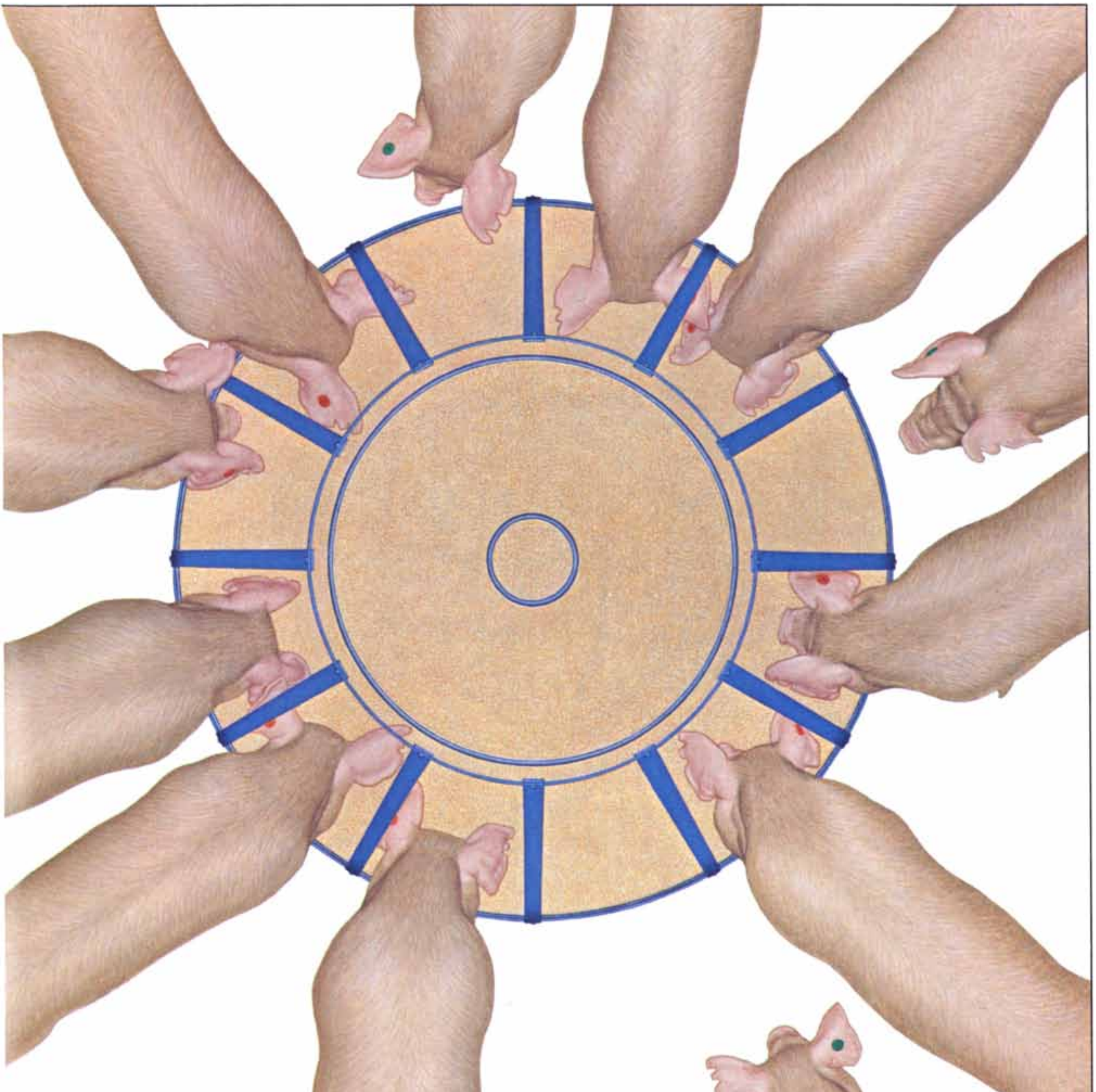
SCIENTIFIC AMERICAN



PORK PRODUCTION

$2.50

May 1983

How Mercedes-Benz harnesses the wind.

This 300 SD measures almost 17 ft. long, weighs 3,780 lbs., affords 108.8 cu. ft. of interior volume. But the wedge-shaped body 1 is so skillfully contoured, that its 0.36 Cd is bettered by no other 4-door sedan sold in the U.S. today.

Brake-generated heat buildup is dissipated into airstream 2 via heavily slotted alloy wheels.

Inlet at windshield base 3 diverts measured airflow to help cool sensitive electronic components under the hood.

Channels integrated with windshield posts 4 intercept rainwater dispersed by wipers, conduct it upward onto roof area 5 and away from side windows.

Outside mirror shell 6 shaped to deflect swirling water, slush from mirror surface.

Fine aerodynamic detailing of cabin exterior 7 keeps wind turbulence and noise at minimum.

Channels set into roof 8 help guide rainwater rearward and away from side window area.

Grooves trap streaming rain and duct it around rear window frame 9 to exit at rear of car.

Ventilation gills 10, placed in a controlled low-pressure area at rear quarter panel, efficiently draw stale air from cabin. Air in cabin is completely exchanged every 20 seconds.

"Washboard" ribbing of taillights 11 helps keep their lenses clean, even in foul-weather driving.

The difference between aerodynamics and Mercedes-Benz aerodynamics.

The Mercedes-Benz sedan sweeps through the stormy night along a rainswept highway. Inside the car, stereo music muffles the rhythmic slap-slap-slap of the wipers and everything is normal.

But outside the car, everything is not normal.

The rain is heavy, but the side windows remain oddly unstreaked.

The surface of the driver's outside mirror is almost free of swirling water.

Sheets of rain should be streaming down and over the rear window glass, but they aren't.

Enough grime is kicked up as the car whisks along to cake the taillights. But the taillights still glow ruby red.

Something strange and wonderful–and aerodynamic–is happening to this Mercedes-Benz.

0.36 not enough

It is happening by design.

There is, of course, no more aerodynamically pure four-door sedan sold in America than the Mercedes-Benz 300SD Sedan. Its 0.36 drag coefficient eclipses many exotic sports cars and coupes in sheer wind-cheating efficiency–a crucial feat, when every gain of 10 percent in aerodynamic efficiency can yield a gain of three to four percent in fuel economy.

But for Mercedes-Benz engineers, it is only step one in the exploitation of aerodynamics. They look beyond the numbers game of low drag coefficients, to the way a moving car and the moving air around it interact. They use the flow of onrushing air, as it passes over and around the car, to perform vital extra functions.

In brief, they put aerodynamics to work. Extending the science of cheating the wind to include the idea of *harnessing* the wind.

Study the diagram at left. As that Mercedes-Benz sedan forges ahead through the storm, aerodynamic science is being used to sluice rainwater away from its side and rear window glass.

The shell of its outside mirror has been aerodynamically shaped to help conduct swirling rain and slush away from the mirror surface itself.

The taillight lenses are deeply ribbed to resemble washboards: not a styling trick but an aerodynamic trick, meant to keep the recessed grooves clean by sheltering them from the vortices of airborne water (and mud and slush) in the car's slipstream.

Aerodynamically stable

Wet weather or dry, aerodynamics helps to exert a constant *downforce* on the front axle of a Mercedes-Benz as it moves along–countering a natural tendency of the front end to "lift" as speeds rise and more firmly bonding the tires to the road.

The car's body side configuration is aerodynamically critical. Correct shaping can help resist the sudden force of violent sidewinds, invisible blows that could jar a car off its true path.

A Mercedes-Benz utilizes aerodynamics to continuously cool specific heat-sensitive electronic components under its hood. To feed fresh air into the cabin while silently extracting stale air. To prevent irritating drafts and pressure imbalances in the occupants' ears, by maintaining a constant level of cabin air pressure at all speeds. To draw exhaust gases away into the airstream behind the car.

A Mercedes-Benz also employs aerodynamic science to help calm its driver's nerves. Even the outside mirror is shaped to generate minimal wind noise; if the driver experiences an almost eerie absence of wind buffeting and howling as he drives along, credit the presence of rigorous aerodynamic detailing.

Mercedes-Benz believes that cheating the wind in the cause of more fuel-efficient driving is good. But Mercedes-Benz believes that both cheating the wind in the cause of fuel efficiency *and* harnessing the wind in the cause of driving efficiency is better.

And that is the difference between aerodynamics and Mercedes-Benz aerodynamics.

Engineered like no other car in the world

Established 1845 **SCIENTIFIC AMERICAN** May 1983 Volume 248 Number 5

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

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

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

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

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

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

Name ____________________

New Address

Street ____________________

City ____________________

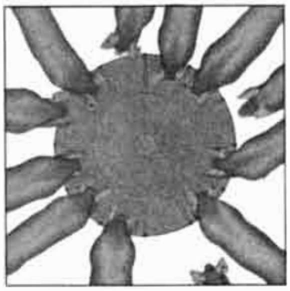
State and ZIP ____________________

Old Address

Street ____________________

City ____________________

State and ZIP ____________________



THE COVER

The painting on the cover shows a rotary feeder for pigs (see "Modern Pork Production," by Wilson G. Pond, page 96). The feeder is one of several in a "finishing operation," that is, a farm that buys pigs when they weigh from 30 to 50 pounds and feeds them until they reach the marketing weight of 220 pounds or more. The feed is a mixture of cornmeal and soybean meal, with supplements of protein, minerals and vitamins, all precisely determined to meet the nutritional needs of pigs of a certain size. On this farm pigs of the same size are fed together regardless of their age. The feed mixture is delivered to the rotary feeders by a conveyor system running overhead. The notches in the ears of the pigs tell where the animals were born; the colored ear tags indicate the auction at which each pig was bought. A strong trend in modern pig farming is toward "farrow to finish" operations, where the pig stays from birth to market.

THE ILLUSTRATIONS

Cover painting by Marvin Mattelson

Page	Source	Page	Source
16–19	Ilil Arbel	97–98	Tom Prentiss
53	United Press International	99	Ilil Arbel
54–57	Walken Graphics	100–102	Tom Prentiss
58–59	General Dynamics Corporation	103	Ilil Arbel
60–61	Los Alamos National Laboratory	105	R. F. Bonifield
63–73	Jerome Kuhl	106–108	Phillip A. Harrington
75	Marten Veenhuis, State University at Groningen	110–114	Ilil Arbel
76	Helen Shio, Rockefeller University (*top*); Eldon H. Newcomb, University of Wisconsin at Madison (*bottom*)	115	Phillip A. Harrington (*top*), Imperial Chemical Industries (*bottom*)
77	Eldon H. Newcomb, University of Wisconsin at Madison	117–125	Allen Beechel
78–83	Alan D. Iselin	126–128	Ilil Arbel
84	Helen Shio, Rockefeller University (*top*); Pierre Baudhuin, International Institute of Cellular and Molecular Pathology (*bottom*)	133	Michel Bur, University of Nancy
		134–135	Michael Holford
		136–138	Laszlo Kubinyi
		145–160	George V. Kelvin
		163	R. F. Bonifield
		164–168	Michael Goodman
		169	Jearl Walker

LETTERS

Sirs:

I read with interest the article titled "Orphans in the Marketplace" ["Science and the Citizen," SCIENTIFIC AMERICAN, January].

I should like to add another orphan disease and drug to your list—this time a success story in that the drug has reached the marketplace. The disease is erythropoietic protoporphyria (EPP), an inherited defect of porphyrin metabolism, in which protoporphyrin accumulates in the patient's blood and skin. The accumulation causes itching and burning sensations of the skin when the patient is exposed to sunlight, in some cases making even brief excursions outdoors during the day extremely distressful, if not impossible. I hypothesized, on the basis of work done by others as well as by me on the function of carotenoid pigments in bacteria, that perhaps the administration of such pigments to patients with light-sensitivity skin diseases might prevent the annoying symptoms associated with these diseases. After successful animal testing I organized a clinical trial to test my hypothesis: we administered the carotenoid pigment beta-carotene to patients suffering from EPP. We found that in the majority of patients the administration of beta-carotene did enable them to spend significantly longer times outdoors before developing the symptoms of light sensitivity than they could before they took beta-carotene. Our findings have been confirmed by other investigators. Hoffmann–La Roche, Inc., which had provided gratis the beta-carotene for our study, then went through the laborious process of having the pigment approved as a new drug by the Food and Drug Administration. Thanks to Hoffmann–La Roche's efforts and the cooperation of the FDA, approval was obtained, and beta-carotene, marketed as Solatene, is now available for physicians to prescribe for the relief of the symptoms of light sensitivity associated with EPP. Judging by the letters I have received from EPP patients who have benefited from the use of Solatene, I can definitely say that the efforts of both the pharmaceutical company and the Government agency are much appreciated.

MICHELINE M. MATHEWS-ROTH, M.D.

Harvard Medical School
Boston, Mass.

Sirs:

Walter C. Leedy, Jr., in his interesting article on the ceilings of certain Gothic cathedrals, expressed in one sentence two popular misconceptions about igloos ["Fan Vaulting," by Walter C. Leedy, Jr.; SCIENTIFIC AMERICAN, February]. He wrote: "The carved stones were then *fitted* together to form the vaulting much as blocks of *ice* are fitted to form an igloo" (italics mine).

Most masonry is constructed so that the boundary between one layer of blocks (called a course) and the next is horizontal. The diagrams of cathedrals in the article show concentric rings of blocks piled on each other in horizontal courses. Some independent method of support such as scaffolding would have been required until the keystone was installed. Scaffolding would not be available to the lone hunter building an igloo in a windy environment.

Virtually every artistic rendering of igloos shows the line between courses to be exactly horizontal. If an igloo were built that way, one horizontal course would be completed, then a new course would have to be started. The hunter by himself would have difficulty supporting the first block of a new course while lifting the second block. This problem would be insurmountable near the top of the igloo dome, where the blocks slope inward.

The Inuit of the Canadian Arctic start an igloo by laying a complete circle of blocks on the snow. These blocks are about three feet long, two feet high and four inches thick. Then about a fourth of the blocks are trimmed down to form a ramp. A notch is created between the sloping ramp and the vertical face of the adjoining untrimmed block. The builder sets the next block into that notch. Since the blocks are roughly rectangular, a new notch is automatically formed for the next block.

Viewed from above, an igloo looks like a spiral; viewed from the side, the lines between courses slope slightly upward with respect to the snow surface (which is seldom horizontal).

When a block has been fitted, it must be firmly tapped once. A surprisingly strong bond is then developed between the block and the rest of the structure. A block will remain in place even in strong winds. I think that much of the strength of an igloo comes from this bond. A completed igloo of blocks four inches thick is strong enough to stand on. One can crawl up the side of a partially completed one.

Ice has several drawbacks as a material for igloos, even though ice blocks would be stronger than snow blocks. Ice would be much harder to shape than snow. Also the essential bond between blocks would not form. Ice blocks could slip out easily unless water were poured into the cracks. Liquid water is scarce at 40 below.

RONALD A. MACNAUGHTON, P.ENG.

Kettleby, Ontario

How neutron scattering entirely new polymer

Exxon Chemical's David Lohse can observe the structure and behavior of polymer molecules.

From easy-care fabrics to the nose cones of space vehicles, the long chain molecular structures called polymers have become, both literally and figuratively, woven into the fabric of 20th century life. Now, Exxon's David Lohse is exploring the characteristics of polymers to extend the potential of these remarkable materials.

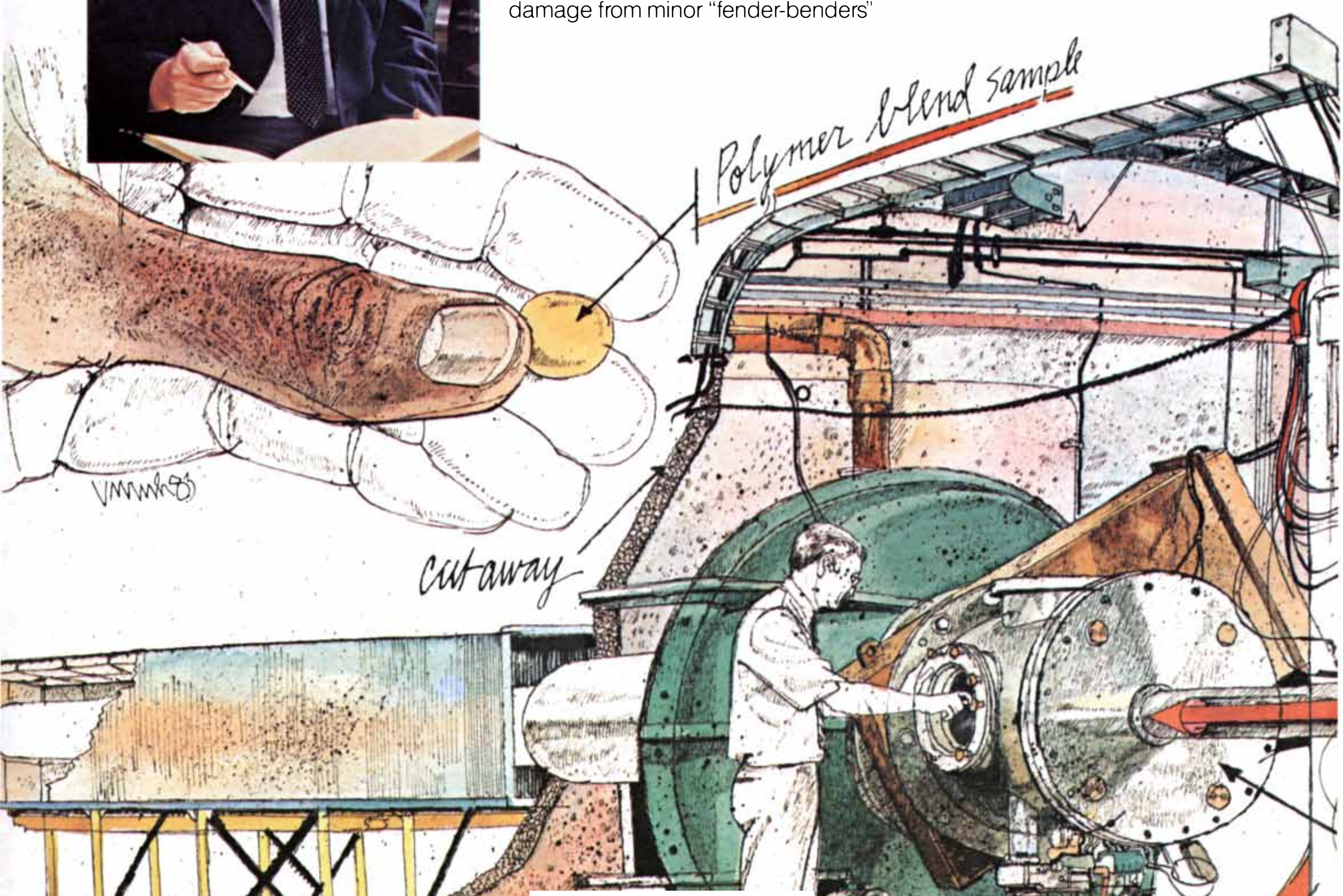
Polymer Blends

Combinations of polymer molecules can produce materials with better properties than those of the individual components. Exxon Chemical developed and is currently marketing, for example, a blend of plastic and synthetic rubber polymer material which has the formability of plastic and the elasticity of rubber. Used in automobile bumpers, this material helps reduce damage from minor "fender-benders" by absorbing impact and then returning to its original shape.

In order to develop more sophisticated blends, David Lohse and others in Exxon Chemical's Long Range Polymer Research Group are developing the use of Small Angle Neutron Scattering (SANS), a technique that subjects these molecules to detailed scrutiny.

Small Angle Neutron Scattering

Conventional electron microscopy techniques allow scientists to study the structure of a polymer blend at room temperature. But the technique, used by Exxon at the National Center for Small Angle Scattering Research (Oak Ridge National Laboratory) and at the National Bureau of Standards in Mary-



is helping Exxon create materials.

land, permits scientists to study polymer molecules at temperatures of 180°C or higher. This is crucial, since it requires high temperatures for the polymer to melt and be processed. In this melt state, the molecules are most mobile and the structure of the blends is formed.

Seeing Polymer Structure by SANS

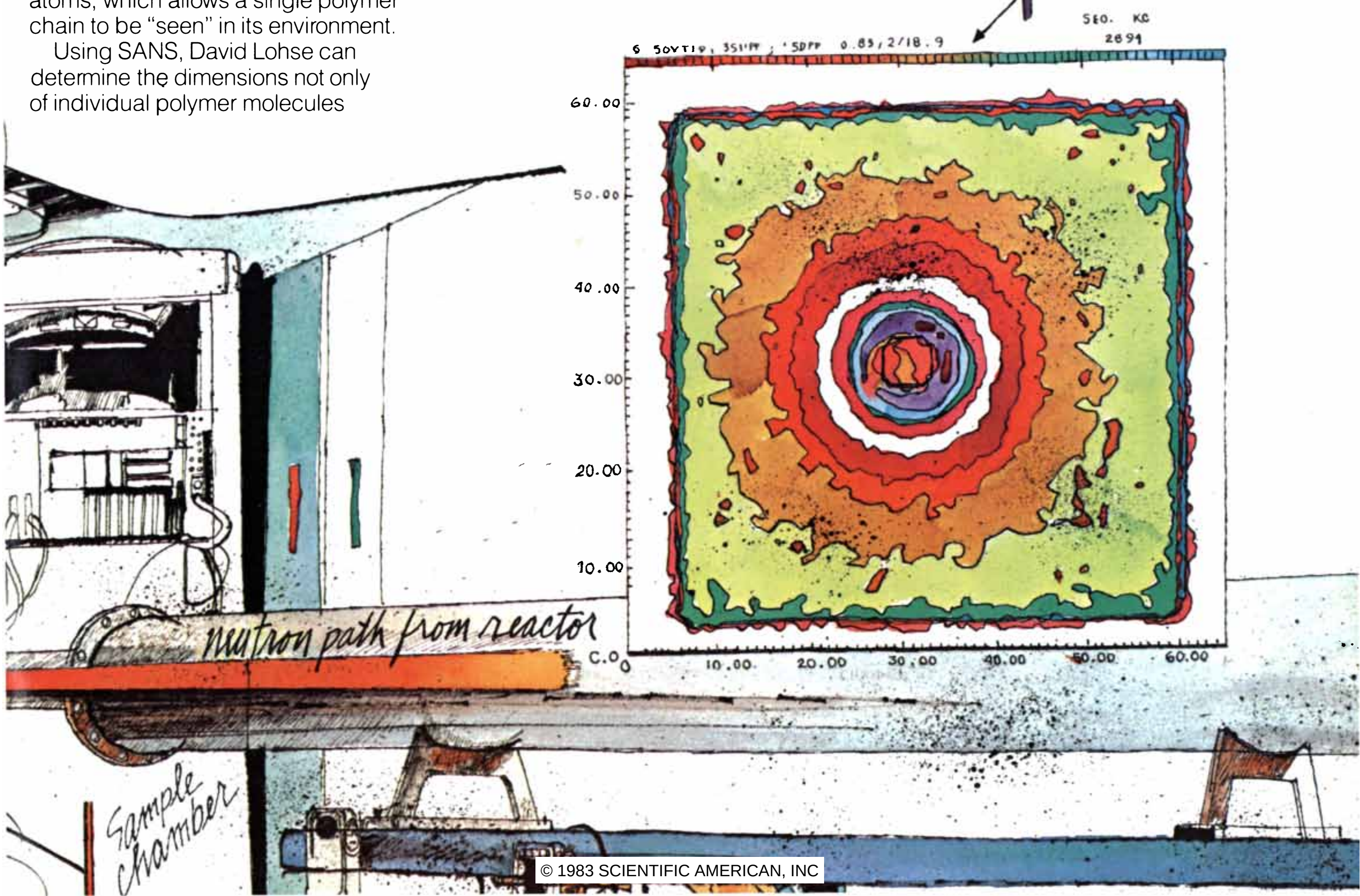
Small Angle Neutron Scattering involves the same basic physics as other types of scattering such as light or X ray. A well-collimated beam of neutrons is directed onto the sample. Some of the neutrons are scattered due to interactions with the atoms in the sample. The angle of the scatter is determined by the size of the molecular structures. Different structures can be labeled by substitutintg deuterium for hydrogen atoms, which allows a single polymer chain to be "seen" in its environment.

Using SANS, David Lohse can determine the dimensions not only of individual polymer molecules (10-100 angstroms) but also the dimensions of the phase domains of the blends (greater than 1000 angstroms). More importantly, while the blends are being heated, he can measure changes in the sizes of the individual molecules and the domains. It is these changes which determine the physical properties of the materials.

The information obtained using SANS will be used to develop new polymer blends that may find important commercial applications in the home, in industry, medicine, science and space.

Exxon Chemical Company

Long Range Polymer Research is just one of many programs of Exxon Chemical Company's Technology Department. A division of Exxon Corporation, Exxon Chemical is the world's ninth largest chemical company. Its research activities are diverse, with 2,000 scientists, researchers and technicians located in 10 countries working in advanced fields of catalysis and polymer science, and on ways to produce chemicals from new feedstocks or from more efficient processes.



50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MAY, 1933: "A unit volume of water has long been regarded as having a definite weight, but E. W. Washburn of the National Bureau of Standards has demonstrated very clearly that water can be as heavy as you make it. Two flasks of pure water, each containing exactly the same volume, were exhibited on a delicate balance, proving very conclusively that Dr. Washburn's water was heavier than any other pure water. This phenomenon is due to the fact that the heavy water was made from heavy hydrogen—or, in chemical parlance, from the isotope of hydrogen that has an atomic mass about twice that of ordinary hydrogen. In nature one out of about 30,000 atoms of hydrogen is double weight. When water is dissociated into hydrogen and oxygen by passing an electric current through it, the hydrogen given off contains more of the lighter hydrogen than normal. When this light hydrogen is burned, it yields a water of lower specific gravity than the water of nature. The water left in the electrolytic cell contains increasing amounts of heavy hydrogen as the electrolysis continues. In this way water that is one part in 1,000 heavier than ordinary water was obtained."

