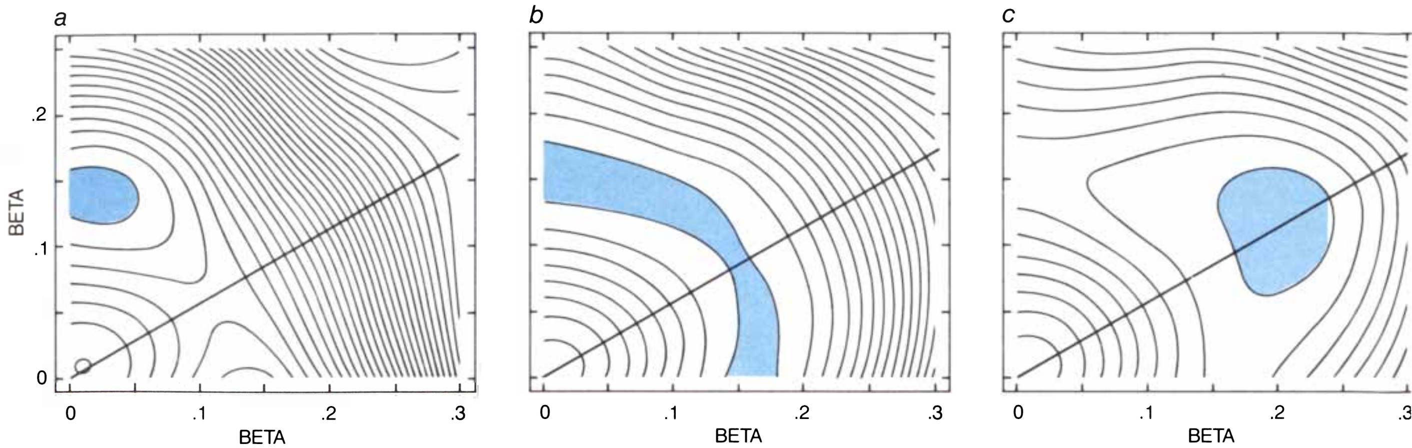


BETA-GAMMA PLANE represents a mathematical shorthand for the range of possible shapes among spheroidal nuclei; each point in the plane corresponds to a value of beta and a value of gamma, and hence to a particular deformation of a sphere. Beta, a number equal to or greater than 0, is a measure of the type of deformation that produces a prolate spheroid by lengthening one of the axes of a sphere at the expense of the other two. Gamma, an angle ranging from 0 to 60 degrees, is a measure of the type of deformation that turns a prolate spheroid into a triaxial spheroid by flattening its circular cross section. If the gamma deformation is sufficiently great, a prolate spheroid is transformed into an oblate spheroid.

The first large-scale effort to study nuclei far from stability by means of an ISOL system was undertaken in the late 1960's at CERN, the European laboratory for particle physics in Geneva, as a collaborative European effort called ISOLDE. The CERN device bombarded its target with a beam of high-energy protons. Then in 1972 the UNISOR (University Isotope Separator at Oak Ridge) began operation at the Oak Ridge National Laboratory. A collaboration organized by Vanderbilt University among investigators from the Federal Government, the state of Tennessee and a number of universities, UNISOR was the first major effort in which a beam of heavy ions served efforts to determine the structure of nuclei far from stability. The ISOLDE and UNISOR consortiums have carried out large-scale, systematic studies that would have been impossible for any one institution.

By now the patterns of energy levels available to a nucleus, stable or exotic, are understood well enough so that the pattern for a newly created exotic nucleus is sufficient to establish that the nucleus is spherical or deformed or that it has multiple "personalities": a nearly spherical shape in some of its states and a deformed shape in the others. The clues to sphericity are easy to find. The stable nucleus nickel 62 offers a good example [*see bottom illustration on next two pages*]. For one thing, the amount of energy required to raise nickel 62 from its ground state to its first excited state is large: 1.17 MeV. Hence the energy has powered a vibration rather than a rotation. (The energy required to power a rotation is typically less than .2 MeV.) This absence of observable rotation is what indicates that the nickel-62 nucleus is more or less spherical. The remarkable fact is that a deformed nucleus can rotate but a spherical one cannot. To put it another way, quantum mechanics dictates that the rotation of a nucleus about an axis of symmetry is undetectable, and in a sphere every axis is a symmetry axis.

The energy-level diagram for a deformed nucleus is more complex than the diagram for a spherical nucleus, but it too has certain telltale characteristics. The diagram for the stable nucleus gadolinium 154 is typical. A gap of only .123 MeV intervenes between the ground state and the first of a series of closely spaced excited states. The excited states represent successively more energetic rotations resembling those of a football tumbling end over end. (In quantum mechanics a prolate, or football-shaped, nucleus can tumble end over end but cannot spin about its long axis. The long axis is a symmetry axis. It is as if a football could be

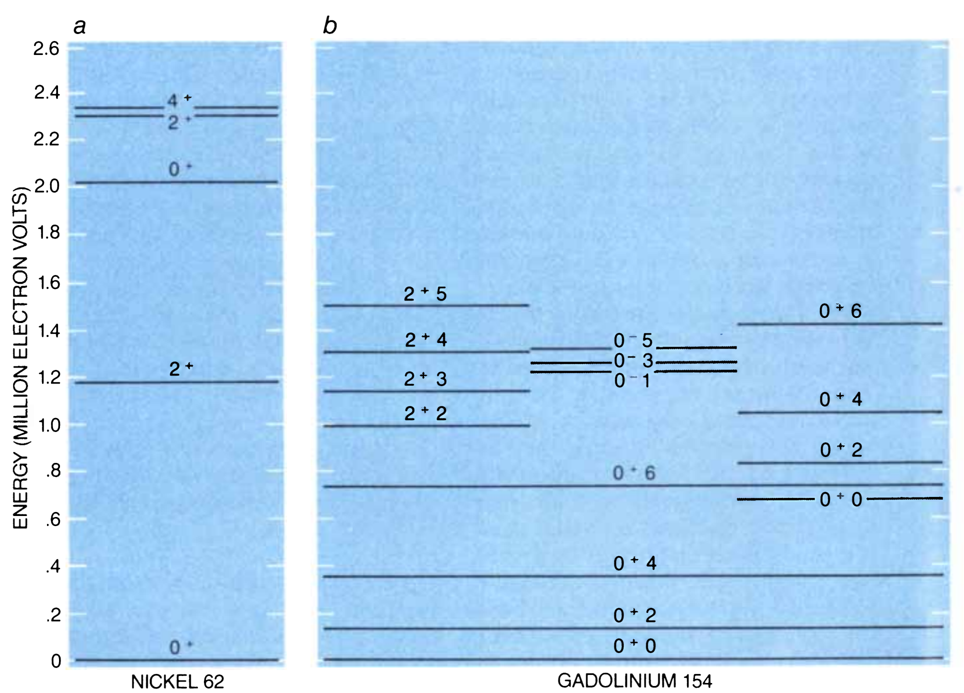


POTENTIAL-ENERGY SURFACES, which chart the energy in a nucleus as a function of its shape, are calculated by assuming that the nucleus could take on any shape. The shape that nature favors is then the one embodying the least amount of energy. Here potential-energy surfaces are charted for the ground state of platinum isotopes ranging from highly stable to fairly exotic, or unstable. The latter contain fewer neutrons than the platinum nuclei found in nature. Each chart follows the scheme set in the preceding illustration: beta increases with distance from the origin, the line for a gamma of 0 degrees rises diagonally and the line for a gamma of 60 degrees forms the vertical axis. Platinum 192 (*a*), with its region of minimum energy (*color*) abutting the vertical axis, is oblate. Plati-

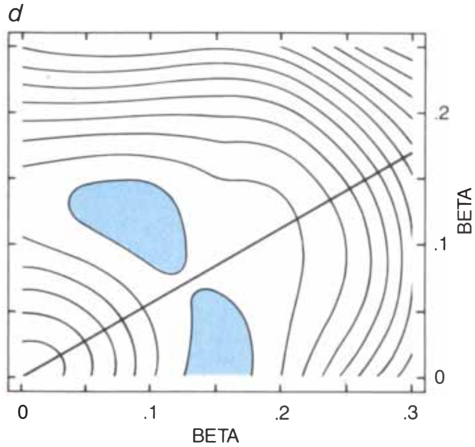
place-kicked but could not be given a spiraling forward pass.) The pattern for gadolinium 154 is completed by bands of vibrational modes of the nucleus, each mode combining with end-over-end rotations.

Then there are nuclei that manifest the newly recognized property of multiple "personalities," or shape coexistence. In this regard the isotopes of mercury are instructive. Stable or exotic, they all have 80 protons, or only two less than 82, a spherical magic number. On that basis one would expect them all to have the widely spaced energy levels characteristic of a near-spherical nucleus. The isotopes mercury 198 through mercury 192 do show such a pattern; indeed, in the early 1970's the only sign of something odd about mercury isotopes was the evidence, discovered at ISOLDE by Ernst W. Otten and his colleagues at the University of Mainz, suggesting a change in properties between mercury 187 and mercury 185. Our investigations, undertaken at UNISOR in 1973, began with mercury 190. It too had widely spaced levels. Then came something unexpected. Experiments at the Lawrence Berkeley Laboratory, at the Chalk River Nuclear Laboratories and at UNISOR revealed that the first excited energy level for the progressively lighter, progressively more exotic mercury isotopes 188, 186 and 184 was indicative of a near-spherical nucleus; in contrast, the higher levels showed a shift to a pattern characteristic of a deformed nucleus. The Berkeley and Chalk River experiments, however, measured the energies of only some of the states available to the nuclei.

Our work at UNISOR filled out the energy-level diagrams. The resulting patterns showed that the nuclei of mercury 188, 186 and 184 do not merely change from one hard shape at low energy levels to another, quite different hard shape at higher levels. Each of the nuclei has two distinct personalities: a full set of states built on a near-spherical ground state and a full set of states, initially spaced at small energy differences and then at increasingly large ones, characteristic of the rotations and vibrations of a hard deformed nucleus. A particular energy level from a near-spherical set is sometimes quite close in energy to a level from a deformed set. The data confirm theoretical predictions, made as early as 1953, that states characterized by different shapes can coexist in a nucleus. The predictions were made by David L. Hill of Vanderbilt and John A. Wheeler of Princeton University, by Walter Greiner of the University of Frankfurt and by Vadim G. Soloviev



ENERGY-LEVEL DIAGRAMS display the quantized amounts of energy a nucleus absorbs or emits in the course of its transitions from one state to another. Each diagram offers clues to nuclear properties such as shape. Nickel 62, a stable isotope (*a*), has an almost spherical nucleus. The large energy gap between the ground state, 0^+, and the first excited state, designated 2^+, indicates that the energy has fueled the vibrations of a sphere. Gadolinium 154, a stable isotope of a rare-earth element (*b*), has a deformed nucleus. The



num 188 (*b*), with its minimum energy ranging through all values of gamma, is a "soft" nucleus, unable to maintain a lasting shape. Platinum 180 (*c*), with its minimum energy straddling the line for gamma equal to 0, is prolate. Platinum 176 (*d*) is triaxial.

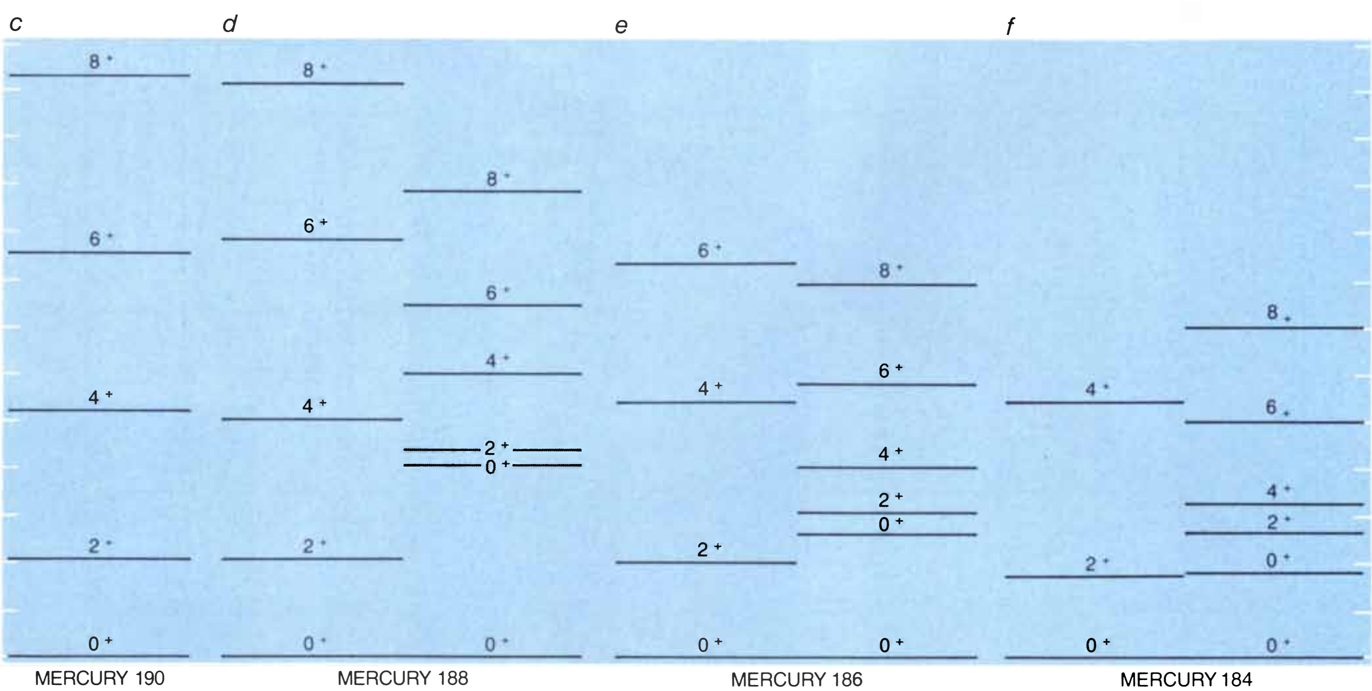
of the Joint Institute for Nuclear Research at Dubna in the U.S.S.R.

Our discovery of shape coexistence in mercury isotopes was accompanied by a similar discovery in selenium 72 and selenium 74. Since then discoveries of shape coexistence have been made even in nuclei that have proton magic numbers, such as tin 116 (50 protons) and lead 196 (82 protons). Shape coexistence is now known to be widespread in the periodic table. The phenomenon constitutes a bridge between spherical nuclei and deformed ones, which had seemed to be separate classes; hence it is proving to be important in developing a unified description of nuclear matter.

As part of the effort to develop such a description, the status of magic numbers—and thus the power of the spherical-shell model—is being evaluated in realms far from those of the stable nuclei found in nature. In this respect a series of remarkable findings have emerged in the part of the periodic table below a *Z* of 50. No nucleus in that region can be very far from at least one spherical magic number: 2, 8, 20, 28, 40 or 50. For that reason our discovery of shape coexistence in selenium 72 (34 protons, 38 neutrons) and selenium 74 (34 protons, 40 neutrons) was quite unexpected. It came as a further surprise when exotic nuclei with abundances of protons and neutrons both near the magic number 40 proved to have the largest ground-state deformations ever observed. A collaboration among investigators from Vanderbilt, Oak Ridge, the University of Cologne, the Central Institute for Nuclear Research near Dresden and the Lanzhou Institute of Modern Physics in China discovered two such nuclei: krypton 74 (36 protons, 38 neutrons) and krypton 76 (36 protons, 40 neutrons). The ISOLDE collaboration found a third: strontium 100 (38 protons, 62 neutrons).

There are at least two justifications for calling these nuclei superdeformed. The extent of a nuclear deformation is specified by the parameter beta, which roughly corresponds to the difference in length between the long axis and the short axis of a nucleus divided by the average between those axes. Near-spherical nuclei have a small value of beta, say .1, whereas typical deformed nuclei have values of from .2 to .25. The beta for a superdeformed nucleus in its ground state is from .35 to .4.

The corroborating justification derives from the energy of the first excited level, which in a deformed nucleus arises (as noted above) from rotation. The greater the deformation, the lower the energy of the first excited state. When the measured values are scaled to allow for the different masses of different nuclei, the deformed actinide nuclei, whose *Z*'s are in the 90's, have first-excited-state energies of 43 to 45 keV (thousand electron volts). Before the discovery of the superdeformed nuclei the lowest value known—42.8 keV—was that of the first excited state of the actinide nucleus plutonium 240. The superdeformed nucleus strontium 100 has a scaled first-excited-state energy of 30 keV; the superdeformed nu-



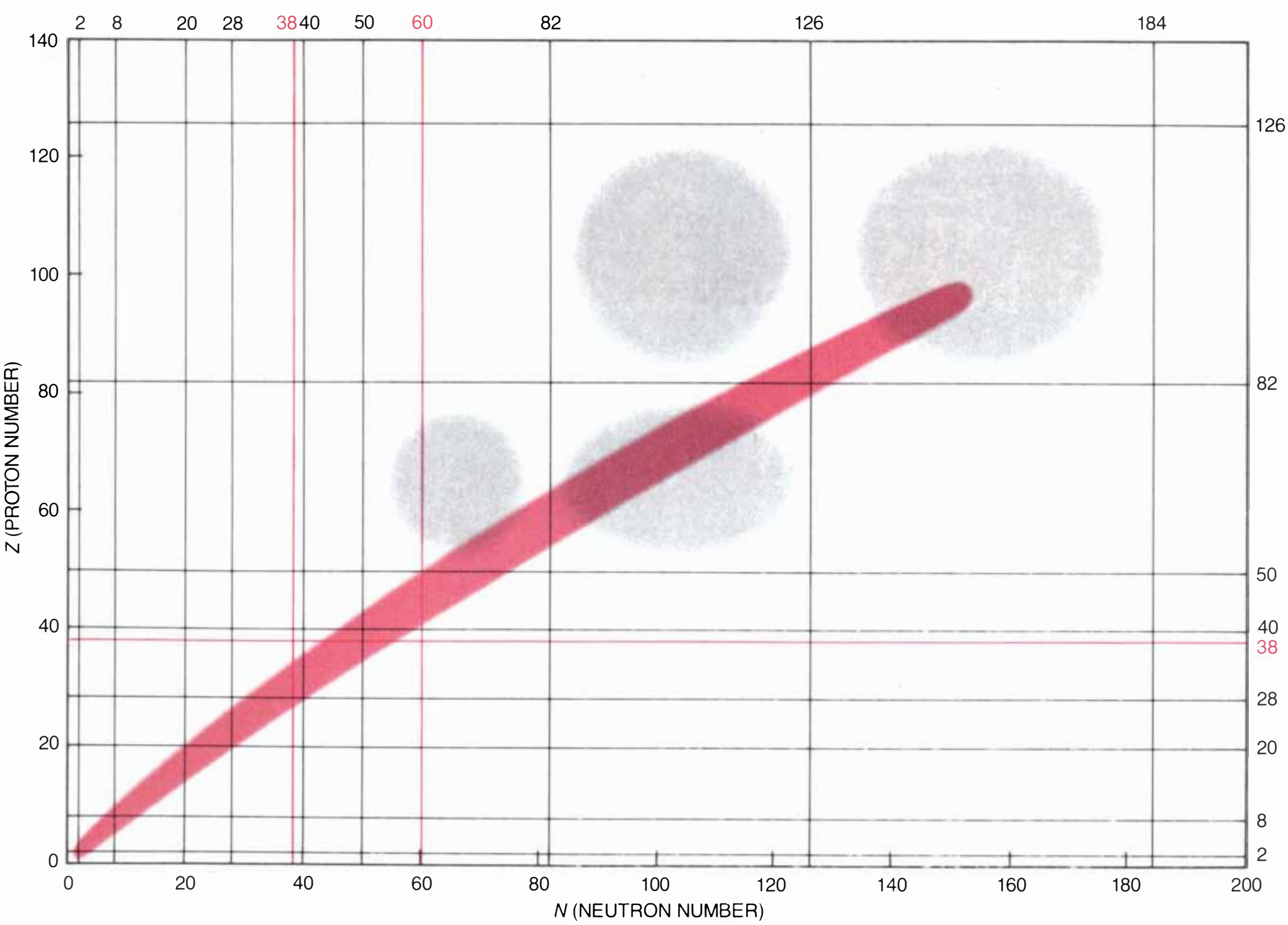
small energy gap between the ground state, 0^+0, and the first excited state, 0^+2, indicates that the energy has fueled the end-over-end rotations of a prolate spheroid. The rest of the pattern is formed by states representing vibrational modes superposed on end-over-end rotations. Isotopes of mercury have a variety of nuclear shapes. Mercury 190 (*c*) shows the widely spaced levels characteristic of a more or less spherical nucleus. In contrast, the progressively more exotic isotopes mercury 188 (*d*), mercury 186 (*e*) and mercury 184 (*f*) have dual "personalities": a set of widely spaced levels revealing a spherical shape and an interleaved set of more closely spaced levels revealing a prolate shape. The numbers accompanying each level are physicists' shorthand for the quantized values of nuclear properties such as spin.

cleus krypton 74 has a scaled first-excited-state energy of 28 keV.

Even as superdeformed nuclei were being discovered, theoretical calculations were predicting their existence. In 1981 Peter Möller of the University of Lund in Sweden and J. Rayford Nix of Los Alamos calculated the minimum potential energy for given numbers of protons and neutrons and derived the masses and shapes for the ground state of more than 4,000 nuclei. The work can be seen as part of a larger effort that had begun in 1955, when S. G. Nilsson of the University of Lund first calculated the energy levels of the individual nucleons in a deformed nucleus, and would continue in 1984, when Möller and Nix, with Ragnar Bengtsson of Lund and Jingye Zhang of the Lanzhou Institute of Modern Physics, redid the calculations, employing a more detailed model of the force between nucleons.

The 1981 predictions nicely corroborated the significance of the established spherical magic numbers. Specifically, the calculations showed that for spherical nuclei significant gaps between successive single-particle energy levels came at the accepted magic numbers. At large deformations, however, a new, unexpected gap appeared at the number 38. As a result a nucleus with an N or a Z near 38 would apparently get what amounts to a push toward deformation. The deformation would surely prevail if the neutrons and protons in a nucleus both favor it; thus Möller and Nix predicted that nuclei near an N and a Z of 38 should be among the most strongly deformed in nature. The superdeformation of krypton 74 (36 protons, 38 neutrons) strongly supports the prediction. For strontium 100 (38 protons, 62 neutrons) the situation is much the same. As one of us (Hamilton) first pointed out, the neutron number 60 is a deformed magic number, provided it is reinforced by another deformed magic number. Thus in strontium 100 as in krypton 74 the superdeformation of the nucleus is a consequence of the reinforcement afforded by proximity to a pair of deformed magic numbers: 38 and 38, or 38 and 60.

The situation, however, is affected by the proximity of the spherical magic number 40 to the deformed magic number 38. In particular, 40 is a spherical magic number for N or for Z if its push toward sphericity is reinforced by a push toward sphericity on the part of the other type of nucleon sharing tenancy of the nucleus. Thus zirconium 90 (40 protons, 50 neutrons) is spherical-double-magic, and so is the recently discovered exotic nucleus nickel 68 (28 protons, 40 neutrons). If there is no supporting push toward sphericity, the deformed number 38 can become dominant. The fact that the Z/N combinations 38/40, 36/40 and 38/60 yield strong deformations indicates that the deformed number 38 is more magic—more important—



MAGIC NUMBERS signify the quantities of neutrons or protons that fill "shells" in a nucleus and bestow endurance on a spherical nuclear shape. Here each spherical magic number is represented by a black line across a chart that plots the proton abundance in nuclei (*vertical axis*) against the neutron abundance (*horizontal axis*). The stable nuclei found in nature occupy a region (*color*) extending diagonally upward from the lightest nucleus, hydrogen 1; the region includes some "spherical double magic" nuclei, such as the nucleus of oxygen 16 (eight protons, eight neutrons). Nuclei with proton and neutron abundances far from spherical magic numbers should be deformed; in certain regions (*gray*) such deformation has been found. Meanwhile the study of exotic nuclei has revealed three deformed magic numbers, represented by colored lines, which bestow endurance on a deformed nuclear shape instead of a spherical one.

than the spherical number 40. Investigators are eagerly searching for the exotic nuclei strontium 76 (38 neutrons, 38 protons) and zirconium 80 (40 neutrons, 40 protons). So far neither has been discovered.

The Vanderbilt group is intrigued by two isotopes of bromine. Bromine 73 (35 protons, 38 neutrons) should be superdeformed and prolate, or football-shaped, whereas bromine 71 (35 protons, 36 neutrons) should mark the first discovery of a nucleus that is superdeformed but oblate, or discus-shaped. The work of Bengtsson, Möller, Nix and Zhang places an *N* and a *Z* of 35 at the center of a region of oblate superdeformation.

In the meantime investigators continue to survey the nuclear landscape. A favored technique is to plot a potential-energy surface, which displays the internal energy of a given nucleus as a function of its shape. The technique assumes that in principle a nucleus, exotic or stable, could adopt any shape, except that each shape requires a certain amount of energy. The nucleus favors the shape that embodies the least amount.

To begin with, one must settle on a mathematical way to characterize the various possible shapes. Imagine a football lying on a table, so that its long axis, which runs from tip to tip of the ball, is horizontal. The football's cross section (a vertical slice through the middle of the ball, equidistant from the tips) is circular. In other words, a football's two short axes are equal and the third axis is longer. Such shapes are prolate spheroids. Now imagine that the long axis shrinks while the circular cross section expands. The result is that the ball's two long axes are equal and the third axis is shorter. In brief, the football becomes a discus. Such shapes are oblate spheroids. Finally, imagine pressing on the football, so that its cross section is more an ellipse than a circle. All three axes now differ in length. Such spheroids are termed triaxial.

How can these distinctions be put in mathematical terms? A system introduced by Bohr and Mottelson is based on two parameters, beta and gamma [*see illustration on page 83*]. Roughly speaking, beta (mentioned above in connection with superdeformed nuclei) is a measure of the deformation that turns a sphere into a football (a prolate spheroid) by lengthening one of its axes. For a perfect sphere beta is 0. Gamma (an angle) is a measure of the deformation that makes a prolate spheroid triaxial by deforming its circular cross section. A gamma of 0 degrees signifies a prolate spheroid. Then, as gamma increases, one of the body's short axes grows at the expense of the others, so that all three axes come to have a different length: the spheroid indeed is made to be triaxial. As gamma approaches 60 degrees the growing axis becomes equal to what used to be the long axis, and so the spheroid becomes flat, like a disk; that is, it becomes oblate.

In sum, each possible shape of a nucleus corresponds to values of beta and gamma, and thus to a point in a pie-shaped swath of what a mathematician would call the beta-gamma plane. The amounts of energy required to force that nucleus into the various shapes define a surface: the potential-energy surface. Suppose a nucleus has the lowest point of its potential-energy surface at a beta equal to 0 and an energy that rises steeply for excursions to shapes away from a beta of 0. Such a nucleus would be spherical in its ground state and any deviation from a sphere would give rise to a restoring force. Hence the nucleus should show vibrations about a spherical shape.

For a second class of nuclei the minimum in the potential-energy surface lies at a nonzero value of beta along the line corresponding to a gamma of 0 degrees or the line for a gamma of 60 degrees. These nuclei are deformed: in their ground state they are prolate (for a gamma of 0 degrees) or oblate (for a gamma of 60 degrees). Their prominent mode of excitation is rotation about a nonsymmetry axis. Such nuclei, however, can vibrate about their equilibrium shape. In turn each mode of vibration can support a band of rotational states.

A third class of nuclei consists of those with a minimum in their potential-energy surface at a beta not equal to 0 and a gamma not equal to 0 or 60—a minimum, that is, in the triaxial region of beta-gamma space. For these nuclei the picture loses clarity and the shape of the nucleus becomes harder to identify. If the nucleus is fairly close to one of the limiting cases represented by the first two classes of nuclei, the pattern revealed by the energy-level diagram may help to determine properties of the nucleus such as the magnitude of its deformation and the nature of its vibratory excitations.

Greiner and his colleagues at Frankfurt have developed an elegant method that derives a potential-energy surface from the measured energy levels of a given nucleus. In turn the potential-energy surfaces are an excellent means of comparing the properties for a sequence of exotic isotopes of a chemical element [*see top illustration on pages 84 and 85*]. Quite recently groups including one from Lund and Oak Ridge have been making systematic calculations of potential-energy surfaces for nuclei in several regions of the periodic table. The predictive power of such work is impressive. For example, shape coexistence was successfully predicted for platinum 176. In addition the calculations now treat the possibility that the constellation of protons and the constellation of neutrons in a nucleus may vibrate with respect to each other, so that the nucleus has distinct deformations for protons and for neutrons. Such modes of vibration were first considered by Greiner, in 1965. The modes exert an important influence on the nuclear magnetic moment, a measure of the flow of electric charge created by motions of protons. The vibrations themselves have now been discovered.

There are two aspects of exotic atomic nuclei we have not done justice to. First, a nucleus can be exotic (it can differ substantially from the nuclei found in nature) in that it has far more angular momentum, a measure of the rotation of the nucleus or of the motions of constellations of its individual nucleons. Investigators at the Lawrence Berkeley Laboratory, the Argonne National Laboratory, Oak Ridge, the Gesellschaft für Schwerionenforschung (GSI, the Laboratory for Heavy Ion Research) in Heidelberg and the Daresbury Laboratory in England have detected gamma rays emanating from nuclear states with angular momenta of 40 units and more. (In quantum mechanics angular momentum is measured in multiples of a quantity called Planck's constant.) The states commonly observed in nuclei have angular momenta of, say, six units. As the angular momentum of a nucleus increases, the permanence of its shape breaks down. In fact, even the best examples of hard spherical nuclei or hard deformed nuclei can undergo a variety of changes. The penultimate result is superdeformation; after that the nucleus splits by nuclear fission.

Second, nuclei far from stability have revealed many new and exotic modes of radioactive decay. In nuclei rich in neutrons (that is, nuclei that have many more neutrons than the stable nuclei of a given *Z*) a beta decay may leave the nucleus in such a highly excited state that it emits a neutron rather than a gamma ray (the usual means of "blowing off" energy). This two-stage radioactivity was first identified by E. T. Booth and John R. Dunning of Vanderbilt and Francis G. Slack of Columbia University. In nuclei extremely rich in neutrons Born Jonson, P. Greggors Hansen and their

"I can't, I'm wearing magnetic underwear."

—Computer backup excuse #685

Doing computer backup really gets under people's skin.

They'll do anything to get out of doing it, because backing up is so utterly boring.

But not doing it can definitely be a matter of getting your education at the college of hard knocks.

We're reminded of the PhD candidate in biology who did hunger research for his doctoral thesis without backing it up.

He had data from a year's worth of injecting and weighing rats stored on the disk. The computer crashed, the rat data was erased from memory, and he had it to do all over.

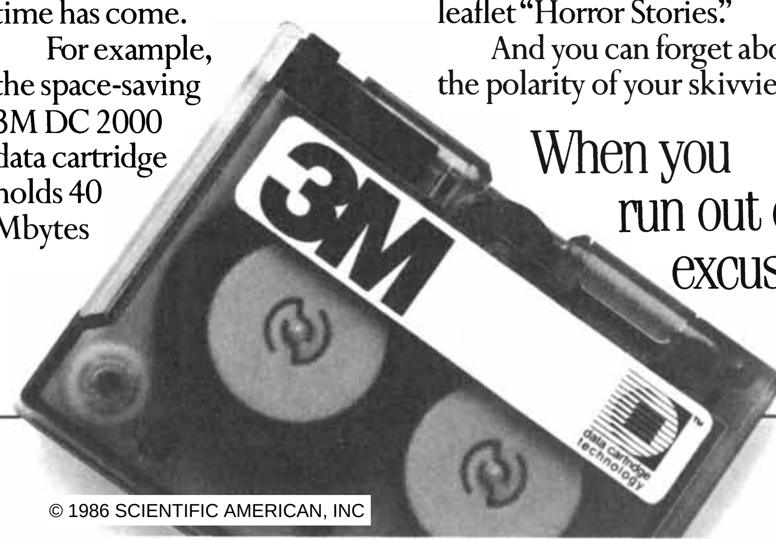
Back to the rat race.

So do your backup. Floppies if you don't have much memory. Or, if you have 5 to 10 Mbytes or over, on data cartridge – a 3M developed technology whose time has come.

For example, the space-saving 3M DC 2000 data cartridge holds 40 Mbytes of information. You'd need 80 floppies to store that much. And it can fit in your breast pocket.

To read more true-to-life "horror stories" of people who didn't back up, ask your computer products dealer for our leaflet "Horror Stories."

And you can forget about the polarity of your skivvies.



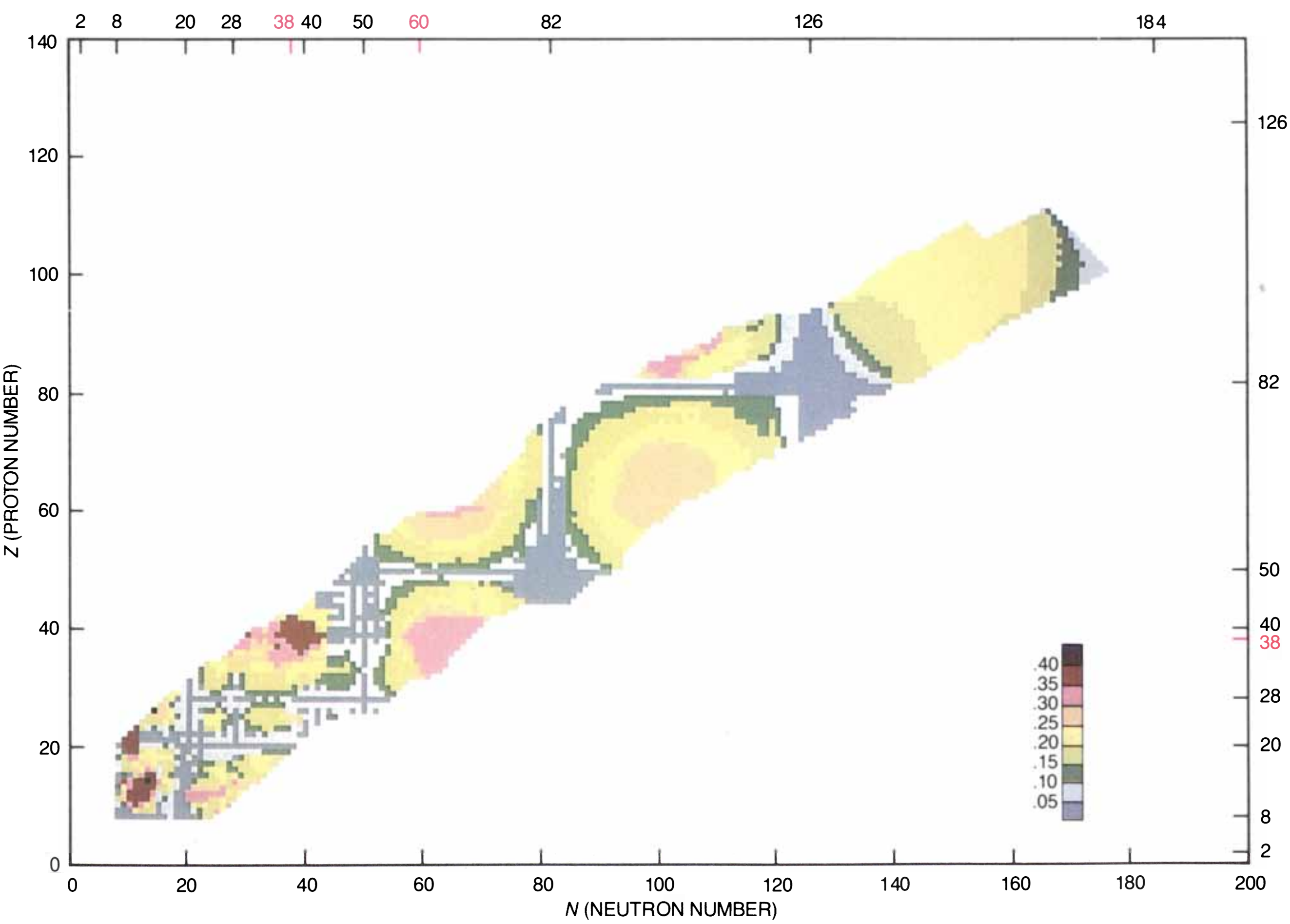
When you run out of excuses.™

colleagues at ISOLDE have discovered beta-delayed two-neutron and beta-delayed three-neutron modes of decay. One-neutron and two-neutron radioactivities from exotic nuclei in their ground state have been predicted but have not yet been seen.

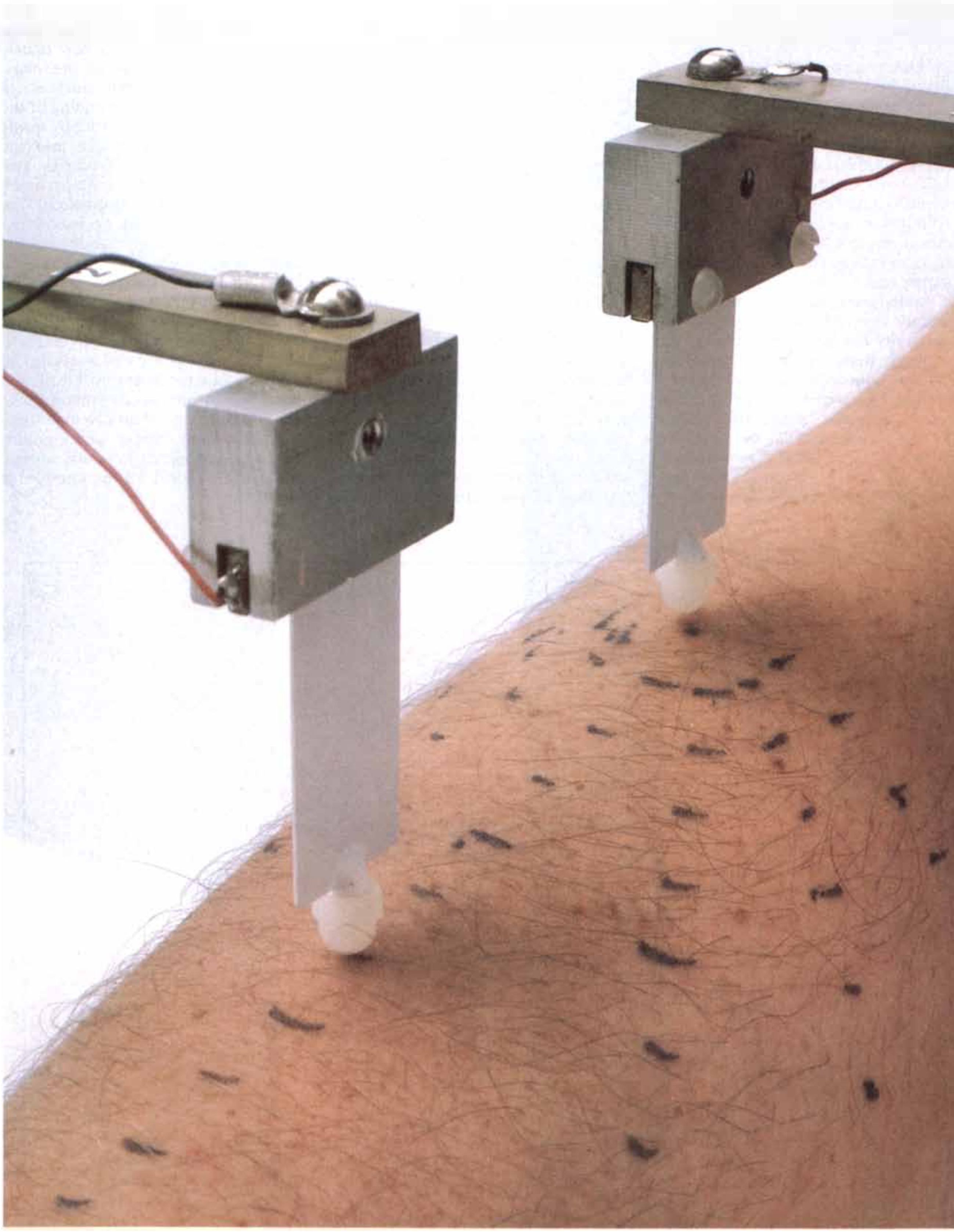
Nuclei rich in protons have also yielded discoveries. V. A. Karnoukhov of the Joint Institute for Nuclear Research at Dubna discovered beta-delayed proton decay. Here the nucleus emits a positron (a subatomic particle identical with an electron except for its charge, which is positive) and then, in an excited state, emits a proton. More recently beta-delayed two-proton radioactivity has been discovered at the Lawrence Berkeley Laboratory. Beta-delayed alpha decay (the emission of a positron followed by the emission of what amounts to a helium-4 nucleus) has long been known; beta-delayed hydrogen-3 decay was discovered recently at ISOLDE. Ground-state proton radioactivity was discovered independently by two groups at GSI in Darmstadt. Two-proton radioactivity has been predicted but has not yet been discovered.

Some of the newly discovered exotic modes of decay are particularly intriguing. Greiner and his colleagues at Frankfurt and at the Central Institute for Physics in Bucharest predicted in 1980 that heavy nucleon clusters such as the equivalent of a carbon-14 nucleus or a neon-24 nucleus could be emitted in radioactive decay if the process left a double-magic daughter nucleus, such as lead 208 (82 protons, 126 neutrons), or a nucleus close to being double-magic. Two years ago the prediction was confirmed. Carbon-14 decay, first detected in radium 223 at the University of Oxford, has been detected in several nuclei, and neon-24 decay has been detected in uranium 232 and protactinium 231.

There is increasing evidence for the existence of such clusters in excited states and even the ground states of some heavy nuclei. The discoveries may lead, therefore, to a new understanding of how nucleons are organized inside the atomic nucleus. In turn an increased understanding of the collective motions of nucleons inside nuclei should yield insight into the more general problem posed by systems made up of subatomic particles. One of the central motivations for probing all aspects of nuclear structure is that the atomic nucleus is the major, if not the only, testing ground in an important intermediate realm of quantum-mechanical many-body problems, namely systems incorporating an intermediate number of particles: too few to be treated by statistical methods but too many to be treated easily one by one. Much remains to be done, both experimentally and theoretically, if the shapes and motions now being observed in exotic atomic nuclei are to yield all the knowledge they promise.



PREDICTED SHAPES for the ground state of 4,023 nuclei ranging from the nucleus of oxygen 16 to a nucleus with 279 nucleons (102 protons, 177 neutrons) are displayed on a chart that plots beta (the parameter representing the departure of a nucleus from sphericity) as a function of neutron number and proton number. Nuclei predicted to deviate the least from sphericity (*gray*) tend to lie at or near the lines traced out by spherical magic numbers. Nuclei predicted to be the most deformed in nature (*brown*) occupy a number of pockets, including one that constitutes a successful prediction of the superdeformation discovered in exotic nuclei such as krypton 74 (36 protons, 38 neutrons). The predictions were made by Peter Möller of the University of Lund in Sweden and J. Rayford Nix of the Los Alamos National Laboratory; their calculations were based on a mathematical representation of the force that binds nucleons.



PIEZOELECTRIC SKIN CONTACTORS designed by the authors for their touch experiments bend momentarily whenever they receive a pulse of electricity. The tip of the contactor rests on the skin, exerting a static pressure that can be regulated by shifting a weight on the other end of the brass lever arm to which the contactor is attached. A pattern of light taps can then be superposed on this "background" pressure by a programmed sequence of electric pulses. In the basic experiment one contactor delivers a warning-and-reference tap (designated P_1) followed .8 second later by another tap (P_2). The other contactor, placed a certain distance away, then administers a third and final tap (P_3). The time interval between P_2 and P_3 as well as the separation between the two contactors can be varied readily. Marks, shown here on the subject's forearm, aid in positioning the contactors for successive measurements.

Space, Time and Touch

Experiments show that slight variations in the timing of taps on the skin can produce wide fluctuations in the taps' perceived locations. These findings help to reveal how the central nervous system works

by Frank A. Geldard and Carl E. Sherrick

There is a parlor game in which one person (the subject, as he would be designated in a psychology experiment) closes his eyes while someone pokes his arm lightly with a pencil. The subject must then indicate where he thinks he has been touched. Almost invariably he points to a site somewhat displaced from the actual spot where contact was made. For more than a decade we have been doing experiments in our laboratory at Princeton University that are in effect controlled variations of that game.

We began with the knowledge that the human senses have adapted variously to space and time and are far from being equally proficient at dealing with these two broad dimensions of the physical world. Although vision has mastered many of the intricacies of space, it is rather sluggish where time is concerned. Consider, for example, the way a motion picture's discrete frames of film, projected onto the screen at a rate of 24 frames per second, dissolve into action perceived as smooth and continuous. Audition is keen at handling time but informs one about space only in an indirect and often faulty manner. A sound originating somewhere in the median plane of the body is almost as likely to be perceived as coming from behind as coming from in front. The senses of smell and taste normally contribute very little to the formation of spatial and temporal distinctions, responding relatively slowly to chemical changes in the environment.

What about touch? It has much to do with both space and time but does not excel in handling either. The skin is better than the ear at making spatial distinctions and better than the eye at making temporal distinctions; on the other hand, it cannot compete with the eye in representing space, nor is it as good as the ear in gauging the time interval between stimuli. Touch, the one sense modality that is widely distributed over the entire body, occupies a middle position in the hierarchy of the senses. In a way that is roughly analogous to what goes on in the retina of the eye, the skin is able to report on where it is touched, on the distance between discriminable stimuli, on what forms or shapes it makes contact with and on whether movement takes place over its surface; yet all these capabilities are accompanied by errors.

It was to gain some insight into the sense of touch, including its characteristic fallibility, that we undertook our experiments. What we did was essentially to vary several features of the parlor-game procedure, among them the number and the timing of the touches. The experiments showed that the amount of error in locating one tap on the skin with respect to a second tap fluctuates widely as the timing of the taps is varied slightly. This result as well as others have revealed some things about the organization of the nervous system; the findings have also defined a phenomenon, which we call saltation, that still remains to be adequately explained.

The basic experiment consists in delivering three distinct taps to the skin of a subject. Piezoelectric skin contactors that flex in response to pulses of electricity are used to generate the taps, which are designated P_1, P_2 and P_3. The initial tap (P_1) precedes the others by nearly a full second; it not only gives warning that two more taps are coming but also serves as a bench mark by which any displacement in the perceived location of the second tap can be detected. The second and third taps (respectively P_2 and P_3) are given in quick succession. The first of the pair is applied at the same spot as the precursor tap P_1 was; the second is delivered by a different skin contactor a certain distance away. Hence P_1 and P_2 define one stimulus site and P_3 defines another; the temporal relations among the three taps involve a relatively long but constant wait between P_1 and P_2 and a short, variable one between P_2 and P_3.

As one might expect, if this experiment is set up so that the taps are applied to a subject's forearm by two skin contactors about 10 centimeters apart, and if the time between P_2 and P_3 is about a quarter of a second or more, it will be found that the subject feels two consecutive taps (P_1 and P_2) at the same position followed by one tap (P_3) at a position removed from the other two. (Because of the normal touch-localization error, the felt positions are not necessarily coincident with the actual places where the pressure was applied, but they are consistently within the normal error radius.) If the interval between P_2 and P_3 is reduced to less than about a quarter of a second, however, something unexpected happens: the subject no longer feels P_2 at its "veridical" position, that is, where he felt P_1. Now P_2 is perceived instead as having been applied somewhere between the positions of P_1 and P_3. The last tap seems to have "drawn" the second one toward it.

The apparent displacement depends on the P_2–P_3 time interval. If the subject is instructed to vary the P_2–P_3 timing slowly from the interval that allows veridicality until he senses P_2/P_3 coincidence, he can tell the experimenter when P_2 seems to have traveled certain simple fractions of the distance between P_1 and P_3. For example, as the interval between P_2 and P_3 is shortened from a quarter to a tenth of a second, P_2 will seem to have moved halfway toward P_3. If the time is further reduced to a twentieth of a second, P_2 is felt to be three-fourths of the way toward P_3; at about a fiftieth of a second P_2 becomes indistinguishable in location from P_3.

It is obvious that this phenomenon is strongly time-dependent. When subjects' performances of the task are averaged, the resulting curve shows that the perceived displacement of the second tap varies approximately linearly with the time between the second and third taps.

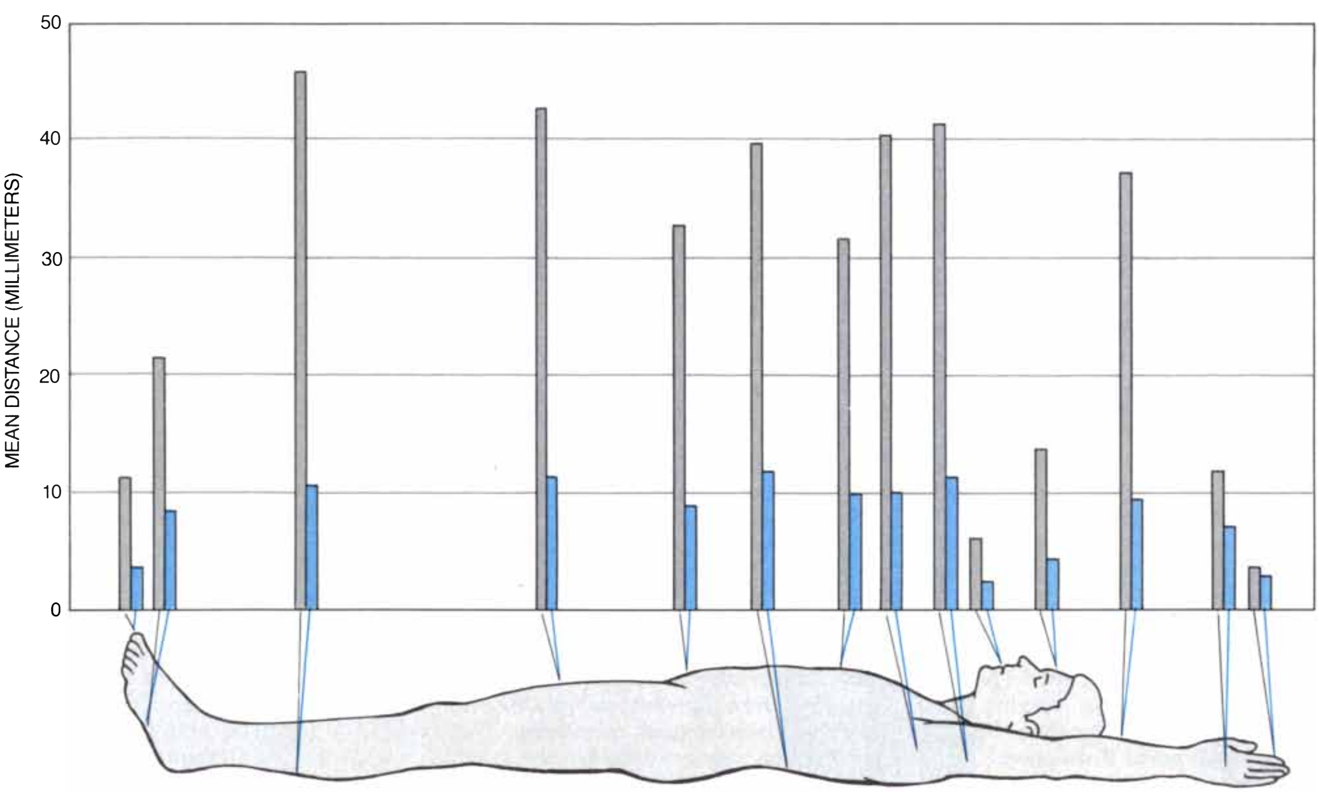
We think of the tap sensations as three beads on a string. The two end ones, those coming from P_1 and P_3, are fixed in position and anchor the string. The sensation aroused by P_2 is found somewhere along the string at a place dictated by how far P_2 is from P_3 in time. If the offset is a fiftieth of a second or less, the tap coincides spatially with the one from P_3; if the timing differs by a quarter of a second or more, the P_2 sensation joins the one from P_1. Because the sensation evoked by P_2 appears to have jumped from where one would expect it (at the P_1/P_2 locus) to where it is actually felt (nearer the P_3 locus), all in the brief time between P_2 and P_3, the effect has been given the name saltation (from the Latin *saltare,* to leap).

In the experiment we have just described it was specified that the two skin contactors were separated on the forearm by approximately 10 centimeters. Does the same thing happen if the two contactors are, say, 20 centimeters apart? The answer is no. It turns out that there is a somewhat limited field over which the saltatory effect operates, and it varies in size and shape from one part of the body to another.

The dimensions of the field in which saltation is observed can be readily established. The contactor delivering the initial pair of taps is held stationary while the P_3 contactor is moved in small uniform steps out along a radius of a circle centered on the P_1/P_2 locus. At each step the observer tests all possible P_2–P_3 time intervals to see whether or not P_2 is felt in a displaced location. When a point on the skin is reached where there is no longer any saltation at any time interval, the series is ended and the distance between the P_1/P_2 contactor and the P_3 contactor is measured. The P_3 contactor is then moved out a bit farther along the radius and a new series of trials is run, this time at successively smaller distances from the P_1/P_2 locus, until saltation reappears. The average of the two measured distances designates a point on the saltation boundary. If the procedure is repeated along several radii, the points can be connected to enclose a field within which saltation always occurs. This is the saltatory area for that particular skin sample.

Such explorations have provided maps of the saltatory area of various body parts: the inner and outer surfaces of the forearm, the chest, the front and back of the thigh, the palm and the index finger. Attempts have been made to explore still other extended skin areas but there are some difficulties. It is not just a matter of setting contactors down and taking readings. The static pressure of the contactors has to be held rigidly constant, and the body motions associated with breathing and circulation often produce great variations in contact force. Special provisions also must be made for the various shapes and skin textures of different body parts.