"On any cloudless, moonless spring evening, as the long western twilight fades, the sky remains brighter in one region than the rest. Rising from the point of sunset is a tapering wedge of faint light. This 'zodiacal light' has been known from ancient times. With good skies it can be seen to run on until its morning and evening halves meet or are joined by a faint 'zodiacal band.' Upon this band, and directly opposite the sun in the heavens, is a brighter spot called the Gegenschein, or zodiacal counterglow. Few if any measures of its brightness have ever been made. This gap in our observational data has recently been filled by C. T. Elvey of the Yerkes Observatory, working with a very delicate photoelectric photometer. The total brightness of the Gegenschein is surprisingly great. Elvey's measures on three nights give an average equal to the light of Arcturus. It is only because the light is so widely spread out and fades away so gradually that we find it difficult to see. The Gegenschein, the zodiacal band and the zodiacal light find a simple explanation in the assumption that a lens-shaped swarm of particles surrounds the sun and extends beyond the earth's orbit, but not very far."

"When Lindbergh flew to Japan, some newspaper comments had it that he was not flying the shortest route from the West Coast to Japan. He flew across Canada and via Alaska and the Aleutian Islands. Lindbergh was perfectly right. Our notions of distances are often based on maps in which distances are distorted. If we look down on a globe from above the North Pole, we see that the shortest way from the United States to China is over the Arctic region. The shortest way from San Francisco to England is not via New York and the North Atlantic but across Canada, Greenland, Iceland and the Faroe Islands. We must revise our notions of the shortest routes to other continents. When the United States is connected by regular airlines to China, Japan, Siberia and Europe, it may well be that these airlines will all go near the Pole. Mankind is never daunted by difficulties if a worthwhile objective is to be obtained, and plane designers, inventors and airplane flyers and operators may eventually transform the wild regions of the North into a busy sea of aerial activity."

SCIENTIFIC AMERICAN

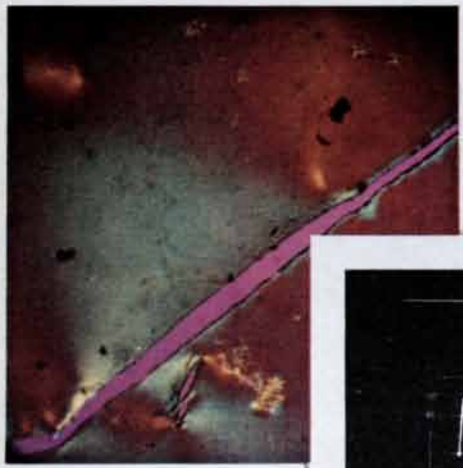
JUNE, 1883: "The liquefaction of oxygen gas and nitrogen, and the freezing of alcohol and sulphide of carbon, are the latest achievements of chemical science. This news comes to us from the laboratory of M. Wroblewski in Cracow, Poland. By the use of liquefied ethylene M. Wroblewski and K. Olszewski obtained the remarkably low temperature of $-136°$ C., equal to $-212.8°$ F. Oxygen gas subjected to about this temperature and compressed to about 25 atmospheres, or 375 pounds to the square inch, was readily liquefied in glass tubes and formed a colorless and transparent liquid, very mobile and resembling carbonic acid. Nitrogen was also liquefied, forming a colorless liquid. Alcohol was solidified at $-130.5°$ C., or $-202.9°$ F., forming a white body. Sulphide of carbon froze at about $-116°$ C., or $-176.8°$ F. The difficulty heretofore experienced in the liquefaction of oxygen and nitrogen has been to obtain a sufficiently low temperature in conjunction with compression. This obstacle now appears to be removed, and a variety of new and valuable observations concerning the nature of gaseous substances can be expected."

"The underground electric railway in London, whose construction has been authorized by act of Parliament, will commence near the north end of Northumberland Avenue opposite the Grand Hotel and pass under that avenue and the Victoria Embankment to a tunnel under the Thames, thence by College Street and Vine Street to Waterloo Station, where it will form a connection with the platforms of the London and Southwestern Railway. A separate approach to the Waterloo terminus of the line will, however, be built at York Road. The line will be double and will be worked by a stationary engine at Waterloo. The cars will run singly and will start as soon as they are filled, like omnibus cars. The journey will occupy about three and a half minutes. A contract for the supply of the electrical plant has been entered into with Messrs. Siemens Brothers, and a tender for the construction of the permanent way in 18 months' time has also been accepted."

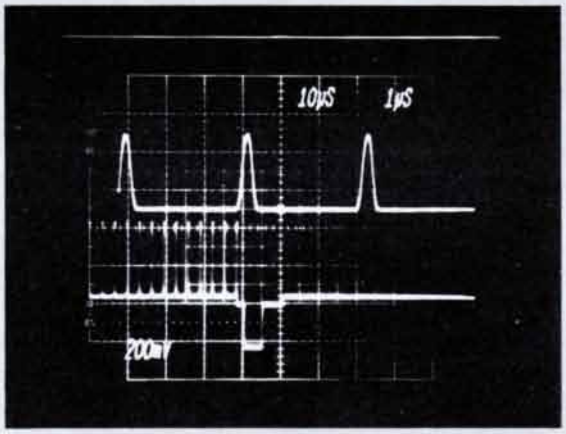
"There are several kinds of torpedoes. The one that is most used in the French navy is called the 'carried' torpedo (*torpille portée*), so named because the torpedo boat literally carries it right under the sides of the enemy's ship. It consists of a cartridge of about 20 kilogrammes of gun cotton, placed at the extremity of an iron rod, 12 meters in length, projecting in a downward direction from the fore part of the boat. The charge is fired by an electric spark by means of an apparatus placed in the lookout compartment. The water spout produced by the explosion sometimes completely covers the torpedo boat, and the latter would be sunk if it were not that all apertures were closed so as to make her a true buoy. The attack on an ironclad is a perilous operation, and one that requires men of coolness, courage and experience."

"The age in which we live is pre-eminently regardful of the rights of man. Constitutions are framed and statutes enacted for the protection of the weak against the possible encroachments of the strong. In an age and a country such as ours the very weakness of mental disease is its safeguard. A major right of the insane is the right to medical inquiry, in lieu of the ordinary trial by jury, into the question of their insanity, before committing them to asylum care and custody—such a thorough, unimpassioned medical inquiry as would certainly reach the true nature and needs of the malady, and such an inquiry is best secured by men competent from experience to investigate the nature of mental disease. An incidental right of the insane is to have proper instruction, in regard to insanity, provided for in the medical schools, and we make this demand for them: that henceforth no medical college shall be chartered that does not provide a chair of psychiatry. The true friends of the insane are in the medical profession, and its members should understand them."

Instant te



Experimental polymer on heating stage at 310°C, polarized light, (65X) on Time-Zero Type 778 film.



Oscilloscope trace of 10 nanosecond event captured on Ultra High Speed Type 612 film.

Landsat weather image on Polacolor ER Type 809 8×10″ film. Recorded by: GOES-West Satellite. Courtesy: International Imaging Systems.

With Polaroid Instant Photography, there's no limit to what you can explore. In seconds you can have color hard copies of computer-generated images that contain more information than you'll find on even the highest resolution raster-scan display. Or capture events not visible to the naked eye. Instantly. Easily. On-the-spot.

It's all possible with our 32 different Polaroid instant professional films. They range in speed up to ISO 20,000/82°. In sizes up to 8 x 10″. They have several different spectral sensitivities. And can be used in the most sophisticated laboratory instruments, as well as in

Instant Technology:

The images on the preceding pages show just a small segment of the range of explorations possible when Polaroid self-developing films are used. They also show only 5 of the 32 different Polaroid instant professional film types available. **1.** Color photomicrograph of polymer taken on Polaroid self-developing Time-Zero Type 778 film (ISO 150/23°) with the SX-70 Microscope Adapter, which features automatic exposure control. **2.** Instant 8×10″ color hard copy or Colorgraph overhead projection transparencies of computer graphics produced in just minutes on Polaroid VideoPrinter Model 8. **3.** Oscilloscope trace taken in just 30 seconds with hand-held Polaroid CU-5 Close Up Camera and recorded on Type 612 Ultra High Speed 3 1/4 × 4 1/4″ black and white film (ISO 20,000/82°). **4.** Polaroid 3 1/4 × 4 1/4″ color or black and white print of 35mm slide made at the touch of a button with the portable Polaprinter Slide Printer. **5.** Instant color 4×5″ or 3 1/8 × 3 1/8″ working size hard copies, and conventional or new Polaroid 35mm Autoprocess* transparencies of computer-generated images can be produced on-the-spot using Polaroid's VideoPrinter Model 4.

1.

2.

3.

4.

5.

Instant Response:

If you have any specific questions on the uses or performance of Polaroid films, or would like information on how to adapt your equipment to accommodate our films, call us, toll-free in the continental U.S.: 800-225-1618. (In Massachusetts, call collect: 617-547-5177.)

For more general information see your Polaroid Photographic Dealer or use the coupon.

POLAROID

Polaroid Corporation, Industrial Marketing Dept. 567, P.O. Box 5011, Clifton, NJ 07012

Please send me appropriate literature. I am interested in: ______

Name: ______

Company: ______

Address: ______

City: ______

State: ______ Zip: ______

Phone: ______

SA-5/83

THE AUTHORS

PAUL F. WALKER ("Smart Weapons in Naval Warfare") is director of the division of education and programs of Physicians for Social Responsibility. After being graduated from the College of the Holy Cross he joined the U.S. Army, where he was a Russian translator and intelligence specialist, among other duties. On leaving the Army he went to graduate school, receiving his M.A. in international relations at the Johns Hopkins University School of International Relations in 1973 and his Ph.D. in political science from the Massachusetts Institute of Technology in 1978. Before taking up his current job he was director of research on arms-control issues for the Union of Concerned Scientists. Walker is one of six coauthors of *Winding Down: The Price of Defense,* published by W. H. Freeman and Company.

GEORGE F. BERTSCH ("Vibrations of the Atomic Nucleus") is professor of physics at Michigan State University. He attended Swarthmore College as an undergraduate, getting his bachelor's degree in 1962. He went on to earn his Ph.D. in physics from Princeton University in 1965. After serving on the faculties of Princeton and the Massachusetts Institute of Technology he moved to Michigan State in 1970. He writes: "Physics has interested me since I was a teenager in New York. When I discovered in graduate school that experimental physics requires long hours of work late at night, I chose theoretical physics for a career. My main research goal is to understand the cooperative behavior of many-particle systems, with nuclei being the subject of most of my work."

CHRISTIAN DE DUVE ("Microbodies in the Living Cell") is a cell biologist who divides his working time among three widely separated jobs. He is founder and president of the International Institute of Cellular and Molecular Pathology in Brussels, Andrew W. Mellon Professor at Rockefeller University and professor and chairman of the department of physiological chemistry at the Faculty of Medicine of the Catholic University of Louvain. He is a citizen of Belgium who was born in England in 1917. He got his M.D. at Louvain in 1941 and remained there to earn a master's degree in 1946. In 1949, employing a method of cell fractionation developed by Albert Claude, de Duve and his colleagues at Louvain discovered the lysosome, an organelle of the living cell. For this work, and for other work on the structure and function of the cell, de Duve (with Claude and George E. Palade) received the Nobel prize for medicine or physiology in 1974. He moved to Rockefeller University in 1962 but kept his professorship at Louvain. He founded the Institute of Cellular and Molecular Pathology in 1975.

WILSON G. POND ("Modern Pork Production") is research leader in nutrition at the Roman L. Hruska Meat Animal Research Center of the U.S. Department of Agriculture in Clay Center, Neb. He got his B.S. at the University of Minnesota in 1952. His Ph.D. in animal nutrition, granted in 1957, is from Oklahoma State University. From 1957 until 1978 he was a member of the department of animal science at Cornell University. In 1978 he went on to the Hruska Center. Pond's main scientific interest is the subject of his article. He has served as a consultant to pork producers in Japan, Italy, Germany, Taiwan and Ecuador as well as the U.S.

J. D. BIRCHALL and ANTHONY KELLY ("New Inorganic Materials") are respectively senior research associate at Imperial Chemical Industries (ICI) and vice-chancellor of the University of Surrey. Birchall writes: "I joined the research staff of ICI in 1957 following a period of working in small companies and a brief period as a self-employed inventor. My first patent was granted at the age of 19. At ICI, I became interested in crystal nucleation and growth. This interest widened to include studies of the behavior of crystalline solids on heating." Birchall also is visiting professor at the University of Surrey. Kelly was graduated from the University of Reading with a B.Sc. in 1949 and received his Ph.D. in physics from the University of Cambridge in 1953. After academic appointments at several institutions, including Northwestern University and the University of Cambridge, he moved to the National Physical Laboratory in Britain in 1967 as superintendent of the Division of Inorganic and Metallic Structure. For part of his tenure at the National Physical Laboratory he was seconded to ICI as part of a task force for encouraging exchange among business, government and universities. Kelly took up his current job in 1975.

PERSI DIACONIS and BRADLEY EFRON ("Computer-Intensive Methods in Statistics") are both professors of statistics at Stanford University. Diaconis' first profession was that of illusionist. He left home at 14 to work with Dave Vernon, a well-known expert in sleight of hand. He spent 10 years traveling as Vernon's assistant before enrolling as an undergraduate at the City College of the City University of New York. He got his B.S. there in 1971 and three years later had earned his Ph.D. in statistics from Harvard University. His research interests include Bayesian statistics and the use of probability to analyze computer algorithms. Diaconis writes that his "background in magic can be seen in papers that point out errors in modern parapsychology and in work on questions such as 'How many times do you have to shuffle a deck of cards until it is close to random?'" Efron writes: "I received my bachelor's degree in mathematics from the California Institute of Technology, came to Stanford in 1960 as a graduate student, received my Ph.D. in statistics in 1965 and have been essentially unmovable ever since. My half-time appointment in the medical school as professor of biostatistics has provided the applied impetus for much of my theoretical work."

MICHEL BUR ("The Social Influence of the Motte-and-Bailey Castle") is professor of medieval history at the University of Nancy. He was trained as a historian and wrote his doctoral dissertation on the formation of the Comté de Champagne. He writes that "because of the paucity of documentary evidence" pertaining to the Middle Ages, much of his work is now archaeological. He has two major current projects. He is director of an inventory of the fortified buildings of the Middle Ages in France, which is being done under the auspices of the Centre National de la Recherche Scientifique. He is also analyzing the results of excavations at a 13th-century *château.* Bur adds: "I am aware that archaeological research rarely becomes part of the discourse of historians, and my preoccupation is to make a bridge between archaeology and history."

ARTHUR T. WINFREE ("Sudden Cardiac Death: A Problem in Topology") is professor of biological sciences at Purdue University. He studied engineering physics as an undergraduate at Cornell University, from which he was graduated in 1965. He continued his study of biology and physics at Princeton University; his Ph.D. in biology (1970) is from Princeton. From 1969 to 1972 he was assistant professor of theoretical biology at the University of Chicago. At the end of that time he went on to Purdue. He writes that three questions guide his current work: "How can we decipher the curious regularities of human sleep-wake timing? What are the organizing centers that form and maintain themselves in excitable media? What is the turbulence called fibrillation in heart muscle and how does it start?" Winfree is currently on sabbatical leave from Purdue as a John Simon Guggenheim Memorial Fellow.

METAMAGICAL THEMAS

Computer tournaments of the Prisoner's Dilemma suggest how cooperation evolves

by Douglas R. Hofstadter

Life is filled with paradoxes and dilemmas. Sometimes it even feels as if the essence of living is the sensing—indeed the savoring—of paradox. Although all paradoxes seem somehow related, some seem abstract and philosophical whereas others touch on life directly. A very lifelike paradox is the Prisoner's Dilemma, discovered in about 1950 by Merrill M. Flood and Melvin Dresher, and later formalized by Albert W. Tucker. I shall present it first as a metaphor and then as a formal problem.

The original formulation in terms of prisoners is a little less clear to the uninitiated, in my experience, than the following one. Assume you possess large quantities of some item (money, for example) and want to obtain some amount of another item (stamps, groceries, diamonds). You arrange a mutually agreeable trade with the only dealer of that item known to you. You are both satisfied with the amounts you will be giving and getting. For some reason, though, the exchange must take place in secret. Each of you agrees to leave a bag at a designated place in the woods and to pick up the other's bag at the other's designated place. Suppose it is clear to both of you that you will never meet or have further dealings with each other again.

Clearly there is something for each of you to fear, namely that the other one will leave an empty bag. Obviously if you both leave full bags, you will both be satisfied, but equally obviously it is even more satisfying to get something for nothing. You are therefore tempted to leave an empty bag. In fact, you can even reason it through with seeming rigor this way: "If the dealer brings a full bag, I'll be better off having left an empty bag, because I'll have got all I wanted and given away nothing. If the dealer brings an empty bag, I'll be better off having left an empty bag, because I'll not have been cheated. I'll have gained nothing but lost nothing either. Thus it seems that *no matter what the dealer chooses to do* I'm better off leaving an empty bag. And so I'll leave an empty bag."

The dealer, meanwhile, being in more or less the same boat (although at the other end of it), thinks analogous thoughts and comes to the parallel conclusion that it is best to leave an empty bag. And so both of you, with your impeccable (or seemingly impeccable) logic, leave empty bags and go away empty-handed. How sad, because if you had both just cooperated, you could each have gained something you wanted to have. Does logic prevent cooperation? That is the issue presented by the Prisoner's Dilemma.

In case you are wondering why it is called the Prisoner's Dilemma, here is the reason. Imagine that you and an accomplice (someone you have no feelings for one way or the other) committed a crime; now you have both been caught and jailed and are apprehensively awaiting trials. You are being held in separate cells with no way to communicate with each other. The prosecutor offers each of you the following deal (and tells you both that the identical deal is being offered to each of you—and that you both know it): "We have a lot of circumstantial evidence on you both. So if you both claim innocence, we'll convict you anyway and you'll both get two years in jail. But if you'll help us out by admitting your guilt and making it easier for us to convict your accomplice—pardon me, your alleged accomplice—why then we'll let you out free. And don't worry about revenge—your accomplice will be in for five years. How about it?" Warily you ask, "But what if we both say we're guilty?" "Ah, well, my friend—I'm afraid then you'll both get four-year sentences."

Now you are in a pickle! Clearly you do not want to maintain your innocence if your partner has confessed, because then you are going to jail for five long years. It is better you should both confess; then you will get only four. On the other hand, if your partner maintains innocence, then the best possible thing for you to do is to confess, since then you go scot-free. Hence at first it seems obvious what you should do: sing. What is obvious to you, however, is equally obvious to your partner, and so now it looks as if you both ought to sing, which means—Sing Sing for four years! At least that is what logic tells you to do. It's funny, since if both of you had been illogical and had maintained your innocence, you would both be in for only half as long. Ah, logic does it again.

Let us now go back to the original metaphor and slightly alter its conditions. Suppose both you and the dealer very much want to have a regular supply of what the other has to offer, and so before conducting your first exchange you agree to carry on a lifelong exchange once a month. You still expect never to meet face to face. In fact, neither of you has any idea how old the other is, so that you cannot be sure how long this lifelong agreement will last, but it seems safe to assume it will last for a few months anyway, and quite likely for years.

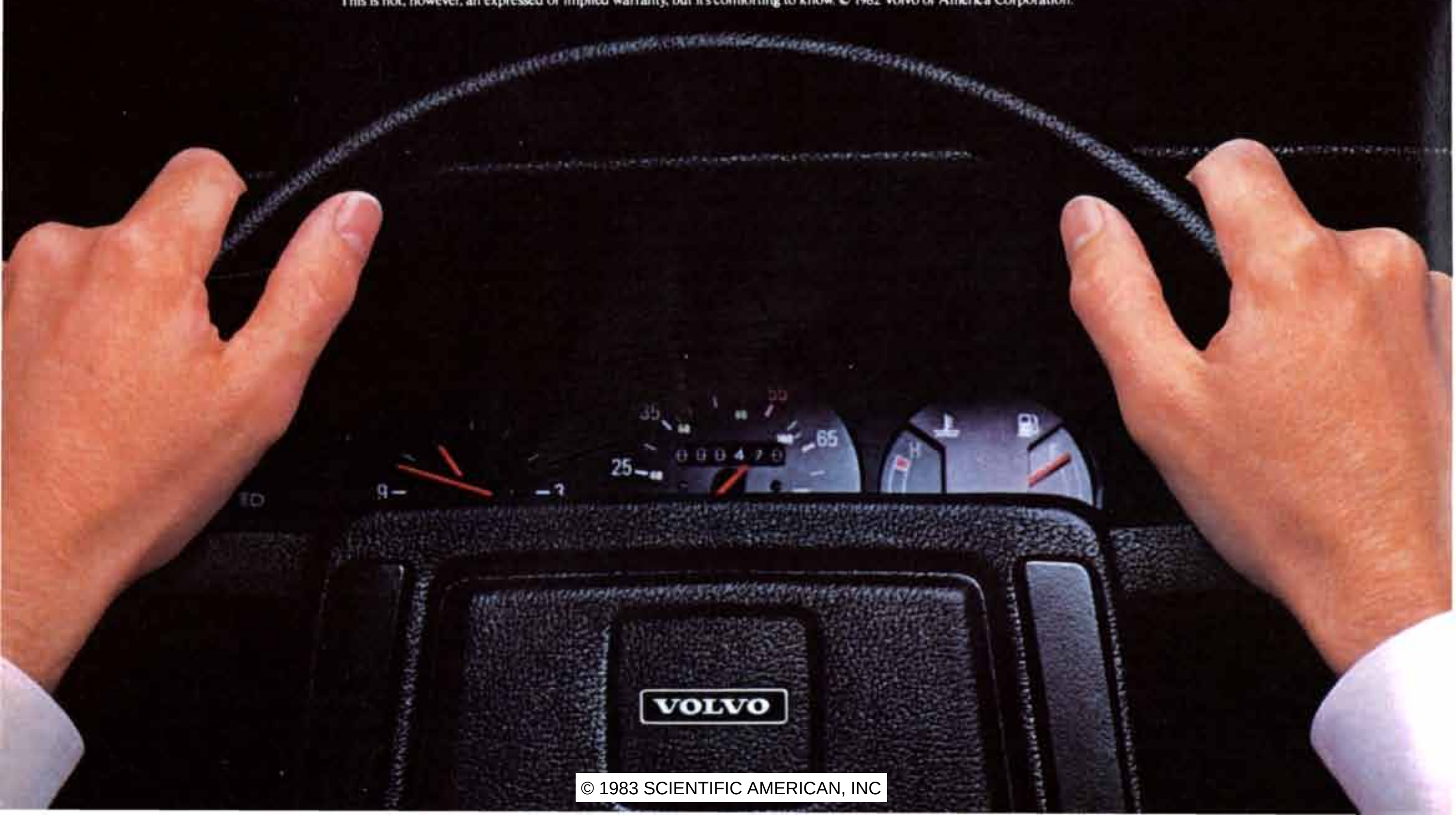
Now, what do you do on your first exchange? Bringing an empty bag seems fairly nasty as the opening of a relationship; hardly an effective way to build up trust. Suppose you bring a full bag and the dealer brings one too. All is bliss—for a month. Then you both must go back. Will your next bag be empty or full? Each month you have to decide whether to "defect" (bring an empty bag) or to "cooperate" (bring a full one). Suppose one month, unexpectedly, your dealer defects. Now what do you do? Will you decide that the dealer can never be trusted again and that from now on you will always bring an empty bag, in effect totally giving up on the entire project? Or will you pretend you didn't notice and continue being friendly? Or will you try to punish the dealer by some number of defections of your own? One? Two? A random number? An increasing number, depending on how many defec-

		Dealer: Cooperates	Dealer: Defects
You	Cooperate	(2,2)	(−1,4)
	Defect	(4,−1)	(0,0)

Where (x,y) means you get x points and the dealer gets y points.

A payoff matrix for the Prisoner's Dilemma


Lake Joseph, Ontario/Canada



VOLVO

tions you have experienced? Just how mad will you get?

This is the Iterated Prisoner's Dilemma. It is a very difficult problem. It can be, and it has been, rendered more quantitative and in that form studied with the methods of game theory and computer simulation. How does one quantify it? One builds a "payoff matrix" presenting point values for the various alternatives. A typical matrix is shown on page 16.

In this matrix mutual cooperation earns both parties 2 points (the subjective value of getting a full bag of what you want and giving up a full bag of what you have). Mutual defection earns you both zero points (the subjective value of gaining nothing and losing nothing, apart from making a vain trip out to the woods that month). Cooperating while the dealer defects stings: you get −1 point whereas the rat gets 4 points. Why so many? Because it is so pleasurable to get something for nothing. And, of course, if *you* happen to be a rat some month when the dealer has cooperated, then you get 4 points and he gets −1.

It is obvious that in a collective sense it would be best for both of you to always cooperate. We are assuming, however, that you have no regard whatsoever for the other person. There is no "collective good" you are both working for. You are both supreme egoists. Then what? The meaning of "egoist" can perhaps be made clear by the following. Suppose you and your dealer have developed a trusting relationship of mutual cooperation over the years, but one day you receive secret and reliable information that the dealer is quite sick and will soon die, probably within a month or two. The dealer has no reason to suspect you have heard. Are you not powerfully tempted to defect all of a sudden, in spite of all your years of cooperating? You are, after all, looking out for yourself and no one else in this cruel world. And since it seems this may well be the dealer's last month, why not profit as much as possible from your secret knowledge? Your defection may never be punished, and at the worst it will be punished by one last-gasp defection by the dying dealer.

The surer you are that this next turn is to be the very last one, the more you feel you must defect. Either of you would feel that way, of course, on learning the other one was nearing the end of the line. This is what is meant by "egoism." It means you have no feeling of friendliness or good will or compassion for the other player; you have no conscience; all you care about is amassing points, more and more and more of them.

What does the payoff matrix for the other metaphor, the one involving prisoners, look like? You will find it at the top of the page.