Nevertheless, some generalizations can be made from the data already obtained. In particular, the shapes assumed by some of the saltatory areas are intriguing. By and large those on the limbs are oval, with the long axis parallel to that of the limb. For the inner and outer surfaces of the forearm as well as for the front and back of the thigh the longitudinal axis is about twice as long as the transverse axis. The palm and the index finger have more nearly round saltatory areas. The chest follows neither rule: here the saltatory area is sharply truncated at the midline of the body. The same thing is found on the forehead, in the center of the back and near the center of the abdomen. In every case an important basic principle applies: saltation never crosses the midline of the body. This says something about the



NORMAL ERROR ASSOCIATED WITH TOUCH can be expressed in two ways: as the average minimum separation necessary for two probes to be sensed as a pair of discrete touches when they simultaneously contact the skin (*black lines*), and as the average separation between the perceived point of contact and the actual point at which a single probe touched the skin (*colored lines*). As can be seen, the accuracy of the tactile sense varies widely from body part to body part, being highest on the fingertips and the lips.

underlying neural events in saltation, because the phenomenon evidently respects the central nervous system's fundamental dichotomy into right and left halves.

So far we have dealt with the simplest case of saltation: two staggered taps not too far apart on the skin, preceded by a warning-and-reference tap. What happens if P_2 consists of multiple stimuli, a series of taps replacing the single one? Provided the series does not take too long to deliver and provided the entire train of taps is delivered at the P_1 site, two responses are noted. In the original experiment the place where the single P_2 tap was felt could be any point along a line drawn from the felt position of P_1 to the felt position of P_3, depending on the P_2–P_3 timing. It is understandable, therefore, that multiple P_2 taps are sensed as being strung out in position between P_1 and P_3: in effect, an entire spectrum of P_2–P_3 intervals is administered at once in the revised experiment. More puzzling, however, is the observation that the saltatory area for multiple P_2 taps is always larger than the area for single P_2's on the same skin surface.

Series of trials (with the same volunteers who served as subjects when we mapped the single-P_2 saltatory areas) were carried out in which the number of P_2 taps was varied from two to 12, keeping the taps a constant 25 milliseconds apart. The P_3 skin contactor was moved progressively away from the other contactor on successive trials, and the number of P_2 taps was changed randomly for each series. The findings show that the number of taps in the P_2 train matters: the more taps there are, the farther away the P_3 contactor can be placed and still produce the saltatory effect.

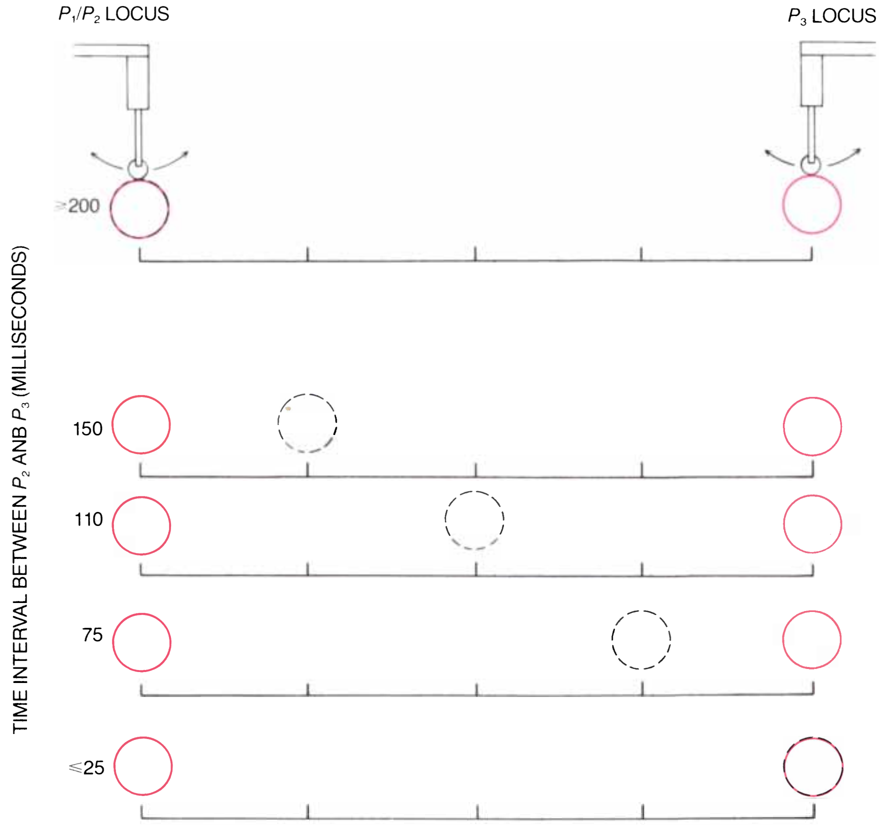
This characteristic of saltation along with the others mentioned poses neuropsychological riddles, the solutions to which can be only guesses. Yet certain modest questions can be answered definitely. For example, we noted earlier that the P_2 tap's inability to cross the body's midline was in accordance with the symmetry of the central nervous system. Can we show that the phenomenon indeed has its seat in the central nervous system and is not generated in the skin?

To investigate this question the surface layers of an oblong strip of skin on the forearm are anesthetized with procaine so that the subject cannot feel taps or even pinpricks on the affected area. At one end of the oblong, just beyond the insensitive area, P_1 and a train of six P_2 taps are applied; at the other end of the insensitive strip a single P_3 tap is delivered after the P_2 train has ended. If mechanisms in the skin are responsible for the saltatory effect, the taps should be felt only at the loci where the stimuli are applied. If, on the other hand, saltation has its source in the central nervous system, there is no reason the taps should not be felt as if they were in their "normal" locations: between P_1 and P_3 on the anesthetized skin.

When the experiment was carried out, the answer was not long in coming. At all appropriate temporal separations palpable P_2 taps hopped straight through the deadened zone. Saltatory leaping is obviously generated centrally, presumably in the brain, and not at the stimulation sites.

Because saltation occurs at the cerebral level, it is reasonable to infer that it may have analogies in the other senses. As a matter of fact, an analogous phenomenon in the visual system has been well established. If two small spots of light, separated from each other by five degrees of vertical angular separation, are flashed simultaneously or in quick succession and are viewed by fixing on a point between 15 and 40 degrees off to the side of the flashes, the subject can often see a "phantom" light somewhere between the two spots. Even if the two veridical light spots straddle the blind spot of the eye, the phantom is still seen between them. It has also been noted that if one spot is red and the other is green, the phantom light looks yellow.

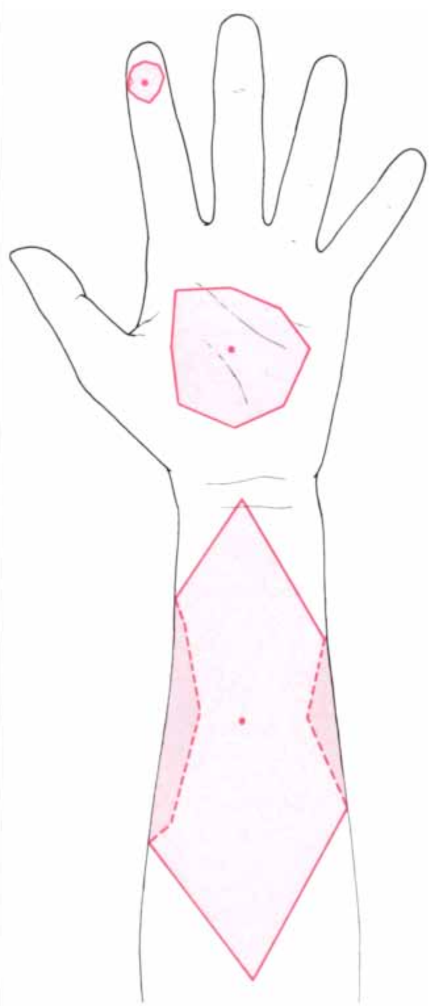
It is also possible to demonstrate saltation in auditory space. Three clicks (corresponding to P_1, P_2 and P_3 in the tactile case) are sounded from two speakers at discriminably different locations, both on the same side of the head and about a meter away from it. When the warning-marker P_1 click and the P_2 click are presented from one speaker, followed 40 milliseconds later by the P_3 click from the second speaker, P_2 sounds as if it had originated between the true locations of the two speakers. As in the case of the skin, varying the time between P_2 and P_3 moves the phantom click back and forth between the two speakers. A similar phenomenon was described by the Hungarian-born physicist and physiologist Georg von Békésy, who believed it accounted for the perceived mislocalization of orchestral instruments in



PERCEIVED LOCATION OF THE P_2 TAP (*broken circles*) **in relation to the perceived location of P_1 and P_3** (*colored circles*) **is approximately a linear function of the time interval between P_2 and P_3 within the range from about a quarter of a second to a fiftieth. This defines the phenomenon of saltation: in spite of the fact that P_2 is delivered by the same skin contactor that delivers P_1, the sensation evoked by P_2 can be far outside the normal touch-error radius of P_1. Given the proper P_2–P_3 timing, the P_2 stimulus seems to "jump" abruptly from the P_1/P_2 contactor locus to a position nearer the P_3 locus before it is sensed.**

concert halls that are plagued with echoes. When a sound reaches the ear by a direct path and an echo of the same sound reaches it by a less direct path milliseconds later, the conditions for auditory saltation, and hence the apparent mislocalization, are readily established.

Given all the facts uncovered by experimentation, we can begin to speculate on the underlying neurophysiology of saltation by distinguishing it from another sensory phenomenon, called phi. Phi is a well-known illusion that was originally observed in studies of visual perception. It can be induced by administering in quick succession two spatially separated stimuli (light flashes in the case of vision, skin taps in the case of touch). If this is done properly, the subject perceives the stimulus as moving smoothly across the visual field or along the surface of the skin.



SALTATORY AREA (*colored region*) of the inner forearm, the palm and the tip of the index finger has been mapped by the authors. These are regions of skin within which the P_3 contactor can be moved from the fixed P_1/P_2 locus (*colored dot*) and still elicit the saltation effect. The broken-line boundary indicates where the saltatory areas wrap around the arm and index finger.

Because the phi illusion occurs optimally with a time separation of 100 milliseconds or less (well within the broad range in which saltation occurs), it might be mistakenly assumed that one phenomenon is a variation of the other. A careful observer would note, however, that there is a significant difference between the two phenomena with respect to their qualitative appearance. Whether in touch or in vision, the essence of phi is the perception of a continuous motion. This distinguishes it from the saltation experience, which is the perception of one or more touches at successive, discrete locations on the skin. Movement in this instance can only be inferred, not sensed.

Of greater significance is the space over which the two tactile phenomena operate. If two properly timed taps are delivered to opposite sides of the forehead, phi can often be experienced; the saltatory phantom tap is never felt, however, because saltation never takes place across the body's midline. It seems clear that the neural mechanism responsible for phi is not the one that causes saltation.

In spite of the many illuminating findings yielded by recent research on the central nervous system, not much more than this can be said with certainty about the neural mechanism that might account for the various manifestations of saltation. For now only hypotheses can be offered.

Studies in monkeys of the somatosensory region, the part of the cortex of the brain that responds to tactile stimuli, have revealed the presence of receptive fields: areas of skin that, when they are touched, produce a discharge of localized neurons in the cortex. Examining the variation in size of receptive fields with location on the body, one is struck by the correspondence with the relative sizes of the saltatory areas. The fields are large on the body and upper limbs but small on the hands and digits. Moreover, when such receptive fields were studied earlier on the thorax of the cat, the sharp boundary at the midline of the body was a notable feature. The finer detail in the spatial layout of the central nervous system, in addition to the gross dichotomy into left and right halves, is therefore also reflected in the saltatory phenomenon.

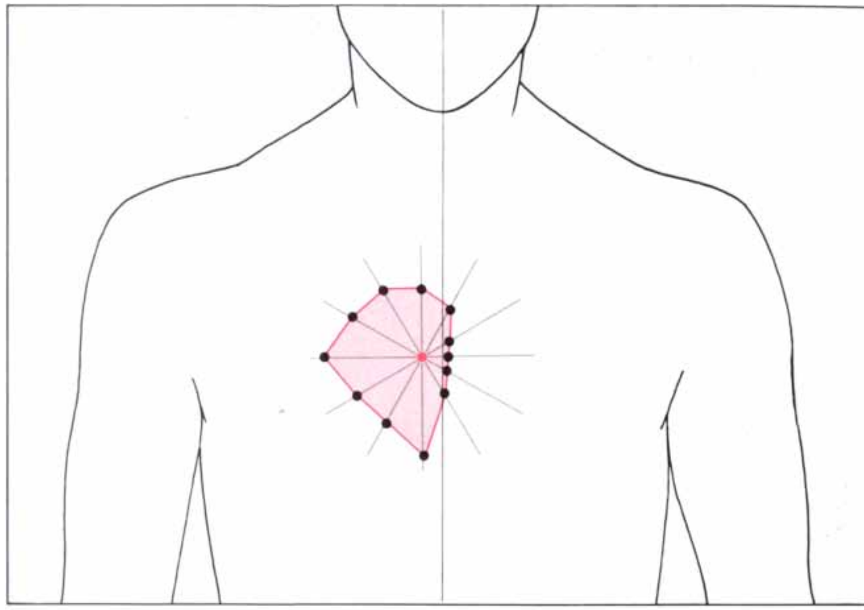
As we pointed out above, when several P_2 taps are inserted between P_1 and P_3, the saltatory area itself seems to enlarge. This may be owing to the cumulative effect of repeated taps, which causes the excitatory activity to expand into outlying receptive fields, thereby enlarging the cortical region that is active in the saltatory process.

Studies of temporal organization in the cortex comparable to those mapping out its spatial arrangement have yet to be done, but an examination of the "backward masking" of patterns in vision, hearing and touch suggests it is closely related to saltation. Backward masking is observed when two patterns, such as speech sounds or letters, appear in rapid succession. The lagging pattern has a surprising effect on the leading one: it does not merely dominate or overcome the first pattern but tends to incorporate it. The fact that the phantom tap produced by P_2 seems to be drawn to P_3 could be viewed as the result of a similar modification of an initial neural signal by a subsequent one. Because neural-activity patterns generated by later taps encroach on ones generated by earlier taps (assuming that the taps are confined to a small time span in a single saltatory zone), it is the later pattern that tends to prevail in consciousness: it is followed by relative silence, or at least by no modifying stimulus.

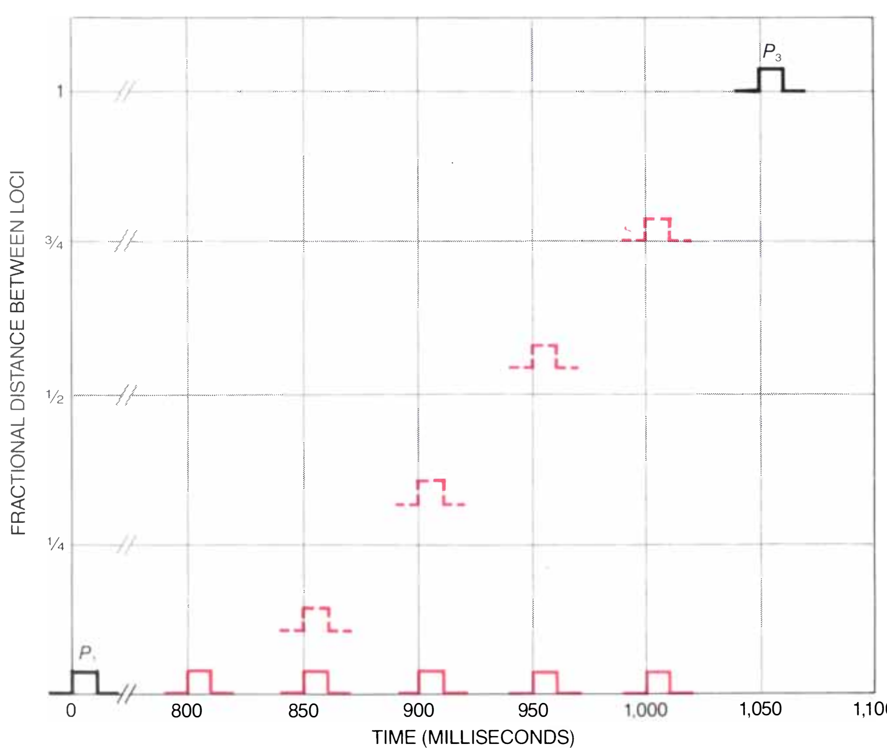
Indeed, close inspection shows that the P_2 tap changes in ways other than in apparent position under the influence of P_3: it loses its "sharpness," and perhaps some of its size and intensity. The closer it is brought to P_3 in time, the more extensive these modifications become. In the scheme of things neural, however, the preeminent quality is location. The loss of spatial identity is what stands out in the experience of the observer as the temporal framework is compressed and the present engulfs the past.

The neuroelectric activity of the cortex reflects the excitation of many nerve cells, since there is great overlapping of receptive fields. As long as the two taps remain within each other's overlapping receptive fields and so trigger the firing of the same cortical neurons, their temporal conjunction produces the backward inhibitory effects; as the taps are registered in separate fields, the confusion produced by their conjoint space-time occurrence in the brain is reduced or eliminated. This suggests that the definite borders of a saltatory area define the boundary of shared receptive fields.

All these conjectures are, to be sure, wholly speculative, yet they serve the purpose of suggesting places and processes that may be fruitfully explored. It is now up to those scientists who look for answers under the skin to determine whether these hypotheses are supportable.



BODY'S AXIS OF SYMMETRY (*black line*) **defines a boundary over which the "phantom" P_2 touch will not jump. The saltatory area of the chest** (*colored region*) **described when the P_1/P_2 locus** (*colored dot*) **is placed a centimeter off center accordingly exhibits a sharp truncation at the midline. Saltation thereby betrays its underlying connection to the central nervous system, which preserves the dichotomy between the body's right and left halves. The gray lines emanating from the P_1/P_2 locus are radii along which the P_3 contactor was moved in order to determine the saltatory area's boundary points** (*black dots*).



MULTIPLE P_2 TAPS (*color*) **that are evenly spaced in time (beginning, as in the single-P_2 case, 800 milliseconds after P_1) are evenly distributed in their perceived locations** (*broken colored lines*) **between the P_1/P_2 locus and the P_3 locus. The position at which each P_2-component tap is sensed reflects the tap's temporal relation to P_3: the closer to P_3 it is in time, the closer to the P_3 locus it is ultimately sensed to be. Multiple-P_2 experiments reveal a more puzzling phenomenon: the saltatory area for trials with multiple taps is always larger than the area for trials with single P_2 taps on the same skin surface. Moreover, the more taps there are in the P_2 train, the greater is the observed expansion of the area.**

Flight Simulation

Simulators are increasingly employed to train pilots for complex tasks. Generating the accompanying visual scene continues to present a challenge to computer technology

by Ralph Norman Haber

Most pilots learn to fly today just as they have for 75 years—with their hands on the controls of a real plane in real flight. They may begin with an instructor sitting beside or behind them, but as soon as they can land safely most of the rest of the training is done solo. Consequently learning to fly has posed increasingly severe problems as the performance capacities of planes increase and as the multiple demands for the pilot's attention escalate.

Whereas a World War II fighter plane could barely exceed the 250-mile-per-hour speed of the race cars at the modern-day Indianapolis Motor Speedway, today's fighters routinely fly two to five times as fast. More important, their maneuverability at the higher speeds means that reaction times for responses have to be incredibly quick. Even at a lazy 600 knots a plane travels more than 1,000 feet per second—a mile every five seconds. Moving at such a speed in level flight at an altitude of 1,000 feet, the plane will collide with the ground in 20 seconds if the ground slopes upward at a 5 percent grade, a slope often difficult to detect from the air. The collision time drops proportionately with altitude and becomes only two seconds for flight at an altitude of 100 feet. Two seconds may still be enough time to avert disaster—except that the pilot may also be listening to a warning from a wingman, or looking behind for a possible attacker, or preparing to attack a target already in sight ahead. Even with only a few of these added tasks 20 seconds may not be enough time to avoid crashing.

How can pilots be trained for such complex tasks without killing them in the process? The preferred solution, introduced during World War II, is the flight simulator, in which the pilot can practice many flying tasks while sitting safely on the ground. He or she can even practice responses to unlikely events, particularly those that might lead to disaster.

A commonly encountered relative of the aircraft flight simulator is the automobile driving simulator. In the automobile simulator the car windshield is replaced by a projection screen on which a roadway scene is displayed. A preprogrammed motion picture presents challenges to the driver, who must react to them. Other cars approach, for example, or the road turns or highway signs and signals appear. The driver's task is to control the car by operating its steering wheel, gas pedal and brake. This type of simulator has a major limitation: the handling of the controls by the driver does not produce a corresponding change in the scene displayed on the screen. Turning the wheel does not allow the driver to see or feel the car turn. Nearly all automobile simulators are of this "open loop" type.

In a "closed loop" system, in contrast, the use of controls does produce changes in the scene. To close the loop in a simulator the scene being presented must actually exist either as a physical model or as a program in a computer. When a physical model is used in a flight simulator, for instance, a moving camera photographs the model from the position in the model occupied by the aircraft and moves as the craft moves. In this way the pilot sees the part of the model over which the plane is traveling, so that if the pilot speeds up, the camera speeds up correspondingly, and if the plane turns, the camera turns too. When the air scene is constructed by a program in a computer, instead of using a physical model, the computer generates an image of that part of the scene directly in front of the plane. The scene changes in relation to the movement of the plane. Both procedures depend heavily on computers.

Flight simulators are now used to train people for each crew position in every type of fixed-wing aircraft, helicopter and spacecraft. I shall focus on pilot training in jet-fighter aircraft because it is here that simulation poses the greatest difficulty. The combined effects of high speed and low altitude make it a technological challenge to change the visual scene and the apparent motion of the simulator realistically in response to the pilot's actions.

The basis of a typical jet-fighter simulator is a real cockpit and seat without the rest of the plane. The cockpit is enclosed by a projection surface on which the visual scene is shown. To simulate the motion of the plane the cockpit is mounted on a platform that can be moved up and down and from side to side as well as tilted. As the pilot operates the controls of the aircraft both the visual-scene content and the sensation of aircraft motion change. The simulator can also reproduce the sensation of atmospheric turbulence.

The cockpit (surrounded by the image-producing devices and mounted on the motion-producing platform) is generally in a room by itself. A separate control room contains the associated computer, from which all aspects of the simulated flight are supervised, monitored and recorded. The simulator is usually monitored by an instructor pilot who is in communication with the pilot in the simulator. The instructor can enact the role of wingman, ground control or mission control, and

SIMULATED LANDING of an AV-8B Harrier II light attack aircraft on the deck of a carrier is demonstrated by a U.S. Marine Corps pilot. The training system, which is known as an Operational Flight Trainer, was developed by the McDonnell Aircraft Company.

he can also manipulate or interfere with the various tasks to be done by the pilot. The instructor can, for instance, move targets, fire surface-to-air missiles, make other planes appear and even attack, suddenly change the weather, cause the plane to malfunction, amend the student pilot's orders and in general produce realistic chaos.

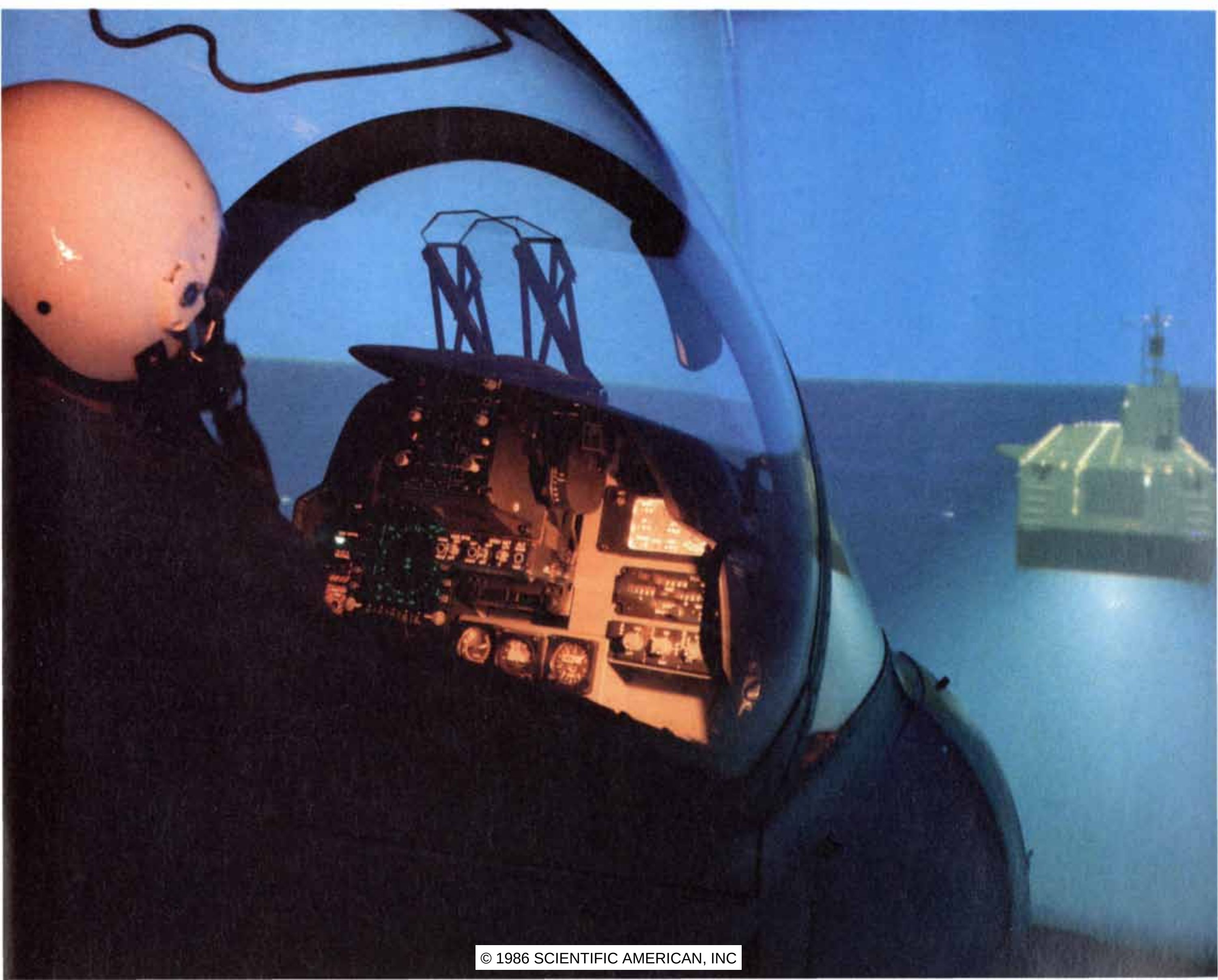
How are such simulations created? The first step is to make a model of the terrain over which the flights are to take place. The content and detail of the model are determined by the training task. The earliest models were of airfields or aircraft carriers, because flight simulators were first employed primarily for practice in takeoff and landing. Other models are of tanker planes in an otherwise empty sky (to provide practice in air-to-air refueling) or of one or more fighter planes (to simulate air-to-air combat). Recent models have included larger stretches of ground over which low-level flight can be simulated. Such terrain can also include targets to be observed or attacked, surface-to-air missiles to be evaded and, of course, natural features to be avoided.

The model begins as a conceptual description, which is currently realized in one of two ways: as a physical board model over which a camera is "flown" or as a computer program capable of generating and displaying an image. In either case the image of the scene is displayed to the pilot on a projection or video screen.

A typical board model might be 30 by 60 feet, scaled to represent about six square miles of terrain. The flight controls operated by the pilot in the simulator cockpit are connected to high-speed motors that move a video camera over the board exactly as the plane is made to fly; the movement of the camera is in scale with the model.

The main advantage of the board model is the tremendous detail it can present—something no computer imagery system is ever likely to duplicate. The effectiveness of the board is limited by how close to the surface the camera lens can be flown. A pilot in a plane on a runway might be 10 feet off the ground, which corresponds to a distance of less than half an inch off a modeled runway constructed at a scale of 300 to 1; this is closer than the lens can be positioned without risk of damage. Moreover, at that small distance normal lenses do not have enough depth of field to simulate the sharp images a pilot would ordinarily see. Newer techniques using laser sensing scanners are being developed to solve both technical problems.

Board models are expensive to construct, cumbersome to alter and limited as to the size they can represent. They are also noninteractive: objects do not move or disintegrate in a puff of smoke when they are attacked. Because board models are small, they are also quickly memorized during repeated practice. They are therefore effec-



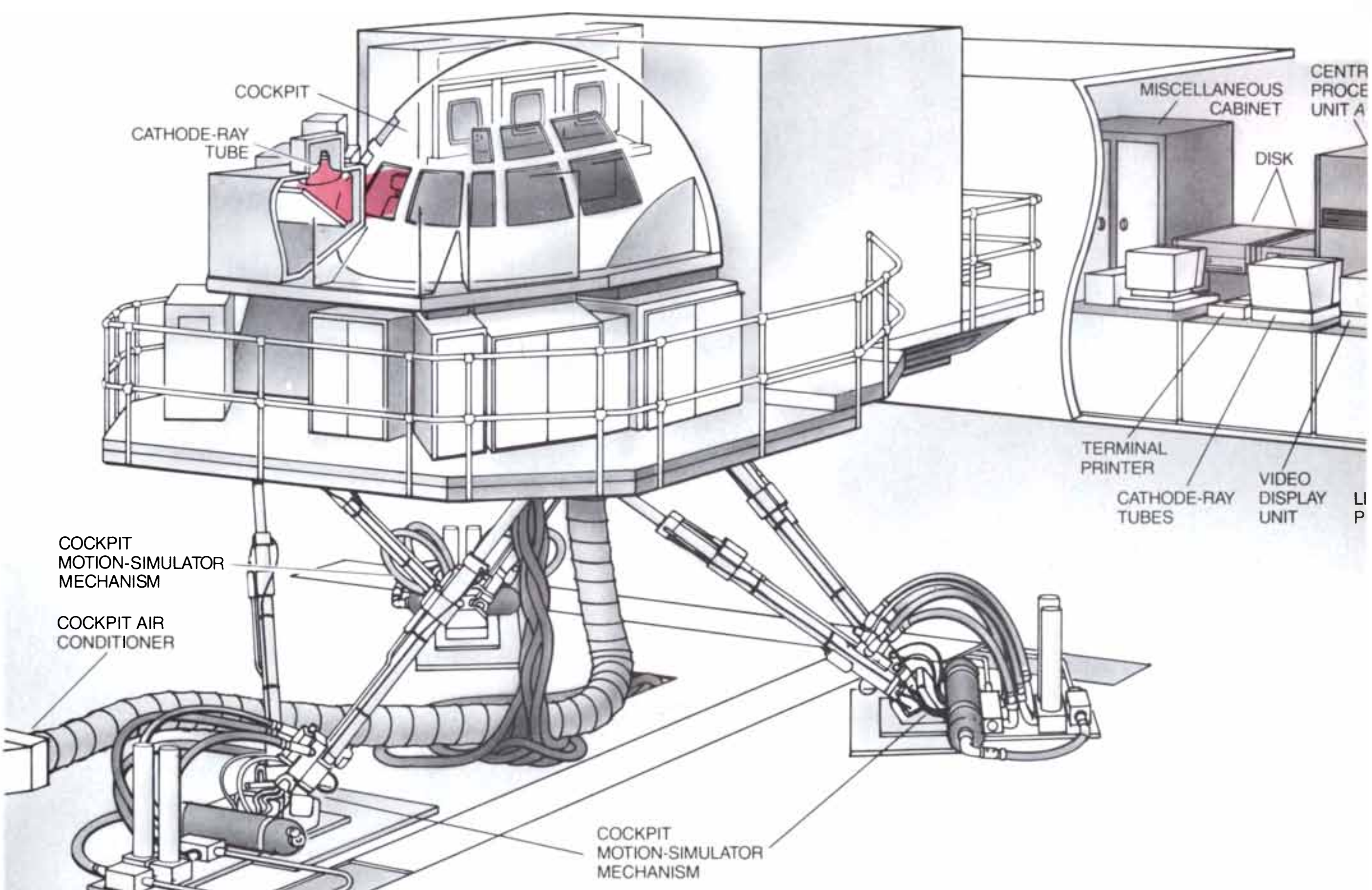
tive only for the simulation of repeated flying tasks near the ground (particularly landing), flight over difficult target areas or flight in which extensive surface detail is required, as it is in hovering practice for helicopters.

Computer-generated imagery is far more versatile, and a growing number of jet-fighter simulators use scenes generated by a computer. The first step is to develop a conceptual model similar to a physical board model, except that its size is limited only by the amount of memory and speed of the computer. The model defines the physical dimension of the terrain, its surface texture and character and any objects and artifacts that might be on its surface.

Once a model is programmed and stored in the computer's memory it is manipulated so that particular parts of it—corresponding to the pilot's view from the cockpit at a given instant—are displayed. This requires the program to compute the location of each part of the model with respect to the position of the plane. The vast majority of the scene is invisible at any one time, including what is behind the plane, what is too far ahead to be visible or what is hidden behind more prominent features in the line of sight.

For terrain features that are visible the program computes the projection appropriate to the angle of regard of the pilot. A pilot flying across a square cornfield planted on level ground, for example, does not see a square image. When it is approached from a distance, the field seen straight ahead is greatly foreshortened, with the far edge shorter than the near one and the two sides converging to a vanishing point at the horizon. When the plane gets closer to the field, the sides of the field elongate in relation to the near and far edges, and the shape becomes less trapezoidal as the sides become more parallel. The field would be projected as a square only if the plane were directly over it and all four sides were simultaneously visible, and that is possible only if the plane is diving at the center of the field. The geometric transformations of a surface as a function of viewpoint are well known, but different computer-generated imagery systems follow different procedures to achieve them.

The main advantage of computer-generated imagery systems so far has been the unlimited size of the terrain they can generate and display. Whereas board models represent only a few miles of ground, computer-generated imagery can encompass the entire vault of the sky, spanning thousands of square miles of every possible kind of terrain, from deserts to mountains. Although in theory a program can be written at any level of detail, in practice the greater the detail is in the surface texture (leaves on trees, markings on buildings, roads or runways, rocks on mountain slopes), the longer it takes the computer to generate any specific instance of the model. Since the program has to generate the model in real time as the plane flies over it, the level of detail is limited by the computer's size and speed and by its "refresh" and "update" rates. This is particularly critical when the plane is flying close to the ground. Even though computer-generated imagery

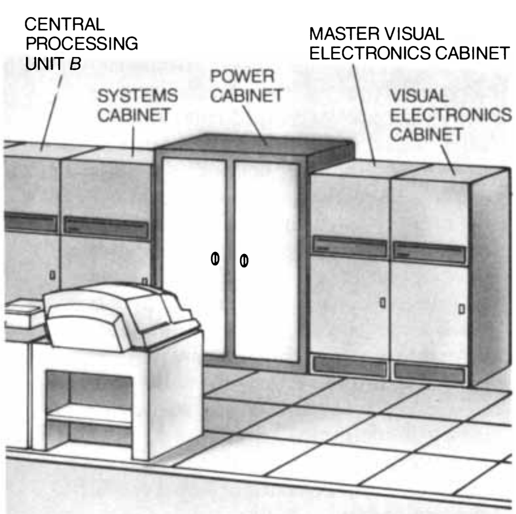


FLIGHT SIMULATOR consists of a cockpit (*left*) and a control room (*right*). The cockpit is enclosed by a projection surface (*shown as a series of cathode-ray tubes*) on which the visual scene is displayed. To simulate motion the cockpit is mounted on a platform that can be moved up and down and from side to side as well as tilted. As the pilot operates the controls of the aircraft both the visual-scene content and the sensation of aircraft motion change. The control room contains the associated computer, from which all aspects of the simulated flight are supervised, monitored and recorded. The simulator is usually monitored by an instructor pilot

systems could in principle generate a scene with as much detail as that of a board model, none of them is currently capable of doing so in real time, particularly the real time demanded by rapid flight at low altitude.

To display the visual scene, whether it is produced by a board model or by a computer-generated imagery system, most simulators rely on a device known as a raster-scan cathode-ray tube, which is similar to a conventional television tube. A simple system appropriate for an airplane with only a narrow field of view to the front requires only a single cathode-ray-tube screen positioned straight ahead in the pilot's line of sight. Most fighter aircraft have a wide field of view, however. In order to create a visual scene for them a number of separate screens are positioned in front of, above, to the sides of and even to the rear of the pilot's head.

A raster-scan cathode-ray tube produces an image through the rapid illumination of discrete points, called pixels, arranged along scan lines. Home

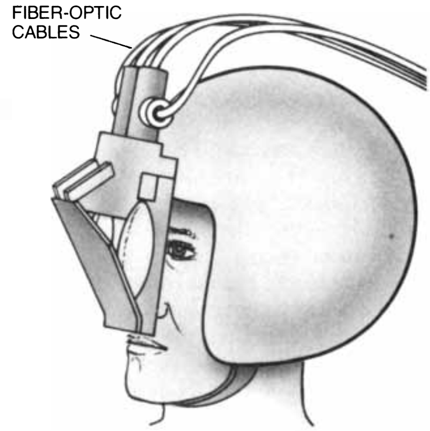


who is in communication with the pilot in the simulator. The illustration is based on the C-130H flight-training simulator, which is under development at the Link Flight Simulation Division of the Singer Company.

television sets in the U.S. have a standard number of 480 horizontal scan lines, regardless of the physical size of the tube's display area. There are 640 separate pixel positions located on each scan line, and each position can be illuminated by an electron gun whose beam moves along the line. To reduce the flickering effect of the sequential illumination of the points, the pixels on the odd-numbered rows are illuminated first, starting at the lower left-hand corner of the tube, going across each of the 240 odd rows and sequentially moving up the screen. Then the process is repeated for the 240 even rows. It takes a sixtieth of a second to do the odd rows and another sixtieth of a second to do the even rows. All the pixels needed to define the scene being displayed are illuminated 30 times per second, which defines the quantity known as the refresh rate for the cathode-ray tube.

For flight simulators the number of scan lines varies from 300 to 1,024 and the number of pixels per scan line varies from 512 to 1,024. Both variations affect the resolution of the tube: the fineness of detail that can be displayed and that can be distinguished by the viewer. For a cathode-ray tube, resolution is determined by the distance between the pixels in relation to the viewing distance. If the number of scan lines and the number of pixels per scan line are held constant, the resolution of a cathode-ray tube decreases as the screen size is enlarged in relation to the distance from which it is viewed. In other words, when a cathode-ray tube has to cover only a small visual angle, its resolution capabilities can be quite high, so that fine detail can be displayed and perceived. If, however, a cathode-ray tube has to cover a wide visual angle, resolution is reduced. To avoid reduced resolution, multiple independent tubes are mounted side by side in the simulator. The F-16 jet-fighter simulator currently uses seven 36-inch tubes, which provide a field of view measuring 300 degrees horizontally and 150 degrees vertically.

The brightness of a scene displayed on a cathode-ray tube is determined both by the light output of the cathode-ray tube itself and by the nature of any optical elements placed between the surface of the tube and the pilot. When a real scene is viewed from a fighter, its details are effectively at optical infinity. Collimating lenses are therefore placed between the tube and the pilot, so that the light rays coming from the tube are made parallel, as if they were reflected from distant objects. The scene generated on each surface of a cathode-ray tube is imaged on a diffusing screen window. Com-



PROJECTION SURFACE can also be incorporated in the pilot's helmet. The displayed image is transmitted from a projector behind the pilot through a pair of fiber-optic bundles. The system has only just been developed; several experimental prototypes are now in the early stages of testing.

pared with a normal room illumination of about 50 candelas per square meter the cathode-ray tubes in the F-16 system would generate about 100 candelas (if the tubes could be viewed directly). The collimating lenses and diffusing screen windows absorb so much light, however, that the maximum light reaching the pilot's eyes is only about one candela—closer to the light level of twilight than to that of high noon. Because the pilot is able to adapt to such a light level easily, there is little difficulty in seeing all the details of the scene. Unfortunately the low light level puts a limit on the contrasts in light possible—a limit that makes it difficult to display objects close to the lower threshold of vision.

Although cathode-ray tubes provide the visual simulation for most of the flight trainers in use today, there are two other display procedures. In one of them the scene is projected onto a panoramic motion-picture screen in front of the pilot. A system of this kind achieves slightly better resolution and brightness than a cathode-ray-tube display but otherwise has similar characteristics.

In the other display procedure the projection surface is incorporated into the pilot's helmet. This system has only just been developed, with several designs still in the testing stage. The displayed image either can come from a pair of miniature cathode-ray tubes also attached to the helmet or can be transmitted from a projector behind the pilot through a pair of fiber-optic bundles. When it is perfected, the helmet-mounted display should prove to be a highly cost-effective visual simulator because it is so small. Although both the computer and the projection equipment are similar to those needed for other types of simulators, no external screen is required or any battery of mammoth cathode-ray tubes: everything fits in the pilot's helmet.

Because it is capable of much greater resolution and brightness, the helmet-mounted display system offers substantial theoretical advantage over other systems. Furthermore, it allows separate visual input into each eye. Research suggests that binocular disparity is useful in processing depth, and that information about depth from relative motion is also processed binocularly, yet no current simulator provides the pilot with this information.

Display and computer-generated imagery systems are not overly taxed when they are used to simulate training tasks such as air-to-air refueling, formation flying or air-to-air combat, since in each such task only one or two planes need be displayed against an otherwise empty sky. Problems in displaying details do occur, however, in the simulation of all aspects of low-level flight, including target detection, air-to-ground attack, visual and radar avoidance by flying close to the ground, and close-in support and reconnaissance.

The problems in displaying scenery for low-level flight arise for two independent reasons: more detail is needed so that the pilot can correctly perceive altitude and maneuvering room (this demands more resolution) and the detail has to be changed faster given the rapid movement of the viewpoint from moment to moment (this demands faster update rates). The need for detail arises because the ground cannot simply be a matrix of uniform checkerboards; empty checkerboards specify little about the undulations of the ground. Without substantial added surface details and objects a pilot has great difficulty telling altitude or judging whether ridges can be cleared or obstacles avoided. If the task is one of hugging the ground as closely as possible in order to avoid detection, then richness of surface texture and detail becomes overwhelmingly important.

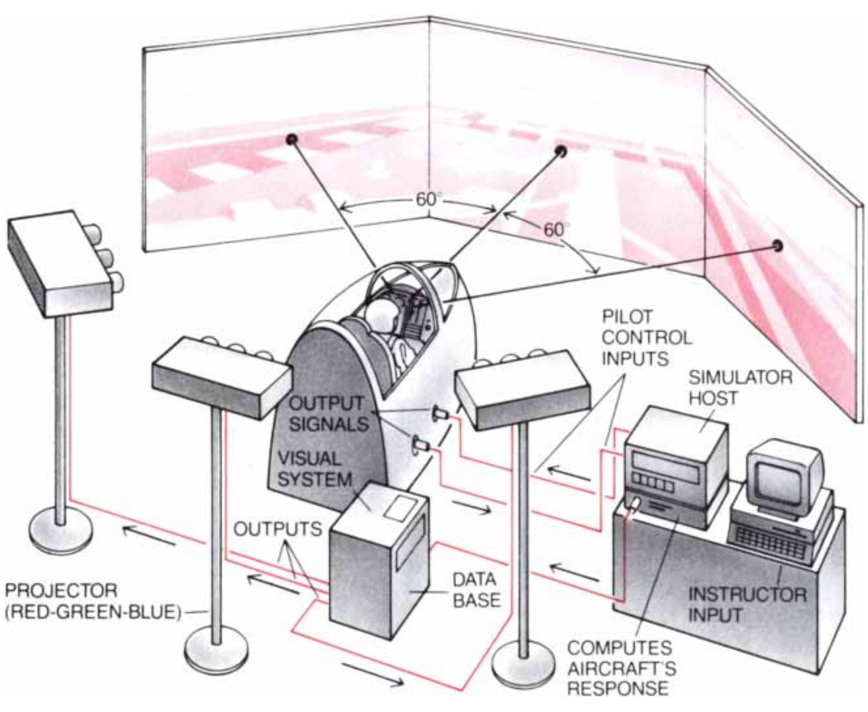
Three solutions have been pursued to provide enough detail for low-level flight. Since most computer systems create features by drawing straight lines or curved edges, one solution has been to demand more edges in real time. This demand is gradually being met. Off-the-shelf systems now have 10 times as many available edges as they did just five years ago, and the industry expects another tenfold increase in the next five years.

The second solution has been to add an extra high-resolution channel to the display system, one that can be moved around depending on the direction of the pilot's gaze. An eye- or a head-position sensor detects the direction of gaze at every moment. Either electronically or mechanically, the area of the scene directly in line with the direction of gaze is displayed with a resolution much finer than that of the rest of the screen. This can be done if, instead of using each cathode-ray tube to cover a 90-degree-square area, the high-resolution channel is fed to a cathode-ray tube that covers a 30-degree-square area—a tripling of resolution. The overlapped area is blanked out of the background channels so that only one image is seen, but the display straight ahead is much sharper.

The third solution depends on deciding which kinds of detail are really necessary, and then on displaying just those. In the long run this approach will be more effective, but the research is just being done, so that only a few answers are available. One general finding appears to be that the different sources of information vary in value depending on the perceptual task facing the pilot. When the plane is flying low and fast, for example, edges used to define surface texture are less important than edges used to define changes in relative position. In other words, the information about relative motion of ground features may be more important than information about ground texture itself.

Low-level flight also strains computer-generated imagery systems because all the edges in the scene must be replotted at a very rapid rate. The refresh rate of a raster-scan cathode-ray tube is 60 times per second: the tube completely rewrites the contents of the scene every 16 milliseconds. Since the scene itself is changing, the computer must update the display to reflect the new positions of every object in the scene as perceived by the pilot. With the need for rapid change, as when the pilot flies low and fast, the available update rates for the entire scene are simply inadequate. Larger and faster computers will help to solve this problem, but more inventive display systems are needed as well.

In actual flight the pilot can see the movement of the plane across the terrain and can feel the changes in gravitational forces on him associated with turns, climbs, dives and atmospheric turbulence. These perceptions arise from visual motion information that comes to the eyes and gravitational-force changes that come to the vestibular organs in the inner ear. Most motion can be perceived by vision alone, so that the inclusion of gravitational information in a simulator is technically not necessary. Since such information is always available to pilots in real flight, however, many flight simulators include a motion platform whose movement stimulates the vestibular organs.



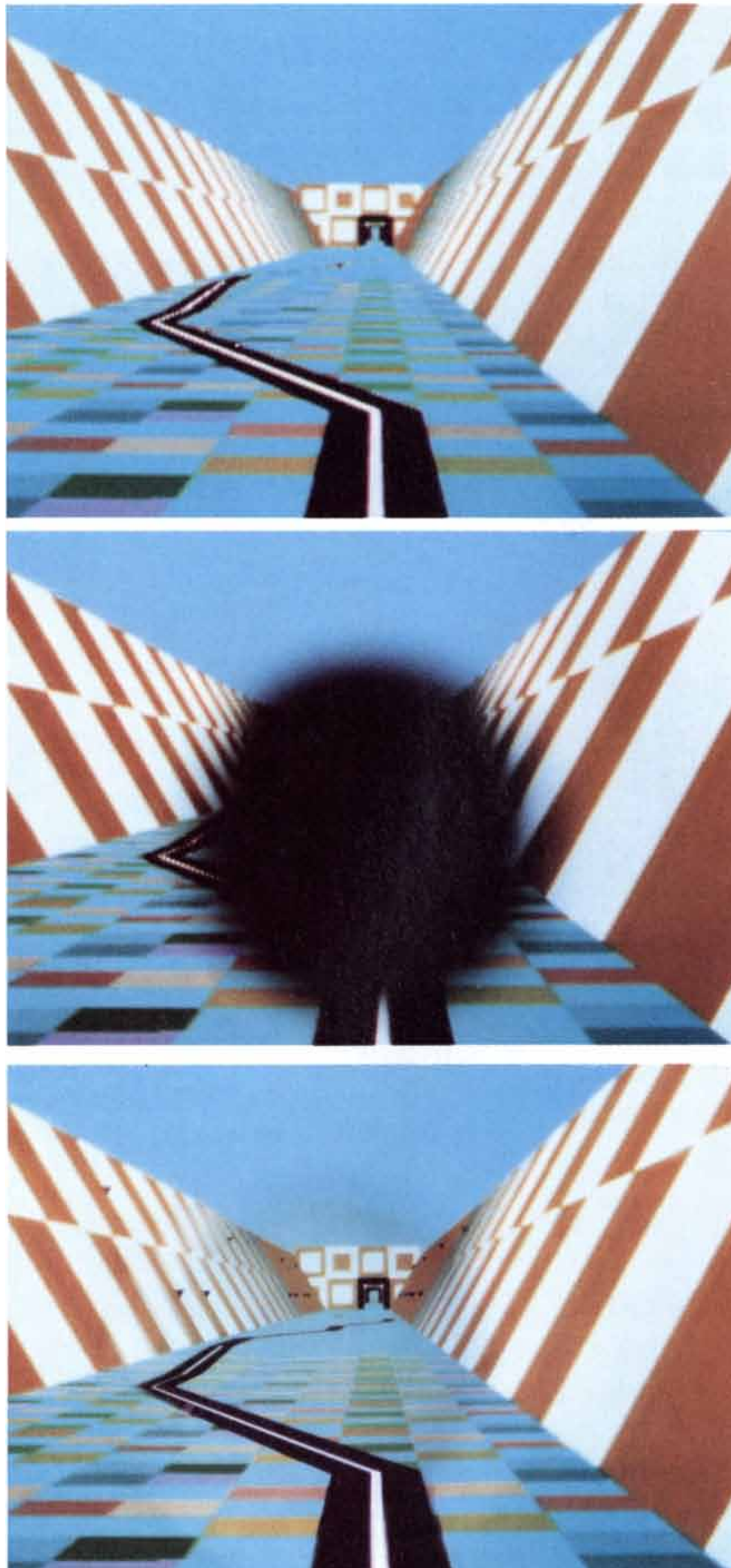
PROJECTION SCREENS offer yet another alternative to cathode-ray tubes as a means of displaying simulated visual scenes. The control room of a flight simulator need not be large; in some instances, such as the one depicted here, even a small microcomputer is sufficient.

When the body is moving at a constant speed, only changes in the relative positions and sizes of objects as detected by the visual system provide information about that movement. The vestibular organs can detect only acceleration and deceleration, which are changes in the rate of motion. The changes can be in a transverse orientation, such as those felt in a car or an airplane when it speeds up or slows down. A well-tuned sports car in a jackrabbit start can produce an additional *g* (a unit of force equivalent to the force of gravity) of transverse force, equivalent to a momentary doubling of the driver's weight during the start. A 727 jetliner creates about 1.5 additional *g*'s on takeoff.

In addition to the transverse forces, changes in motion along a vertical dimension can also be felt, as in the starting or stopping of an elevator. Whenever a plane begins to climb or to bank into an inside turn, gravitational forces push the body down in the seat (positive *g*'s), whereas beginning a dive lifts the body out of the seat (negative *g*'s). In a fighter plane turning provides the strongest change in motion information. Just as the weight of a bucket at the end of a rope increases as it is swung around faster, so the plane and its pilot become heavier as the plane makes tighter turns at higher speeds.

Such changes are also expressed in terms of *g* forces. A roller coaster can produce up to about +4 *g*'s of lift during its sharpest turn; an F-16 jet fighter can exert more than +11 *g*'s of lift on the pilot (and plane) during a steeply banked level turn. This is equivalent to increasing the pilot's weight from, say, 200 pounds to 2,200 pounds during the turn; a four-pound helmet suddenly weighs more than 40 pounds. Although +11 *g* turns are made only rarely, turns producing from +6 to +9 *g*'s are made routinely.