		Your accomplice	
		Stays mum	Sings
You	Stay mum	(−2, −2)	(−5, 0)
	Sing	(0, −5)	(−4, −4)

Where $(-x, -y)$ means you get x years in jail and your accomplice gets y years.

A payoff matrix for actual prisoners in the Prisoner's Dilemma

The equivalence of this matrix and the preceding one is clear if you add a constant—namely 4—to all terms in this one. Indeed, we could add any constant to either matrix and the dilemma would remain essentially unchanged. And so let us add 5 to this one to get rid of all negative "payoffs." We get the "canonical" Prisoner's Dilemma payoff matrix at the bottom of the page.

The number 3 is called the "reward for mutual cooperation," or R for short. The number 1 is called the "punishment," or P. The number 5 is T, the "temptation," and 0 is S, the "sucker's payoff." The conditions that make a matrix represent a Prisoner's Dilemma situation are these:

$$(1) \qquad T > R > P > S$$

$$(2) \qquad \frac{T + S}{2} < R$$

The first condition simply makes the argument go through for each of you that it is "better for me to defect no matter what my counterpart does." The second one simply guarantees that if the two of you somehow get locked into out-of-phase alternations (that is, "You cooperate, I defect" one month and "You defect, I cooperate" the next), you will not do better—in fact, you will do worse—than if you were cooperating each month.

Well, what would be your best strategy? It can be shown quite easily that there is no universal answer to the question. That is, there is no strategy better than all other strategies under all circumstances. Consider the case where the other player is playing ALL D: the strategy of defecting on every round. In that case the best you can possibly do is to defect every time, including the first time, yourself. On the other hand, suppose the other player is using the Massive Retaliatory Strike strategy, which means "I'll cooperate until you defect and thereafter I'll defect forever." In that case if you defect on the very first move, you will get one T and all P's thereafter until one of you dies. If you had waited to defect, you could have benefited from a relationship of mutual cooperation, amassing many R's beforehand. Clearly that bunch of R's will add up to more than the single T if the game goes on for more than a few moves. This means that against the ALL D strategy the best counterstrategy is ALL D, whereas against Massive Retaliatory Strike the best counterstrategy is "Always cooperate unless you learn that you or the other player is just about to die, in which case defect." This simple argument shows that how you should play depends on the other player.

The entire concept of the "quality" of a strategy takes on a decidedly more operational and empirical meaning if one imagines an ocean populated by dozens of little organisms swimming around and playing Prisoner's Dilemma over and over with each other. Suppose each time two such organisms encounter each other they recognize each other and remember how previous encounters have gone. This enables each one to decide what it wants to do this time. Now, if each organism is continually swimming around and bumping into the others, eventually each one will have met every other one numerous times and therefore all strategies will have been given the opportunity to interact with one another. By "interact" what is meant is certainly not that any one organism knocks any other one out of the ocean, as in an elimination tournament. The idea is simply that each organism gains zero or more points in each meeting, and if enough time is allowed to elapse, each organism will have met with every other one about the same number of times, and now the only question is: Which organism has amassed the most points?

It does not help organism X if it has "beaten" organism Y, in the sense that X gained more from interacting with Y than Y gained from interacting with X. That kind of "victory" is irrelevant here. What matters is not the number of "victories" rung up by any organism but the organism's *total point count*—a number that measures its overall viability in this particular "sea" of many strategies. It sounds nearly paradoxical, but an or-

		Player *B*	
		Cooperates	Defects
Player *A*	Cooperates	(3,3)	(0,5)
	Defects	(5,0)	(1,1)

Canonical Prisoner's Dilemma payoff matrix ganism could lose many, indeed all, of its individual skirmishes with other organisms and still come out the overall winner.

As the image suggests, the situation is relevant to questions in evolutionary biology. Can totally selfish and unconscious organisms living in a common environment come to evolve reliable cooperative strategies? Can cooperation emerge in a world of pure egoists? In a nutshell, can cooperation evolve out of noncooperation? If it can, it has revolutionary import for the theory of evolution, because many critics of evolutionary theory have maintained that this was one place where it was hopelessly snagged.

As it happens, it has now been definitively demonstrated that such cooperation can emerge, and it was done through a computer tournament conducted by Robert Axelrod of the Department of Political Science and the Institute for Public Policy Studies at the University of Michigan at Ann Arbor. More accurately, Axelrod first studied by means of a computer tournament the ways cooperation evolved, and when general trends emerged, he was able to perceive the underlying principles and to prove theorems that established the facts and conditions of the rise of cooperation out of nothing. Axelrod has written a remarkably thought-provoking book on his findings: *The Evolution of Cooperation,* soon to be published by Basic Books. Furthermore, he and William D. Hamilton, an evolutionary biologist, have worked out and published many of the implications of these discoveries for evolutionary theory. Their work has won much notice—including the 1981 Newcomb Cleveland Prize, which is awarded annually by the American Association for the Advancement of Science for "an outstanding paper published in *Science.*"

There are really three aspects of the question "Can cooperation emerge in a world of egoists?" The first aspect is: How can cooperation get started at all? The second is: Can cooperative strategies survive better than their noncooperative rivals? The third is: Which cooperative strategies will do best, and how will they come to predominate?

To make these issues vivid let me describe Axelrod's tournament and its astonishing results. In 1979 Axelrod sent out invitations to a number of professional game theorists, including people who had published studies of the Prisoner's Dilemma, telling them that he sought to pit many strategies against one another in a round-robin Prisoner's Dilemma tournament, with the overall goal for each strategy being to amass more points than any other strategy. He asked for strategies to be encoded as computer programs that could respond to the *C* (cooperation) or *D* (defection) of another player, taking into account the remembered history of previous meetings with that same player. A program should always reply with a *C* or a *D,* of course, but its choice need not be deterministic. That is, consultation of a generator of random numbers was allowed at any point in a strategy.

Fourteen entries were submitted to Axelrod, and he introduced into the field one more program, called RANDOM, which in effect flipped a coin (computationally simulated, to be sure) on each move: cooperating if heads came up, defecting if tails came up. The field was a rather diverse one, consisting of programs ranging from as few as four lines to as many as 77 (of the computer language Basic). Every program was made to engage every other program (and a clone of itself) 200 times. The tournament was actually run five times in a row, so that pseudoeffects caused by statistical fluctuations in the random-number generator would be smoothed out by averaging.

The program that won was submitted by an old hand at the Prisoner's Dilemma: Anatol Rapoport, a psychologist and philosopher at the University of Toronto. His was the shortest of all programs submitted, and it is called TIT FOR TAT. TIT FOR TAT has a very simple tactic, which one might call a tacfortic: Cooperate on move 1; thereafter do whatever the other player did on the previous move. That is all. It sounds outrageously simple. How in the world could such a program defeat the complex stratagems devised by other experts?

Well, Axelrod maintains that the game theorists in general did not go far enough in their analysis. They looked "only two levels deep," when to do better they should have looked *three* levels deep. What precisely does this mean? Axelrod takes a specific case to illustrate his point. Consider the entry called JOSS (submitted by Johann Joss, a mathematician in Zurich). JOSS's strategy is quite similar to TIT FOR TAT's in that it begins by cooperating, always responds to defection by defecting and *nearly* always responds to cooperation by cooperating. The hitch is that JOSS uses a random-number generator to help it decide when to pull a "surprise defection" on the other player. JOSS is set up so that it has a 10 percent probability of defecting right after the other player has cooperated.

In playing with TIT FOR TAT, JOSS will do fine until it tries to catch TIT FOR TAT off guard. When JOSS defects, TIT FOR TAT retaliates with a single defection, whereas JOSS "innocently" goes back to cooperating. Thus we have a *DC* pair. On the next move the *C* and the *D* will switch places, since each program echoes the other's latest move, and so it will go: *CD,* then *DC, CD, DC* and so on. There may ensue a long reverberation triggered by JOSS's *D,* but sooner or later JOSS will randomly throw in *another* unexpected *D* after a *C* from TIT FOR TAT. At this point there will be a *DD* pair, and that determines the rest of the match. Both strategies will now defect forever. The echo effect resulting from JOSS's first attempt at exploitation and TIT FOR TAT's simple punitive act leads ultimately to complete distrust and lack of cooperation.

This may seem to imply that both strategies are at fault and will suffer for it at the hands of others, but in fact the one that suffers from it most is JOSS, since JOSS tries the same trick on partner after partner, which in many cases leads to the same type of breakdown of trust, whereas TIT FOR TAT, never defecting first, will never be the initial cause of a breakdown of trust. Axelrod's technical term for a strategy that never defects before its opponent does is "nice." TIT FOR TAT is a nice strategy, JOSS is not. It should be noted that "nice" does not mean that a strategy *never* defects. TIT FOR TAT defects when it is provoked, but that is still considered being nice.

Axelrod summarizes the first tournament: "A major lesson of this tournament is the importance of minimizing echo effects in an environment of mutual power. A sophisticated analysis must go at least three levels deep. First is the direct effect of a choice. This is easy, since a defection always earns more than a cooperation. Second are the indirect effects, taking into account that the other side may or may not punish a defection. This much was certainly appreciated by many of the entrants. But third is the fact that in responding to the defections of the other side one may be repeating or even amplifying one's own previous exploitative choice. Thus a single defection may be successful when analyzed for its direct effects, and perhaps even when its secondary effects are taken into account. But the real costs may be in the tertiary effects when one's own isolated defections turn into unending mutual recriminations. Without their realizing it, many of these rules actually wound up punishing themselves. With the other player serving as a mechanism to delay the self-punishment by a few moves, this aspect of self-punishment was not perceived by the decision rules....

"The analysis of the tournament results indicates that there is a lot to be learned about coping in an environment of mutual power. Even expert strategists from political science, sociology, economics, psychology and mathematics made the systematic errors of being too competitive for their own good, not for-

12 flights a week are 12 too many, Qantas.

Once the koala roamed The Outback unrestrained. The eucalypti were lush. Life was serene.

Then came Qantas.

And on Qantas came people. Business people. Tourist people. Too many people.

Today Qantas brings more people to Australia than any other airline in the world. Because Qantas has more flights to Australia than any other airline.

Qantas flies every day from San Francisco. Five days a week from Los Angeles. Twice a week from Vancouver.

Which means they're coming into Sydney daily. And into Melbourne daily. And all too often into Brisbane, Perth, Townsville and Darwin.

We demand Qantas cut back this unreasonable schedule. Unite with us.

Koalas Against ... an Airline.)

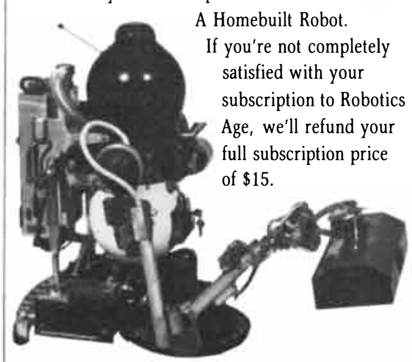
giving enough and too pessimistic about the responsiveness of the other side."

Axelrod not only analyzed the first tournament; he even performed a number of "subjunctive replays" of it, that is, replays with different sets of entries. He found, for instance, that the strategy called TIT FOR TWO TATS, which tolerates two defections before getting mad (but still only strikes back once), *would* have won, had it been in the lineup. Similarly, two other strategies he discovered, one called REVISED DOWNING and one called LOOK-AHEAD, would have come in first if they had been in the tournament.

The summarized lesson of the first tournament seems to have been that it is important to be *nice* ("Do not be the first to defect"), and *forgiving* ("Do not hold a grudge once you have vented your anger"). TIT FOR TAT has both qualities.

After this careful analysis Axelrod felt that significant lessons had been unearthed, and he was convinced that more sophisticated strategies could be concocted based on the new information. Therefore he decided to hold a larger computer tournament. For this tournament he not only invited all the participants in the first round but also advertised in computer-hobbyist magazines, hoping to attract people who were addicted to programming and who would be willing to devote a good deal of time to working out and perfecting their strategy. To each person who entered Axelrod sent a detailed analysis of the first tournament, along with a discussion of the "subjunctive replays" and the strategies that would have won. He described the strategic concepts of niceness and forgiveness that seemed to capture the lessons of the tournament, and also strategic pitfalls to avoid. Naturally each entrant realized that all the other entrants had received the same information, so that everyone knew that everyone knew that everyone knew that....

There was a large response to Axelrod's call for entries. Entries were received from six countries, from people of all ages and from eight different academic disciplines. Rapoport entered again, resubmitting TIT FOR TAT (and was the only one to submit it, even though it was explicitly stated that people could enter any program they wanted). A 10-year-old entered, as did one of the world's experts on game theory and evolution: John Maynard Smith, professor of biology at the University of Sussex, who submitted TIT FOR TWO TATS. Two people independently submitted REVISED DOWNING. All told 62 entries were received, and generally speaking they were of a considerably higher degree of sophistication than those in the first tournament. The shortest was again TIT FOR TAT; the long-



est was a program from New Zealand, consisting of 152 lines of FORTRAN. Once again RANDOM was added to the field, and with a flourish and a final carriage return the horses were off! Several hours of computer time later the results came in.

The outcome was nothing short of stunning: TIT FOR TAT, the simplest program submitted, won again. What is more, the two programs submitted that had won the subjunctive replays of the first tournament now turned up way down on the list: TIT FOR TWO TATS came in 24th, and REVISED DOWNING ended up buried in the bottom half of the field. This may seem horribly nonintuitive, but remember that a program's success depends entirely on the environment in which it is swimming. There is no single "best strategy" for all environments, so that winning in one tournament is no guarantee of success in another. TIT FOR TAT has the advantage of being able to "get along well" with a great variety of strategies, whereas other programs are more limited in their ability to evoke cooperation.

Axelrod puts it this way: "What seems to have happened is an interesting interaction between people who drew one lesson and people who drew another lesson from the first round. Lesson One was 'Be nice and forgiving.' Lesson Two was more exploitative: 'If others are going to be nice and forgiving, it pays to try to take advantage of them.' The people who drew Lesson One suffered in the second round from those who drew Lesson Two."

Hence the majority of participants in the second tournament really had not grasped the central lesson of the first tournament: the importance of being willing to initiate and reciprocate cooperation. Axelrod feels so strongly about this that he is reluctant to call two strategies playing against each other "opponents"; in his book he always uses neutral terms such as "strategies," "players" and so on. He even does not like saying they are playing *against* each other, preferring "with." I have tried to follow his usage in this article, with occasional departures. One striking fact about the second tournament is the success of "nice" rules: of the top 15 finishers only one (which came in eighth) was not nice. Amusingly, a sort of mirror image held: of the bottom 15 finishers only one was nice.

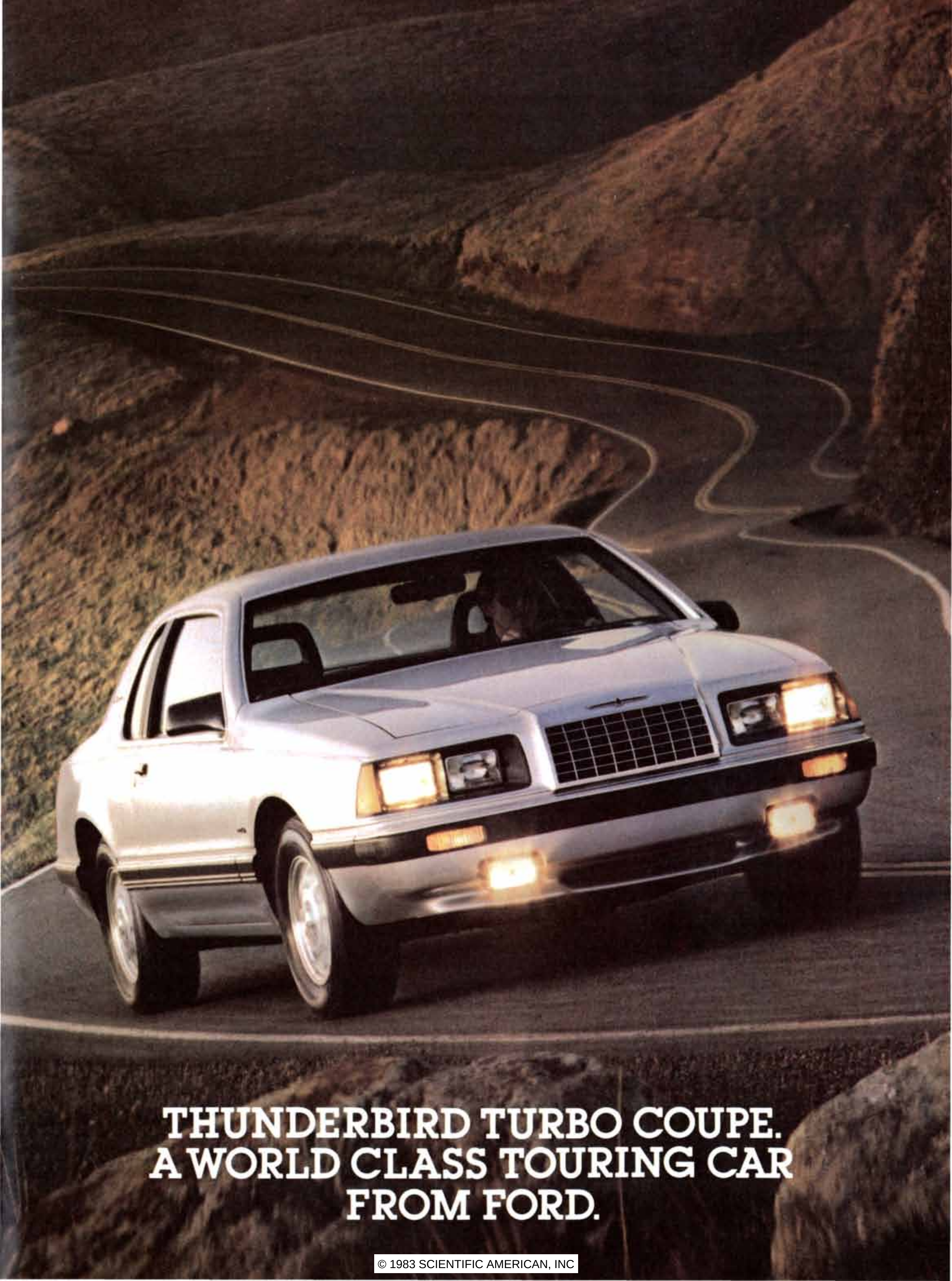
Several non-nice strategies featured rather tricky probes of the opponent (sorry!), sounding it out to see how much it "minded" being defected against. Although this kind of probing by a program might fool occasional opponents, more often than not it backfired, causing severe breakdowns of trust. Altogether it turned out to be very costly to try to use defections to "flush out" the other player's weak spots. It proved to be more profitable to have a policy of cooperation as often as possible together with a willingness to retaliate swiftly against any attempted undercutting. Note, however, that strategies featuring massive retaliation were less successful than TIT FOR TAT, with its gentler policy of restrained retaliation. Here forgiveness is the key, since it helps to restore the proverbial "atmosphere of mutual cooperation" (to use the phrase of diplomacy) after a skirmish.

The overall lesson of the first tournament was in essence "Be nice and forgiving." Apparently, however, many people just could not get themselves to believe it; they were convinced that with cleverer trickery and scheming they could win the day. It took the second tournament to prove them wrong. And from the second tournament a third key strategic concept emerged: that of *provocability*—the idea that one should "get mad" quickly at defectors and retaliate. Thus a more general lesson is "Be nice, provocable and forgiving."

Strategies that do well in a wide variety of environments are called by Axelrod "robust," and it seems the strategies with "good personality traits"—that is, nice, provocable and forgiving strategies—are sure to be robust. TIT FOR TAT is by no means the only possible strategy with these traits, but it is the canonical example of such a strategy and is astonishingly robust.

Perhaps the vividest demonstrations of TIT FOR TAT's robustness were provided by various subjunctive replays of the second tournament. Of all the types of replay that Axelrod tried (nearly all of which were won by TIT FOR TAT) undoubtedly the most significant and ingenious was the "ecological tournament." Such a tournament consists not only of a single subjunctive replay but also of an entire cascade of hypothetical replays, each one's environment determined by the results of the preceding replay. In particular, if you take a program's score in a tournament as a measure of its "fitness," and if you interpret "fitness" to mean "number of progeny in the next generation," and finally if you let "next generation" mean "next tournament," then what you get is that each tournament's results determine the environment of the next tournament—and successful programs become more numerous in the next tournament. This type of iterated tournament is called ecological because it simulates ecological adaptation (the shifting of a fixed set of species' populations according to their mutually defined and dynamically developing environment) as contrasted with the mutation-oriented aspects of evolution, where *new* species can come into existence.

As one carries an ecological tournament through generation after genera-


THUNDERBIRD TURBO COUPE.
A WORLD CLASS TOURING CAR
FROM FORD.

THUNDERBIRD TURBO COUPE

Turbo Coupe is a fuel-injected, turbocharged, five-speed Thunderbird with sophisticated running gear, performance compound tires, and seats that are more than a place to sit. A driver's car.

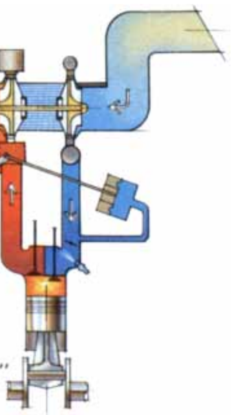
SOPHISTICATED POWER, JUDICIOUSLY APPLIED.

The engine.

An efficient, high-performance cross-flow head, overhead cam four-cylinder displacing 2.3 liters.

The engine employs five main bearings, forged aluminum pistons, high-temperature alloy valve materials, oil cooler and tuned intake manifold.

The turbocharger.



A "blow through" turbocharger is positioned upstream from the injectors and throttle plate.

Unlike other turbocharging arrangements, this "blow through" system pressurizes on demand. Which in turn gives the driver crisp, immediate throttle response. At full boost, around 4,600 engine RPM, the engine achieves 145 horsepower based on SAE Standard J-1349. More power than most of today's V-8s. A triumph of technology over brute force.

The electronic fuel injection system with EEC-IV.

The fuel injection system, fed by an electric fuel pump, is port type—with an injector ahead of each intake valve for precise metering.

The system is run by a computer called EEC-IV. A fourth-generation, state-of-the-art microprocessor-based engine control system capable of thousands of operations per second.

At one extreme, it will immediately set up air/fuel mixture and timing for a quick cold-start. At the other extreme, it will help the engine push to its safe limits. In between, it senses by the millisecond what you are asking the car to do, and balances the engine's air/fuel mixture and ignition timing for optimum power, response and efficiency.

The 5-speed transmission.

The free-revving powerplant is mated to a 5-speed gearbox. The ratios are carefully calibrated so there is a proper gear for any turn, any straightaway.

Accomplished drivers will appreciate the pedal design. The brake and accelerator pedals are positioned to facilitate true heel-and-toe downshifting.

BALANCE AND POISE ON THE ROAD.

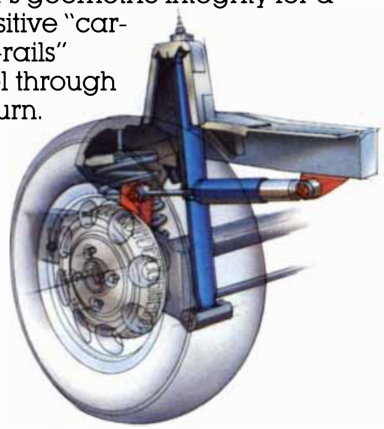
Aerodynamics/air flow management.

Turbo Coupe has one of the lowest drag coefficients of any four-passenger high-performance car in America. But more than that, Thunderbird uses the wind to help hold itself firmly on the road. It reduces lift on the hood for directional stability. Reduces lift on the rear for cornering agility. The smooth lines even lower wind noise inside the cabin. This is the difference between mere aerodynamics and what we call "air flow management."

Gas-filled struts and shocks for ride and handling.

The front struts and rear shocks on Turbo Coupe contain a low pressure charge of nitrogen. This suppresses foaming of the hydraulic fluid due to rapid bumps or exuberant driving. This in turn means the ride doesn't have to be overly stiff in normal driving to be stiff enough when you, or the road, demand.

Plus two additional rear shocks for pure handling.
Turbo Coupe has a second set of rear shocks mounted in a rather unique and interesting position . . . longitudinally between the rear axle and the body. These help maintain the car's geometric integrity for a positive "car-on-rails" feel through a turn.



Performance tires.
Turbo Coupe mounts Goodyear Eagle P205/70HR 14 performance tires on lightweight cast aluminum wheels. The tires feature an aggressive tread pattern, performance compound rubber, and low profile design for a wide footprint. Result: excellent traction and cornering power on wet or dry pavement.

MAN. MACHINE. UNITY.

Seats you sit in. Not on.
Turbo Coupe's front seats reward the demanding driver. (1) The under thigh support adjusts. (2) The side thigh support adjusts. (3) The seat back angle adjusts. (4) The lumbar support adjusts with (5), an infinitely adjustable pneumatic actuating pump. In addition, the (6) firm side supports help hold you in place. And unlike some other high-performance seats, you don't have to be an accomplished acrobat to get in and out. Considering that a Turbo Coupe is capable of .81G in a turn, seats that support you laterally as well as up and down are essential.

Command centers.
There are three. The driving center puts steering wheel, light controls and shift lever right where your hands want them to be. The environment center clusters radio and climate control a comfortable reach away. The console center nestles electric controls for windows, seats and mirrors in a pod next to your right thigh. The car feels as if it flows from your fingertips.

IN SUM.

An engine so efficient it's rated at 145 horsepower on one hand and 33 estimated highway, (22) EPA estimated miles per gallon on the other.* Superb handling with no surprises. And a feel that blends man with machine in a thoroughly harmonious way. A driver's car in every respect.

ONE FINAL BIT OF INFORMATION.

Fewer than 10,000 Turbo Coupes are scheduled to be built this first model year. Now you know what you should do. Quickly.

Call 1-800-772-2100 for more information.

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

Get it together—Buckle up.