To produce the same sensations on the pilot's vestibular system the entire cockpit of the simulator must be moved physically to imitate the changes in speed or altitude of the aircraft. The National Aeronautics and Space Administration has built centri-

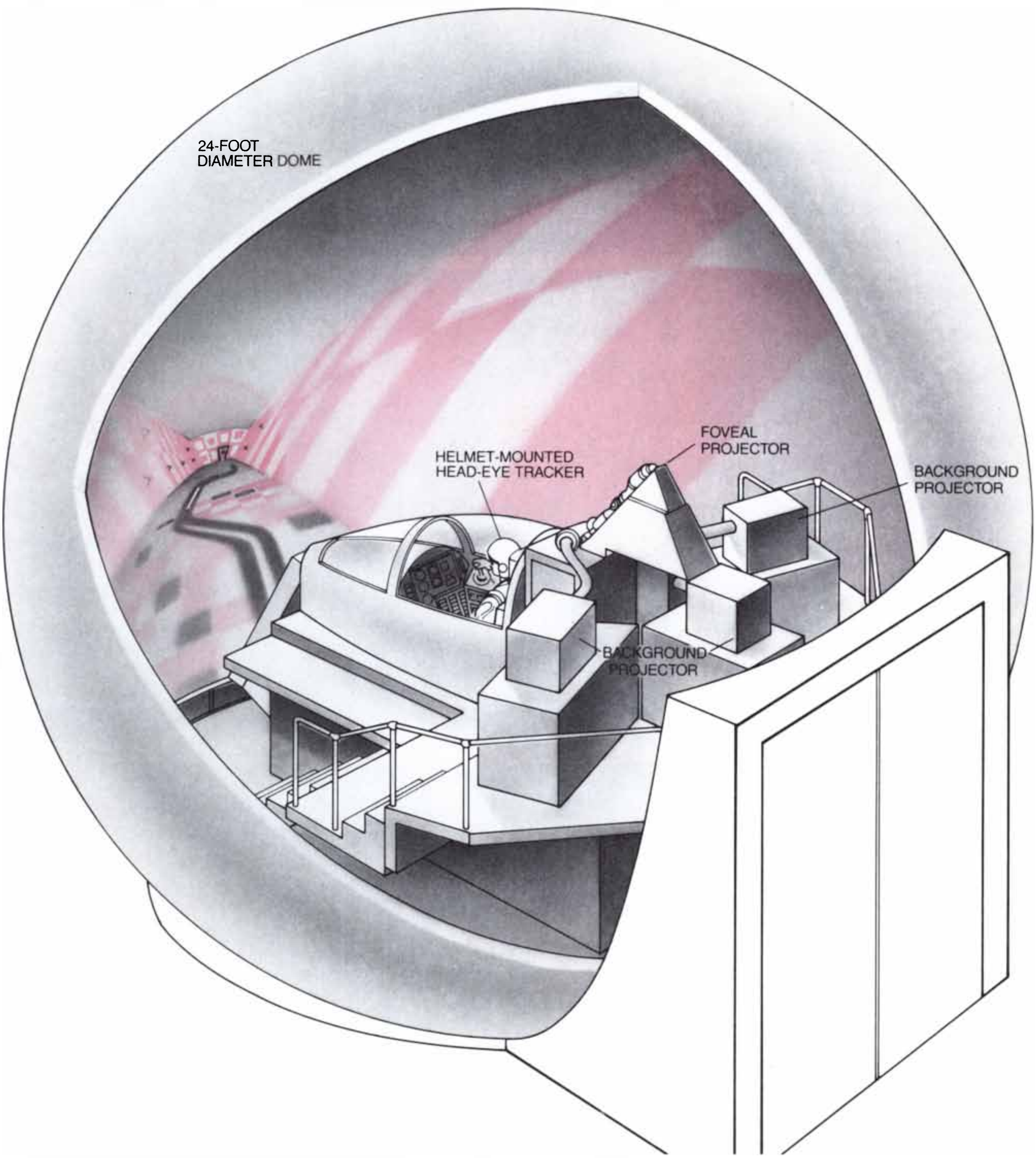


GREATER DETAIL in the visual scene of a flight simulator can be provided by adding an extra high-resolution channel to the display system. An eye- or a head-position sensor detects the direction of the pilot's gaze at every moment. The area of the scene directly in the line of sight is blanked out of the background channels (*middle*) and displayed with greater detail by means of the high-resolution channel (*bottom*). Only one image is projected; the amount of visual detail is greatest in the forward direction.

fuges to simulate the forces, which can exceed +10 *g*'s, that an astronaut experiences during the lift-off and return of a rocket. It is theoretically possible to mount an entire fighter cockpit in such a centrifuge, but that would be astronomically expensive and is probably not necessary.

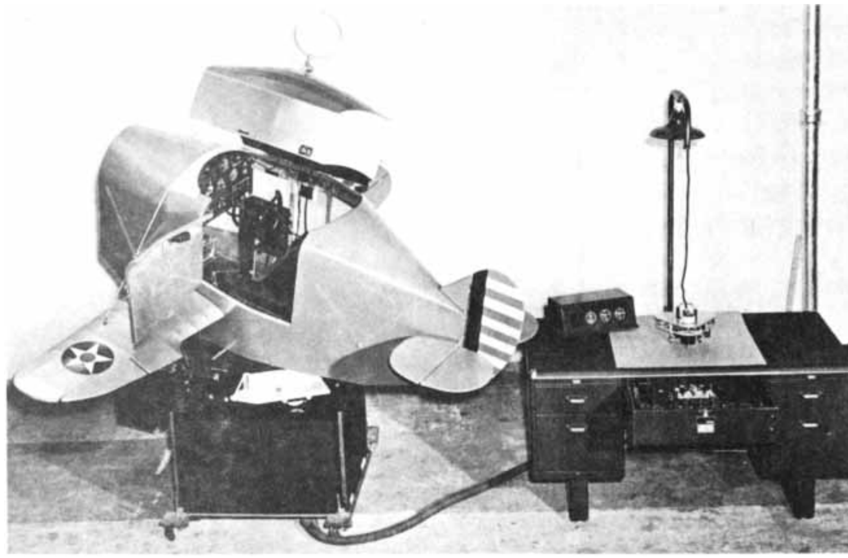
Instead the cockpit is mounted on a platform that is moved about by a set of hydraulic pistons. The distance of travel is rarely more than two feet in any direction, and the simulated forces are limited to 1.8 *g*'s, enough to provide only "onset" cues to the beginning of a change in motion. The system certainly provides more motion information than no system at all, but there is little evidence that it improves flying performance in the simulator. In training for instrument flying (necessary in fog or clouds), however, even the minimal motion cues provided by the motion platform are compelling.

A relatively inexpensive and ingenious method of simulating positive *g* forces is the inflatable *g* suit. A *g*



DOME-PROJECTION SIMULATOR under construction at Link will provide a continuous image over a field of view measuring 270 degrees horizontally by 138 degrees vertically. Known as ESPRIT (for eye-slaved projected raster inset), the system will endow the visual area in the trainee's line of sight with greater detail than is received by the remainder of the visual area (*see illustration on preceding page*). The background projectors will display the peripheral images; the foveal projector will display the area in the line of sight.

suit covers the abdomen and legs and acts much like a blood-pressure cuff around the lower parts of the body whenever positive *g* forces are experienced. This both increases blood pressure and retards the flow of blood away from the upper parts of the body, particularly the head. Even in a simulator without any motion platform, if the plane's motion would normally subject the pilot to positive *g* forces, the *g* suit inflates, providing a cuing sensation of a turn or a climb.



EARLIEST FLIGHT SIMULATOR, the Link trainer, was used during World War II. Demands by air forces have been the chief incentive for the development of simulators.

Flight simulators can obviously be justified in savings in fuel and maintenance for aircraft. Whereas it costs about $5,000 per hour in fuel and maintenance to fly an F-16 jet fighter in training, it costs less than a tenth of that per hour to fly the F-16 simulator. More important, it costs nothing to crash the simulator, and all risk is removed that the plane and the pilot might be lost in an accident.

The strongest argument for simulators involves neither cost nor safety, however. Rather, simulators provide the opportunity to ensure training effectiveness in ways that are not possible in real aircraft in actual flight. Simulators can, for example, provide practice for responses to highly unlikely and potentially disastrous events. When a DC-10 lost an engine during takeoff in 1979, its crew had no experience with the sudden changes in stability and lift resulting from the uneven loss of power and damage to the wing. Although it is not realistic to practice such an event in a real DC-10, it is easy to simulate the effects of such changes on the aerodynamics of the DC-10 and let pilots learn to handle the emergency. In the simulator a wide variety of emergencies, failures and accidents can be experienced and countermeasures perfected.

The rate of learning can also be improved with simulated flying. Research has shown that some flying tasks are mastered more quickly and brought to a higher level of skill if the subcomponents of the task are practiced separately or in an order different from the natural one. It has been shown, for instance, that pilots learn to land better if they first practice the final approach without having to fly the earlier stages of the approach.

Simulated training also allows much greater concentration of trials. In air-to-air combat an hour's flying time may allow only three or four engagements, even though each one lasts for only a minute. In the simulator 20 to 30 such engagements can take place in an hour. As another example, in actual landing practice 95 percent of the time is spent circling and taxiing. All of this can be eliminated in the simulator, so that 20 times as many landings can be made per hour of flight time in the simulator.

Some flight-training simulators are equipped with "freeze and replay" capabilities. If the instructor sees the pilot making an error, the position of the plane can be frozen and the error can be discussed and analyzed before flight is resumed. After the flight ends the entire sequence can be repeated with the computer piloting the plane just as the pilot had done previously.

The simulated scene can also add information that is not in the real scene to aid the pilot in developing flying skills or distance perception. In low-level flight, for example, pilots often have trouble at first in judging the height of objects or ground contours. To help them, artificial cues to height can be added until the pilot develops an internal scale. Then the cues can be gradually deleted from later practice.

The experiences of the major air forces in World War II, in Korea, in the Middle East and in Vietnam suggest another potential purpose for simulated training. Losses in combat are concentrated almost exclusively among pilots with five or fewer combat missions; if the fifth mission is survived, the probability of surviving the remaining ones is more than 95 percent, regardless of how many additional missions are flown. These figures suggest that if all pilots could be given the equivalent of their five combat missions before they face the enemy, losses could be minimized drastically, perhaps altering the outcome of the engagement or war.

Flight simulators could also facilitate comparisons of different training programs—comparisons that would otherwise be costly, risky and difficult to make. The advantages and disadvantages of each program could be examined in the course of training in the simulator. It would become possible to determine which components of existing training programs are effective and which are not and to test new ideas and procedures. This is particularly advantageous when the procedure under examination is only one part of a long training sequence.

A final justification for simulators goes beyond flight training altogether. Simulators of the kind described here are the best existing devices with which to carry out basic research on many aspects of visual perception. Is stereoscopic vision useful at 100 feet? How much information does a moving person get from peripheral vision? Can peripheral information be processed automatically without interfering with the need to focus on targets that lie straight ahead?

Experimental psychologists have asked these questions before, but up to now most research has relied on either impoverished laboratory presentation of stimuli or stationary displays and stationary observers. Flight simulation allows far better control over the presentation of moving stimuli while allowing for movement through or over the scene. It is obvious that answers to these questions can help to improve the design of simulators. More important, however, they can significantly advance understanding of how even earthbound human beings perceive their environment.

The Arthropod Cuticle

This complex covering accounts for much of the adaptive success of the arthropods. As its structure and chemistry become clearer, so do the properties allowing it to provide protection and support

by Neil F. Hadley

The phylum Arthropoda exceeds all other groups of animals in number of species and diversity. In fact, arthropods, which include the crustaceans (crabs, lobsters, isopods), insects (wasps, bees, ants, beetles), arachnids (spiders, scorpions, ticks, mites), centipedes, millipedes and several lesser groups, account for approximately 80 percent of all known animal species. They have also radiated to occupy every conceivable type of ecological niche. After the first arthropods made a success of aquatic habitats, members of the phylum eventually became not only the first animals to make the land their home but also the only invertebrates to fly and to successfully adapt on a large scale to life on land.

A key factor in the eminence of the phylum, whose members possess a segmented body, jointed limbs and a strong external skeleton, is the cuticle. This material serves as the skeleton, covers the body surface and appendages and, in many insects, forms wings. It also lines the tracheal tubes, various gland ducts and sense organs and the anterior and posterior regions of the digestive tract. The cuticle provides support, a place for muscles to attach and a barrier between the animal and the environment. Indeed, one of the cuticle's most critical functions, limiting water loss from the animal, enabled the arthropods to make a successful transition from an aquatic to a terrestrial existence. The cuticle can be rigid or soft. For example, many crustaceans and beetles have a thick, rigid, armorlike cuticle, but soft-bodied larvae of many arthropods have a thin, more flexible exoskeleton.

Because the cuticle is so important to the evolution and continued survival of the arthropods, investigators the world over have studied its features for decades. The effort has progressed remarkably in the past 15 years, owing in large part to advances in microtechnologies. These technologies have, in varying degrees, now answered such questions as: What is the basic fine structure of the cuticle? What is the chemical composition and architectural arrangement of the major cuticular divisions and how do these features affect cuticular function?

Early studies done with light microscopes uncovered the two major structural divisions: a thin top layer known as the epicuticle and a thicker layer known as the procuticle. These layers rest above the epidermis, a single layer of cells that originally secrete the cuticle. Together the cuticle and epidermis form the integument, or outer covering, of the arthropod.

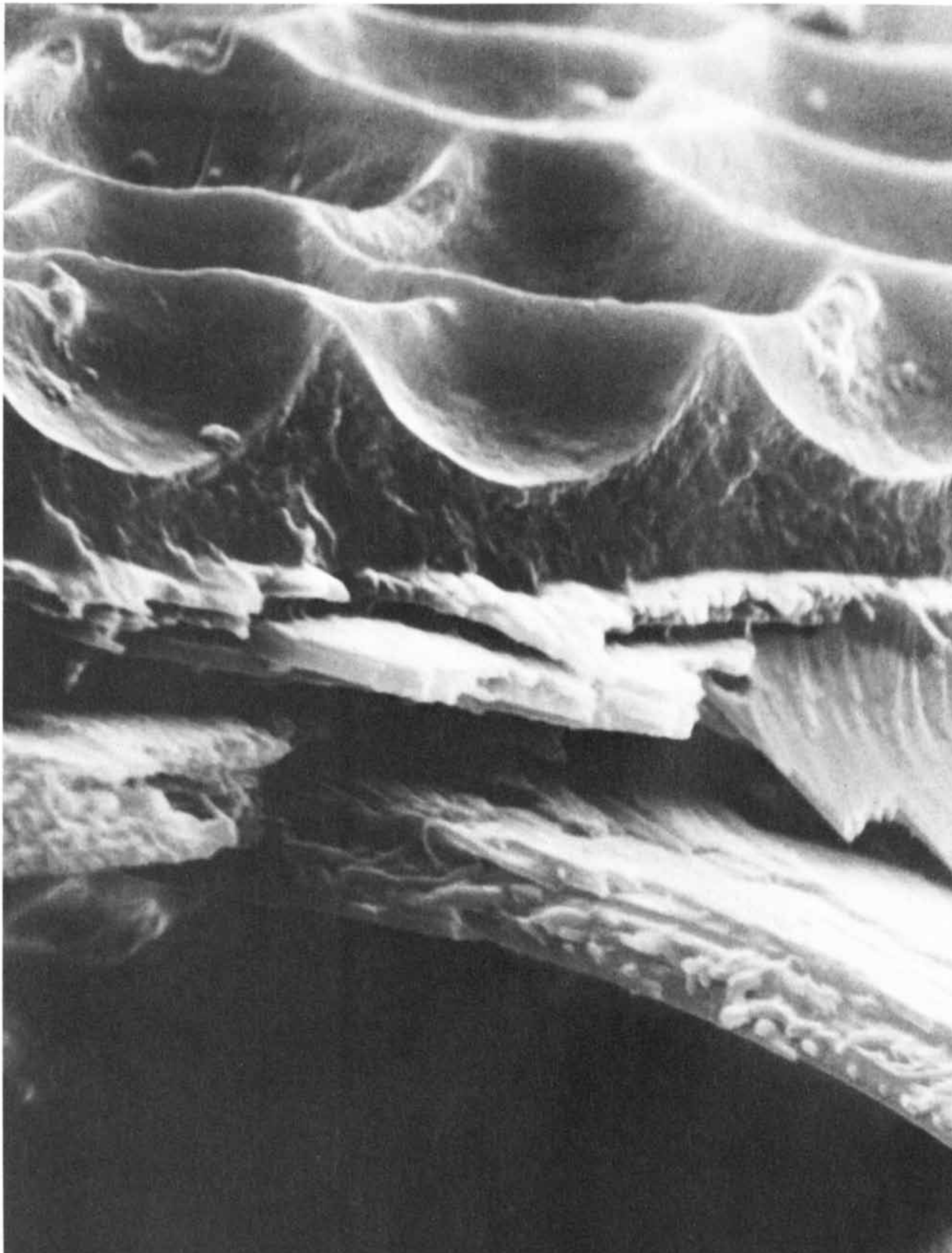
Observation of the fine structure of the two basic cuticular divisions had to await the development of electron microscopes and specialized methods for preparation of specimens for viewing. Among those who have applied such tools, particularly the transmission electron microscope, are Anthony C. Neville of the University of Bristol, Michael Locke of the University of Western Ontario, Barry K. Filshie of the Commonwealth Scientific and Industrial Research Organization in Australia and Yves Bouligand of the École Normale Supérieure in Paris.

The epicuticle, which is important to the ability of most arthropods to retain water, contains at least two sublayers in spite of its being only from .1 micrometer to three micrometers thick. It may have as many as five sublayers in some species, although the cuticulin layer on top and the inner epicuticle, which itself accounts for more than 90 percent of the entire epicuticle, appear to be the ones that are universally present.

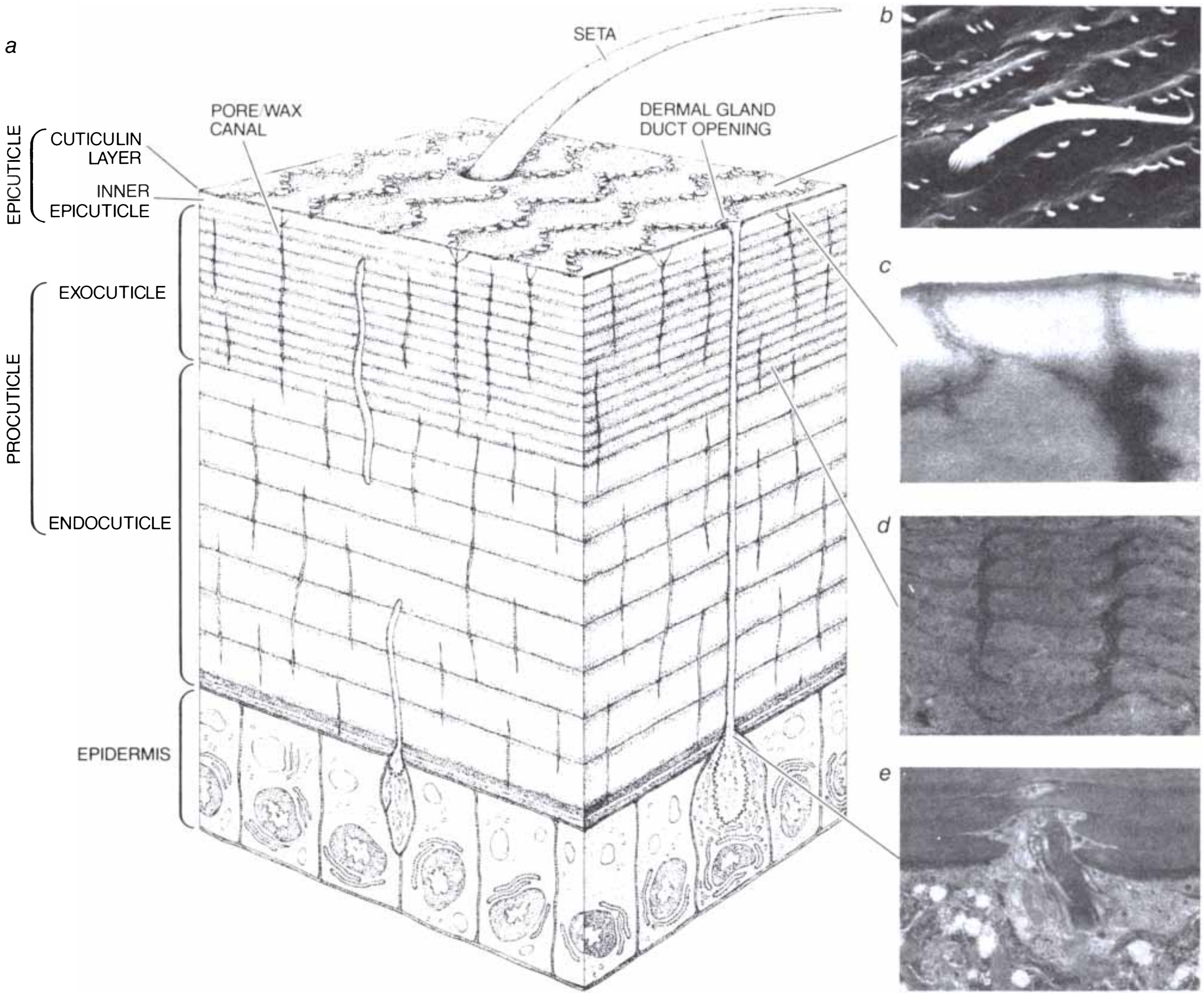
The surface of the epicuticle is particularly fascinating. Viewed with a scanning electron microscope, which produces three-dimensional images, the surface shows a surprising degree of microsculpturing. A common pattern consists of repeated polygons, often hexagons, with slightly raised margins. Such a relatively flat surface typically makes the animal glossy and often appears in species that sport warning coloration. Often microscopic structures of various shapes—such as knobs, spines or folds—partially or completely obliterate the polygons. These microstructures reduce the reflection of light from the surface and can give the arthropod a lusterless appearance; they are commonest in insects that depend on concealment for protection from predators.

The other major cuticular division, the procuticle, makes up the bulk of the integument and ranges in thickness from 10 up to several hundred micrometers. The procuticle accounts for the shape and strength of the cuticle. It typically divides into a relatively thin exocuticle—a pigmented layer at the top—and a thicker endocuticle. In some instances two other sublayers form: the mesocuticle, between the exocuticle and the endocuticle, and the deposition layer, between the epidermis and the cuticle.

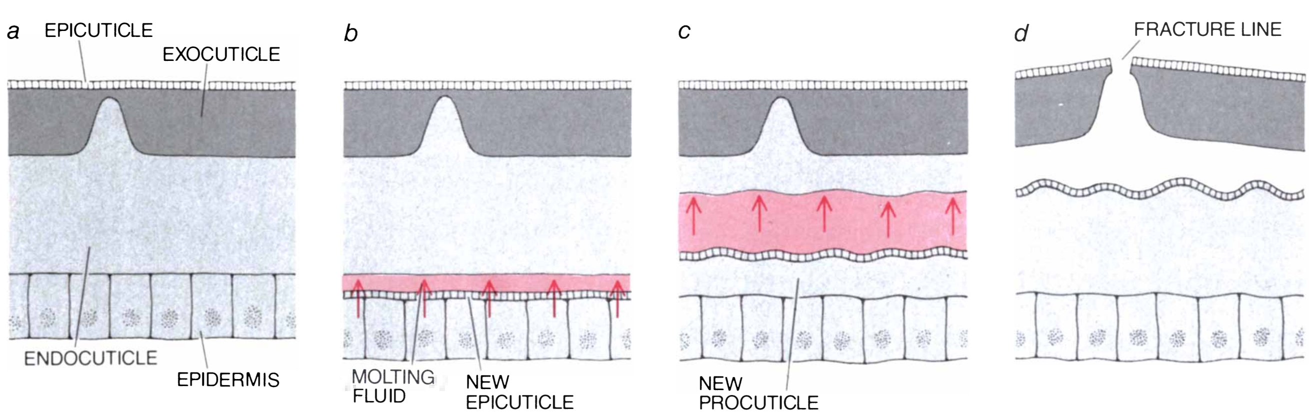
Vertical channels penetrate the various cuticular layers. The largest of these, the dermal gland ducts, extend from sacs in the epidermis to the surface, where their openings are often spaced at regular intervals. The material secreted by these ducts has yet to be firmly identified. The other channels, complexes of pore canals and wax canals, are narrower and more numerous than dermal gland ducts. The pore canals emerge from the epidermis and join wax canals at the junction of the exocuticle and the epicuticle; the wax canals terminate at or near the surface of the cuticulin layer. The complex of pore and wax canals is thought to be the route by which lipids ("waxes") travel from the epidermis to the epicuticle. In crustaceans pore canals also probably transport calcium



CUTICLE covering the wing of a tiger beetle has the layering that is characteristic of the arthropod cuticle. The specimen was frozen and fractured, and a cross section is enlarged some 7,200 diameters in this scanning electron micrograph made in the author's laboratory. Underlying the raised polygonal shapes on the surface (*top*) are the epicuticle and the exocuticle, which appear as a single dark area at this magnification. Under these layers is the endocuticle (*jagged white sheet*); its organization resembles that of plywood. The white arc below the wing cuticle is the endocuticle-lined upper wall of a large tracheal duct, and the black area is the duct's cavity.



BASIC FEATURES of the arthropod cuticle are evident in the cuticle of the common house cricket (*a*). The major divisions, the epicuticle and the procuticle, rest above a single layer of epidermal cells (the epidermis). The procuticle is often further divided into an exocuticle and an endocuticle. The cuticle is penetrated vertically by dermal gland ducts and narrower pore canals that join wax canals near the top of the cuticle. At the right four micrographs from the author's laboratory show (*top to bottom*) a seta and many smaller protrusions (microtrichia) on the surface (*b*); wax canals joining a pore canal at the junction of the epicuticle and exocuticle (*c*); pore canals (*dark areas*) twisting their way through the exocuticle (*d*), and a dermal gland duct emerging from a gland in the epidermis (*e*).



NEW CUTICLE IS FORMED in stages. Before the arthropod molts (*a*), or sheds its old cuticle, a new cuticle begins to form. First the original cuticle separates from the epidermis, which secretes fluid containing molting enzymes into the newly created space and also secretes a new epicuticle (*b*). Then the fluid digests the old endocuticle while the epidermis secretes a new, undifferentiated procuticle (*c*). When the fluid completely digests the old endocuticle, much of the fluid is reabsorbed, the original cuticle cracks (*d*) and the animal emerges. Soon after, the new exocuticle may stiffen and the epidermis may add material to parts of the fresh cuticle.

to the endocuticle and exocuticle; calcium and other inorganic materials harden the cuticle.

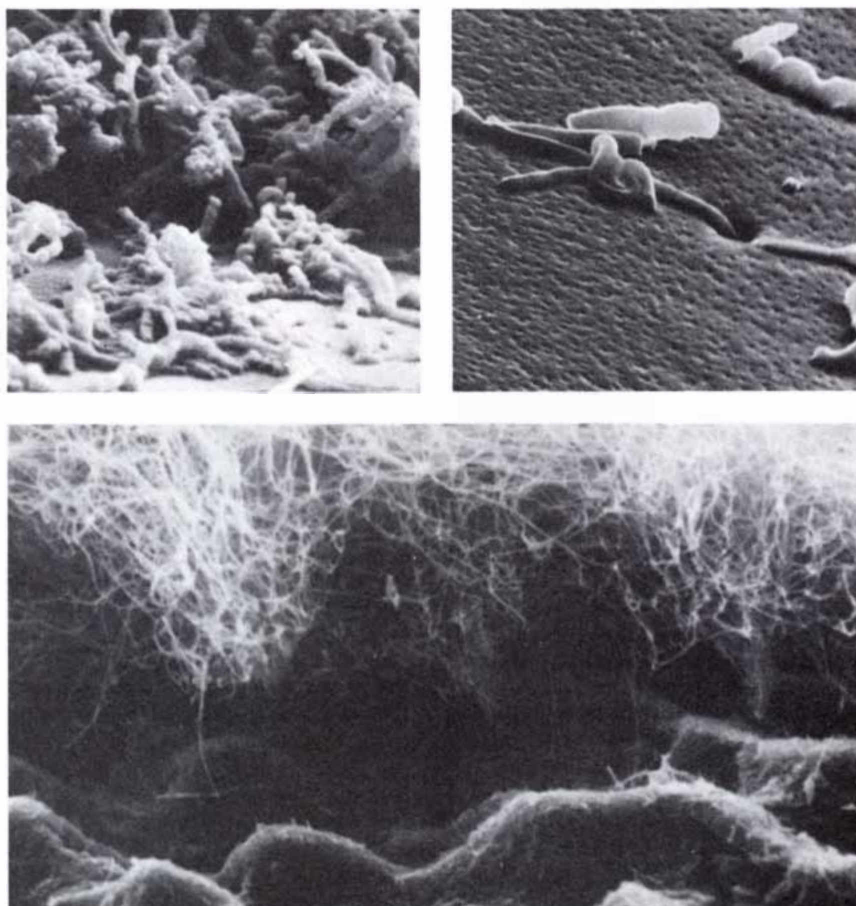
Of course, the cuticular structure undergoes change during ecdysis, or molting—a process characteristic of arthropods and one that is essential if the animal is to grow. Before the old cuticle is shed the epidermal cells secrete a new, soft epicuticle. (It is the epidermal cells that probably create the polygonal shapes that ultimately appear on the surface; the ridges of the polygons seem to correspond to the edges of the cells.) Once the epidermis secretes the epicuticle, enzymes pass through the new cuticulin layer and begin to digest the old endocuticle. When the endocuticle is partially digested, the epidermis secretes a soft procuticle. Once the old endocuticle is fully digested, the animal emerges from the old outer shell and swells with water or air to stretch the cuticle. Within less than 24 hours the new exocuticle stiffens. Later some arthropods may add calcium, lipids or other substances to the new cuticle.