HAVE YOU DRIVEN A FORD...LATELY?

Ford

tion the environment gradually changes. At the beginning poor programs and good programs alike are equally represented. As time passes, the poorer ones begin to drop out as the good ones flourish. The rank order of the good ones, however, may now change, because their "goodness" is no longer being measured against the same field of competitors.

Thus success breeds ever more success, but only provided that the success derives from interaction with other similarly successful programs. If, in contrast, some program's success is due mostly to its ability to milk "dumber" programs for all they are worth, then as those programs are gradually squeezed out, the exploiter's base of support will be eroded and the program will suffer a similar fate.

A concrete example of ecological extinction is provided by HARRINGTON, the only non-nice program among the top 15 finishers in the second tournament. In the first 200 or so generations of the ecological tournament, as TIT FOR TAT and other successful nice programs were gradually increasing their percentage of the population, HARRINGTON too was increasing its percentage. This was a direct result of HARRINGTON's exploitative strategy. By the 200th generation or so, however, things began to take a noticeable turn. Weaker programs were beginning to become extinct, which meant there were progressively fewer dupes for HARRINGTON to profit from. Soon the trend became apparent: HARRINGTON could not keep up with its nice rivals. By the 1,000th generation HARRINGTON was as extinct as the dodoes it had exploited.

Axelrod summarizes: "Doing well with rules that do not score well themselves is eventually a self-defeating process. Not being nice may look promising at first, but in the long run it can destroy the very environment it needs for its own success."

Needless to say, TIT FOR TAT fared spectacularly well in the ecological tournament, increasing its lead ever more. After 1,000 generations not only was TIT FOR TAT ahead but also its rate of growth was greater than that of any other program. This is an almost unbelievable success story, all the more so because of the absurd simplicity of the "hero." One amusing aspect of it is that TIT FOR TAT did not defeat a single one of its rivals in their encounters. This is not a quirk; it is in the nature of TIT FOR TAT. TIT FOR TAT cannot defeat anyone; the best it can achieve is a tie, and often it loses (although not by much).

Axelrod makes this point very clear: "TIT FOR TAT won the tournament, not by beating the other player but by eliciting behavior from the other player that allowed both to do well. TIT FOR TAT was so consistent at eliciting mutually rewarding outcomes that it attained a higher overall score than any other strategy in the tournament.

"So in a non-zero-sum world you do not have to do better than the other player to do well for yourself. This is especially true when you are interacting with many different players. Letting each of them do the same as or a little better than you is fine, as long as you tend to do well yourself. There is no point in being envious of the success of the other player, since in an iterated Prisoner's Dilemma of long duration the other's success is virtually a prerequisite of your doing well for yourself."

Axelrod gives examples from everyday life in which this principle holds. Here is one: "A firm that buys from a supplier can expect that a successful relationship will earn profit for the supplier as well as the buyer. There is no point in being envious of the supplier's profit. Any attempt to reduce it through an uncooperative practice, such as by not paying your bills on time, will only encourage the supplier to take retaliatory action. Retaliatory action could take many forms, often without being explicitly labeled as punishment. It could be less prompt deliveries, lower quality control, less forthcoming attitudes on volume discounts or less timely news of anticipated market conditions. The retaliation could make the envy quite expensive. Instead of worrying about the relative profits of the seller, the buyer should worry about whether another buying strategy would be better."

Like a business partner who never cheats anyone, TIT FOR TAT never beats anyone, yet both do very well for themselves.

One idea that is amazingly counterintuitive at first in the Prisoner's Dilemma is that if the other player is unresponsive, the best possible strategy for you to follow is ALL D. It might seem some form of random strategy might do better, but that is completely wrong. If I have laid out all my moves in advance, then playing TIT FOR TAT will not do you any good, nor will flipping a coin. You should simply defect on every move. It does not matter what pattern I have chosen. Only if I can be influenced by your play will it do you any good to cooperate.

Fortunately in an environment where there are programs that cooperate (and whose cooperation is based on reciprocity) being unresponsive is a very poor strategy, which in turn means ALL D is a very poor strategy. The single unresponsive competitor in the second tournament was RANDOM, and it finished next to last. The last-place finisher's strategy was responsive, but its behavior was so inscrutable that it *looked* unresponsive. And in a more recent computer tournament conducted by Marek Lugowski and me in the Computer Science Department at Indiana University three ALL D's came in at the very bottom (out of 53), with a couple of RANDOM's giving them a tough fight for the honor.

One way to explain TIT FOR TAT's success is simply to say it elicits cooperation by means of friendly persuasion. Axelrod spells this out as follows: "Part of [TIT FOR TAT's] success might be that other rules anticipate its presence and are designed to do well with it. Doing well with TIT FOR TAT requires cooperating with it, and this in turn helps TIT FOR TAT. Even rules that were designed to see what they could get away with quickly apologize to TIT FOR TAT. Any rule that tries to take advantage of TIT FOR TAT will simply hurt itself. TIT FOR TAT benefits from its own nonexploitability because three conditions are satisfied:

"1. The possibility of encountering TIT FOR TAT is salient;

"2. Once encountered, TIT FOR TAT is easy to recognize; and

"3. Once recognized, TIT FOR TAT's nonexploitability is easy to appreciate."

This brings out a fourth "personality trait" (in addition to niceness, provocability and forgiveness) that may play an important role in success: recognizability, or straightforwardness. Axelrod chooses to call this trait clarity, and he argues for it with clarity: "Too much complexity can appear to be total chaos. If you are using a strategy that appears random, then you also appear unresponsive to the other player. If you are unresponsive, then the other player has no incentive to cooperate with you. So being so complex as to be incomprehensible is very dangerous." How rich these comments are in their implications about social and political behavior!

In a letter to Axelrod, Rapoport cautioned against overstating the advantages of TIT FOR TAT; in particular he believes the strategy is on occasion too harshly retaliatory. It can also be persuasively argued that on other occasions TIT FOR TAT is too lenient. Certainly there is no evidence that TIT FOR TAT is the ultimate strategy. Indeed, as has been emphasized repeatedly, the very concept of "best" is incoherent, since all depends on the environment. In the tournament at Indiana University several strategies like TIT FOR TAT did better than pure TIT FOR TAT. They all shared, however, the three critical "character traits" whose desirability had been clearly revealed by Axelrod's earlier analysis of the important properties of TIT FOR TAT. They were simply a little better at detecting nonresponsiveness, and when they were convinced the other player was unresponsive, they switched over to an ALL D mode.

In his book Axelrod takes pains to spell out the answers to three fundamental questions about the evolution of cooperation in a world of raw egoism. The first concerns *initial viability:* How can cooperation get started in a world of unconditional defection—a "primordial sea" swarming with unresponsive ALL D organisms? The answer (whose proof I omit here) is that an invasion by small clusters of cooperating organisms, even if they form a tiny minority, is enough to give cooperation a toehold. One cooperator alone will die, but small clusters of cooperators can arrive (by means of mutation or migration, say) and propagate even in a hostile environment, provided they are defensive like TIT FOR TAT.

The second fundamental question concerns *robustness:* What type of strategy does well in unpredictable and shifting environments? We have already seen that the answer to this question is any strategy possessing the four fundamental "personality traits" of niceness, provocability, forgiveness and clarity. This means such strategies, once established, will tend to flourish, particularly in an ecologically evolving world.

The final question concerns *stability:* Can cooperation protect itself from invasion? Axelrod proved that it can indeed. In fact, there is a gratifying asymmetry to his findings: although a world of "meanies" (ALL D strategies) can be penetrated by cooperators in clusters, a world of cooperators *cannot* be penetrated by meanies, even if they arrive in clusters of any size. Once cooperation has established itself it is permanent. As Axelrod puts it: "The gear wheels of social evolution have a ratchet."

The term "social" does not mean these results necessarily apply only to higher animals that can think. Clearly four-line computer programs do not think, and yet it is in a world of just such "organisms" that cooperation has been shown to evolve. The only "cognitive abilities" needed by TIT FOR TAT are recognition of previous partners and memory of what happened the last time with a particular partner. Even bacteria can do this, by interacting with only one other organism (so that recognition is automatic) and by responding only to the most recent actions of their "partner" (so that memory requirements are minimal). The point is that the entities involved can be on the scale of bacteria, small animals, large animals or nations. There is no need for "reflective rationality"; indeed, TIT FOR TAT could be called "reflexive" (in the sense of being as simple as a knee-jerk reflex) rather than "reflective."

For people who think that moral behavior toward others can emerge only when there is some horrendous and totally external threat (of fire and brimstone, say) or some soothing promise of heavenly reward (such as eternal salvation) the results of this research must give pause for thought. Axelrod captures the entire idea in one sentence: "Mutual cooperation can emerge in a world of egoists without central control, by starting with a cluster of individuals who rely on reciprocity."

There are so many situations in the world today where these ideas seem of extreme relevance—indeed urgency—that it is tempting to draw all kinds of morals. In the later chapters of his book Axelrod offers advice about how to promote cooperation in human affairs, and at the end the political scientist in him cautiously ventures some broad conclusions about global issues, which are a fitting way for me to conclude:

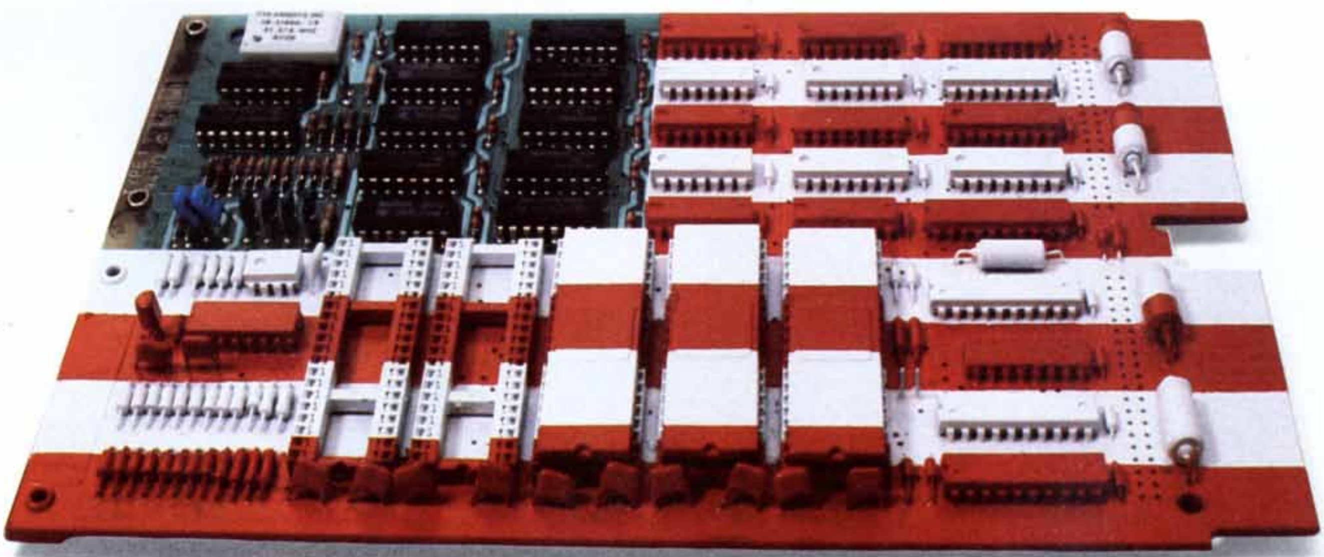
"Today, the most important problems facing humanity are in the arena of international relations where independent, egoistic nations face each other in a state of near anarchy. Many of these problems take the form of an iterated Prisoner's Dilemma. Examples can include arms races, nuclear proliferation, crisis bargaining and military escalation. Of course, a realistic understanding of these problems would have to take into account many factors not incorporated into the simple Prisoner's Dilemma formulation, such as ideology, bureaucratic politics, commitments, coalitions, mediation and leadership. Nevertheless, we can use all the insights we can get.

"Robert Gilpin [in his book *War and Change in World Politics*] points out that from the ancient Greeks to contemporary scholarship all political theory addresses one fundamental question: 'How can the human race, whether for selfish or more cosmopolitan ends, understand and control the seemingly blind forces of history?' In the contemporary world this question has become especially acute because of the development of nuclear weapons.

"The advice given in this book to players of the Prisoner's Dilemma might also serve as good advice to national leaders as well: Don't be envious, don't be the first to defect, reciprocate both cooperation and defection, and don't be too clever. Likewise, the techniques discussed in this book for promoting cooperation in the Prisoner's Dilemma might also be useful in promoting cooperation in international politics.

"The core of the problem is that trial-and-error learning is slow and painful. The conditions may all be favorable for long-run developments, but we may not have the time to wait for blind processes to move us slowly toward mutually rewarding strategies based upon reciprocity. Perhaps if we understood the process better, we could use our foresight to speed up the evolution of cooperation."

Advanced Technology Centers



Strategies for Corporate Growth

Innovation, development, and mature manufacturing at the edge of knowledge need different kinds of support. Where to find that support is basic to corporate strategy.

Peter J. Brennan and
Development Counsellors International, Ltd.

"During the next five years, Michigan will be the home for venture capital."

That's what Dwight Carlson, founder and president of Perceptron, a Michigan machine vision manufacturer, has to say about Michigan's massive new capitalization initiatives.

Thanks to the liberalization of retirement fund usage, and the establishment of the Michigan Economic Development Authority and the Michigan Certified Development Corporation, businessmen like Mr. Carlson will find more than five hundred million new dollars in Michigan. Michigan offers new financing for venture capital, land acquisition, machinery, construction and conversion. A state research fund provides grants for Michigan-based entrepreneurs to develop their products.

"Michigan has improved my chance for success and allowed me to build here," says Carlson. "I thought I'd face a lot of red tape and programs with no substance. But in less than three months, I've negotiated a substantial investment from the state retirement fund along with some of the top venture capital firms in the country."

Innovative capital programs like these, along with Michigan's exceptional research and educational facilities and its skilled labor force, move Carlson and colleague James Swartz of Adler & Company, a leading venture capital firm, to say: "We're excited about the positive action being taken in Michigan to enhance the growth of technology firms."

Let the consultants in Michigan's Office of Economic Development help you Say Yes with financing, location and state-financed labor training. To find out more, call toll free: 1-800-I SAY YES (in Michigan: 1-800-292-9544) or simply write on your business letterhead and mail to:

Albert A. Bogdan, Director
Office of Economic Development
Michigan Department of Commerce
Dept. 10583, P.O. Box 30225
Lansing, Michigan 48909

Dwight Carlson, President, Perceptron.

Attend: Bio-Connections '83, July 14 and 15, 1983.

Business Opportunity in Applied and Exploratory Biology
Michigan State University, East Lansing, MI
For information contact [illegible] amber of Commerce,
Box 1[illegible]340.

Innovation, development, and mature manufacturing at the edge of knowledge need different kinds of support. Where to find that support is basic to corporate strategy.

It was 1938. David Packard was a young engineer, newly graduated from Stanford University, working for one of the world's largest technology firms. His employer was General Electric Company in Schenectady, New York, a pleasant and long-established industrial city a continent away from the sunny Santa Clara Valley, between the bay and the sea.

The Valley had two main exports—fruit and college graduates. At Stanford, the late Professor Frederick Emmons Terman, a tireless promoter of his students, of Stanford, of the Santa Clara Valley and of the beginnings of electronic technology embodied then in radio, fretted over the loss of his brightest students. The Professor found them jobs, too often somewhere else like Schenectady. When he could, he found financing and research contracts to keep his graduates in California, close to home.

Thus William Hewlett stayed on to do graduate work. His friend, David Packard, married to a Californian, could be persuaded to return on a fellowship. Hewlett invented an audio oscillator and Terman, himself a transplanted Midwesterner, suggested Hewlett and Packard try marketing it. In early 1939, they did, from the garage of Packard's rented house.

From Orchards to Technology

Thus did the Valley begin to change to an economy based on new technology. It was not planned. There were no consultants' studies. No one was persuaded to shut down an uneconomic plant somewhere else and move it to cheaper California. No economic development organization promoted Palo Alto to all parts of the country.

Rather, as is true for most successful advanced technology centers, a local, possibly unique, combination of people, facilities and circumstances combined to yield an unforeseen result.

Stanford had excellent electrical engineering and physics departments, poised at the edge of practical applications for electronics technology. In Fred Terman, the school and the locale had a teacher, an organizer, a man of vision and determination, a superior technologist with a firm grasp of the real world. He was the essential industry-university-community linkage.

David Packard, a Coloradan, was prepared to go to the University of Colorado but ". . . the time I spent in Palo Alto in the summer of 1929 convinced me that I should apply to Stanford."

Like most human adventures, the process begins at a point no one then present can identify. Through a series of circumstances and serendipities, it gathers momentum. The phenomenon acquires shape, becomes large enough to have its own center of gravity. Like a black hole it begins to attract and coalesce new ingredients. A new-technology center appears amid the prune yards, on the swampy land above a salt bed, on the prairie, strung out along a highway . . .

Most important, the world notes that the new technology center is a net gain for both its region and the nation as a whole. Like a seed dropped on the forest floor, it has germinated, grown to maturity, gains its subsistence from roots much deeper than the surrounding shallow plants to which it gives shelter and sustenance. In time, it drops its own seeds, which grow to form a grove.

Cultivating Technology Business

We long ago learned how to farm trees so that we no longer depend on accidents of nature for timber. We can analyze after the fact what prompted a center of advanced technology to germinate and grow in the wild. We have been less successful in transplanting those conditions so that we can start and grow new technology centers wherever we want.

Technology is the key to increasing productivity of both capital and labor. A producer made most efficient by computer-aided design (CAD), state-of-the-art instrumentation and control, electronic data processing for inventory and materials management and modern telecommunications is not as vulnerable as his less technology-intensive, less efficient competitor.

Advanced technology is thus vital to the competitiveness of older industries, which will continue to exist and expand. Whether they reindustrialize in the United States or migrate entirely to other parts of the world largely depends on how well they adapt new technology to their own ends. Survival also hangs on how well the high-technology industries themselves fill the needs of their non-high-technology so-called smokestack industry customers.

State governments recognized long ago that advanced technology is the key to future prosperity. The economic vitality of Silicon Valley and the Minneapolis-St. Paul area are well known. The virtual rebirth of the Massachusetts economy based on seven high-technology industries that often started out in the abandoned factories of long-gone smokestack industries has not gone unnoticed. Neither has the symbolism. Recent economic history, however, has greatly concentrated the attention of both state governments and businesses.

Traditional regional, state or area development programs have aimed at painting an attractive picture for any kind of industry. Development officers largely were marketers rather than developers. Their sales message was and often still is financial incentives. The long-term objective was an increase in local and state tax bases and simultaneous increases in employment.

The strategy works well enough in a rising economy when addressed to traditional non-high-technology industries, whose requirements are based on the products they make, the raw materials they need and the markets they serve. However, if a rising tide raises all boats, it is equally true that a falling one exposes the rocks and impales a few of the larger vessels, which will never float again. More than any other in the past twenty years, the 1979-83 recession exposed the rocks in the American industrial economy.

High Tech Is Different

The strategy does not work as well when addressed to high technology industries. Measured against the factors that traditional industries must consider, high technology industries are virtually site-independent. They can set up just about anywhere. And they do, all too often.

Area development officers must ask: who are the Hewlett-Packards of today and what do they need?

Companies that help corporations choose sites for new facilities have started to find those answers. Robert M. Ady, executive vice president of The Fantus Company, a Chicago-based site location firm, says: "When we deal with a traditional industry, the company's short list of preferences will all be in the same region—say Texas, Oklahoma and Louisiana. But when we deal with high technology companies, the preferences will be scattered across the country—California, Texas and North Carolina."

States, regions and localities have started to recognize this fundamental difference between high technology and traditional industries. Further, they recognize that there are differences within the high-technology industries themselves. The most technology-intensive of all industries, guided missiles and spacecraft (Standard Industrial Classification [SIC] 376) has some of the aspects of a traditional industry. At its manufacturing level, this industry deals with big items that need plenty of space. Among the largest single manufacturing facilities in the world are aircraft plants in Texas, Kansas, and coastal Washington.

Hi Tech—a Definition

According to the Bureau of Labor Statistics, a high technology manufacturing firm is one in which engineers and scientists comprise more than five percent of the total work force. That, however, is too loose a definition. It includes transportation equipment in general, for example, of which aerospace is a part, and chemicals, of which pharmaceuticals are a part.

Advanced technology industries are better defined as those that require high levels of continuing innovation and whose markets can change overnight. These firms typically have 10 percent or more scientists and engineers. Firms working at the leading edge of technology have 15 percent or more engineers and scientists on staff.

Only six industries qualify. These and their SIC codes are: Pharmaceuticals (283), Computers (357), Semiconductors (367), Communications (367), Aircraft (372), and Instruments (381 through 384).

Instruments includes medical, controlling and scientific instruments.

One must also include the service sectors that are so increasingly important in the U.S. economy. More than most manufacturing sectors, such service industries as finance, banking, communications, software development, insurance, medical services and data-processing are fast-growing users of advanced-technology products and the driving force behind the rapid commercialization of new ones. Commerce is more and more a market for industry—but commerce basically services industry.

A high-technology company typically goes through three distinct phases. Fantus's Robert Ady defines these as the theory-driven, product-driven and market-driven stages. A successful innovative and productive company will exist on all three levels. However, each phase requires a different set of circumstances at its beginning. As each phase becomes a continuing level, it requires still another set of conditions to prosper and grow.

The Three Drives

First is the initial scientific discovery and the follow-on work that converts the discovery to a commercial product. A lone and brilliant inventor, a Hewlett or an Edison, may well make the initial discovery working in his basement far from any known technology center. He may even carry the discovery to the point of commercial viability. But as the frontiers of knowledge have expanded, wresting new secrets from science and moving them up the scale to commercial practicality has become an expensive effort.

The second product-driven stage entails development and first-level manufacturing. Once a scientific idea has been proved and its practical applications divined and defined, a company will exploit it. The company may be a new one formed out of the basic research group that developed the idea, or it may be a large established one seeing opportunity in new technology.

The classical view in the United States is that small, entrepreneurial firms bring most new technologies to market. And in fact, most new jobs produced in the U.S. economy come from small companies exploiting new technologies. However, large companies with large R&D budgets and extensive facilities spend far more on R&D than do small new firms. So do universities, which, says the National Science Foundation (NSF), perform half of all basic research in the U.S.

During the product-driven stage, a company depends on its source of basic science and technology, which may be the company itself. More likely, though, the source will be a nearby university, research institute or large technology-oriented corporation from which the company's founders came. As the company expands and increases its R&D expenditures, it will become less directly dependent on external technology. Ideally, it will itself become a technology source in a growing technology center.

Outgrowing Home Base

As the growing firm's manufacturing capacity expands, its need for space and other services grows. Manufacturing remains closely linked to research and development sources. Consequently, if manufacturing and R&D cannot be at the same place, any new site must be within easy reach of the R&D center. The manufacturing site must have locally available nearly the same level of technology as does the home site. To attract and hold engineers and scientists, the new location must offer similar amenities.

Few people, not Hewlett, not Packard, not Samuel N. Irwin, deliberately move to a location specifically to found a company. "Companies start where people are," says Mr. Irwin, founder and president of Irwin International in Ann Arbor, Michigan. The founders may be there because of a university, an employer, research foundation, government installation or were born there. Having started the firm, local factors keep the founders in the area, at least through initial success and early expansion.

Local factors may enable a newborn company to grow to a certain point in its product-driven phase. However, this stage is a critical time in a company's life. Further expansion may produce new needs that the local economy cannot fill. In its early years, any fast-growing company is locked in to its location by its need for cash to finance growth. It must generate most of that cash internally. Venture capitalists may fund a company only through startup. Equity markets want to see a track record. Conventional bankers prefer more established, less risky ventures.

When the Company Should Move

When a company outgrows the local supply of brain power, cannot attract the people it needs, encounters too much red tape in continuing financing, it moves to an area where the community understands and is prepared to serve its needs.

Or a company may outgrow its manufacturing space. Its first satellite is likely to be nearby in an area that offers most or all of the benefits of the headquarters location. Since the com-

pany can now choose its new site, management will look more closely at factors the founders could not control.

Taxes Are Important

There is a misconception that tax structure is important only to large-scale, large-employment smokestack traditional industries. To justify this view, people point to California and Massachusetts, the top advanced-technology states but far from the bottom of the list on overall tax burden.

Not so. "Taxes are very important to small, expanding high-technology companies," says Dr. Robert Premus, Dayton-born staff economist of the Subcommittee on Monetary and Fiscal Policy of the Joint Economic Committee. Premus is the author of a recent report entitled "Location of High-Technology Firms and Regional Economic Development," based on a survey of 691 companies.

"Small companies in particular rank the tax burden high," said Dr. Premus in an interview. "Cash flow is critical. Local taxes can take cash from a company when it most needs it." At the same time, they are locked into their location. By the time they can afford to move, the tax burden has become secondary. They may even move to a more heavily taxed jurisdiction.

The founder of one high-technology company explained that paradox. William C. Norris, chairman of Control Data Corporation in Minnesota, said that the State of Nebraska's willingness to increase both the tax base and rate was a powerful factor in the company's decision to locate a new facility there. "Nebraska demonstrated that it meant to provide the tangible and intangible facilities and services that (we sought)," said Norris. "Without such a change in philosophy," the Nebraska native continued, "we likely would have gone elsewhere."

The Market-Driven Facility

The third stage identified by Robert Ady is the mature manufacturing or market-driven level. The company and its products have now grown to a point where costs are paramount. The company has outgrown its older facilities. Processes and products are mature. Transportation, local financial incentives and tax structure, availability of low cost and trainable labor all assume greater importance in the site selection decision.

The market-driven firm's products are high technology. But manufacturing is essentially an assembly process. The technology and science is in the maintenance of the manufacturing systems and in quality control of raw materials and finished items. Engineers and scientists are a small percent of total employment.