Differing chemistries in large part determine the functional differences of the two major cuticular divisions. For example, the epicuticle is rich in lipids, although these chemical compounds are present to some extent throughout the cuticle. Lipids seem to provide the arthropod's primary barrier against water loss, but they apparently contribute little to the shape or strength of the cuticle. The procuticle, in contrast, consists primarily of chitin embedded in a protein matrix. Chitin, which is widely thought to be absent from the epicuticle, is a polysaccharide (a complex of linked sugars) similar to the fibrous cellulose that is a structural component of plants. Together chitin and protein make up 90 percent of the organic content of the cuticle and also provide much of its structural integrity. Other cuticular components include pigments, acids known as phenols and various inorganic materials.

Because no cuticular function is likely to be more important than water retention, investigators have given considerable attention to the chemical composition of cuticular lipids and to determining how these substances—particularly those associated with the epicuticle—prevent dehydration. Lipids are present in the epicuticle as free molecules, protein-bound molecules and bulk deposits on the surface. Extraction of lipids from whole bodies or from the molted cuticles of terrestrial arthropods usually yields a complex mixture of molecular classes.

Hydrocarbons, which contain long

DESERT BEETLE (*Onymacris rugatipennis*) **at the left is covered by a wax bloom; the beetle at the right (same species) has no bloom. Wax blooms are bulk deposits of lipid (wax) that may cover the entire cuticular surface and give the animal a frosted appearance. They help to prevent water loss and also can provide camouflage.** ***O. rugatipennis*** **is one of many tenebrionid beetle species found in the Namib Desert. The photograph appears through the courtesy of Elizabeth McClain of the University of the Witwatersrand in Johannesburg.**



WAX DEPOSITS shown in these micrographs from the author's laboratory are examples of the varied shapes that wax-bloom components can take. Bulk wax deposits are said to form a wax bloom when they project from the cuticular surface and are randomly oriented. The bloom of a wood-boring beetle (*top left*) **consists of twisting, tubular secretions. Individual secretions emerging from a pore on the surface** (*top right*) **can be seen when most of the bloom is removed by lightly brushing the surface. The bloom of a desert tenebrionid beetle** (*bottom*), **in contrast, consists of fine filaments forming a mesh that covers the cuticular surface** (*bumpy material*). **The "boundary layer" of air between the cuticle surface and the bottom of the filaments combines with the waxy mesh to provide the tenebrionid beetle with a water barrier that is among the most effective such barriers reported for any animal.**

chains of hydrogen-bound carbon atoms, are often the most abundant lipid types. A preponderance of hydrocarbons is significant because they generally interact poorly with water. The hydrocarbons that have been found are *n*-alkanes, or "normal" alkanes, which are paraffinlike molecules composed of straight chains that are saturated (they contain only single bonds between the carbon atoms); *n*-alkenes, which are also straight chains but are unsaturated (they contain a double bond between the carbon atoms), and branched compounds with one or more methyl groups ($-CH_3$). Oxygenated derivatives of *n*-alkanes accompany the hydrocarbons, albeit usually in smaller, and sometimes trace, amounts. These derivatives may include wax esters, free fatty acids, alcohols and ketones. Cyclic, or ring-shaped, lipids may also be present; sterols—particularly cholesterol—are the commonest. Of all these lipids *n*-alkanes and wax esters are generally the most hydrophobic, or water-repellent. In addition long chains of any lipid tend to be more hydrophobic than shorter chains.

Like the chemical makeup of lipids in the cuticle, the distribution of lipids could be expected to influence the barrier properties of the epicuticle. Unfortunately transmission electron microscopy has proved to be only mildly successful at revealing the distribution. There are several reasons for the difficulty. One is that removing waxes by current methods does not seem to affect the thickness or shape of the cuticular divisions from which the waxes are drawn; therefore the presence of lipids in the divisions of a given epicuticular cross section cannot be inferred by monitoring structural changes resulting from lipid extraction. Moreover, although workers can clearly see lipids in the upper cuticle when they examine stained cuticle with a light microscope, they cannot confirm the presence of lipids—or analyze them in detail—with a transmission electron microscope unless special precautions are taken; most conventional techniques for preparing cuticular samples dissolve the very lipids one hopes to reveal.

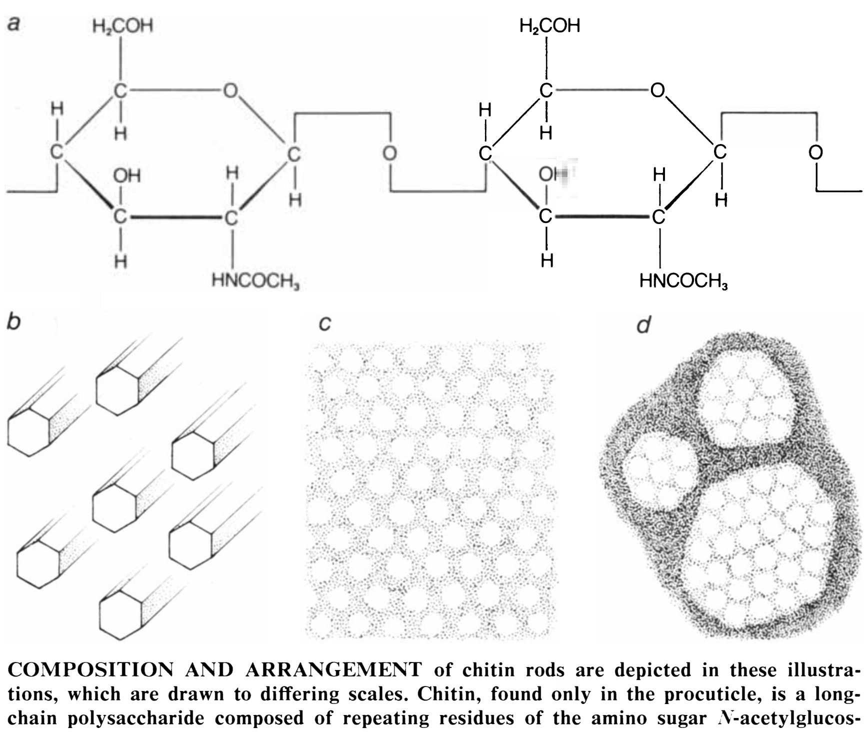
Some progress has been made, however. In 1975 Sir Vincent Wigglesworth of the University of Cambridge developed a way to view lipids in thin sections of cuticle by exposing the tissue to myrcene, an unsaturated hydrocarbon ($C_{10}H_{16}$), between applications of the fixative osmium. It is believed that myrcene becomes incorporated into the existing lipids and that osmium, which both fixes and stains the tissue, binds to the myrcene. With this technique Wigglesworth revealed lipid in the epicuticular channels of the assassin bug, *Rhodnius prolixus*.

In spite of difficulty confirming the presence of cuticular lipids with transmission electron microscopy, several investigators, myself included, have found evidence for the importance of these lipids, and of long-chain molecules in particular, to water retention by the arthropod. My students and I have shown, for instance, that insects and arachnids exhibiting low rates of water loss or living in environments that favor water loss often have the greatest quantities of cuticular lipids and also the greatest quantities of hydrocarbons or other lipids composed of long-chain, saturated molecules.

My own research seeks correlations between the effectiveness of the cuticular water barrier and the composition, shape and distribution of cuticular lipids. In reality all cuticular layers and several types of chemical compounds probably contribute to the water-resisting property of the arthropod cuticle, but my investigations have focused on the lipids within and on top of the epicuticle because they appear to be the main barrier to water loss in most species.

Much of the evidence for the presence of epicuticular lipids, and the importance of specific types of lipids to waterproofing, comes from studies of species that bear bulk wax deposits on the surface, particularly species that have wax blooms. A wax bloom is a coating formed by lipid deposits that project outward from the surface and are randomly oriented. Like plants that have wax on the surface, an animal with a wax bloom often looks as if it has been frosted with powder. Wax blooms appear in a variety of insect orders, including Lepidoptera (moths and butterflies), Coleoptera (beetles), Hymenoptera (wasps and bees), Homoptera (cicadas and aphids) and Hemiptera (true bugs).

Among the investigators who have described wax blooms are R. D. Pope and Jennifer M. Cox of the British Museum of Natural History (London). They have applied scanning electron microscopy to the study of aphids and mealybugs. This approach has provided much information about the shapes of wax extrusions and the channels from which they are secreted. Pope and Cox found that the extrusions in most species have shapes and contours similar to the pores that secrete them. Mealybugs, for instance, extrude curled wax filaments from



COMPOSITION AND ARRANGEMENT of chitin rods are depicted in these illustrations, which are drawn to differing scales. Chitin, found only in the procuticle, is a long-chain polysaccharide composed of repeating residues of the amino sugar *N*-acetylglucosamine; a fragment (*a*) is shown. Chitin molecules arranged in two or three rows form crystalline, hexagonal rods (*b*), shown here in a highly idealized form. When a cross section (cut from top to bottom) of an insect procuticle is photographed with an electron microscope and magnified a million times (*c*), the ends of the chitin rods appear as translucent areas (*white*) surrounded by protein (*light shading*). When a section of crab cuticle is similarly photographed and magnified (*d*), the protein-bound chitin rods appear to be grouped into bundles having various diameters. The spaces between the bundles (*dark shading*) are occupied both by protein and by the mineral calcium. Calcium increases the hardness of the crab's cuticle.

three- and five-holed tubular canals and produce long, hollow filaments.

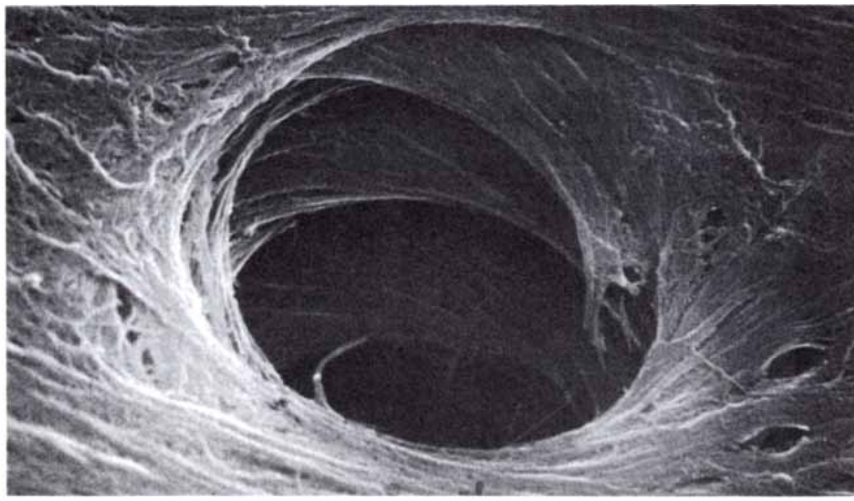
Wax secreted by certain species, though, does not retain its original shape. I found that the desert beetle *Cryptoglossa verrucosa* secretes surface wax as an amorphous blob from tubercles, or miniature nodules, on the cuticle. On exposure to low humidity the wax breaks up into many slender filaments (.14 micrometer in diameter) that radiate from the tip of the tubercle and ultimately connect with other filaments to form a dense mesh over the entire beetle.

The arthropods with the most spectacular—and perhaps the most extensive—wax blooms are the tenebrionid beetles of the Namib Desert in southwestern Africa. (Tenebrionid beetles are hard-bodied, generally dark insects that are common in deserts worldwide.) Elizabeth McClain of the University of the Witwatersrand in Johannesburg and Mary K. Seely of the Desert Ecological Research Unit of the Council for Scientific and Industrial Research in Namibia have found that wax blooms cover all or part of the body surface in 26 of the 46 species studied, often forming intricate patterns and assuming a variety of colors.

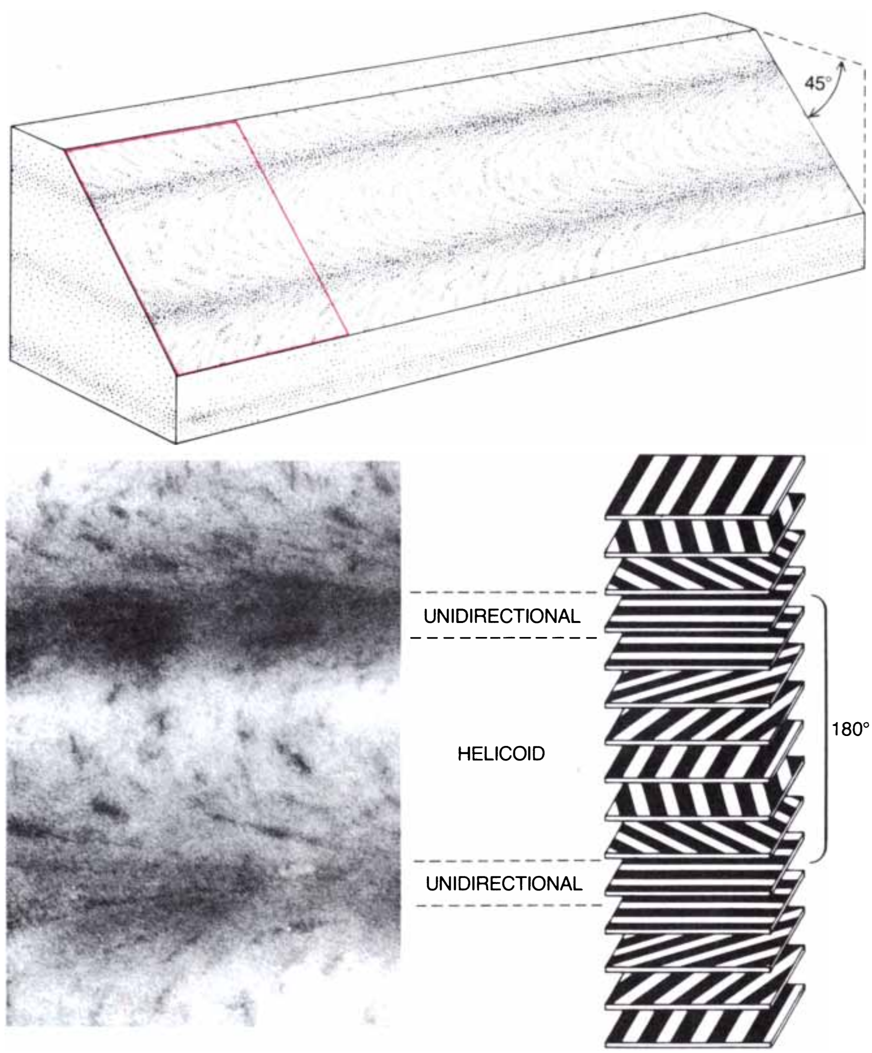
In a joint study McClain, Seely and I found that the percentage of species having wax blooms increases as the cool, foggy desert on the coast gives way to the hotter, drier inland desert. We also found that in a species inhabiting both regions the amount of bloom and its distribution over the integument increase as the climate becomes hotter and drier. Analyses of the lipid components of the blooms showed that hydrocarbons predominate, with the majority being the most water-repellent type: *n*-alkanes.

My students and I have also shown that individual animals from many arthropod groups can increase the amount of wax on their surface during times when conditions in the environment become particularly dry. Moreover, they can alter the composition of surface lipids when seasonal or laboratory-induced conditions dictate the need for increased water conservation. For instance, in summer tenebrionid beetles of the species *Eleodes armata* produce a greater quantity of hydrocarbons than they do in winter; moreover, a higher percentage of the hydrocarbons are long-chained. We also found that when beetles taken from a winter climate are acclimated to summer temperatures and humidities, their pattern of hydrocarbon production becomes similar to that of beetles in the summer.

The contribution of lipids to water



DERMAL GLAND DUCT (*dark circle*) **is surrounded by chitin fibers in this micrograph of the endocuticle of the scorpion *Hadrurus arizonensis*, photographed from the underside and enlarged roughly 11,000 times. The helicoid arrangement of fibers allows the cuticle to withstand stress that might otherwise lead to cracking. The micrograph is from Barry K. Filshie of the Commonwealth Scientific and Industrial Research Organization in Australia.**



SECTION OF ENDOCUTICLE from the thorax of a cricket, sliced at a 45-degree angle to the surface (*top*), is magnified about 50,000 times in the electron micrograph (*bottom left*). The alternating dark and light bands correspond respectively to unidirectionally and helicoidally arranged layers of chitin rods. The model (*bottom right*) illustrates rotation of sheets of chitin rods through 180 degrees; one full 180-degree turn constitutes one lamella.

retention may be the most important function of surface wax deposits, but the deposits also serve the arthropods in several other ways. Surface waxes can be a deterrent to predation. For example, the sticky, waxy threads that coat the surface of larvae of many ladybird beetles form a physical barrier to attacks by potential predators and to infestation by parasites. Elaborate colors and patterns created by certain surface waxes may also protect animals by camouflage or mimicry or both. Surface wax may even save certain arthropods from themselves; for instance, the curled filaments of wax on the surface of aphids and mealybugs probably prevent those animals from being trapped by their own viscous honeydew, or secretions, as well as that of their neighbors.

Temperature control is another potential function of the surface lipids. Light-colored wax uniformly distributed over the body surface of desert-dwelling arthropods increases the reflection of solar radiation, which in turn delays the rate at which the body temperature of small insects will approach lethal levels; this enables the animals to prolong their activity on the desert surface. Surface waxes may also reflect potentially damaging ultraviolet light, which could partially account for the high surface densities of cuticular waxes found in many insects and arachnids living at high elevations.

Lipids may also deflect water that might otherwise enter the animal from the environment. For instance, it has been suggested that the thick wax coat of aphids repels water during rainstorms. In these animals, which do not seek shelter during storms, the superficial waxes may act as a two-way barrier to the movement of water across the integument. Finally, mounting evidence indicates surface waxes are a source of important chemical signals and visual cues that elicit mating or other behavior.

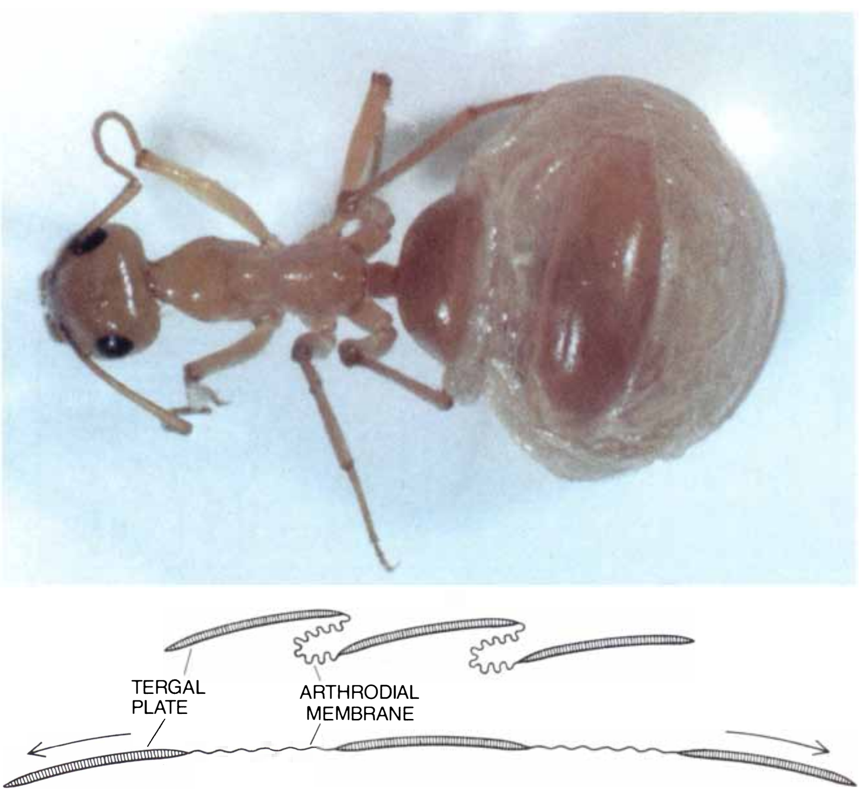
Cuticular lipids are certainly crucial to the survival of most arthropods. Of what value are the more abundant components—the chitin and protein of the procuticle? Together these two substances give the cuticle its mechanical properties, which have been likened to those of such strong two-phase materials as fiberglass. The chitin is bundled into rods that provide tensile strength but are flexible. The protein binds strongly to chitin and serves as the matrix in which the chitin rods are embedded, thereby cementing the rods together and reducing the cuticle's flexibility. In combination the chitin and protein enable the cuticle to bear a great deal of stress without cracking or tearing. Remarkably, the strength of the chitin-protein complex even compares favorably with steel: the tensile strength of steel is only 10 times that of cuticle when samples with equal surface area are tested.

This great durability of the cuticle is largely a function of both the chemical properties and the arrangement of chitin and protein in the procuticle. Chemically chitin is a long, straight chain of ring-shaped glucose derivatives connected by covalent bonds. Adjacent chitin chains are held together by hydrogen bonds. Extensive hydrogen bonding is particularly important because it contributes greatly to chitin's stability and tensile strength. Chitin typically accounts for between 25 and 50 percent of the dry weight of the cuticle, although estimates vary quite widely.

Compared with what is now known about chitin, much remains to be discovered about the composition and molecular arrangements of cuticular proteins. Investigators have, however, discovered that the protein in the procuticle can undergo sclerotization, a process that stiffens the cuticle and also causes the protein to become water-insoluble and usually darkened. Stiffness is widely believed to result from the linkage of adjacent protein chains by bridges composed of phenolic compounds or other compounds; it has been suggested, however, that water loss, which promotes bonding between protein chains, may alone account for stiffening.

Sclerotization, a process that occurs only in the exocuticle, is not reversible and its extent varies greatly among arthropod types. Whereas the cuticle of the tergites, or plates, covering the dorsal segments of caterpillars is only slightly sclerotized, the cuticle of mandibles (appendages for chewing) of certain beetles is highly sclerotized. Such sclerotization, along with the deposition of zinc and sometimes of manganese, contributes to the astounding ability of the beetle mandibles to bite through such metals as lead and copper.

Electron microscopy and X-ray diffraction studies have now yielded insight into the complex arrangement of chitin and protein that accounts for much of the cuticle's extraordinary strength. Some 18 to 25 individual chitin molecules grouped in two or three rows form a single crystalline rod called a microfibril. The protein that is bound to chitin probably links with the chitin at the periphery of the microfibrils and is thought to be part of the



HONEY ANT stores nourishment for the colony in its swollen abdomen, which is covered by a highly expandable, soft cuticle (*top*). Soft cuticle, more appropriately termed arthrodial membrane, is commonly found between plates and between appendage segments. As is shown at the bottom, it remains folded until the plates or segments move; then it unfolds.

general protein matrix. Chitin microfibrils are oriented parallel to the surface of the epicuticle and are arranged in sheets. Collections of sheets compose lamellae, or thin plates, within both the exocuticle and the endocuticle. In some animals, such as crabs, calcium and other substances are deposited between chitin rods.

The arrangement of sheets within each lamella is of particular interest because the sheets form a helix. Yves Bouligand first discovered this in crustaceans. Anthony Neville later described a helicoid structure in insects as well. In the exocuticle each sheet lies in a plane parallel to the sheet above it but is rotated counterclockwise at a slight but regular angle; a full 180-degree rotation constitutes one lamella. In the endocuticle, arthropods may deposit microfibrils that are helicoid throughout, but more often they lay down stacked sheets of chitin rods in one direction between partial or complete helixes. The lamellae in the endocuticle are often thicker than those in the exocuticle.

Combined with the microfibrils oriented in single directions, the helicoidally arranged microfibrils give the cuticle strength in much the same way as the layers of plywood make that material strong. Angled layers reduce the formation and spread of cracks that could otherwise cause the cuticle to fail under stress. Friedrich Barth of the University of Frankfurt has nicely demonstrated this property in spider cuticle. Instead of propagating, cracks in spider cuticle are deflected along the protein matrix between fibers and are stopped there.

Neville demonstrated in the early 1960's how the alternating layers of helicoid and unidirectional microfibrils are laid down. He showed that locusts deposit sheets in a single direction during the day and in a helicoid shape at night. Neville also showed that daylight and darkness, rather than the passage of time, control such deposition. By rearing locusts in constant light or darkness, he stimulated the insects to produce either totally unidirectional or helicoid cuticles. Other studies have since demonstrated similar daytime and nighttime variations in many insects; beetles are a notable exception. Counting the layers, which are visible under a light microscope, provides a reliable determination of the age of most adult insects.

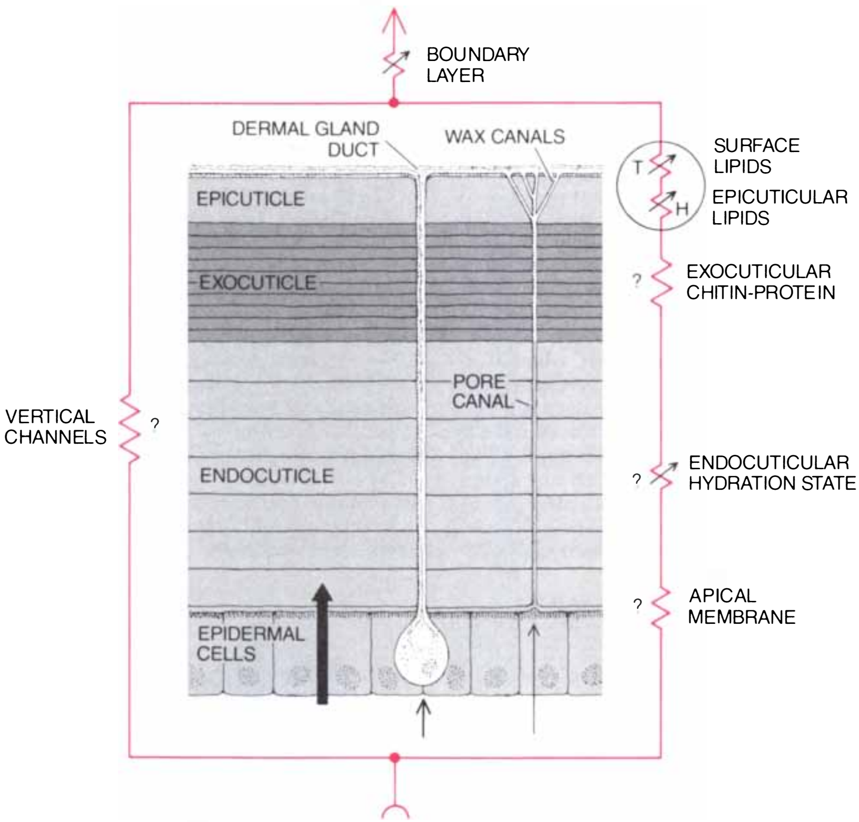
If all cuticle is constructed similarly, why is some cuticle rigid and some flexible? Variations in types of protein, in the relative proportion of chitin to protein and in the degree of sclerotization can each affect the degree of stiffness, elasticity and stretchability of cuticle in different animals and in different parts of the body. Two types of specialized cuticle immediately come to mind: the one containing the rubberlike protein resilin, and "soft" cuticle that joins plates or appendage segments. Both are much less stiff than the armor frequently found on the dorsal surface of beetles.

Resilin is found in regions where cuticular elasticity is necessary for function, such as in the wing ligaments of locusts, in the energy-storage pads that facilitate jumping in fleas and in the claws of scorpions, where extensor muscles are absent. The uniquely elastic material is composed of coiled polypeptide chains that are linked at their sides by amino acids. Networks of coiled resilin molecules in cuticle straighten when tension is applied and, even after prolonged stretching, return to their normal positions when tension is released.

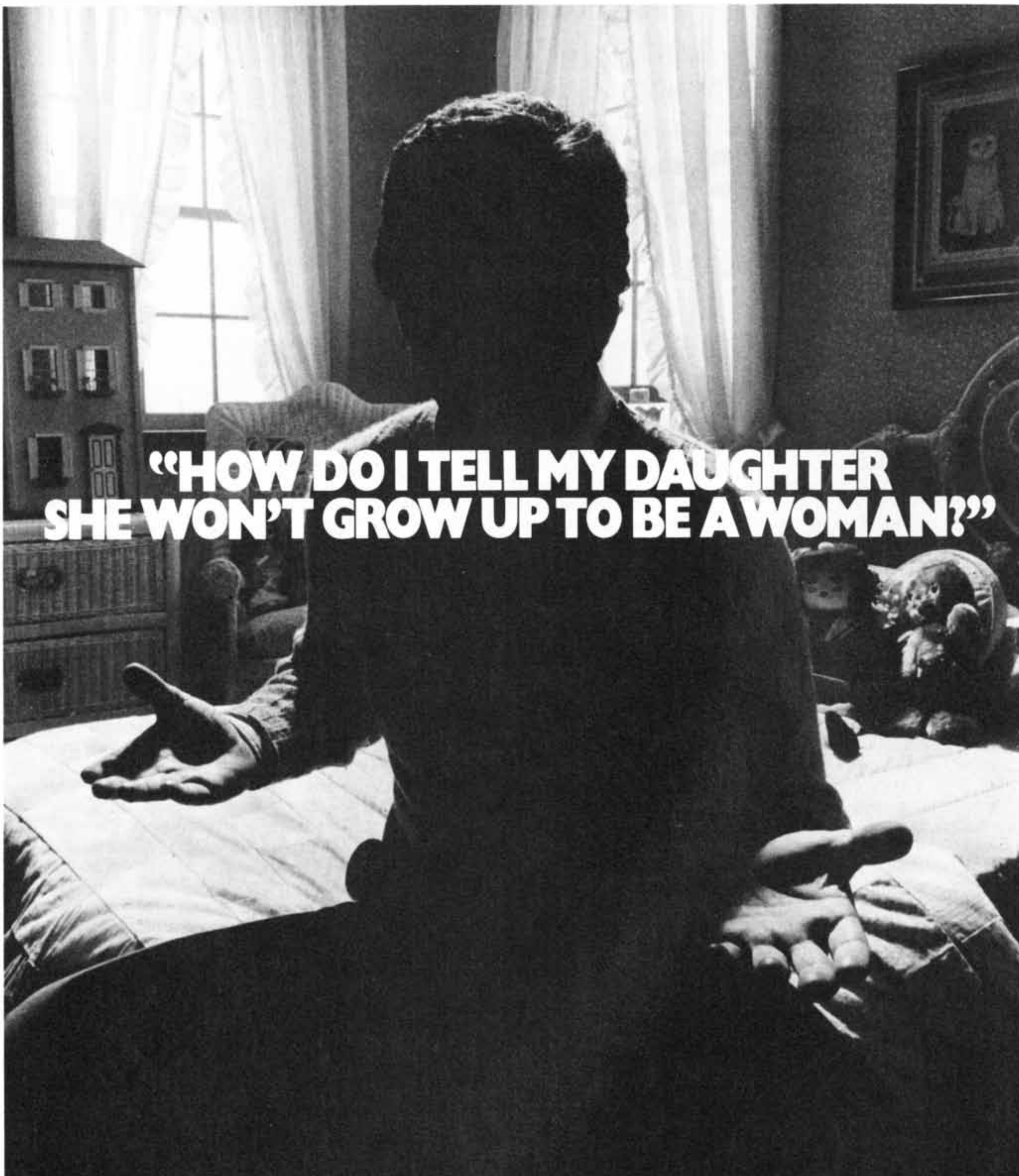
Soft cuticle, which is more appropriately termed arthrodial membrane, is highly flexible and stretchable, largely because its proteins are not sclerotized. Arthrodial membrane also contains more chitin than stiff cuticle does. Such a combination of unstiffened protein and chitin allows the membrane to bend and deflect easily and yet sustain relatively high tension. Sexually mature female locusts exhibit extreme expansion of soft cuticle; their abdominal intersegmental membrane can stretch up to 15 times its original length, enabling the female to lower her abdomen deep into the hole she digs for her eggs. Similarly, the abdominal surface area of the queen termite increases many times to accommodate the developing egg mass. Although not all molecular interactions involved in stretching are known, it is believed that chitin microfibrils are oriented in the direction of the stretch and that matrix proteins of soft cuticle are relatively reactive to water.

Whereas a great deal has been learned about the structural, chemical, mechanical and functional properties of the arthropod cuticle, several questions remain unanswered. Investigators have yet to determine the mechanisms by which various cuticular components cross the upper membrane of the epidermal cells and become part of the appropriate cuticular layers. The functions of the dermal glands and the chemical composition of their secretions are unclear. No one yet fully understands the workings of sensory structures that project from the cuticle, particularly the receptors that are believed to respond to chemicals in the environment or to changes in temperature and humidity. More remains to be learned about the proteins in which chitin is embedded and about the physical and chemical changes the proteins undergo when they become sclerotized.

Investigators are also seeking to better explain the nature and mode of functioning of the cuticle's waterproofing barrier. Clearly lipids associated with the epicuticle offer the principal resistance to water loss, and this resistance is supplemented by the thin, unstirred layer of air lying immediately above the cuticle surface. Investigators know less about the contributions made to this water barrier by the exocuticle, endocuticle and underlying epidermal cells. Also not known are the locations of specific lipid types within the epicuticle and the ways in which each type interacts with other lipid and nonlipid cuticular components to influence water retention. Studies that measure water loss from the cuticle of intact arthropods by means of miniature capsules attached to the surface, and that measure water loss across excised pieces of cuticle, are now under way. The results of the experiments may help to answer some of the above questions and, combined with continued morphological and biochemical investigations, should ultimately enable investigators to provide a more complete description of the structure and functioning of the arthropod cuticle.



BARRIERS TO WATER LOSS across the arthropod integument, which consists of the cuticle and the epidermis, are found throughout the cuticle. Water flow through the integument is represented as an analogue of an electric circuit (*colored line*). In diagrams of electric circuits, resistance symbols (*zigzag shapes*) indicate that something opposes the flow of an electric current through a circuit; here the symbol means that the element listed next to the symbol opposes water movement. Arrows through the symbols show that the resistance to water loss varies over time; letters enclosed by circles indicate that the barrier function of the element is influenced by temperature (*T*) and humidity (*H*). Question marks mean that the element possibly contributes to water resistance. In spite of the series of barriers, some water diffuses across the apical (upper) membrane of the epidermis (*thick arrow*) and across the layers of the cuticle. Although an endocuticle having a low water content tends to resist such diffusion, lipids within the epicuticle and deposited on the cuticular surface have been found to provide the principal resistance to water loss. The thin layer of air immediately above the cuticle surface (boundary layer) supplements the resistance provided by the integument. Some water may also cross the cuticle by way of the dermal gland ducts and the pore-and-wax-canal complexes (*thin arrows*). The ducts and the canals, however, are lined by compounds that may impede water movement. Hence the channels may also contribute to the resistance to water loss that is provided by the cuticle.



The thought of losing a child is unbearable for any parent. But I don't want your sympathy. I want your attention.

My child has Cystic Fibrosis. Every year 18,000 children are born with it, and every day three children die from it.

It takes more children's lives than Multiple Sclerosis, juvenile diabetes, and Muscular Dystrophy combined.

In spite of these bleak statistics, Cystic Fibrosis has yet to receive the attention it desperately needs. And this has to change. Because if it doesn't, the research that's needed to find a cure or a treatment for this disease will remain years behind other childhood diseases.

And that doesn't give children with Cystic Fibrosis much hope.

So how will I tell my daughter she won't grow up to be a woman?

If enough people help, maybe I won't have to.

"I WANT TO GIVE MORE THAN SYMPATHY."

Parents for Cystic Fibrosis Research,
P.O. Box 705, Plandome, N.Y. 11030

☐ Please send me more information.

☐ My $____ contribution is enclosed.

Name ________________
Company ________________
Address ________________
City ________________
State ________ Zip ________
Phone ________________

HELP CURE CYSTIC FIBROSIS.

THIS AD WAS WRITTEN BY A PARENT WHO HAS A CHILD WITH CYSTIC FIBROSIS AND WAS SPONSORED BY FRIENDS FOR "PARENTS FOR CYSTIC FIBROSIS RESEARCH."

THE AMATEUR SCIENTIST

Exotic patterns appear in water when it is freezing or melting

by Jearl Walker

If you look closely at a piece of ice, you will find several puzzling features. Why is it filled with bubbles and with tubes that resemble wormholes? Why do the tubes vary in diameter, sometimes in a regular pattern? What determines the distribution of the bubbles? When an ice cube begins to melt, why does it sputter and give off air and water? If a cube is held in bright sunlight, why do bubbles, hexagonal figures and fernlike figures develop inside it? The answers to these questions lie in an examination of how ice crystals form as water freezes.

Water freezes only after it has first cooled below the standard freezing point of zero degrees Celsius. In that state it is a supercooled liquid. The freezing process begins with a nucleating agent. In most cases the agent is a snow crystal or an impurity in the water. Molecules of water attach themselves to the agent, initiating a crystal structure. As additional water molecules join the crystal, it grows until it meets another crystal or the boundary of the container.

When the surface has frozen, the freezing process moves into the liquid. Freezing requires the removal of heat from the water to cool it to the freezing point and then to lock the atoms into the crystal structure. The heat passes through the ice layer by conduction, and then the air above the layer carries the heat off by conduction or convection.

If the water is pure (an ideal case), the temperature is lowest in the ice. At the interface between the ice and the liquid the temperature is the freezing point of water. In practical cases the water contains impurities that lower the freezing point to an extent depending on the concentration of the impurities. As water turns into ice at the freezing interface, the impurities are carried deeper by diffusion. Because diffusion operates quite slowly, the concentration of the impurities is highest somewhat below the freezing interface.

This situation is called constitutional supercooling. It is unstable enough to cause the freezing interface to move downward in projections rather than as a plane. When constitutional supercooling proceeds at a low rate, the projections are hexagonal cells of ice separated by water. At higher rates of constitutional supercooling the projections resemble fern leaves. The pattern, which is called dendritic, consists of sharply pointed leaves branching into the water in directions governed by the orientations of the ice crystals along the freezing interface. Adjacent crystals that are in different orientations produce branches extending in different directions. It is possible that none of the projections will be perpendicular to the freezing interface.

Constitutional supercooling favors the development of projections. Imagine a small initial projection developing from some chance deposit of ice. The projection has several advantages over any plane area at the interface. It extends into the region where the water is supercooled most. The increase in surface area due to the growth of the projection improves the removal of impurities and heat. Hence water begins to freeze onto the projection, extending it deeper into the liquid.

The basic geometry of an ice crystal is a hexagonal plate. The plane of the plate is called the basal plane; its perpendicular axis is called the *c* axis. When the crystal first forms, it grows much faster in the basal plane than it does parallel to the *c* axis. If the axis is vertical, the crystal grows horizontally as a plate. If the axis is horizontal, the crystal grows into either a horizontal needle or a vertical plate. The vertical plate is unstable and may rotate to the horizontal. Alternatively, it may be locked into position if it expands enough to meet adjacent crystals before it rotates.

Evidence for a series of parallel vertical plates can sometimes be seen on the top surface of ice. As the plates form in the liquid surface, they buoy upward because ice is lighter than liquid water. When the surface freezes, the tops of the plates remain as narrow parallel ridges slightly elevated above the rest of the surface.

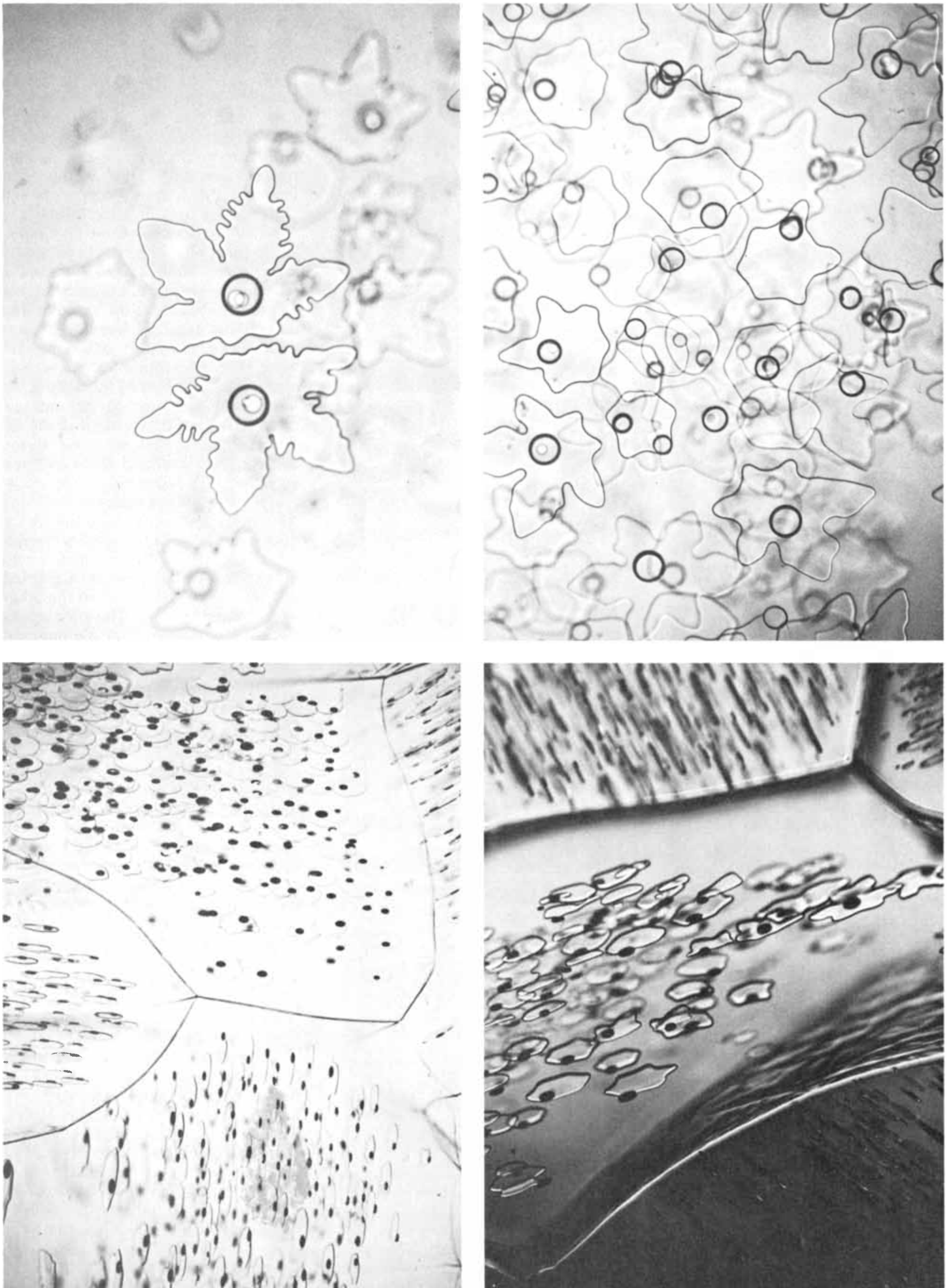
A sample of ice is likely to contain many crystals initiated by many nucleating agents. You can distinguish the crystals with the aid of polarized light in a procedure outlined by Robert A. Laudise and Robert L. Barns of the AT&T Bell Laboratories. Build a wood stand two feet high or less with open ends and an open top [*see lower illustration on page 117*]. Line the bottom with white cardboard to reflect light. Put a sheet of glass on the top. Shine bright light on the cardboard through each open end.

Position a polarizing filter on the glass. Place a thin section of ice on the filter. Place on the ice another filter with its polarizing axis perpendicular to that of the lower filter. (If the filters might be harmed by the ice, sandwich them between glass slides and tape the edges.) The different crystals within the ice show up in various colors or as gray areas. The colors change if you rotate the ice about the vertical axis of the filters. A crystal with a vertical *c* axis (along your line of sight) remains a gray of unchanging intensity. A crystal with a *c* axis well off the vertical changes in intensity as it is rotated; if it is thin enough, it varies in color.

You can detect a small deviation between the *c* axis and your line of sight by means of a clear glass marble placed between the ice and the top filter. The marble should touch the filter but not the ice. When you look through the top filter and the marble, you will see a dark cross. It is centered on the marble if the *c* axis is aligned with your line of sight; with increasing misalignment the cross becomes increasingly off-center.

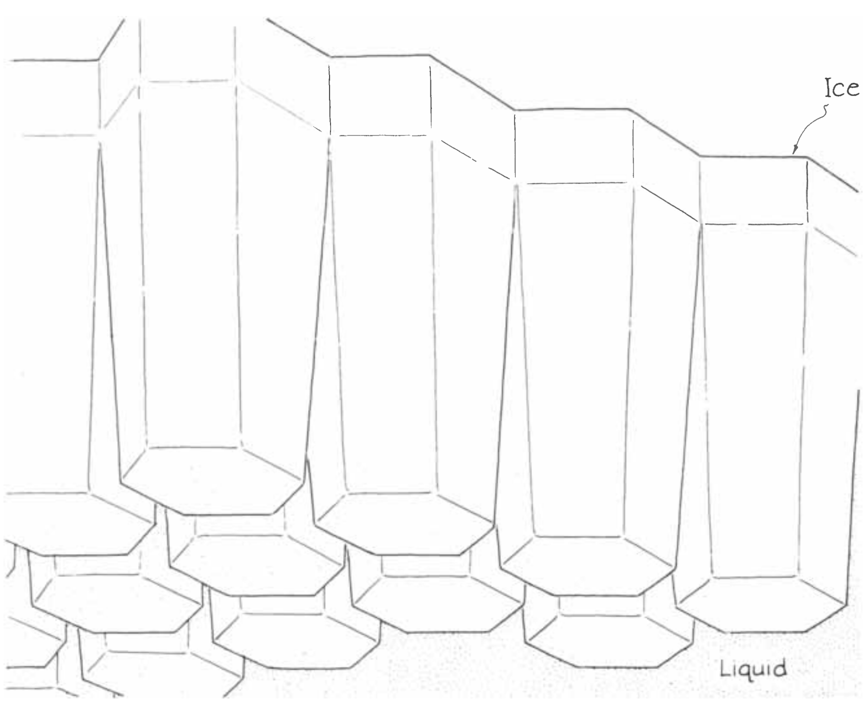
Thin sections of ice can be prepared by placing a chunk of ice on a flat metal plate at room temperature. The plate quickly conducts heat to the ice, which starts to melt. When the bottom of the ice is smooth, turn the piece over to smooth the other side. If you have a large chunk, try melting it in a heated frying pan.

Ice crystals are usually not perfect, because the atoms are laced with dislocations that spoil the crystal structure. The boundaries between adjacent crystals also include misfits among the atoms. To reveal the lines and bound-



Tyndall figures in ice

A gallery of possible Tyndall figures



The cellular advance of a freezing interface

aries on a surface of an ice cube I direct the thermal radiation from my desk lamp onto it. After a few minutes I remove the lamp and examine the ice surface. Atoms along the dislocation lines and crystal boundaries melt and evaporate more readily than atoms elsewhere on the surface, leaving narrow grooves. I avoid heating the ice so fast that the surface becomes rough and the lines are difficult to follow.

When I leave an ice cube on a poorly conducting surface, it gradually warms in the room air. When the cube reaches the melting point, zero degrees C., thin tubes develop through the cube along fracture lines or boundaries between crystals. My name for these tubes is wormholes. Water and air bubbles move along them from the interior of the cube to the surface; there the water spews out and the air gurgles through the meltwater on the surface. You can hear the water and air leaving the wormholes, particularly if the cube is on a plate.

I study this activity by putting an ice cube on an observation platform similar to the one made by Laudise and Barns. A layer of glass is supported on one side by a box and on the other side by a baking dish. The glass slopes toward the dish so that melting water collects in the dish. The cube does not slide down the glass because it rests against several mounds of glue attached to the glass. Light from my lamp reflects from white paper below the glass and travels through the ice.

Within a wormhole air bubbles glisten, displaying a fairly sharp contrast with the ice. Regions of water show poor contrast. The water and air are apparently pushed along the holes by pressure within the cloudy center of the cube. As the holes expand in diameter with the melting of ice along their sides, the train of air bubbles and water moves through them more slowly.

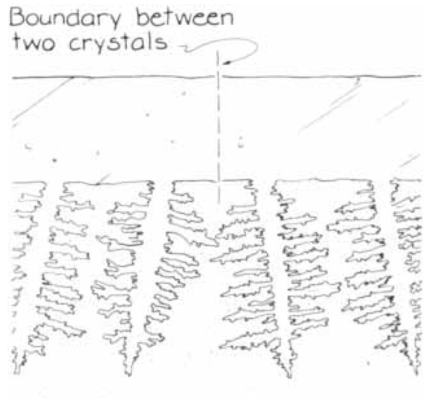
The ice cubes I make from tap water contain many small bubbles and tubes filled with air. As ice advances inward from the six surfaces of a cube, air is forced out of solution and collects in the bubbles. When a bubble forms on the bottom of the cube, it usually breaks free and floats up to the top layer of ice. There the freezing interface may advance around it, leaving it nearly spherical. Instead the bubble may pick up air as the interface advances; this is how a tube forms. The tube runs approximately parallel to the direction of advance and thus generally toward the center of the cube. Eventually the tube ends as the interface is able to pass around it, cutting off the supply of air from the remaining liquid.

The development of a tube depends on two competing factors: the rate at which the interface advances and the rate at which dissolved air diffuses to the site of a tube and joins the air in it. The advance of the interface depends on how quickly heat is conducted through the ice from the interface to the outer surface. As the interface moves deeper into the cube, the ice through which the heat must be conducted is thicker and the conduction rate is therefore lower. Moreover, as the amount of water in the cube decreases, the diffusion of air to the bubble inside a tube may increase. The changes allow most tubes to widen toward the interior of the cube. Small twists are evident in the narrowest sections at the outer ends, presumably because the interface does not move at a uniform rate.

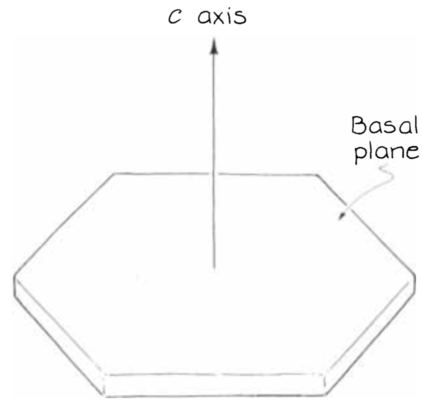
Many tubes vary periodically in diameter. This variation probably results from variations in the advance of the interface due to the cooling cycle of the freezer. The air in my freezer varies in temperature between −6 and −14 degrees C. When the air is in the warmer phase, the conduction of heat through the ice is slow and so is the advance of the interface. The diffusion of air to the tube increases the diameter of the bubble at its mouth, thereby increasing the diameter of the tube. When the air is coldest, the conduction of heat is fast and the interface advances faster than air can diffuse to the bubble at the mouth of the tube; the tube gets narrower. The periodicity of these variations in tube diameter is complicated by many other factors, including the increasing distance in the ice through which the heat must be conducted.

Tubes that form in the ice advancing from the bottom are vertical and narrow, varying little in diameter. This uniformity is due to two factors. The conduction of heat through the bottom layer of ice is usually insensitive to variations in the air temperature in the freezer. In addition the bubbles forming at the mouth of these tubes break away when they reach a certain size, leaving all the tubes with about the same diameter.