This third stage plant will have different needs than the earlier stage one. This new facility is still not tied much to natural resources, nor overly dependent on market proximity. Such site-independent plants can serve markets in North America as easily from Taiwan or Malaysia as they can from Tennessee. Beyond minor input by local managers, all technology is generated elsewhere. New technology arrives daily, by telephone and computer network or encased in the latest model of manufacturing systems equipment—but it comes from somewhere else.

Such are the plants that site development managers seek. They are clean, reasonably kind to the environment. Their demands on the local infrastructure are easily met if there is adequate electrical power and water. They provide jobs for local people without placing great demands on the local services and amenities. They do not bring in large numbers of people who are accustomed to more than the community provides.

However, since these plants are essentially assembly operations, they commonly do not have the same ties to their location as does the company's technology base. To quote Dr. Robert Premus's Congressional report: "The survey . . . indicate(s) that high-technology companies are 'footloose' . . . access to raw materials . . . markets and transportation are not major locational determinants. Nor are . . . water . . . energy . . . and climate important determinants . . . high-technology companies are drawn more to highly specialized resources such as labor skills and education and to factors that make it easier to attract and maintain a skilled labor force, most notably State and local taxes . . ."

The survey also indicated that most high-technology companies prefer an urban to a rural environment. The centripetal effect of an urban-centered location brings in more people and companies and encourages technology transfer. Indeed, if a university in the traditional sense is a community of scholars, then a high-technology center may itself be a university—a community of scholarly companies.

Dynamics of High-Tech Industry

Most states historically understood little about the dynamics of high-technology industry. To attract such companies, they did little analysis and less development. In effect, the state rounded up what already existed and packaged it attractively for itinerant industry. The package did little for resident businesses and less still for nascent ones.

Miles Friedman, Executive Director of the National Association of State Development Agencies (NASDA), notes the changes in the attitudes of state development people. "It is still true that many state agencies see it as their jobs to move plants from other states to theirs. But a common business complaint is that the state agencies ignore the businesses they already have. Now there is big emphasis on in-state development and the incubation of new businesses."

Some states and localities years ago tried to develop conditions that would permit advanced technology centers to flourish. Not until 1981 did the National Governors Association (NGA) establish a Task Force on Technological Innovation. Local generation of high-technology industry is clearly an idea whose time has come.

Partially as a result of the NGA's initiative, many states from Maine to Hawaii established high-technology study groups. Eventually, some states will take all possible steps to establish and strengthen the structural elements needed to breed high-technology industry. Some few may decide that their future economies will depend on maintaining the major base of their present economies. The states' new awareness of high technology can help direct technological resources to appropriate economic sectors, whatever their place on the technology spectrum.

The Governors' Survey

In late 1982, the NGA Task Force surveyed every state governor on organization of state efforts, economic

1714-C

WANTED.

PARTNERS TO HELP PENNSYLVANIA FORGE THE FACTORY OF THE FUTURE.

You have a dream, a vision of the coming technology.

Its seeds were planted by the likes of Asimov, Clarke and H.G. Wells. Its growth nurtured by every teacher, class, professor and experiment that ever challenged you. And inspired you.

In Pennsylvania we, too, have a dream. The creation of an environment in which our universities, industries and state government join forces in a partnership to help scientists, engineers and entrepreneurs like you forge your dreams into the technologies of the future.

Much of that environment already exists. As you read this, more than 150,000 Pennsylvania engineers and scientists are developing all facets of advanced technology. Robotics. Microelectronics. CAD/CAM. Biotechnology.

There are now more major research universities in Pennsylvania than any other Northeastern state. And those schools are supported by corporations whose R&D contributions are among the highest in the nation.

Last year we awarded almost nine million dollars in low interest loans to advanced technology companies.

We also created the catalyst that will make our dream a reality.

The Ben Franklin Partnership.

It combines the talents and resources of our universities, small and large corporations, and state government to further stimulate advanced technology research and development. Millions of dollars in grants will provide funds for everything from basic product research to specialized work force training and sophisticated market planning.

As thorough as this partnership is, it lacks one essential element.

You. Your talent. Your vision.

Send the coupon to Walter Plosila, Deputy Secretary for Technology. Or call him at (717) 787-3003. You'll learn more about the quality of commitment, as well as the quality of life you will find here.

In Pennsylvania, we share your dream.

Come, be partner to ours.

Mail To: Department of Commerce, 433 Forum Building, Box 2001, Harrisburg, PA 17120

Name ______ Title ______

Company ______

Address ______

Telephone ______

ADVANCED TECHNOLOGY HAS A FRIEND IN PENNSYLVANIA

 Dick Thornburgh, Governor James O. Pickard, Secretary of Commerce incentives, local university-industry cooperation and worker training programs. The survey report in July 1983 will have two sections. One organized by state will describe state policies and programs in technological innovation. The other, organized by program category, will make it easy to compare state approaches to particular aspects of economic development and technology, by various methods.

Among the early findings: at least 11 states have appointed task forces or boards as overall policy-determining bodies on technology. These are California, Illinois, Iowa, Kansas, Michigan, Maryland, New Jersey, North Carolina, Ohio, Pennsylvania and Virginia. Most set these up in 1982, though North Carolina was way ahead of its time, having established its Board of Science and Technology, chaired by the Governor, in 1963.

Twenty-two states have advisory groups outside the state government. Their scope and linkage to the government varies greatly. Georgia, for example, recently set up the Advanced Technology Center at Georgia Institute of Technology. The Governor and legislature created the Center in 1980, but it is not a government body.

State money and private-sector matching funds finance the Maine Development Foundation, which the state legislature enabled in 1977. A governor's advisory committee proposed the cooperative business and government approach in 1975. In Minnesota, twenty-eight prominent people established last year the Minnesota Wellspring. The governor is honorary chairman. New Hampshire, which has developed a high-technology service center based on publishing and computer software around the town of Peterborough, has the Center for New Hampshire's Future. This is a private organization established in 1979.

The not-for-profit Indiana Corporation for Science and Technology was authorized by the General Assembly in 1982. The Governor appoints a 24-member board representing the private, public and educational sectors. California has the Commission on Industrial Innovation; Maryland its Governor's Advisory Council; Michigan the proposed High Technology Corporation; New York the Science and Technology Foundation; Pennsylvania the Governor's Council on Science and Technology; and South Carolina a proposed Industrial Research Board. The City of Chicago and The State of Illinois, together with Chicago universities, have created an "Illinois Technology Partnership."

The Early Starter

Few programs were as early or as ambitious as North Carolina's. As George Herbert, president of North Carolina's Research Triangle Institute describes it, the state in the mid-fifties was near the bottom in per capita income and too dependent on "old-line segments of its economy: agriculture, textiles, tobacco manufacturing, furniture and brick and tile." At the same time, the state ranked among the top ten in numbers of colleges and universities. But for their graduates there were few local opportunities.

North Carolina looked enviously at the centers in Massachusetts and California and noted that the desired industry existed in areas of strong research concentration. Further, most important research centers lived in the shadow of major graduate-level universities. The state had at least three such schools close together: Duke at Durham; the University of North Carolina at Chapel Hill; and North Carolina State University at Raleigh.

The state in 1956 decided to establish a research park in the triangle formed by the universities, thus the Research Triangle. The concept was not an instant success, nor did the founders expect it to be. As Mr. Herbert says, "They knew they were working for returns that would not be truly significant for 15, 20 or 25 years." And he cautions newcomers to the high technology stakes, "This is a reality, too often ignored by many of the groups that visit us today looking for a quick fix for 1983's economic woes."

No one really knows the optimum size and mix of the critical mass that turns a high-technology manufacturing center into a truly innovative one. Few doubt that it takes many years to reach that self-igniting concentrated mix of research centers, universities and theory-oriented industry.

The planners first established the Research Triangle Foundation to acquire land and develop the 5,700-acre park. They also set up the Research Triangle Institute in close association with the universities. None of these is a state agency.

Despite an early capture of a major industrial laboratory in 1959, others were slow to follow. By 1965, there were only nine laboratories with a total employment of 1,000. But then things began to pick up. IBM bought 400 acres. The National Institutes of Health established its National Institute of Environmental Health Sciences. Another major industrial laboratory moved in and the project was off and running.

The most recent addition is the state-sponsored Microelectronics Center of North Carolina, established in 1980. The legislature appropriated the first one million dollars for startup, and the state is providing the first 24.4 million dollars for ongoing support.

North Carolina's experience has many lessons for others. But it is not yet clear that the state has managed to produce an industrial rival to Silicon Valley or Route 128. The location is high on everyone's short list when looking at possible new locations. The universities and RTI itself are well-reputed basic research centers. But the bubbling ferment of entrepreneurial activity that is the true mark of a self-generating advanced-technology center is not yet evident.

A Downtown Research Park

Another research center started on a different premise but at about the same time is the University City Science Center in Philadelphia, Pa. There, the University of Pennsylvania and Drexel University were situated in deteriorating neighborhoods. The universities, business and community conceived The Science Center as a way to improve the city's technology base and reverse urban blight. Under the auspices of some 28 member institutions that own it, the Center began in 1964 in a renovated building. It is now an urban research center covering some 19 acres with over seventy science-based organizations in its nine buildings. Some of these organizations are new to the city while, more important, the Center encourages others to remain there.

The University City Science Center is so far the nation's only downtown research park, though Detroit and Wayne State University in Michigan are starting a similar concept in the Detroit-based Metropolitan Center for High Technology.

The University City Science Center has a relatively small Research Institutes Division. The Research Triangle Institute is a much larger, free-standing organization similar to such older institutions as Battelle Memorial Institute at Columbus, Ohio; Stanford Research Institute in Menlo Park, Calif.; Midwest Research Institute in Kansas

City, Kansas; IITRI in Chicago; Southwest Research Institute in San Antonio, Texas, and many more.

For smaller companies and even larger ones faced with scientific questions that they have neither time nor staff to handle themselves, the research institutes are an important resource. As do most companies, the research institutes often specialize in technologies that reflect their geographical location and the interests of their major clients.

Subdivisions for Technology

Clean high-technology industry usually does not need large and heavy installations. One can develop industrial parks for such industry just as one develops residential sub-divisions for people. A well laid out industrial or research park with room for expansion can be an important factor in site selection, particularly for a smaller, high-technology company making its first move out of its basement.

Successful developers present a balanced facility. If the park itself does not contain a research institute and university campuses, these will be within easy reach. The park will also have conference facilities, hotels, shopping centers and restaurants. A large development may be a completely planned community.

The Huron Center near Ann Arbor, Mich., is such a real estate development. A joint venture between Mitsubishi and Morgan Stanley, the Center is a 393-acre multi-use development eight miles from the University of Michigan and 25 miles from Detroit. When completed, it will contain residences as well as research labs, hotels as well as light industrial plants.

Utah, recently in the news for the artificial heart work at the University of Utah Medical Center, has a growing advanced-technology center in what one publicist refers to as "Bionic Valley," near Salt Lake City.

Montgomery County, Maryland, capitalizes on the nearby concentration of Government medical research facilities embodied in the National Institutes of Health, Bethesda Naval Hospital and many medically-oriented companies in the area. The county has established the 232-acre Shady Grove Medical Park, which has reserved 145 acres for medical science-related businesses and institutions.

Also in Montgomery County is the

planned city of Columbia with both residences and work places for people in high technology. Columbia mirrors to some extent the planned city of Reston across the river in Fairfax County, Virginia, another site for much high-technology industry.

The Planned City

An ambitious development that has learned much from Columbia, Reston and the Research Triangle is The Woodlands near Houston, Texas. This project of the Mitchell Energy Company is a wholly integrated 25,000-acre residential, research and light manufacturing community. The developers intend that everyone who works in the Woodlands can afford to live there. Several thousand people already do. In that, the community differs little from the traditional concept of the small, self-contained town, which it is.

The research forest has four elements. There will be a 400-acre campus for the University of Houston. The Texas Medical Center Inc., has 150 acres for a research campus. One hundred acres belong to the Houston Area Research Center (HARC), a research institute under the auspices of Texas A&M, Rice University and the University of Houston. Some 1,300 acres are set aside for high-technology businesses and their suppliers.

While developers take pristine land and turn it into parks, in many parts of the country suitable sites lie fallow. These are deactivated military bases. Though these sites are frequently off the beaten path, companies can turn their existing infrastructure to commercial use at low cost. Many were air bases. In Chippewa County, Michigan, for example, market-driven firms can set up on 41 acres of lighted concrete.

Other states and regions hope to use what they already have to generate new technology industry. Illinois, for example, has the Argonne National Laboratory, Fermilab, the Illinois Institute of Technology Research Institute (IITRI), the many fine universities around Chicago and the University of Illinois at Champaign-Urbana. A technology corridor extends westward from Chicago toward Fermilab. The area has an enormous base in existing medical and electronics technology. It has some of the more active and venturesome venture capitalists. But it has lacked the university-industry-community links that have been so important to theory-based entrepreneurial activity elsewhere, which explains recent Task Force development of significant education-based local cooperation.

The GOCO Connection

Many localities with government high-technology installations now see that these facilities are part of the community's technology base. They employ thousands of scientists and engineers as well as equipment and instrumentation that only the wealthiest of industries and few universities can afford. These are the GOCO (Government-Owned, Contractor-Operated) facilities. Much of their technology is in the public domain.

A government high-technology facility linked to local universities creates a high-technology center all its own. With proper encouragement, availability of capital, extended relationships with local businesses, GOCOs become a factor that expanding high technology companies must consider as well, as potent centers for local development.

Tennessee, recognizing the enormous science and technology base at the Union Carbide-operated Oak Ridge National Laboratory and the University of Tennessee a few miles west at Knoxville, has undertaken a major program to develop a technology corridor. Tennessee is not widely thought of as a high-technology area. Yet there are over 2,000 Ph. D. level professionals in the Oak Ridge area alone, including the largest concentration of doctorate-level biologists in the world.

Similar GOCOs and even GOGOs (Government-Owned, Government-Operated) facilities across the country provide ready-made advanced technology centers to which aspiring companies might well attach themselves or from which new ones might spring. Dayton, Ohio, for example, rightly considers Wright Patterson Air Force Base, the Air Force's major technical center, to be a community high technology resource.

The benefits to California of such GOCOs as the Jet Propulsion Laboratory and the Livermore Laboratory are substantial. Florida's technology has gained from the space flight operations at Cape Canaveral and the industries that have sprung up to serve them.

NASA's contribution to the technology base in Houston is incalculable. Joseph P. Loftus, Director of Technical Planning at NASA, Houston, recalls the moon-shot days when thousands of eager engineers and scientists flocked to Houston to be at the leading edge of something exciting. "I see 15,000 engineers and scientists who once worked here and stayed. They have effected an enormous transfer of material and management technology from the very leading edge where NASA works into the everyday business of the region."

New Mexico has university-operated Los Alamos National Laboratory and AT&T-run Sandia National Laboratory. These and other GOCO facilities of the Department of Energy work in many economically significant areas such as geothermal research, synfuel, solar and new and renewable energy, as well as nuclear fission and fusion.

The Federal government's primary laboratory for western coal and lignite research is at Grand Forks, North Dakota. It is now owned by its former contract operator, the 100-year-old University of North Dakota, and renamed the University of North Dakota Energy Research Center. It will continue to handle government projects but will also conduct programs for other sponsors. The Center could be the nucleus of a technology center.

University-Industry Linkages

Successful technology transfer from research to commerce generally results from good linkages between the universities and industry. Such linkages do not always exist. In maintaining something of an ivory tower attitude toward the real world, some universities have forbidden their faculties to engage in commerce. But as manufacturing technology approaches levels of science once found only in the academic research laboratory, and as only industry and government can afford the equipment needed to probe the outer reaches of knowledge, more universities participate in the marketplace.

At the first meeting of the NGA task force, Dr. George A. Keyworth, Science and Technology Adviser to the President, remarked about the problem: "The resistance of the flow of technology from our research laboratories to the marketplace is abysmal," he said.

"There was a decoupling of universities and the business community," says Robert Premus. "The universities sought Federal money for their programs, which were for Federal purposes. That reduced the flow of technology into U.S. industry. But the government cut back support to universities. We see the universities getting anxious and talking to the

Advertisers in this special SCIENTIFIC AMERICAN advertising supplement will be pleased to provide additional information about their services:

- Commonwealth of Pennsylvania, Department of Commerce
- Consumers Power Company
- Glenrothes Development Corporation
- Huron Center
- Industrial Development Board for Northern Ireland
- New Jersey Department of Commerce & Economic Development
- Public Service Electric & Gas Company
- State of Michigan, Office of Economic Development
- State of New Mexico, Economic Development Division
- State of Tennessee, Department of Economic & Community Development
- State of Utah
- State of Vermont

Interested readers may contact advertisers directly or may write: Manager—Industrial Sections, SCIENTIFIC AMERICAN, 415 Madison Avenue, N.Y., NY 10017.

ADVANCED TECHNOLOGY CENTERS—USA

"Michigan—Technology That Works," September 1982.

"Oregon: Environment for Technology," May, 1982.

"Advanced Technology in Illinois," September 1981.

"Advanced Technology Center Santa Clara Valley, California," March 1981.

"Dayton: Formula for High Technology," January 1981.

"Technology of Maryland," December 1980.

"Minnesota: Technology Wellspring," October 1980.

"Advanced Technology Center Fairfax County," September 1980.

For information about obtaining reprints, write:

Manager-Industrial Sections
Scientific American
415 Madison Avenue
New York, N.Y. 10017

corporate community again."

A respected university has value to an expanding firm, because it supplies a talented labor pool, is an educational and technology resource for the company's staff, and because it attracts a continuing flow of talent. Indeed, basic to technology transfer is the flow from other university-centered areas of people who like birds carry seeds of technology sown elsewhere.

Many people recognize the problem of attracting development support from academia and are doing something about it. Solutions range from organized industry-university linkage organizations through business incubation programs at universities to loosening the rules so that faculty may more easily talk to business.

The University of Wisconsin at Madison has had the University-Industry Research Program (UIR) since 1966. UIR helps develop relationships between UW faculty and industry and helps industry and faculty identify programs, facilities and people with mutual research interests.

The Oregon Graduate Center, a non-university degree granting organization, performs a similar function. The Center has been an important factor in the state's success in attracting spinoff advanced technology companies.

In Pennsylvania, the aptly named Ben Franklin Partnership program concentrates on patents. College and university patent policies cover new developments through the advanced technology centers. Policies concern licensing in the state, royalties and other fees to support advanced technology centers.

Incubating New Businesses

At Carnegie Mellon University in Pittsburgh, Prof. Dwight M. Baumann has been trying a somewhat different approach—the incubator. He set up the Center for Entrepreneurial Development at Carnegie Mellon in 1971. Since a common feature of successful advanced technology centers is a growing cluster of entrepreneurial companies passing on through several generations, Prof. Baumann wanted "to see if we could create a cluster."

Prof. Baumann's Pittsburgh Center has been responsible for 16 first-generation high-technology companies and 9 second-generation ones. One sold for $6 million. Prof. Baumann acknowledges the importance of venture capital, but not necessarily where the entrepreneur is: "Venture capital follows innovation—not precedes it," he points out.

Where The Money Is

Most states recognize that venture capital and follow-on financing are important. Some have innovative programs to channel funds to new companies. Connecticut has a unique concept in the Connecticut Product Development Corporation (CPDC). This publicly-funded, state-chartered corporation's mandate is to invest in innovative new products. CPDC borrows from the state to share the costs of development with the company. CPDC earns royalties and repays the state treasury. The funding is not a loan, but neither is it equity. On a successful product, CPDC can recoup five times the development funding. If the product fails, CPDC loses its money, so far about 3 percent of its total investment.

Michigan has begun to move into areas where private capital has been absent. The state has venture capitalists, notably Doan Resources of Midland, but they are not sufficient to the demand. So the state legislature passed new laws that allow state retirement funds to make equity investments in Michigan businesses. Over $350 million is thus available to high technology companies.

Under new legislation proposed by Governor Ted Schwinden, Montana will invest 25 percent of certain tax proceeds in new and expanding Montana firms, about 13 million dollars the first year. Some 20 percent of the state's own investment portfolio will be in new or expanding Montana firms, about 140 million dollars per year. The state will also create a private-sector venture-capital Montana Development Credit Corporation.

Maryland takes a somewhat more conventional approach. The Development Credit Corporation of Maryland (DCCM) does not take an equity position; rather it lends money to fledgling firms that have shown some evidence of managerial ability to operate at a profit.

Different States, Different Styles

As the competition to attract advanced-technology industry intensifies among advanced technology centers, the states and regions have begun to resemble the firms they seek. Economic development departments are becoming analogous to profit centers. They raise money, mount marketing efforts, devote more attention to such quality controls as cutting out bureaucratic red tape and honestly analyzing themselves to isolate specific benefits and competitive advantages.

The mode and philosophy of economic development varies greatly from state to state, as the NGA survey makes clear. The governor himself, however, is the chief economic development officer. States Fantus's Robert Ady "The Governor is the ultimate spokesman. He makes a profound impression by his personal involvement in the economic development process."

Some states maintain a relatively low-key, low-budget effort. There may be a small office buried in another department or attached to the Office of the Governor. Cities and counties may have only the business-funded local chamber of commerce.

Others rival many countries. A well-funded central organization may have a cabinet-level director and control of a huge budget for everything from advertising to financial aid. Such an organization may well have branch offices in other states and in foreign countries.

Some states that believe they have a special focus maintain relatively small organizations. Delaware, for instance, seeks high-technology industry but is more oriented toward technology-using service industries like banking and finance, a route North Dakota has also taken. Delaware is one of only four states with a court devoted entirely to business law, the Court of Chancery. Other states, says Delaware, can change their systems to match. But nothing can match Delaware's eighty-year history of case law in business affairs.

A few states have publicly-funded municipal and county development organizations that are nearly as prominent or more so than the state organization. A gain for the county or city is of course a gain for the state. But once a county or city organization learns of a prospect, the rivalry and competition can be as fierce as any between nations.

Utilities Eager to Help

Companies need adequate and dependable utilities. Among the factors that influenced the government's decision to install Fermilab's huge particle accelerator in Illinois, for example, was Commonwealth Edison's ability to satisfy enormous instantaneous power demands at a reasonable price.

Dependable power is seldom a problem in the United States. In any case, utilities are not a deciding factor for most high-technology companies. However, in many states the utility companies are a valuable resource in another way. They are literally wired in to their state or region. As investor-owned companies, they are invaluable information sources.

Sometimes the utilities do more state promotion than the states. As John H. Maddocks, General Manager for Area Development at New Jersey's Public Service Electric and Gas Company puts it: "We're part of New Jersey. The vitality of our company is related to the economy of the state."

Michigan's Consumers Power Company gives prospective Michiganders extensive information. Not only does the company help select a site, its training program for Michigan communities shows them how to retain industry and attract new ones.

A new company or an expanding one thus has many places to turn in seeking those factors that will best suit its unique needs. Company management, particularly when looking for new sites for market-oriented plants, cannot easily rule out any location. It is a truism that every place will uniquely match someone's needs. And, to quote Florida's promotional motto, "People like to work where they like to live."

Venture Capital—High-Tech Nutrient

Most government financial incentives have little to do with getting the shaky embryonic firm with a great idea out of the basement, garage or university laboratory and into the marketplace. That function is truly the province of venture capital. A few states have made some provision to help guide state funds into venture funding. But it is not yet clear what form these initiatives will eventually take.

They might be repaid out of profits to an extent beyond the initial sum. Such a form, similar to that of the Connecticut Product Development Corporation, is a grant whose reimbursement is contingent on the success of the enterprise. Maximum reimbursement is limited to five times the amount of the original grant.

Venture capital is equity money invested in the early stages of a company's life to help it grow beyond startup. While the initial money will have come from principals and their families, venture capital generally comes from outsiders, who take shares in the company and will become part owners.

Relations between venture capitalists and company founders are complex. They can be difficult depending on how much control the venture capitalist demands and the founder is willing to yield. (For an excellent discussion on this point, see "Raising Venture Capital: An Entrepreneur's Guidebook," published by the New York accounting firm Deloitte Haskin & Sells.)

In the Last Analysis

A management's decision on where to put the new high-technology manufacturing plant can rest on intangibles. Among the factors Bob Premus's survey found was the company founder's place of birth. As says Colorado-born David Packard, "I have always had a high regard for the University of Colorado and it may have something to do with my interest in our location in that part of the country."

Hewlett-Packard's first expansion beyond Palo Alto was to Colorado.

Managements want to increase profits. Mature managers will consciously examine every facet of their company's position within its own evolving industry and the company's own stage of development. Those in the theory-driven phase may also make all the right moves more out of instinct than reasoned strategy—one reason so many company founders find themselves ousted as the company enters the product- and market-driven phases.

But instinctive or planned, any high technology firm must encounter, consider and deal with these many factors governing the right location at the right time in the company's history. No advanced technology business in today's fast-changing markets can stand entirely alone.

A U.S.-based high technology company in the theory- or product-driven stage would not usually locate its operations overseas. There are exceptions, especially in bio-medical technology. An overseas location can allow such a company to bring a product to market sooner under perhaps more relaxed regulatory procedures than those in the U.S.

When a high-technology product reaches the market-driven commodity stage, however, and costs and market access outweigh most other factors, management must look as closely at overseas sites as domestic ones. A company will usually consider off-shore manufacturing if it also has or expects to develop markets abroad. However, foreign production occasionally makes economic sense even when all output goes to the U.S. market.

Many countries are eager to attract high-technology manufacturers. Countries want these industries for the same reasons the states and localities of the U.S. do. They provide jobs, help upgrade local skills and eventually increase the tax base. More important in the view of many countries is the potential for technology transfer. Sadly, the latter expectation is usually a vain one unless a country already has a strong technology base of its own.

That has not deterred most of the world's nations from competing for high-technology industry. Just as have the several states, many have set up special economic development agencies designed to bypass much government red tape and deal directly and efficiently with high-technology companies. Some of these such as the Northern Ireland Industrial Development Organization (NIIDO) are virtually autonomous and empowered to deal with new companies on a range of issues far beyond those that most domestic development agencies can handle on their own authority.