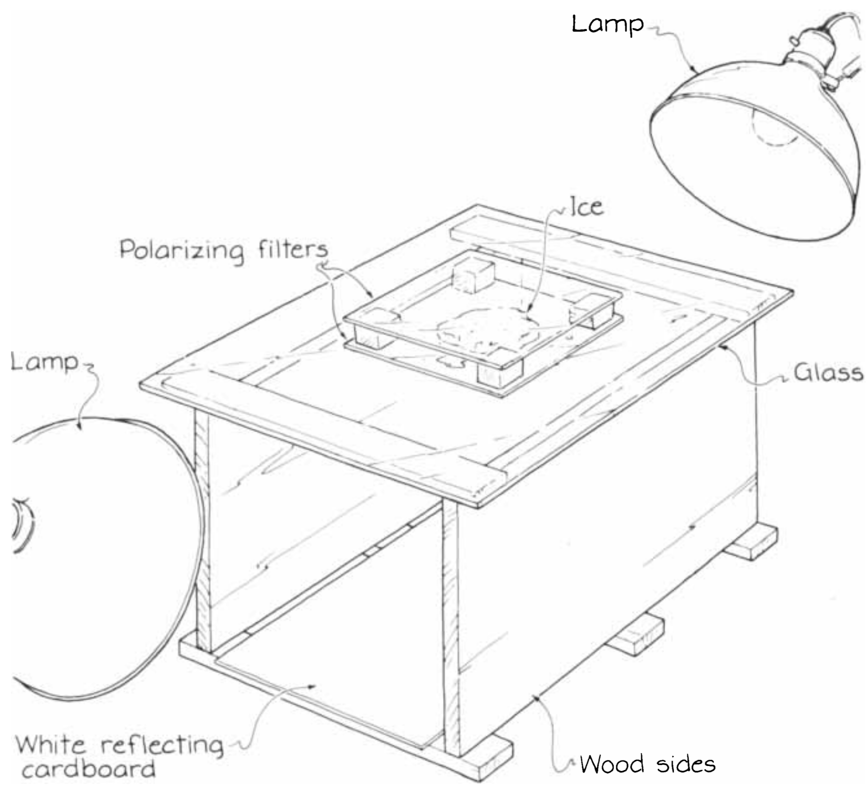
When the freezing process reaches the center of the cube, the formation of air bubbles and the entrapment of impurities in the ice become extensive. The center is a cloudy white rather than transparent because the bubbles and impurities scatter all the light reaching the center. This region is generally thinner than it is wide because freezing progresses faster from the top surface than it does through the sides or bottom of the ice cube.



Dendritic growth of two crystals



Geometry of an ice crystal

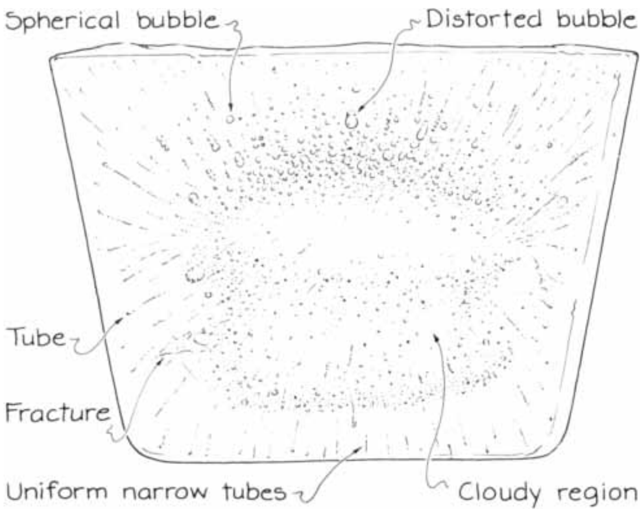


The observation platform made by Robert A. Laudise and Robert L. Barns

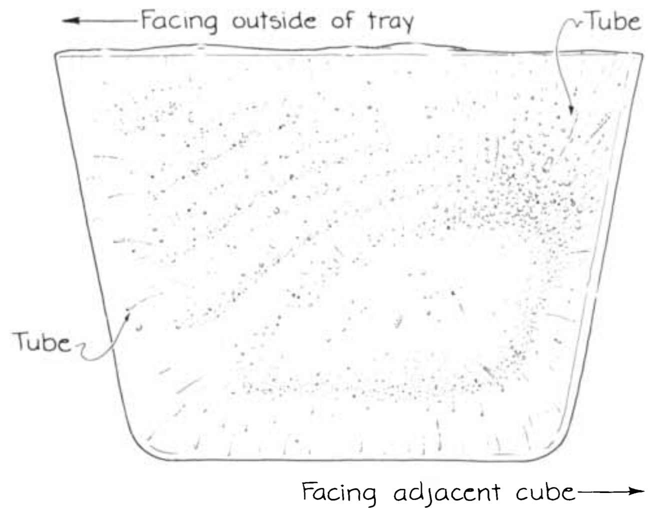
The pressure in the bubbles and the expansion of the water as it freezes often rupture the ice cube. To produce cubes almost free of bubbles and to decrease the possibility of rupture, make ice cubes from distilled water that you have boiled for at least five minutes to eliminate the dissolved air.

The ice cubes at each end of my ice tray have four or five layers of cloudy regions that tilt from the horizontal toward a corner of the tray. The layers are probably due to the cooling cycle of the freezer. When the air temperature in the freezer is low, the rapid conduction of heat through the ice enables the freezing interface to move quickly into the liquid from the top and the sides touching the tray. The rapid advance traps air and impurities in a mesh before they can escape into the liquid. The warmer phase of the cooling cycle slows the advance of the ice, allows the bubbles and impurities to escape and yields clearer ice. The orientation of air bubbles near the cloudy layers reveals the advance of the ice from several directions.

If ice is illuminated with bright sunlight, it can melt internally even if its surface remains frozen. Internal melting was first reported in 1858 by John Tyndall, a British physicist who is remembered for his studies of acous-



Pocketss of air in an ice cube



Layered cloudy regions in ice

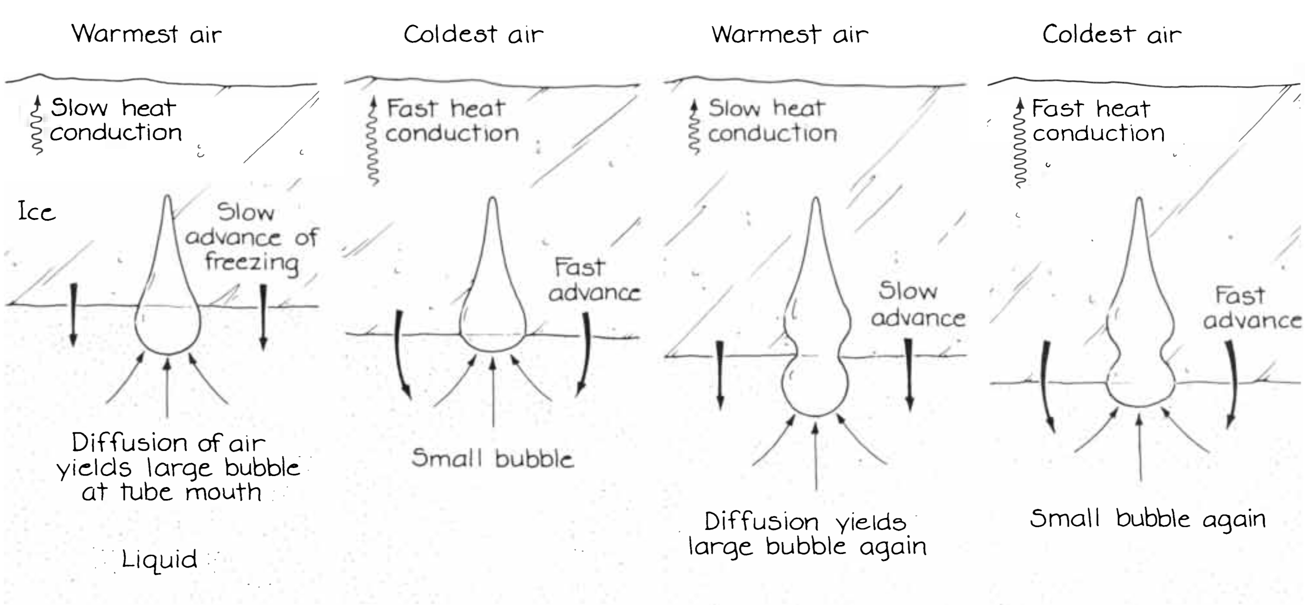
tics. He reasoned that thermal radiation in the sunlight melts certain spots within the ice. Since ice decreases in volume when it melts, a small bubble of water vapor forms in the liquid. Light scattering by the bubbles makes the ice sparkle with bright points.

Tyndall found that the thin spots of internal melting can take on a variety of designs. The commonest type is oval. More rarely there are spots displaying the hexagonal symmetry of a snowflake. Other spots resemble fern leaves and are described in terms of a leaf and the twig to which it is attached. Tyndall called all the spots liquid flowers, but today they are known as Tyndall figures.

Tyndall found the figures just below the surface of a frozen pond illuminated by strong sunlight. He could readily see the largest figures, which were several millimeters long; smaller ones required a magnifying glass. Most of the figures were parallel to the surface of the pond. Tyndall suggested they favor this orientation because the freezing interface of the water and ice is always approximately horizontal.

In 1964 Keiji Higuchi of Hokkaido University advanced a different view. He said the symmetric Tyndall figures lie in planes parallel to the basal plane of the ice crystal in which they form, whereas the fernlike Tyndall figures lie in planes perpendicular to the basal plane, that is, parallel to the *c* axis of the ice crystal. Hence the orientation of the figures with respect to the pond's surface depends on the orientation of the crystal when it freezes. Higuchi also reported that the leaves of the fernlike figures lie at an angle of approximately 45 degrees to the *c* axis. In some figures the leaves are perpendicular to the twig; in others the angle is smaller. What determines the angle is still not known.

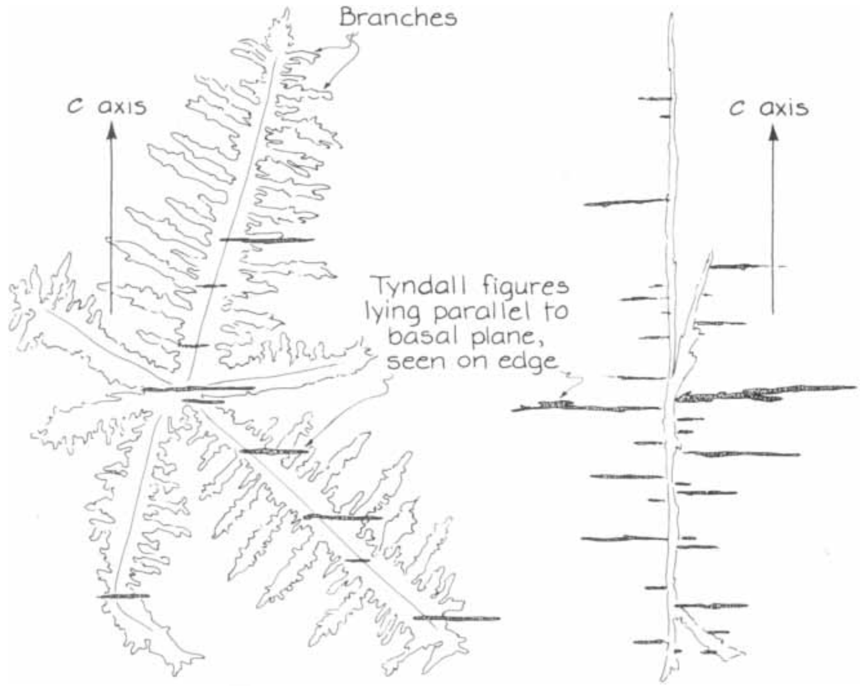
All Tyndall figures probably develop where there are defects or impurities in the ice. The nonuniformity of such places increases the absorption of thermal radiation, bringing about melting. Nevertheless, the mechanism that gives rise to the formation of a Tyndall figure is poorly understood. Charles A. and Nancy C. Knight of the National Center for Atmospheric Research in Boulder, Colo., speculated that the warming process first yields vapor. Thereafter the ice that borders the vapor cavity melts as the Tyndall figure evolves. The fernlike figures im-



How a tube in melting ice can vary in diameter ply that the melting process can be the reverse of the normal dendritic freezing process.

The Knights also discovered a type of Tyndall figure that lacks a vapor bubble. At sites where the ice is compressed on opposite sides of a plane along which the ice has fractured, warming causes melting that forms thin, curved lenses of water. Each lens lies approximately parallel to the basal plane of the crystal. As a lens develops, the melting interfaces move away from the fracture plane in opposite directions. The water molecules along the interfaces move somewhat toward the plane because of compression. Thus there is no net change in volume that would require a bubble to form.

The Knights grew ice crystals by freezing distilled water in a bucket. Rectangular plates measuring several centimeters on a side were then left for five minutes in air at normal room temperature so that they warmed to the melting point. Then each plate was hung in front of a 1,000-watt quartz-iodine lamp partially surrounded by a reflector. Thermal radiation from the lamp produced the Tyndall figures.



Fernlike Tyndall figures observed by Keiji Higuchi

When an ice plate was irradiated at the maximum intensity, clouds of tiny Tyndall figures appeared throughout the bulk of the ice, spreading at a rate of about two centimeters per second. Each element in the cloud was a hexagonal pocket of water along with a vapor bubble. The pockets were not connected. Then why did they appear together? The Knights suggest they are generated mechanically as a result of the strain that accompanies melting.

Shinji Mae of the University of Nagoya in Japan studied the production of more conventional Tyndall figures at the grain boundaries in ice he grew from distilled, degassed and deionized water. To examine a specimen he allowed it to warm in room air until he saw veins of water forming along the boundaries between adjacent crystals. The ice was then at the melting point. He next focused the light from a small lamp onto an area a few millimeters wide within the ice.

Among the set of Tyndall figures produced by the lamp was one that resembled a flower with 12 petals. How a twelvefold symmetry can develop from the basic sixfold symmetry of an ice crystal is not known. Mae also observed fernlike Tyndall figures. They were similar to the figures Higuchi observed to lie in planes perpendicular to the basal plane of the ice crystals.

I studied Tyndall figures in ice cubes frozen in my freezer after I had boiled the water to reduce the amount of air dissolved in it. Before examining a cube I let it warm to the melting point in room air. At the start its surface clouded with condensation drops, but after a few minutes the surface was coated with a layer of meltwater that made the surface transparent again. By then water and air were surging through wormholes. I put the ice on my glass observation platform and examined it with a jeweler's glass.

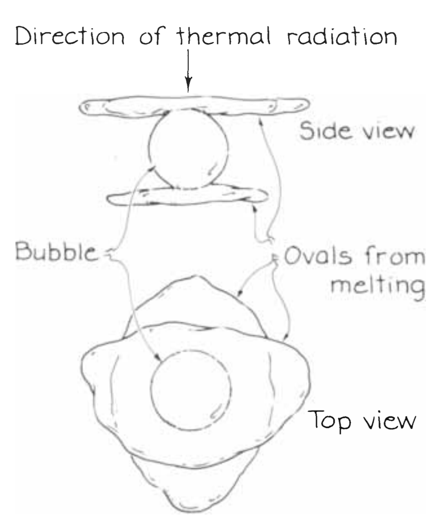
In my first experiments I heated ice cubes with my desk lamp for a few minutes. Most of the internal melting resulted in simple ovals, but I also spotted a few fernlike figures and one bright, magnificent hexagonal figure. Suspecting that the heat from the lamp was insufficient, I resorted to a 1,500-watt quartz heater and held a cube in front of it, draping a thick cloth over my hand as a heat shield. After a few minutes I inspected the ice for Tyndall figures.

I was rewarded immediately. The melting ice contained hexagonal stars and fernlike structures as well as ill-defined figures with sharp perimeters. I found the figures by first bringing a bubble into focus and then carefully adjusting the height of the lens until the associated melting region was in focus. Simultaneously I adjusted the lighting or the position of the ice cube on the glass to improve the contrast of the melting region.

The figures were quite thin. When I viewed them edge on from another angle, I found they were thin lenses of water. The interesting figures were often temporary, evolving into thin ovals. Perhaps I disrupted their growth when I moved the ice to the observation platform.

The heat source also melted ice around air bubbles already frozen in the ice. When I looked down through the top of the ice cube, I saw each bubble surrounded by an oval from melting. Often a bubble had two overlapping ovals. One oval was at the top of the bubble and the other at the bottom, as could be determined by looking at them from another angle.

More can be learned from observations of ice. I should like to hear from any reader who can discover a better explanation of the bubble and tube formations.



Melting around bubbles frozen in ice

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

MICROCOMPUTER DISPLAYS, GRAPHICS, AND ANIMATION. Bruce A. Artwick. Prentice-Hall, Inc., 1984.

40 GREAT FLIGHT SIMULATOR ADVENTURES. Charles Gulick. COMPUTE! Publications, 1985.

POTENTIAL NEW CROPS

CHEMICALS FROM BIOMASS: PETROCHEMICAL SUBSTITUTION OPTIONS. E. S. Lipinsky in *Science,* Vol. 212, No. 4502, pages 1465–1471; June 26, 1981.

NEW OILSEED CROPS ON THE HORIZON. L. H. Princen in *Economic Botany,* Vol. 37, No. 4, pages 478–492; October–December, 1983.

DEVELOPMENT OF NEW CROPS FOR INDUSTRIAL RAW MATERIALS. L. H. Princen and J. A. Rothfus in *Journal of the American Oil Chemists' Society,* Vol. 61, No. 2, pages 281–289; February, 1984.

VERY LARGE STRUCTURES IN THE UNIVERSE

THE LARGE-SCALE STRUCTURE OF THE UNIVERSE. P. J. E. Peebles. Princeton University Press, 1980.

SUPERCLUSTERS. J. H. Oort in *Annual Review of Astronomy and Astrophysics,* Vol. 21, pages 373–428; 1983.

A POSSIBLE 300 MEGAPARSEC FILAMENT OF CLUSTERS IN PERSEUS-PEGASUS. David J. Batuski and Jack O. Burns in *The Astrophysical Journal,* Vol. 299, No. 1, Part 1, pages 5–14; December 1, 1985.

INNER SPACE OUTER SPACE. Edward W. Kolb, Michael S. Turner, David Lindley, Keith Olive and David Seckel. The University of Chicago Press, 1986.

ANTI-IDIOTYPES AND IMMUNITY

IDIOTYPE NETWORKS IN HEPATITIS B VIRUS INFECTIONS. R. C. Kennedy in *Current Topics in Microbiology and Immunology,* Vol. 119, pages 1–13; 1985.

VACCINES UTILIZING INTERNAL IMAGE ANTI-IDIOTYPIC ANTIBODIES THAT MIMIC ANTIGENS OF INFECTIOUS ORGANISMS. Ronald C. Kennedy, Gordon R. Dreesman and Heinz Kohler in *BioTechniques,* Vol. 3, No. 5, pages 404–409; 1985.

ANTI-IDIOTYPIC ANTIBODIES AS IMMUNOGENS. *Immunological Reviews,* No. 90; April, 1986.

ANTI-IDIOTYPIC ANTIBODY VACCINE FOR TYPE B VIRAL HEPATITIS IN CHIMPANZEES. Ronald C. Kennedy, Jorg W. Eichberg, Robert E. Lanford and Gordon R. Dreesman in *Science,* Vol. 232, No. 4747, pages 220–223; April 11, 1986.

THE STRUCTURE OF MOUNTAIN RANGES

ACTIVE TECTONICS OF TIBET. Peter Molnar and Paul Tapponnier in *Journal of Geophysical Research,* Vol. 83, No. B11, pages 5361–5375; November 10, 1978.

CONSTRAINTS ON THE STRUCTURE OF THE HIMALAYA FROM AN ANALYSIS OF GRAVITY ANOMALIES AND A FLEXURAL MODEL OF THE LITHOSPHERE. H. Lyon-Caen and Peter Molnar in *Journal of Geophysical Research,* Vol. 88, No. B10, pages 8171–8191; October 10, 1983.

GRAVITY ANOMALIES AND FLEXURE OF THE LITHOSPHERE AT MOUNTAIN RANGES. G. D. Karner and A. B. Watts in *Journal of Geophysical Research,* Vol. 88, No. B12, pages 10449–10477; December 10, 1983.

SEISMICITY, FAULT PLANE SOLUTIONS, DEPTH OF FAULTING, AND ACTIVE TECTONICS OF THE ANDES OF PERU, ECUADOR, AND SOUTHERN COLOMBIA. Gerardo Suárez, Peter Molnar and B. C. Burchfiel in *Journal of Geophysical Research,* Vol. 88, No. B12, pages 10403–10428; December 10, 1983.

EXOTIC ATOMIC NUCLEI

POTENTIAL ENERGY SURFACE MODEL OF COLLECTIVE STATES. M. Seiwert, J. A. Maruhn and W. Greiner in *High-Angular Momentum Property of Nuclei,* edited by Noah R. Johnson. Harwood Academic Publishers, 1983.

NUCLEAR SHAPES AND SHAPE TRANSITIONS. T. Bengtsson, P. Möller, J. R. Nix and Jing-ye Zhang in *Physica Scripta,* Vol. 29, No. 5, pages 402–430; May, 1984.

MAGIC NUMBERS, REINFORCING SHELL GAPS AND COMPETING SHAPES IN NUCLEI FAR FROM STABILITY. J. H. Hamilton in *Progress in Particle and Nuclear Physics,* Vol. 15, pages 107–134; 1985.

ADVANCES IN STUDIES OF NUCLEI FAR FROM STABILITY. J. H. Hamilton, P. G. Hansen and E. F. Zganjar in *Reports on Progress in Physics,* Vol. 48, No. 5, pages 631–708; May, 1985.

SPACE, TIME AND TOUCH

THE CUTANEOUS "RABBIT": A PERCEPTUAL WORLD. Frank A. Geldard and Carl E. Sherrick in *Science,* Vol. 178, No. 4057, pages 178–179; October 13, 1972.

SENSORY SALTATION: METASTABILITY IN THE PERCEPTUAL WORLD. Frank A. Geldard. Lawrence Erlbaum Associates, 1975.

SALTATION IN SOMESTHESIS. Frank A. Geldard in *Psychological Bulletin,* Vol. 92, No. 1, pages 136–175; July, 1982.

FLIGHT SIMULATION

VISUAL ELEMENTS IN FLIGHT SIMULATION. John Lott Brown in *Aviation, Space and Environmental Medicine,* Vol. 47, No. 9, pages 913–924; September, 1976.

THE INTERPRETATION OF VISUAL MOTION. Shimon Ullman. The MIT Press, 1979.

COMPUTER IMAGE GENERATION FOR FLIGHT SIMULATORS. B. J. Schachter in *IEEE Computer Graphics and Applications,* Vol. 1, No. 4, pages 29–68; October, 1981.

THE ARTHROPOD CUTICLE

FINE STRUCTURE OF THE CUTICLE OF INSECTS AND OTHER ARTHROPODS. Barry K. Filshie in *Insect Ultrastructure,* Vol. 1, pages 281–312; 1982.

CUTICLE ULTRASTRUCTURE WITH RESPECT TO THE LIPID WATERPROOFING BARRIER. Neil F. Hadley in *The Journal of Experimental Zoology,* Vol. 222, No. 3, pages 239–248; September 1, 1982.

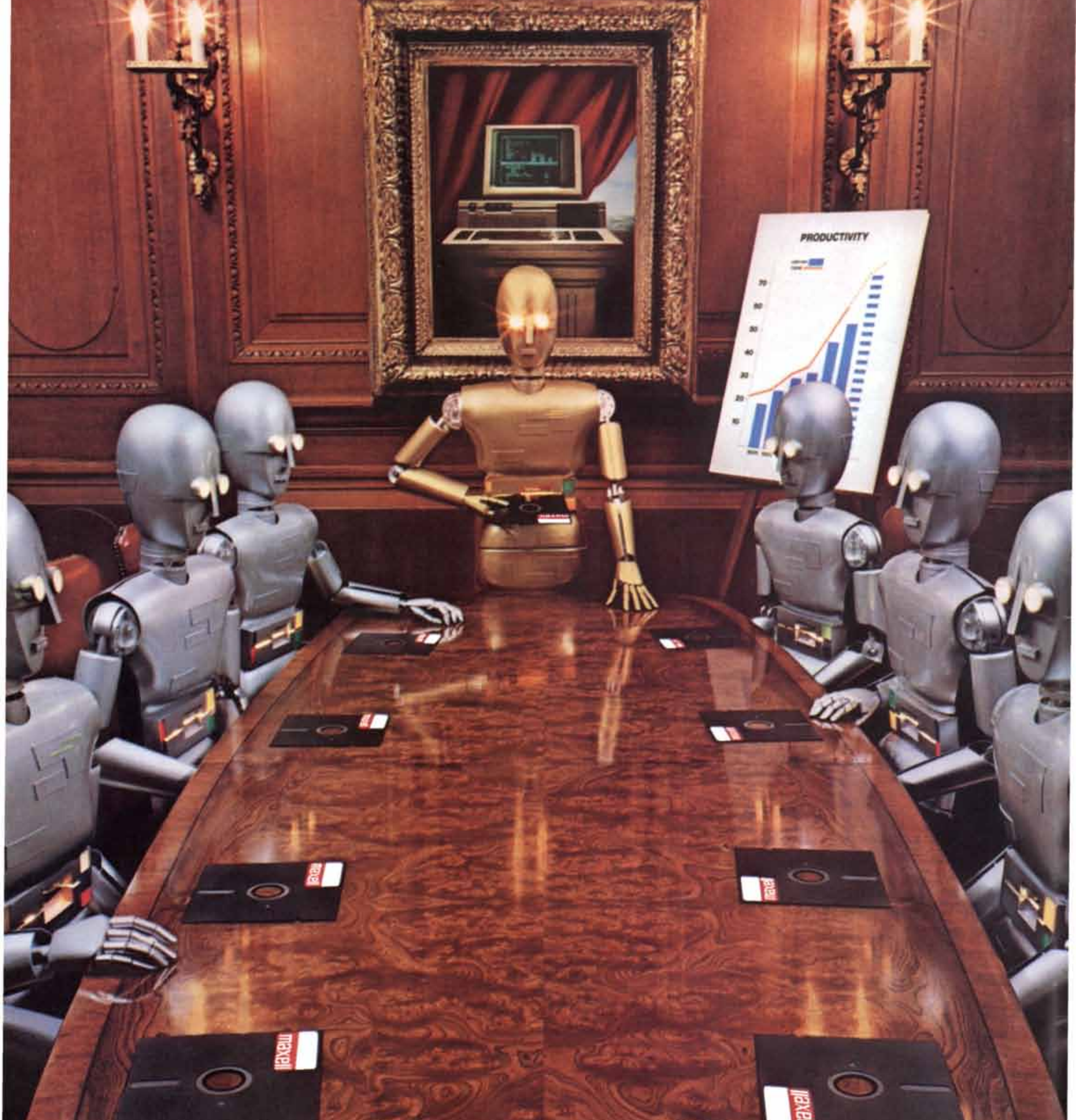
BIOLOGY OF THE INTEGUMENT, VOL. 1: INVERTEBRATES. Edited by J. Bereiter-Hahn, A. G. Matoltsy and K. Sylvia Richards. Springer-Verlag, 1984.

THE AMATEUR SCIENTIST

THE FREEZING OF SUPERCOOLED LIQUIDS. Charles A. Knight. D. Van Nostrand Company, Inc., 1967.

SUPERHEATED ICE: TRUE COMPRESSION FRACTURES AND FAST INTERNAL MELTING. Charles A. Knight and Nancy C. Knight in *Science,* Vol. 178, No. 4061, pages 613–614; November 10, 1972.

TYNDALL FIGURES AT GRAIN BOUNDARIES OF PURE ICE. Shinji Mae in *Nature,* Vol. 257, No. 5525, pages 382–383; October 2, 1975.



When computers get down to business, they move up to Maxell.

Maxell was first to provide you with a 5¼″ high density floppy disk. Just another example of how we keep you a step ahead.

maxell®
THE GOLD STANDARD

Maxell Corporation of America, 60 Oxford Drive, Moonachie, NJ 07074