Overseas organizations like NIIDO offer as incentives the same range of financial and personnel benefits as do the U.S. organizations. And as do the U.S. states, other development agencies such as the Industrial Development Authority of Ireland (IDA), the French, German, Italian, Danish, Luxembourg, Malaysian, Singapore, Sri Lankan and more will try to match those incentives. Indeed, as agents of sovereign governments, not governmental units, these overseas organizations can compete at an awesome level, one few states could match.

Though location per se is seldom the primary factor for a market-driven facility, it can be important if the market is an export one and the company has extensive overseas interests.

Northern Ireland gives ready access to the European market and particularly to the British and British-dependent markets from within the political boundaries of the United Kingdom. Ireland is an English-speaking jumping off place for all Europe from within the Common Market. Luxembourg is a money center. Austria is the gateway to Central and Eastern Europe and Hungary is the stepping stone.

Caribbean dependencies of European countries, such as the Netherlands Antilles, allow favored access to the Common Market from bases closer to home. Special agreements allow other countries like Trinidad and Tobago such access too. Puerto Rico and the U.S. Virgin Islands confer tax benefits no state can match. A high-technology firm can serve markets in Asia better from Sri Lanka or Malaysia.

Most nations have active investment-promotion programs. In some, as in many U.S. states, local units such as Glenrothes in Scotland work in parallel with the national agencies. As in the states, the degree of skill, resources and commitment each nation brings to its promotion effort varies enormously.

Profit is no longer a dirty word in many countries. Foreign investment, once shunned, is now sought. But old attitudes die hard. As one skeptical participant in a recent investment promotion conference said: "Politicians propose; bureaucrats dispose. What have you done about your bureaucrats?" The more successful organizations like NIIDO and IDA were designed to shortcut bureaucratic channels. For lack of that authority, some nations' efforts remain hobbled despite the best intentions.

Numerous investment promotion agencies have branches in the United States. Reflecting their independence, these are not connected to embassies or consulates. Other nations treat development as a sideline for diplomatic personnel.

At the urging of a staff member, Adly abd el Meguid of Egypt, the United Nations Industrial Development Organization (UNIDO) in 1978 set up a program to help governments develop their own investment promotion capacity. The U.S. Overseas Private Investment Corporation (OPIC) helped fund the project. Several countries run their promotion efforts out of UNIDO in New York. Sri Lanka was one of the first and more successful graduates of that program. The country's success reflects location and commitment.

THE REAL NORTHERN IRELAND:

Martin Malone is one reason electronics investors score big in Northern Ireland.

Martin Malone knows his way around both the sports circuit and the printed circuit.

A semi-professional Irish League-footballer in off hours, on the job Martin captains a team of 110 printed-circuit technicians as works manager for Irlandus Circuits Ltd. in Craigavon.

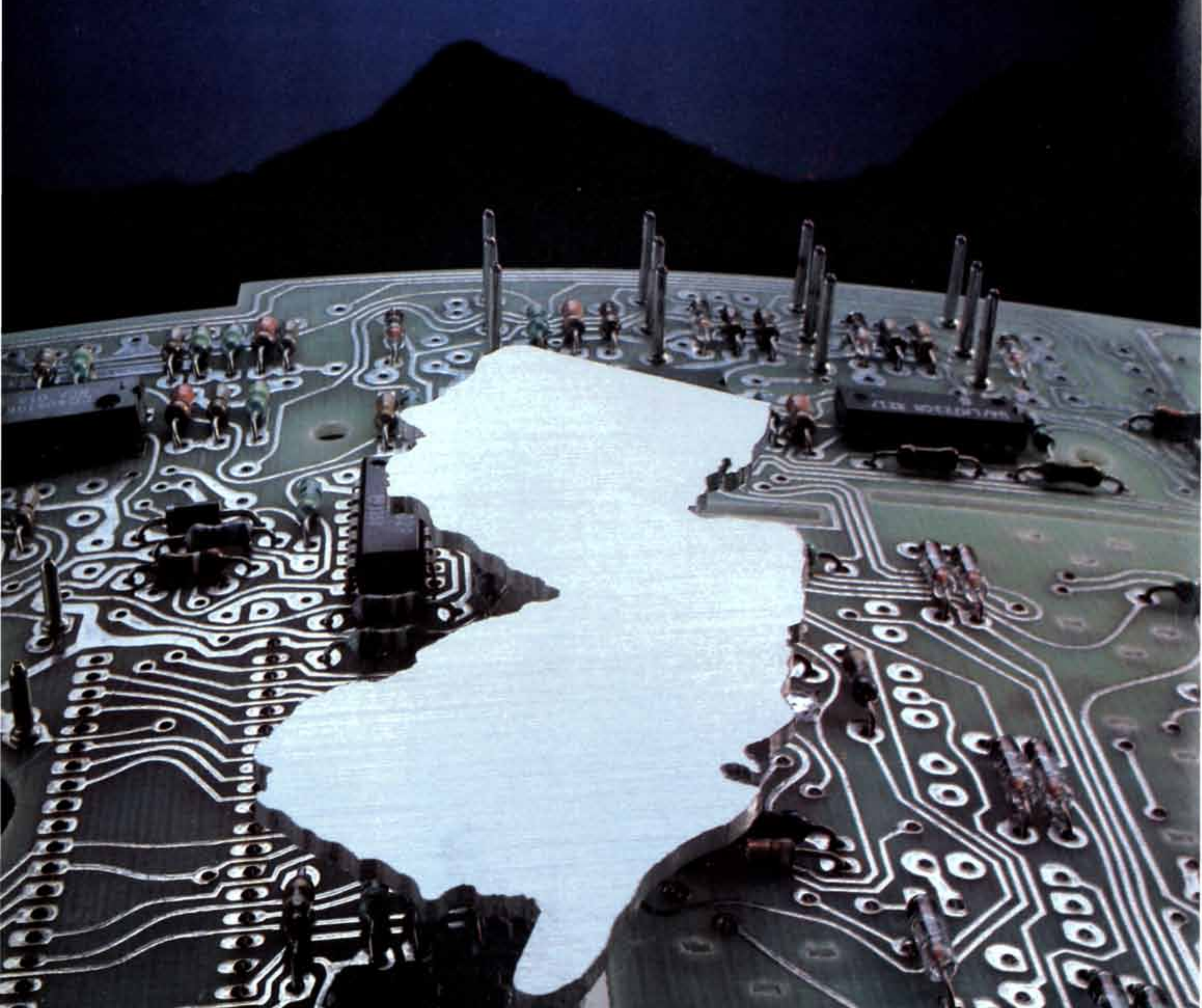
Because our universities and technical colleges turn out a plentiful and welcome supply of specialists in electronics engineering and computer sciences. Northern Ireland is an ideal place for any state-of-the-art company to introduce high technology.

Besides our skilled work force, other good reasons to invest in Northern Ireland are an enviable labor relations record, excellent R&D support programs, top-grade transportation and communications links for easy access to the Common Market and other parts of the world, and investment incentives which have been judged the best overall in the EEC.

If you have any doubts about how smoothly business continues to function today in Northern Ireland, ask anyone who's been here recently. Better yet, come see for yourself. For further information, contact Ian Walters or George Forster at the Industrial Development Board for Northern Ireland, New York office, 150 E. 58th St., New York, NY 10155. Or call (212) 593-2258.

Northern Ireland works.
A visit will convince you.

IDB Northern Ireland



BOOKS

Mineral origins of life, the senses, image processing, an ideal monastery, armaments

by Philip Morrison

GENETIC TAKEOVER AND THE MINERAL ORIGINS OF LIFE, by A. G. Cairns-Smith. Cambridge University Press ($29.95). Once upon a time there were those who held that a self-duplicating system might in principle lie beyond scientific understanding. A machine that could build a simple device, say a pair of scissors, is not too complex to conceive. If the machine were asked to build more complex mechanisms, however, it would need to be more complex than they are. A machine able to build itself, which plainly would be no simple mechanism, might then be a logical impossibility, a vicious regress into infinity.

It was John von Neumann, engaged by what he had newly come to expect of the digital computer, who before 1948 brilliantly countered that rather vague argument. The core of his account bears repeating: It is not hard to imagine a machine able to correctly assemble any finite set of parts drawn from a large store. The machine is complicated, to be sure, but it can be assembled out of the same parts. It builds by following taped instructions. Call it a universal tool machine, to which it is an analogue of sorts. Now code the recipe for the tool machine itself onto a long tape. Such a tape and tool machine can then together self-reproduce without logical troubles, given a few more items to allow copying and the like. The tape has taken over all the instructional complexity; it will be long indeed, but the tool machine has reached its plateau of complexity once it is able to carry out an arbitrary sequence of taped orders. It has only to work away for a long time.

We now recognize the profundity of von Neumann's logical example. It describes all life. What is the living cell but a long tape of DNA informing a tool machine? It is all constructed by linking rather simple monomers into long sequences that fold themselves into the molecular structures and catalysts of life. Replication proceeds exactly as envisioned, with the tape calling up a new tape and a new machine out of an existing machine that studiously scans the ancestral tape.

High on the agenda of molecular biology is an account of the origins of such a system. The problem stands unsolved. It is clear that some small length of early tape needs to call up a working tool, to make up another little segment of tape and so on. Here trouble comes. It is the very separation of DNA and the functioning of protein that lies at the heart of the von Neumann logic, and just that distinction makes an initial coupling hard to find. All the wonderful codes and adapters are fine once they are evolved but look implausible as early elements.

This innovative book by a Glasgow chemist expands on an idea he first put forward in 1966. Although the book (which is accessible, he says in his preface, "if you enjoy reading, say, *Scientific American*") does not mention the von Neumann argument, it flourishes within a gap in it, a gap not so much in logic as in the implications drawn from the strict original proposition. The universal tool machine is a nice contrivance if you expect to oversee the evolution of that intricately woven fabric of diversity that now is life on the earth. In the beginning, however, no such flexibility and versatility need be provided. It is enough to have some crude self-replicator, if only there is future opportunity of passing over to a von Neumann scheme, with its unendingly evolving recipes for varied actions that remain chemically distinct. At the outset no such clean distinction need have existed; the plan and the operating devices might have been much more alike.

The primary organisms might have been fabricated at a much coarser level. Today the DNA stores key instructions, not in one or two atoms per bit, presumably the limit of fine-grained instruction, but in molecular aggregates of hundreds of atoms. Early in the course of evolution the miniaturization may have been much less advanced. Today the organic monomers of life are not found outside the biosphere; originally the material basis of information storage for replication may have been inorganic, mineral, abundant in the earth before any life quickened. Today the key sequences are stored in a fundamentally one-dimensional array, the double helix of DNA, wonderfully spooled and reeled in the chromosome dance. In former days the store might have been two-dimensional, some defect pattern spread across a crystal surface layer, able to promote its own reproduction, albeit only approximately.

Dr. Cairns-Smith starts out with a few chapters in the vein of these arguments: definitions, general questions, even a pleasant dialogue in which he concedes that most of his colleagues in the field see his proposals as having "all the qualities of a bad theory"—easy and inviting reading. About half of the book explains the most unfamiliar and positive part of these ideas, the nature of clays, microcrystalline material that forms out of the weathering of common rocks in water solution. The pictures here are as instructive as the text; the fibers and wormlike lengths of many-layered silicate microcrystals offer a very different view of crystal growth. The point is made that in this domain it does not seem unlikely that defects or even accidental linkings of crystallites might confer a selective growth advantage. A simple scenario, taken as an example only, imagines special clays that might "swim upstream to lay their eggs." It is not absurd; consider that a particular configuration might grow well in some interstice in a wetted sandstone. It might turn flocculent after growth, to drift off and seed itself downstream normally. Downstream, however, the environment might change for the worse. Perhaps some of the clay complexes become flocculent once the concentration of sodium ions drops; those individuals break off only at that time, a clear sign of backflow, to seed more survivably in fresher waters upstream. Given self-reproduction, this seems a plausible example of how marginal natural selection might work on colloids.

After many ingenuities, particularly around the ways in which folding and twinning and surface contact might give rise to broadly inheritable surface pattern among crystallites, the book closes with a tale of revolution. Carbon atoms took over and in the end cut loose from their silicate substrate to establish cells. This part of the story is still more conjectural; it calls for detailed biochemistry. It nonetheless has one strength: the membranes of the living cell are a good deal more evocative of a legacy where they are coupled to a layer of clay than they are of any floating loop of DNA.

Right or wrong, this audacious, informal, instructive book is a scientific thriller. It ends in challenge: the answer to these questions will be given not by tighter argument but by finding out how to make some kinds of primary organisms in the laboratory. That result is not close, but it can be foreseen; the first hints of surface pattern reproduction in clays may already have been found, as long ago DNA was shown to replicate in the test tube. The parallel is cautionary; this is no fleeting puzzle but a riddle grander than the Sphinx ever put.

THE SENSES, edited by H. B. Barlow and J. D. Mollon. Cambridge University Press ($59.50). That the human senses are to be seen as physical instruments is an insight as familiar as it is profound, and the opening chapter of this compact and up-to-date summary directly applies the metaphor. Plainly it makes a difference whether the laboratory example is the camera obscura of the 17th-century philosophers or the digital computer with video input of today. Fourier synthesis and signal-to-noise ratios inform many chapters; they are elucidated graphically and well without formal mathematics. One experimental result put into this context speaks eloquently of unity. A histogram plots the number of nerve impulses recorded at synapses in the spinal cord but originating in a single sensory terminal in the skin. No stimulus was applied: the noisy nerve fiber fired anyhow, with a mean rate of six impulses per tenth of a second. A moderate stimulus doubled the rate of firing. In the two cases the observed firing rate fluctuated at random by about the same amount, plus or minus three pulses per tenth of a second. Thus both the signal and the signal-to-noise ratio doubled. The error rate for the stimulus would be about one in six times; for a judgment to be made at odds of one in 100 the signal would need to be twice again as strong. This is neuroengineering.

A recent experiment displays the response of a single retinal rod cell (it is one taken from a toad retina, an order of magnitude longer than a mammalian rod) to single photons. The current flow to the rod caused by the absorption of light by rhodopsin was recorded for 40 weak flashes. At the time of the flash the response duly appears, except that many bumps are missing. The Poisson statistics hold well because the events are single-quantum ones. Our own color-detecting retinal cones, much shorter and less absorbing than the color-blind rods, need many photons if significant current is to be seen above the background noise. Now, cone color response is shifted about 10 percent redward of that of the more sensitive rods; it is hard to resist the inference, although it is far from proved, that the higher thermal noise faced in the cones because of their lowered energy barrier has called for a reduction in sensitivity. Rods are fit for detecting faint images at the periphery; cones, for acute study when there are plenty of photons. The design of the retinal surface is understandable.

The eye is also, of course, a classical lens system; the most frequent clinical intervention is for lens improvement. The refractive model of the eye is here, and with it one sees a strict measure of performance. The measured spread of a bright-line image is not much broader than the diffraction-limited theoretical best to be had with a given pupil size and color of light. The eye, however, is no mere camera, not even one with a remarkable dual film response: the "sensitive long-latency, slow-exposure monochromatic system (the rods) with a less sensitive, shorter-latency, brief-exposure color system (the cones)," both speeding up automatically as the light level rises. The eye perpetually scans. That foveal spot offers five times the peripheral resolution over a central region only a few degrees across. We carry a detailed survey across the entire visual field quite unconsciously in fifth-of-a-second hops, neither aware of the motion nor bothered by the piecewise detail in the scene.

Such subtleties, along with stereo vision, edge and feature detection, and color perception constant under wild changes in the color of natural illumination, belong to higher centers. Some lie in the retinal interconnections, some close to the cortex itself. There is the computer hard at work in us all, open to laboratory study only since that metaphor itself became realized within the past generation.

About half of this unusually well-written and well-coordinated account by a dozen British research workers (plus one Michigander) is spent on the eye. Another fourth of the volume deals with hearing, presenting a good mix of the familiar with the brand-new. Nothing is said explicitly of hints of regenerative circuitry in the cochlea, although the unexpectedly sharp tuning of that wet system draws attention.

Last and least, although perhaps freshest, is the treatment of the other senses: touch in skin and viscera, the system of the inner ear that is our version of inertial guidance, the chemical channels of smell and taste. There is good evidence of a descending control outward from the brain along the pathways usual for inflowing pain. Biochemical study is rapidly identifying specific chemical mediators and antagonists in service throughout this intricate alarm system.

One remarkable result almost a decade old is the induction of vertigo by drinking heavy water. The interpretation is that the stuff entering the fluid of the semicircular canals increases the fluid density. The slight mismatch in buoyancy then makes the delicately wired diaphragm that blocks the canal sensitive to its position under gravity instead of to motion alone, the normal state. A stagger results, akin to that proverbial after unwise intake of alcohol but with the apparent turning being in the opposite direction. Alcohol may act merely to lessen the density of the canal fluids. The result seems almost too pat to be true, even though it does account for such a specific effect from such a simple substance as ethyl alcohol.

Throughout the mind's instruction is under study with oscilloscopes and signal generators; issues of nature and nurture, of perception and cognition arise almost as naturally in the laboratory as in the physiological details cited. The topic of this contemporary, ambitious and reflective treatment transcends medicine.

DIGITAL IMAGE PROCESSING: A SYSTEMS APPROACH, by William B. Green. Van Nostrand Reinhold Company Inc. ($34.50). It is nothing new to regard an image as a large array of numbers. But this book, quite accessible to a general reader with a limited tolerance for algebraic equations, provides an up-to-date overview of the consequences of regarding that array as an invitation to manipulate. First the artists, then the cartographers, then the photographers and now the computer engineers have used the image as raw material. Of course, eye and brain have always done no less, so to speak in secret, but it takes the digital computer, nothing if not explicit, to make clear how much everything we see is a piece of applied epistemology.

Two fine pictures of crescent Mars face each other, products of the cameras of fast-approaching *Viking Orbiter 2*. One shows the Martian polar cap white below a beige quarter planet mottled with detail. The computer has already been at work. As the spacecraft moved in, three successive views were geometrically registered in spite of steady changes in size and aspect. The accuracy of fit is about as good as possible, to a single picture element, or statistically even better. The technique is powerful; several small features common to all the images are chosen as reference points. The statistical correlation between the numerical images is then computed for a spread of trial alignments in the approximate neighborhood of the recognizable feature. The highest correlation is chosen; repeating the procedure for several tie points, a transformation can be numerically defined to register the distinct images sometimes with "subpixel alignment."

Each of the three Martian views was shot through a different color filter. The camera response was known to be distorting. It had been calibrated by recording a uniform field at different intensities. The corrections were then applied to each image, to yield a set of three single-color images. These were then extrapolated to provide red, green and blue images with bandwidths matched to the requirement of the color film.

So far the task has been remedial. Now for enhancement. The Vidicon camera has collected close to a million numbers per image, each number in the eight-bit range from 1 to 256. The computer takes a census. The head count by intensity of pixel in each of the three colors is graphed right on the margin of the picture. These image histograms are basic tools for the work. The three bland profiles, however, do not mean much; we can see no easy outcome. We need a theory. It is the standard theory of three-color reproduction. The three intensities at one pixel—red, green and blue—define a point in the familiar curvilinear color triangle and also a total intensity.

Now the color space can be represented anew. The three coordinates yield one number to represent intensity, a second to represent the hue (regarded as an angle turning from the green apex around through the other primaries) and a third to represent saturation, increasing with distance out from the white point within the color triangle. These three numbers can be computed for each pixel, and again a census is reported. One stretches the intensity distribution in a simple linear way, to take full advantage of the film latitude, in this case no large effect. The hue and saturation populations are left unmodified.

The result is "the best true-color rendition of the planet that can be achieved" with the system: not much color. The facing page shows the same scene, the same size, the same features. A watercolorist, we imagine, has been at work. Subtle reds and even some greenish tints have been added tastefully to enhance the interest of the beige ball. That was not, of course, the origin of the comparison image. It is a controlled, objective product, even though it looks like a painting. Its hue distribution was spread out by a factor of seven and its saturations by a factor of three, holding their average values fixed. This is the Martian surface seen by an eye of superhuman abilities; color nuances have become bold.

Most of the technical volumes in this field concentrate on the mathematical aspects of the algorithms that can economically carry out such schemes on pixels by the million. William Green has instead chosen to tell this tale in the large, emphasizing the chain of operations, the techniques in general, the choices and the tradeoffs. Illustrative comparisons such as the one outlined above abound, and with them the block diagrams of the systems required. Green summarizes the novel camera-equivalent devices—in satellites, in the medical center and in astronomy—that acquire the pictures, the schemes by which the pictures can be digitized, the growing technology of easy display and access to images in enormous numbers, and the all-important details of their management, such as careful labeling and filing schemes of high flexibility. Images so elaborately treated are misleading without their clinical history, unless long familiarity and "ground truth" are at hand to provide the interpretation out of experience itself.

Many other paths to enhancement of image are described, including more subjective stretches of iterated contrast and the simple color substitutions that are now familiar, particularly in the pages of *Scientific American*. There are new experiments with thematic maps, where a space is defined, say, by the colors of known crops and each pixel is given a coded color in accordance with some measure of its position in the color space. The result looks like a colorful map of remarkable detail, and it would not mislead a canny viewer as the processed photograph can. We recognize at once that maps are partial representations based on understandings, that is, on theories, although we have not quite learned that no image at all, not even the field of view of every waking moment, is in fact much different.

The references here include books and the periodical literature, as is usual. They extend, however, to rich sources of images of earth and sky and to sources of software for image processing, available from a NASA agency at the University of Georgia called COSMIC. Green, a senior figure in the field at no advanced age, "was fortunate to spend ten years at [the Jet Propulsion Laboratory] working in the finest image processing facility in the world." The examples reproduced in the book support his claim, although we may wonder about the inaccessible contractors to the National Reconnaissance Office.

THE PLAN OF ST. GALL: IN BRIEF, by Lorna Price. University of California Press ($55). Some three years ago there appeared a work of intellectual and visual reconstruction as beautiful as it was learned, a "study of the architecture and economy of, and life in, a paradigmatic Carolingian monastery." Its devoted authors, Walter Horn and Ernest Born, had spent some 15 years at the task, and they were able to express their findings in a form worthy of the scholarship, a classic of our day, acclaimed for its encyclopedic richness no less than for its perfection as a printed and illustrated book.

At the same page size, but only a tenth as thick, this new book presents (at roughly a fifth the price) a sample and a summary of the parent work, all in vermilion, black and parchment. The author was the editor of the main volumes, and her explanatory text is set in a design by Professor Born.

What is here is perhaps best described as a chain of projections linked across 11 centuries. In about A.D. 820, when Louis the Pious held the throne of his father Charlemagne, a wonderful plan was made in his palace at Aachen for the ideal structure of a monastery of some 300 souls devoted, under the rule of St. Benedict, to the work of God. It honored and framed the work of the hands—and of hens and oxen and water mill too—that jointly sustain the Opus Dei, under abbot and king. That original plan is lost; what we hold is a single work of art, done in red on a pieced-together parchment almost a meter square, devotedly traced from the original by two men who worked in the scriptorium of Reichenau before A.D. 830. Since then the copy has been held in the monastic library of St. Gall, above Lake Constance. At about the end of the 12th century an unknown man there wrote on the blank side his pious life of St. Martin; he must have admired the old plan, because he was careful not to cut the parchment, although he erased one of the 40 ground plans it bore.

Then the document vanished, although it was still safe on the shelf because it was catalogued as a life of the saint. In 1604 it was returned to the stream of knowledge, its hundreds of detailed legends published without any drawings. In the mid-19th century scholarship touched the plan of St. Gall, but not until 1952 was a facsimile published. As a part of the observation of the 1,200th anniversary of the Carolingian reign the two Berkeley scholars were charged with a study of the plan. Their massive work emerged in time to universal praise; this, its "overview," preserves beautifully a terse narrative and many illustrations prepared for an exhibit of their findings, which will tour a dozen American galleries between 1983 and 1985.

Two major tasks sprang from the plan: to project a third dimension out of each of the careful ground plans neatly composed in elegant rectangularity and to discover the size and scale the planners had intended. The shelters, gardens, bathhouses, cloisters, infirmaries and cemeteries were given full form by the intensive use of comparisons, often archaeological ones. The Church of the Plan, never built, is rendered in quiet nobility, along with its cloister and refectory. The functional structures were not modeled on the classical; indeed, they "descend from a house type" that was common among the Frankish people. The present St. Gall monastery is only of minor help, because there "no stone of Carolingian origin stands today above ground level." Fires, enlargements and the sacking of the site in 1712 have disturbed the signal in the stone. In fact, although Abbot Gozbert depended on the idealized plan as a guide to his A.D. 830 rebuilding of the monastery, then rather seedy, the hilly site could not fully accommodate the paradigm; the adjustments were major.

Searching out the scale revealed a surprise: the design was governed by an unexpectedly harmonious system. The architects followed a rigorous "progressive dichotomy," the modules they employed spanning no less than eight powers of two from great nave to burial plot. Among all these consistent four-square relations some circles appear: the two apses of the church are semicircles, and its two high detached circular towers, without bells, are imposing announcements of "a Fortress of God." A couple of 40-foot circular hen houses, an "exemplary solution" for a busy community that daily needed the eggs of hundreds of birds, are the least expected element of this vivifying display, an extraordinary medieval union of the workaday and the exalted.

WHAT KINDS OF GUNS ARE THEY BUYING FOR YOUR BUTTER? A BEGINNER'S GUIDE TO DEFENSE, WEAPONRY, AND MILITARY SPENDING, by Sheila Tobias, Peter Goudinoff, Stefan Leader and Shelah Leader. William Morrow and Company, Inc. ($15.95). HOW TO MAKE WAR: A COMPREHENSIVE GUIDE TO MODERN WARFARE, by James F. Dunnigan. William Morrow and Company, Inc. ($14.50). In the guns-or-butter title of this pair an academic team has assembled a well-written nonmathematical primer of the nature of modern war, both nuclear and conventional, with such matter-of-fact grimnesses as the following. Old, densely built cities, particularly if there is much wood in them, support fire storms, given a large enough initiating source. That is what Hamburg, Dresden and Tokyo shared, to say nothing of Hiroshima. American suburbs, however, are built with material an order of magnitude less combustible, only two pounds per square foot; they will not sustain a fire storm. Nevertheless, "older American cities, [such as] Baltimore and Philadelphia, would surely burn."

The bulk of the argument is presented in historical context, with a chapter that starts with sticks and slings, another that carefully traces the rise of nuclear weapons and their strategies, and another on efforts toward arms control. A long chapter treats of today's conventional weapons, from rifles to aircraft carriers. The text is narrative and informal, and the authors have quite successfully managed to dispel jargon, although the reader will learn what an M-16 rifle and a P-3 marine patrol aircraft do (and look like). Simple description fills the symbols with content, but it is not easy to talk about such organisms without their names. (An unusually "long aside" takes about seven rather lighthearted pages to give some rationale for Pentagonal nomenclature, and there is a longer annotated vocabulary of these terms of art as well.)

If half of the book is military history and technology, half is rather cool politics: structure. The defense experts, the military, the Congress, the defense contractors great and small (the great ones are listed) are paraded in review, both as individuals and as classes, in a persuasive chronicle of how the fast-growing 60 billions or so spent on weapons procurement annually (the figure is up again since the book was written in 1982) are claimed and used, in "the single most complex process by which policy is made in our society."

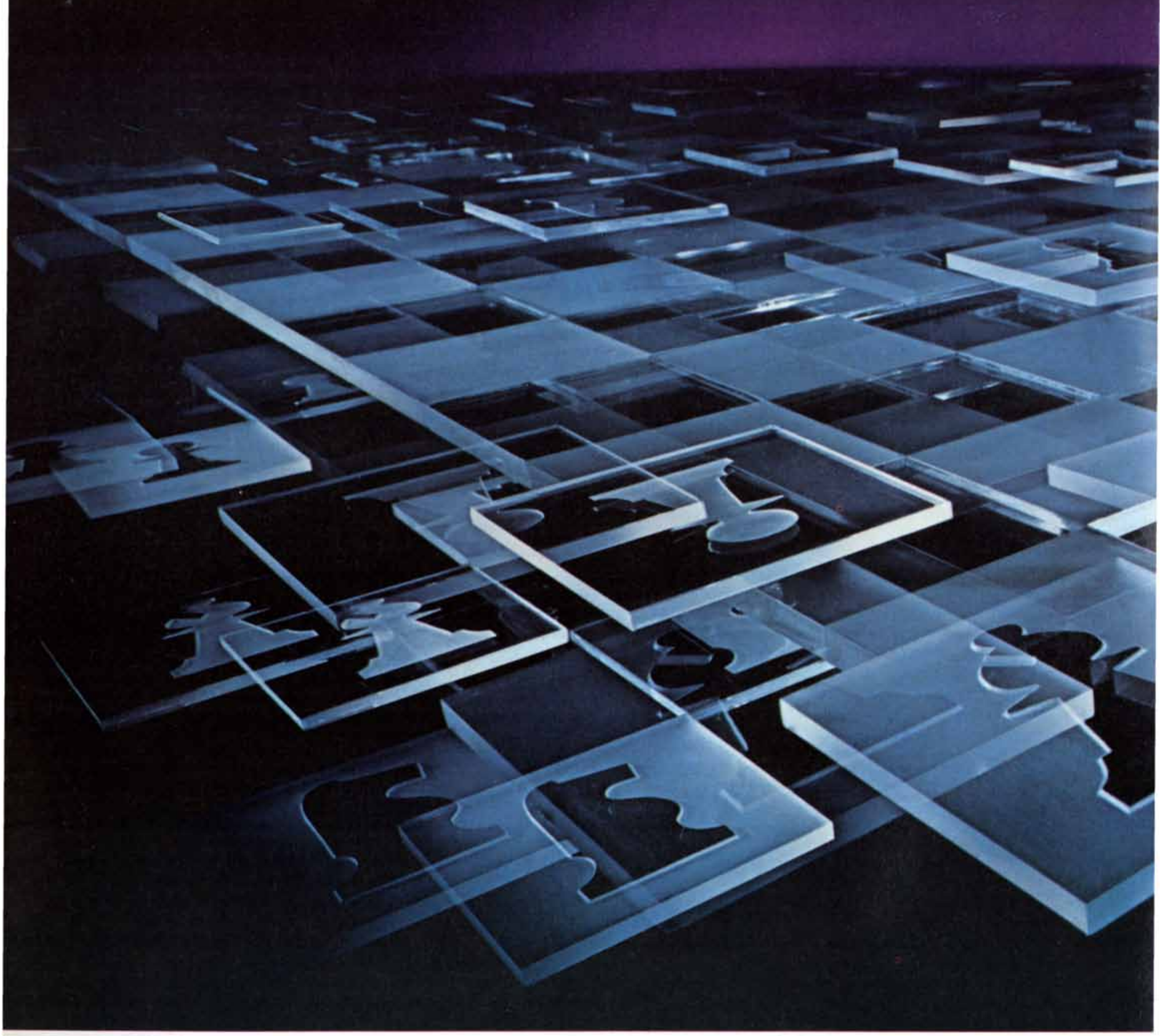
A similar tale is told for arms control—a shorter story—and the book ends with a look at issues of the decade: NATO, a bigger navy, the Rapid Deployment Force, the embattled MX, high-tech v. low, and more. The volume shares the breezy tone of its title, but it is in fact an extensive and serious answer, evenhanded overall, although the authors are plainly engaged and here and there show a claw, if not a tooth. The Mine Warfare Command, they reflect, is starved for skilled people. "It lacks glamour. You simply plant it in the water and go away. It does its work in your absence. No glory there." Over a wide span of material the book is accurate; here and there a technical misunderstanding creeps in, as it does for the beginnings of thermonuclear weapons in the two superpowers. This team of authors nonetheless knows how to inform serious beginners. "We don't want to tell you what to think," they write, "but what to think about."

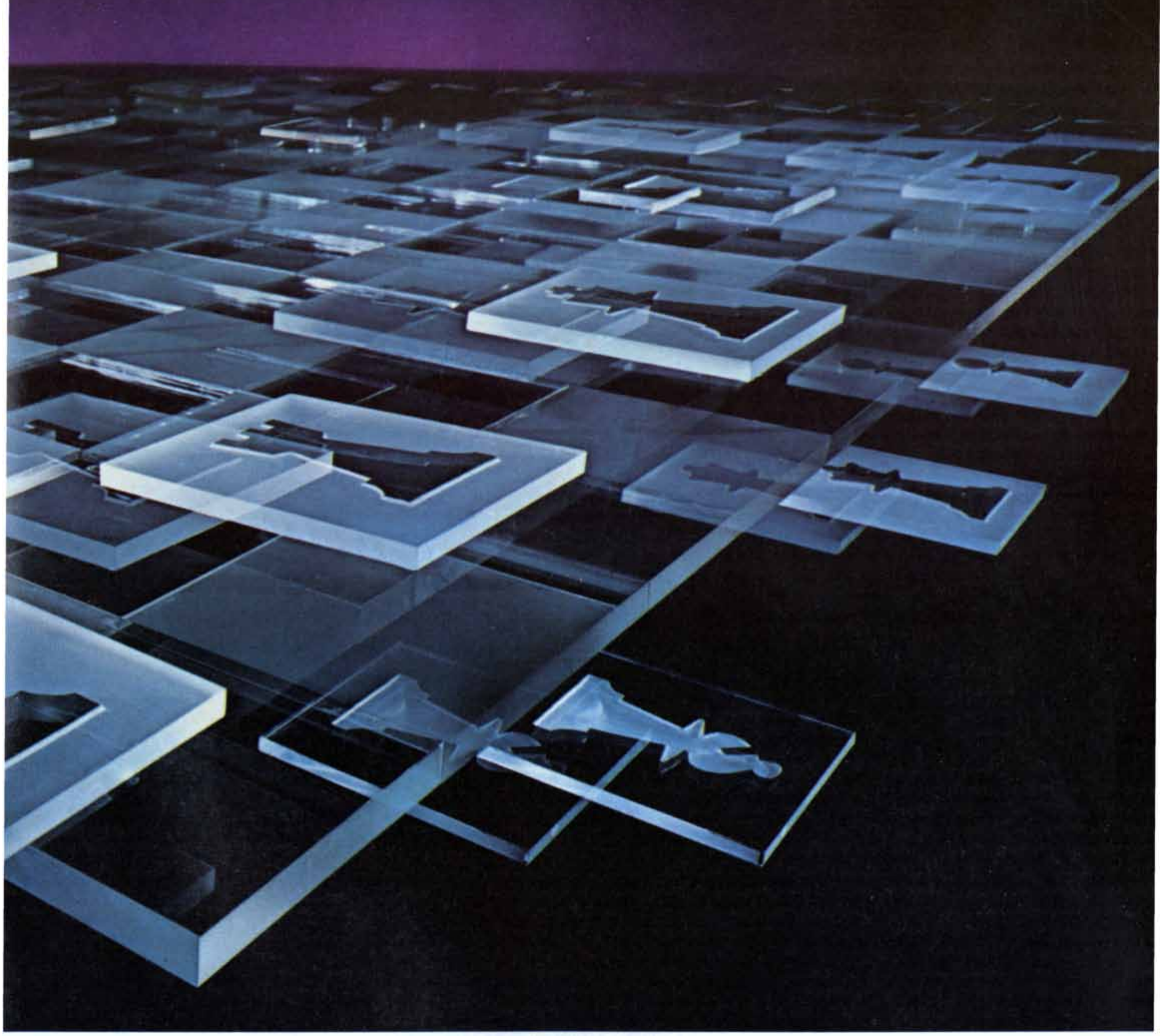
How to Make War, the second of these books, is again an overview of war and weapons today, with a title that carries the full flavor of the somewhat ironic text. James Dunnigan's vantage is a unique one. He makes war, professionally, not in real life but in simulation, as the author and publisher of games and simulations of combat and campaigns through history. His aim is an appraisal of the means and nature of war today, and he treats explicitly those high adversaries, the U.S. and the U.S.S.R., with the other nations at least sketched in.

The book is really an enfolding of two works, both first-rate for readers with some experience. About a third of it consists of a long series of tables of data (typographically hard to read) along with terse, defining glosses and enlargement. The dozen-odd tables include the expected lists of the characteristics of, say, battle tanks (the Russian tanks are in "massive numbers . . . effective yet expendable") or strategic weapons, complete with kill factors. They include as well the supply needs of divisions and of aircraft units, the estimated combat losses according to branch of service, the ground density of personnel in units of the American and Russian forces, and the itemized costs of raising an infantry division. These tables surely derive from the handbooks available for the war games of the author. They are not documented or even dated; they appear to be of the epoch 1980, conscientiously compiled from annuals and field manuals.

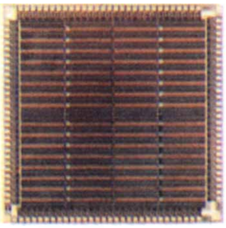
The remaining text is an informal, personal ("Imagine yourself in a suburban town that has been abandoned . . . ") set of brief background essays on war by land, air and sea, on war with special weapons, on logistics and transport and on the human factors less easy to enumerate. The essays read well; the informality and the point of view are reminiscent of the reflective talk of soldiers. They do not neglect the dilemmas of the discipline of making war. Technical analysis and the lessons of history alike claim audience, although they frequently conflict. Only a few pages are spent on the choices of lawmakers and citizens; the book views a soldier's world, with the forces, enemies and allies given, but it allows for change.

Out of all this observation a few conclusions are presented rather as dry aphorisms, although they are well if implicitly argued. One year of a major conventional war as now planned would cost the U.S. more than $1 trillion. Most of this is the cost of air war above land and sea, of which most is the cost of current "smart" munitions. Production would need to rise thirtyfold over a couple of years to make it all possible. A short war is the planners' hope—perhaps the planners' myth. On naval warfare: "The only thing that will defeat the West at sea is complacency or bankruptcy from building too many ships." On nuclear doctrine: "Both Russia and the United States have about ten times more nuclear warheads than they need to guarantee mutual destruction. In such a situation, nuclear 'tactics' becomes irrelevant."





The engineers work interactively with the system, assigning logical functions to devices on the chip. Then, using software, they route conductors—deposited strips of metal—to connect the logic gates. If a pathway becomes too crowded, another layout is automatically provided. And another. Until they find the series of moves that works.



This Hughes integrated circuit is a high-density gate array with two-level metal interconnections. Actual size is only 0.3 inches square.

They can think about the problem, and not worry about the computer, with microcode-assisted IBM software called VM/CMS. "Our people find VM ideal for interactive computing," Tien says. "As many as 40 design engineers make demanding use of the 4341 at once. Response time is excellent. The full-screen editor speeds up programming. And the executive language is simple, yet powerful."

The 4341 is an excellent example of IBM technological leadership. Internal data paths, and the arithmetic and logic unit, are 64-bits wide, built of 64K-bit chips that IBM has been mass producing since 1978. Real memory goes to 16 megabytes. The multiple, semi-autonomous processing units use high-density, large-scale-integration technology.

IBM offers engineering and scientific users extensive support: consultants, education and access to professionals. Tap into our many years of experience.

For an informative brochure, write Dr. Jack W. Hugus, IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, NY 10604.

IBM

Skipper Dennis Conner and his America's Cup crew may rule the Seven Seas...but nothing covers the land in between like the Jeep® Wagoneer Limited. And that's why Wagoneer has been made the Official Vehicle of the America's Cup defenders.

Like the boats they sail, the Wagoneer was designed to perform to the limit, even under adverse conditions. They count on the security of Wagoneer's remarkable Selec-Trac 2-wheel/4-wheel drive system - especially when the dry land is anything but dry.

It's also a comfort to know that in a Wagoneer Limited the classic luxuries are standard. From its genuine leather upholstery to its premium sound system, quality is never compromised.

And because it does so many things without compromise, the Jeep Wagoneer Limited has developed a singular reputation as the perfect vehicle for that *other* quarter of the earth's surface.

Jeep® Wagoneer Limited. The Ultimate Wagon.

FROM AMERICAN MOTORS CORP.

Established 1845 SCIENTIFIC AMERICAN May 1983 Volume 248 Number 5

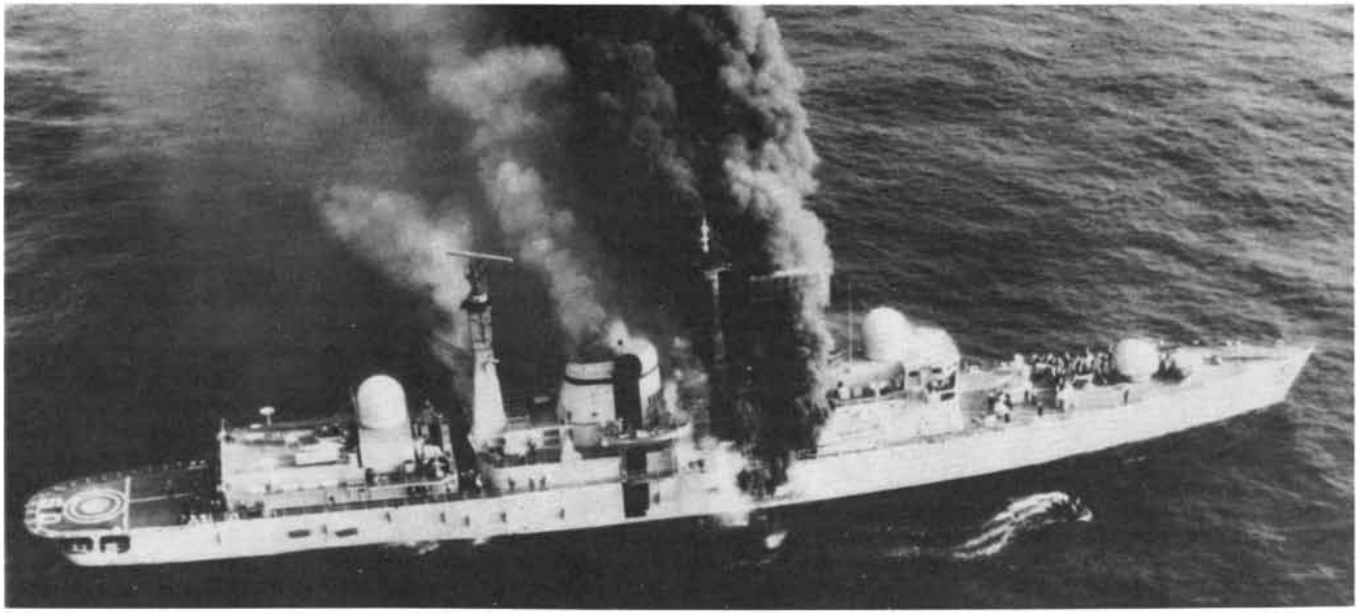
Smart Weapons in Naval Warfare

The impressive performance of precision-guided weapons in the Falkland Islands war calls into question current plans to add more large, expensive surface ships to the U.S. fleet

by Paul F. Walker

Last year's war between Britain and Argentina over control of the Falkland Islands was viewed with particular interest by military planners and others concerned with evaluating the effect of modern precision-guided weapons on the future of naval warfare. Of the 114 aircraft and 10 ships lost by both sides during the two-month conflict more than half were put out of action by such "smart" weapons. Included in the losses were 61 Argentine aircraft shot down by surface-to-air or air-to-air missiles, three British ships destroyed by air-to-ship or surface-to-ship missiles and one Argentine ship (the superannuated heavy cruiser *General Belgrano*) sunk by submarine-launched, acoustic-homing torpedoes. In addition nine British aircraft, 44 Argentine aircraft, three British ships and three Argentine ships were destroyed by more conventional means, such as gunfire, aerial bombing and accidents.

From the viewpoint of naval force planning the most significant loss was that of the British destroyer H.M.S. *Sheffield*. The *Sheffield* was hit shortly after 2:00 P.M. on May 4 by a French-built Exocet missile launched from a distance of about 20 miles by an Argentine navy Super Étendard fighter-bomber (also made in France). At the time of the attack the ship was on picket duty southeast of the Falklands and just west of the main British fleet. The sea-skimming, radar-homing missile struck the *Sheffield* amidships about six feet above the waterline, starting a fire that quickly disabled most of the destroyer's operating systems, including its fire-fighting apparatus. Five hours after the attack,

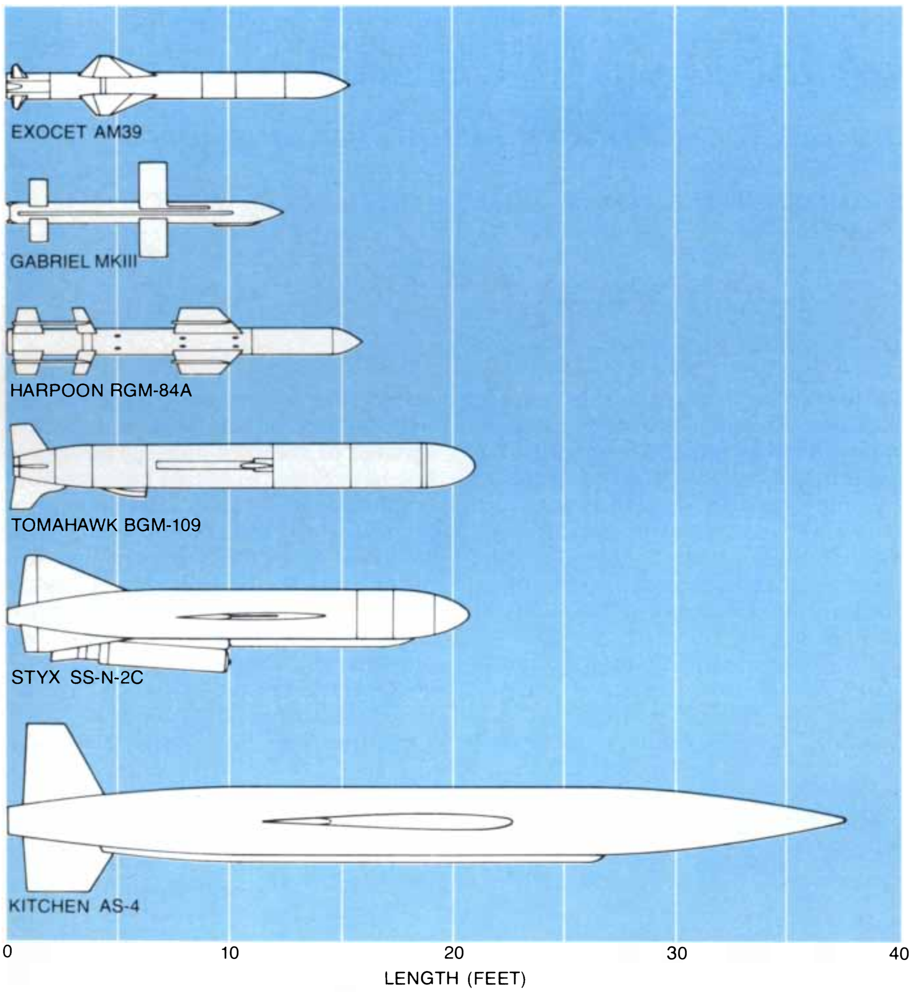


H.M.S. "SHEFFIELD" BURNS UNCONTROLLABLY after being struck by an air-launched Exocet missile during last year's Falkland Islands war. The comparatively small, sea-skimming, radar-homing "smart" missile was fired from a distance of about 20 miles by a low-flying Argentine Super Étendard fighter-bomber. (Both the missile and the aircraft were made in France.) The 4,000-ton British Type 42 destroyer, which was equipped with Sea Dart antiaircraft missiles and other modern fleet-defense systems, was hit amidships about six feet above the waterline. This aerial photograph was made before the order was given to abandon ship. Crew members not involved in fighting the fire can be seen standing on the fore and aft decks. A small inflatable launch moves away from the ship after coming alongside for a close inspection of the damage. The burned-out hulk of the *Sheffield* was later scuttled by the British. Twenty sailors died in the attack.

with the flames spreading dangerously close to the ship's ammunition stores, the order was given to abandon ship. The burned-out hulk was later scuttled after an unsuccessful attempt was made to tow it from the battle scene. Twenty British sailors died in the attack.

The destruction of the *Sheffield* was an impressive display of the cost-effectiveness of modern precision-guided weapons. The *Sheffield* was a highly rated, 4,000-ton destroyer, built about 10 years ago at a cost of approximately $50 million. It was armed with some of the most advanced defensive systems available anywhere today. Yet it was defeated in one stroke by a comparatively small, medium-range missile costing a few hundred thousand dollars.

What lesson, if any, can be drawn from this incident? Does the advent of smart antiship missiles such as the Exocet signal the end of an era in naval warfare, as some have suggested? Are large, expensive surface warships now obsolete—easy targets for a new generation of cheap but deadly precision-guided weapons fired from aircraft, small patrol boats, submarines or shore batteries? Or, as others have argued, are there equally sophisticated defensive measures in the offing that could make such an incident the exception rather than the rule? A review of the technological basis of this debate could not be more timely, in view of the current plans of the Reagan Administration to enlarge the U.S. fleet by more than 100 new warships, including two nuclear-powered aircraft carriers projected to cost more than $3.5 billion each.



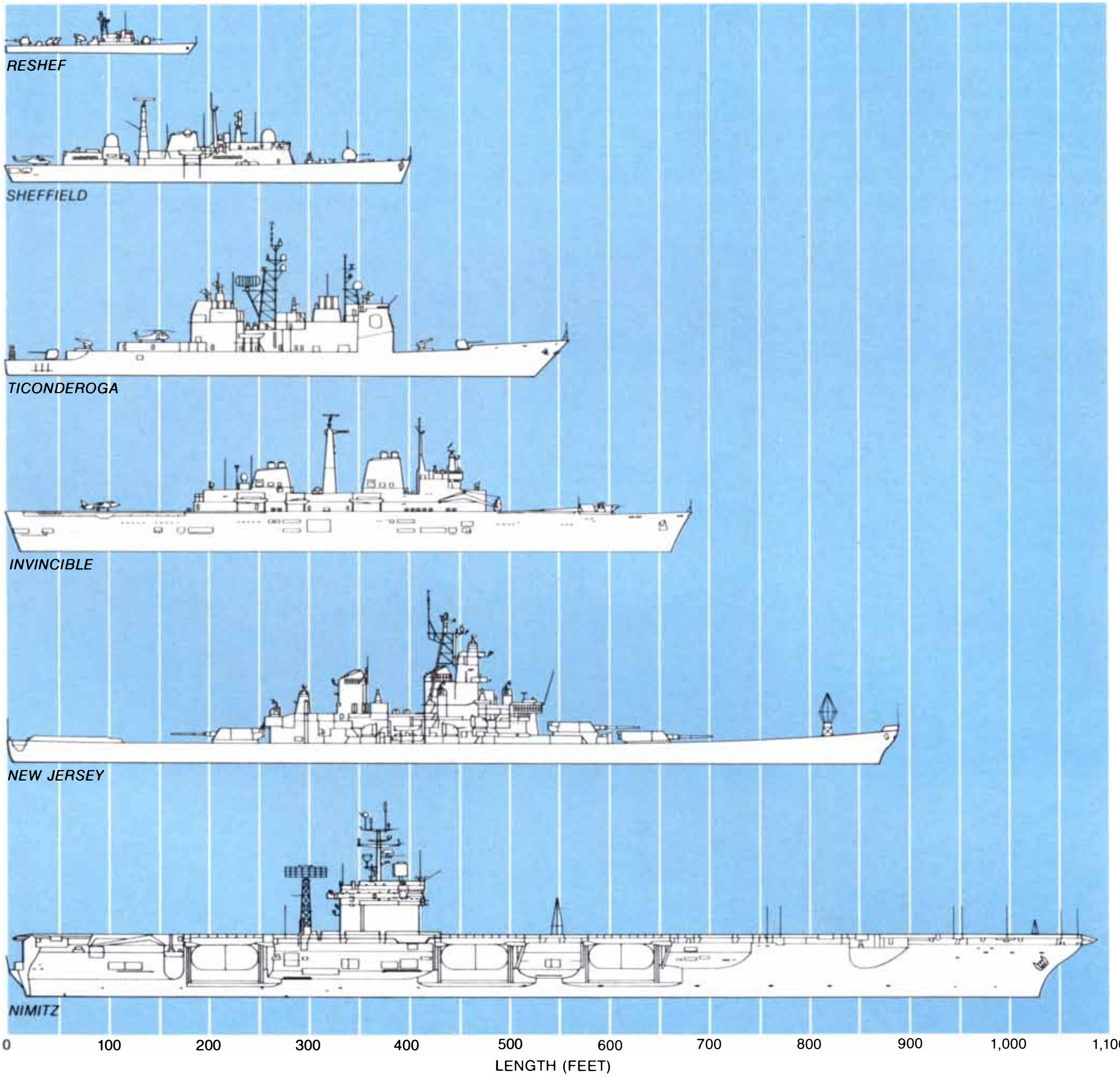
ANTISHIP MISSILE	YEAR DEPLOYED	RANGE (KILOMETERS)	SPEED (MACH NUMBER)	TYPE OF WARHEAD	ESTIMATED COST (1982 DOLLARS)
EXOCET AM 39 (FRANCE)	1979	50+	.93	NON-NUCLEAR	?
GABRIEL MKIII (ISRAEL)	1979	60+	.73	NON-NUCLEAR	$400,000
HARPOON RGM-84A (U.S.)	1977	90	.90	NUCLEAR OR NON-NUCLEAR	$1 MILLION
TOMAHAWK BGM-109 (U.S.)	1982	450	.74	NUCLEAR OR NON-NUCLEAR	$3 MILLION
STYX SS-N-2C (U.S.S.R.)	1965(?)	80	.90	NON-NUCLEAR	?
KITCHEN AS-4 (U.S.S.R.)	1962	300+	2+	NUCLEAR OR NON-NUCLEAR	?

REPRESENTATIVE ANTISHIP MISSILES are drawn to scale in the illustration at the top and are characterized in the table at the bottom. Missiles such as the Exocet are designed to be launched from airplanes, helicopters, ships and land-based installations; the particular model shown here, the air-launched Exocet AM 39, is the one that destroyed the *Sheffield*. The Russian ship-to-ship missile referred to by Western intelligence sources as the Styx was used by the Egyptian navy in 1967 to sink the Israeli destroyer *Elath*. The large Russian missile designated AS-4 is carried by the Backfire bomber; it can be armed with either a conventional high-explosive warhead or a nuclear warhead. Mach numbers are multiples of speed of sound.

The Exocet missile is described by its manufacturer, the French company Société Nationale Industrielle Aérospatiale, as an extremely reliable weapon, "capable of destroying any type of surface vessel" with a "kill probability" of more than 90 percent. Compared with other missiles of its type the Exocet is not very big; it is about 15 feet long and weighs 1,500 pounds, including its 350-pound high-explosive warhead and a full load of fuel. The missile is powered by a two-stage solid-fuel rocket and has a maximum range of between 25 and 40 miles, depending on how it is launched. Different models are designed to be launched from airplanes, helicopters, ships and land-based installations. The Exocet was introduced in 1972 and has been operational in the French navy since 1978; it has also been acquired by at least 17 other nations. So far it has been mounted on more than 180 ships, including the British navy's newest frigate, H.M.S. *Beaver* (which was not ready in time for the Falklands war).

In a typical air launch of a smart antiship missile such as the Exocet the pilot of the attacking aircraft flies low over the water under the defender's radar coverage until it is time for the missile to be launched. The aircraft then "pops up" briefly to get a bearing on the target, dropping below radar coverage again as the coordinates of the target are fixed in the missile's microelectronic memory. The missile is released in a "fire and forget" mode, flying the last leg of its journey under its own power and for the most part under the control of its own inertial-guidance system and radio altimeter; the latter device enables the missile to skim the surface of the water at a height of a few feet. As the missile approaches the target it rises slightly to scan the horizon and lock its terminal homing radar onto the target. The missile's warhead is designed to first penetrate a ship's hull and then detonate.

The Exocet's target in this case, the



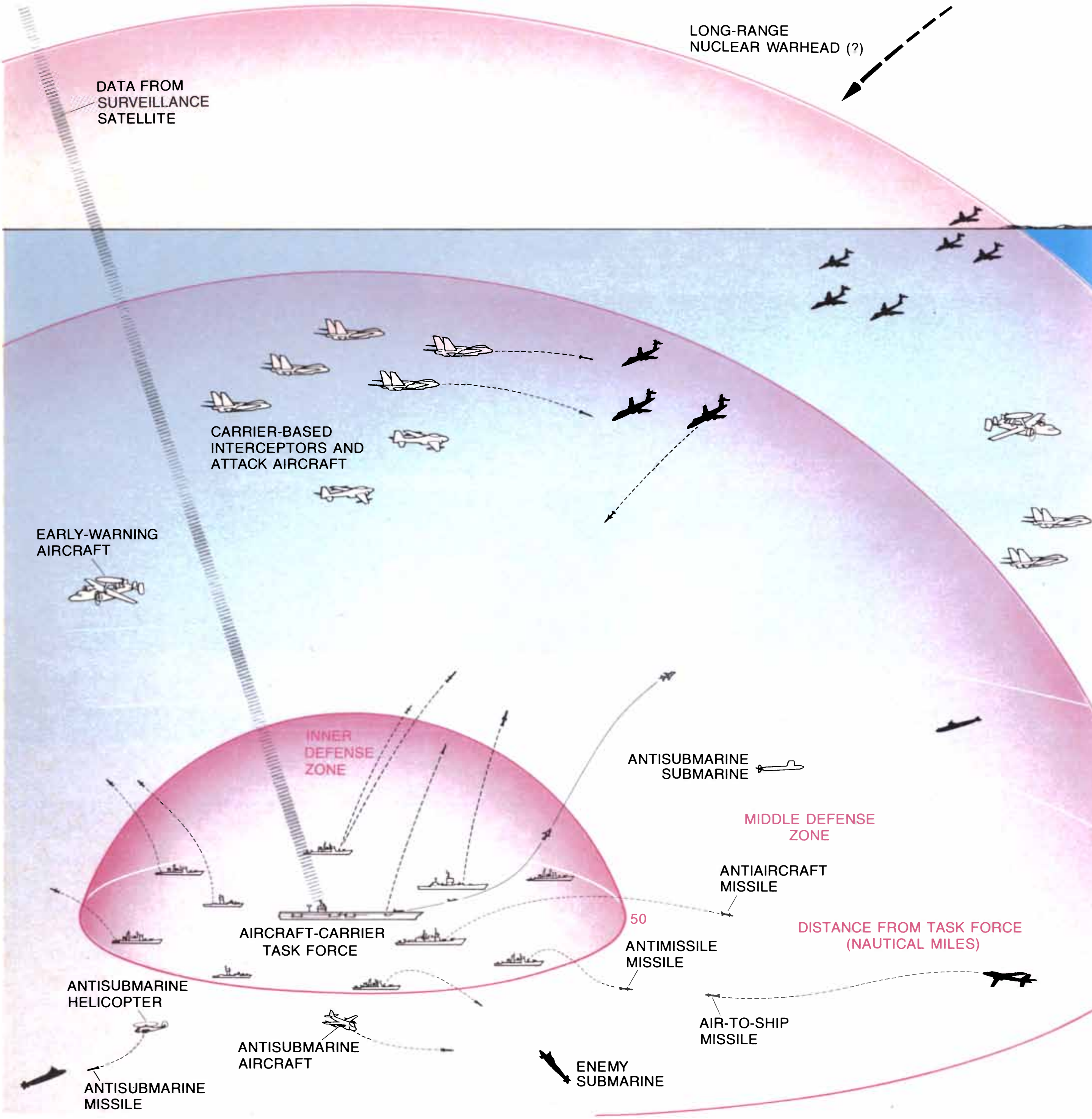
SHIP	YEAR DEPLOYED	DISPLACEMENT (TONS)	CREW	GUNS	MISSILES	AIRCRAFT	ESTIMATED COST (1982 DOLLARS)
RESHEF (ISRAEL)	1973	450	45	2 76 MM. 2 20 MM.	5 GABRIEL 4 HARPOON	0	$50 MILLION
SHEFFIELD (BRITAIN)	1975	4,000	270	1 115 MM. 2 20 MM.	20 SEA DART	1 HELICOPTER	$150 MILLION
TICONDEROGA (U.S.)	1983	9,000	330	2 127 MM. 2 20 MM.	68 STANDARD 8 HARPOON 20 ASROC	2 HELICOPTERS	$1 BILLION
INVINCIBLE (BRITAIN)	1980	20,000	1,300	0	22 SEA DART	5+ AIRPLANES 9 HELICOPTERS	$525 MILLION
NEW JERSEY (U.S.)	1943 (REACTIVATED 1983)	58,000	1,600	9 406 MM. 12 127 MM. 4 20 MM.	16 HARPOON 32 TOMAHAWK	4 HELICOPTERS	$326 MILLION (MODERNIZATION ONLY)
NIMITZ (U.S.)	1975	91,000	6,300	3 20 MM.	24 SEA SPARROW	80+ AIRPLANES 10+ HELICOPTERS	$3.5 BILLION

REPRESENTATIVE WARSHIPS are depicted at the top and characterized at the bottom. The Israeli fast-attack craft of the *Reshef* class are typical of the hundreds of small missile-launching vessels that are being introduced into many of the world's navies. The U.S.S. *Ticonderoga* is the first of 17 advanced fleet-defense cruisers planned for the U.S. Navy. "Jump jet" aircraft carriers such as the *Invincible* have replaced large-deck carriers in the British fleet. The battleship U.S.S. *New Jersey,* which was built during World War II, was reactivated earlier this year. The Reagan Administration currently plans to build at least two more large aircraft carriers of the *Nimitz* class.

Sheffield, was one of 10 Type 42 destroyers built over the past decade for service with the British North Atlantic fleet. It and its sister ships were designed to serve as launching platforms for the new Sea Dart antiaircraft missile; their primary mission is to provide air defense for the naval forces of the North Atlantic Treaty Organization (NATO). Each Type 42 destroyer is outfitted with 20 Sea Dart missiles, a Vickers 115-millimeter automatic gun on the foredeck, two Oerlikon 20-millimeter guns aft and a helicopter armed with four Sea Skua antiship missiles on the rear deck.

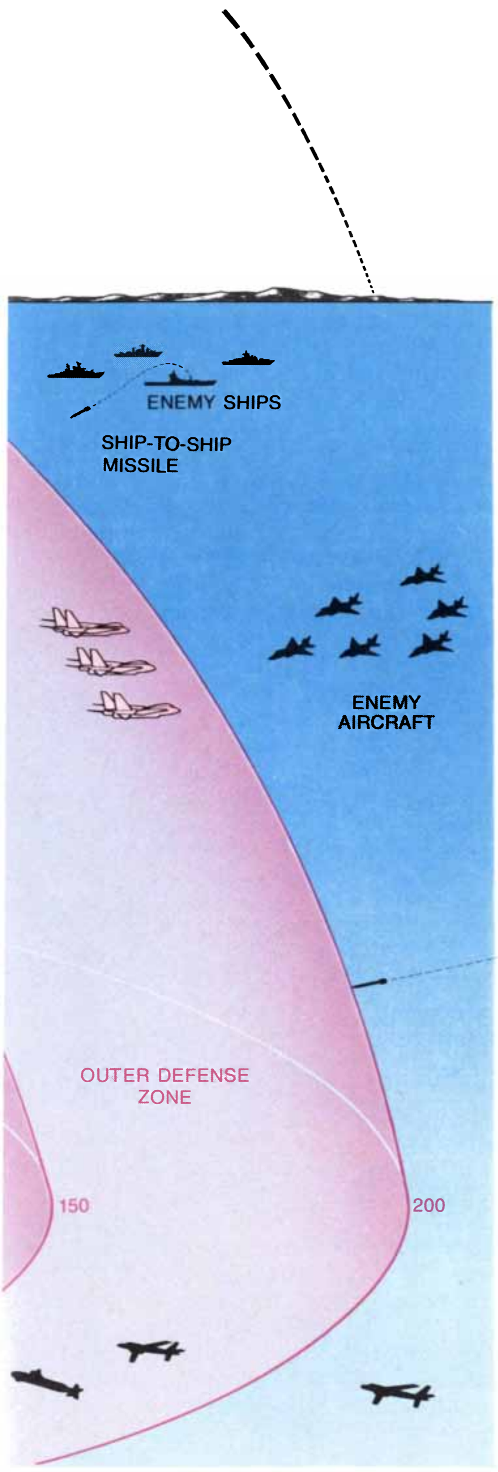
The performance of the Type 42 destroyers against attacking aircraft flying at high or medium altitudes is considered to be very good; indeed, the Sea Dart system shot down eight aircraft during the Falklands war. In the May 4 attack on the *Sheffield,* however, the attack was launched from below radar coverage, and by the time the Exocet was detected visually it was too late for the ship to respond adequately either with antimissile gunfire or with passive countermeasures.

The *Sheffield* had not been outfitted with another advanced British defensive



ZONE DEFENSE designed to protect a U.S. aircraft-carrier task force against a variety of possible attacks is depicted in this idealized scene. In the outer zone attacking aircraft, submarines and surface ships are detected and identified by surveillance satellites or by carrier-based early-warning aircraft; the attackers are intercepted by carrier-based fighters, attack aircraft and antisubmarine-warfare (ASW) aircraft and submarines. In the middle zone attacking aircraft, missiles and submarines are opposed by surface-to-air missiles and antisubmarine missiles launched by ships or helicopters. In the inner zone "point defense" systems such as rapid-fire guns and

missile, the Seawolf, which is intended specifically as a close-in, fast-reaction antimissile missile for just such sea-skimming threats. Two British frigates, H.M.S. *Broadsword* and H.M.S. *Brilliant,* were armed with the Seawolf and used it effectively in the Falklands conflict, bringing down five Argentine aircraft while they were engaged in low-level attacks. It is questionable, however, whether the *Sheffield* would have survived even if it had been armed with Seawolf missiles; the ship did not have its active search radar operating at the time of the Exocet attack in order not to interfere with communications. (Furthermore, according to a recent report, a passive radar system on board the *Sheffield* did detect the approaching missile's homing-radar beam, but the ship's computer had been programmed to identify the Exocet as a "friendly" missile and so did not sound the alarm.)



high-speed antimissile missiles serve as last-resort interceptors. The three defense zones are delineated here somewhat arbitrarily; actually there is considerable overlap between them. Ships and aircraft are not drawn to scale.

The Argentines fired only six Exocet missiles in the course of the Falklands conflict. Four of them hit targets, although perhaps not their intended ones. The first missile, which struck the *Sheffield,* was probably intended for the nearby aircraft carrier H.M.S. *Hermes;* the two ships apparently were confused on the Super Étendard's radar screen. The second Exocet, launched almost simultaneously from an accompanying Super Étendard, passed by the *Sheffield,* falling into the sea short of the *Hermes.* Three weeks later the converted merchant ship *Atlantic Conveyor,* which had delivered a load of Harrier "jump jets" to the British forces and was ferrying supplies and helicopters between the islands, was hit and destroyed by two Exocets, also launched from Super Étendards. A fifth Exocet attack five days later failed. The sixth Exocet, launched from a shore battery on June 11, struck the guided-missile destroyer H.M.S. *Glamorgan,* which was severely damaged but remained afloat.

The 67 percent success rate of the Exocet is remarkable, given the newness of the French missile system and aircraft to the Argentine forces and the much greater overall military power of the British forces. It is worth noting in this context that the Argentines had some difficulty arming the Exocets properly in time for the war. When the first missile penetrated the *Sheffield,* its warhead did not explode; because of the missile's abbreviated flight, however, it contained a substantial amount of unexpended fuel, which exploded and ignited some of the ship's fuel. It was largely a matter of chance that the missile struck in a particularly vital spot near the ship's center of operations.

The success of the Exocet in destroying a target hundreds of times its value has been widely interpreted as demonstrating the impending vulnerability of all surface warships. Nevertheless, some military analysts draw a quite different conclusion. British naval spokesmen, for example, cite the *Sheffield* incident in support of their call for an accelerated buildup of the British surface fleet. They point out that the *Sheffield* was designed for an area-defense role in coordination with large NATO fleets and the highly capable air cover of the U.S. aircraft-carrier task forces in the North Atlantic. British naval officers criticize the retrenchment in their country's defense spending over the past decade, pointing out that less than 20 years ago there were five large-deck aircraft carriers in the British navy. The last British carrier of this type, H.M.S. *Ark Royal,* was retired in 1978, leaving naval air cover dependent on the much smaller *Hermes*-class and *Invincible*-class jump-jet aircraft carriers.

The U.S. Department of the Navy takes a similar view. Secretary of the Navy John F. Lehman, Jr., for example, maintains that it should come as no surprise that one lone ship on picket duty, without adequate antimissile defenses and without adequate air cover, was struck by a homing missile of the Exocet's capabilities. He adds that "if an American ship had been the target of the Argentine attack, it would not have gotten anywhere near kill range."

On this point Lehman is probably right. The U.S. deploys its naval forces, now numbering almost 500 warships, around 13 large-deck aircraft carriers. The current plans call for increasing the number of ships to more than 600, to be organized into 15 carrier task forces and four battle groups. (Each of the latter formations is to be centered on a refitted World War II–vintage battleship.) A carrier task force is a formidable fighting group. A *Nimitz*-class, nuclear-powered aircraft carrier weighs approximately 90,000 tons, four times the displacement of either of the British carriers involved in the Falklands fighting (and more than 20 times the displacement of the *Sheffield*). The U.S.S. *Nimitz* has a flight-deck area of more than 4.5 acres and carries a total of about 100 aircraft: jet fighters, jet bombers, surveillance aircraft, antisubmarine aircraft (including helicopters), utility helicopters and rescue helicopters. At least two-thirds of the aircraft are intended to defend the fleet against attacking aircraft, surface ships and submarines. The remaining 30 or so are there to "project power" (military parlance for attacking the enemy). In addition the task force includes a number of other ships—cruisers, destroyers, submarines and supply ships—whose mission is to protect and maintain the task force for periods of up to six months at sea.

The defense of a U.S. carrier task force is divided into three major zones. In the outer zone submarines and carrier-based aircraft are the main interceptors; in the middle zone ship-launched and helicopter-launched missiles predominate; in the inner zone short-range, "point defense" systems such as rapid-fire guns and antimissile missiles are the weapons of last resort. Any attacking system—aircraft, missile, ship or submarine—must succeed in penetrating all three zones in order to reach the command ship: the aircraft carrier.

A carrier of the *Nimitz* class typically has four E-2C early-warning aircraft equipped with long-range radar. The E-2C's fly over or near the task force at an altitude of between 25,000 and 30,000 feet, in effect extending the "vision" of the task force from about 30 nautical miles out to more than 200 miles. The E-2C is said to be capable of detecting, identifying and tracking more than 600 airborne and surface targets at a time and of directing more than 40 fighter planes from their combat-air-patrol flight paths toward individual targets.

The chief carrier-based combat aircraft in the U.S. Navy today is the F-14, a high-performance, swing-wing jet that can fly at speeds greater than Mach 2 (that is, more than twice the speed of sound). The F-14 is armed with eight precision-guided air-to-air missiles and a 20-millimeter cannon. Aircraft on combat air patrol normally "loiter" between 150 and 200 miles from the carrier, and they are reportedly able to detect bombers at a range of 170 miles or cruise missiles at 65 miles. The radar system on the F-14 that launches the Phoenix missiles can track as many as 24 targets simultaneously and can coordinate an attack on as many as six at a time at a maximum range of 60 miles. In short, the outer defense zone of a U.S. carrier task force is quite well equipped to handle multiple threats at a distance of 200 miles or more from the carrier.

The second zone of defense is based primarily on the surface-to-air missiles of the other ships in the task force. Most U.S. cruisers, destroyers and frigates now carry surface-to-air missile launchers loaded with Talos, Terrier or Tartar missiles. The newest antiaircraft missiles in the U.S. fleet, called Standard 1 and Standard 2, are just entering service. They are described as being capable of intercepting attacking aircraft at a range of between four and 85 miles and at an altitude of up to 80,000 feet. Time is of vital importance here, with a few minutes being the maximum time available to intercept a supersonic aircraft or missile in the middle defense zone.

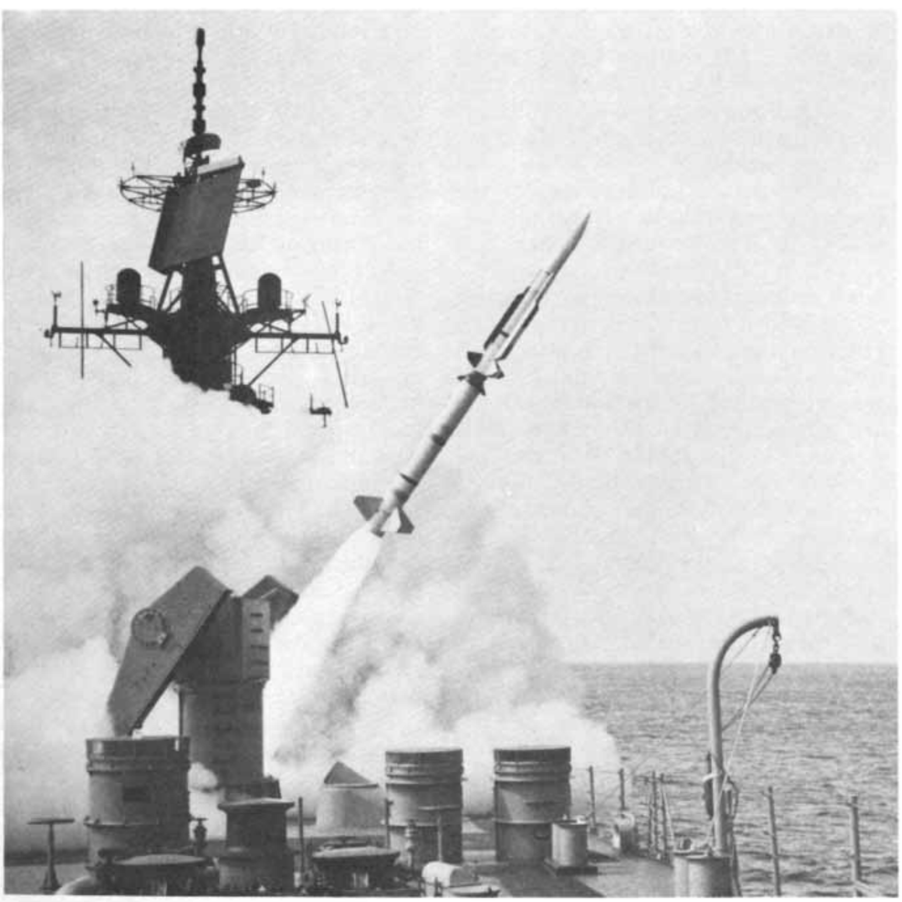
If an incoming aircraft or missile manages to penetrate the first two defense zones, several options remain. In the final few miles of flight the attacker can be destroyed by the older antiaircraft guns still present on most ships, but in the majority of cases it will be opposed as well by modern point-defense systems. The newest point-defense missile on board U.S. ships is the NATO Sea Sparrow, a 12-foot supersonic missile with a radar-homing guidance system. The newest point-defense gun is the Phalanx, a six-ton, fully automatic Gatling gun with six rotating barrels that is reported to be capable of firing a "wall of bullets" at a rate of 50 per second. The bullets have a uranium core that makes them extremely dense and gives them high penetrating power. They are directed at the approaching target by a "closed loop" system that simultaneously tracks the projectiles as well as the target, adjusting the aim of the gun until the target is hit. If a Phalanx gun had been on board the *Sheffield,* it might have been able to stop the Exocet; the British apparently think so, since they have just installed two such guns on their newest aircraft carrier, H.M.S. *Illustrious.*

In addition to such active point-defense systems a variety of passive countermeasures are also available; they include evasive maneuvers, electronic jamming of the attacker's radar, floating infrared decoys and the dispersal of "chaff": clouds of metallic strips that confuse the radar-based guidance systems of the incoming missiles. Both chaff and decoys were used successfully by the British in the Falklands fighting.

In the continuing effort to upgrade its ship defense the U.S. Navy earlier this year commissioned its first *Ticonderoga*-class guided-missile cruiser, which is particularly well equipped to defend an aircraft-carrier task force. The Navy has proposed building 17 of these $1 billion ships, each outfitted with the new Aegis combat-control system, as many as 68 Standard 2 missiles, 20 ASROC antisubmarine missiles, two Phalanx guns and eight Harpoon antiship missiles, along with other weapons. The heart of the Aegis system is the Spy-1 phased-array radar, which is capable of sweeping the horizon in one-fourth the time of its predecessors. The system is designed to detect and track incoming objects in greater numbers and in more extreme circumstances (high and low altitudes, high and low speeds, large and small radar cross sections) with unprecedented rapidity, flexibility and firepower.

U.S. fleet defense thus appears to be highly capable. Indeed, had such a carrier task force sailed into combat in the Falklands last year, few if any Exocets would have reached their targets. Yet one's confidence in fleet defense must be tempered, particularly in view of certain operational limitations, past experience and a consideration of the potential foe.

The operational limitations of fleet defense are numerous. Assuming an attack is detected by patrolling E-2C aircraft, the maximum number of fighter aircraft a carrier can launch is 24;



NEW SURFACE-TO-AIR MISSILE, called the Standard 2, is shown being test-fired from a rail-type launcher on the deck of the guided-missile cruiser U.S.S. *Wainwright.* The Standard 1 and Standard 2 missiles are intended respectively to replace the Tartar and the Terrier missiles as the primary medium-range and extended-range antiaircraft missiles of the U.S. fleet.

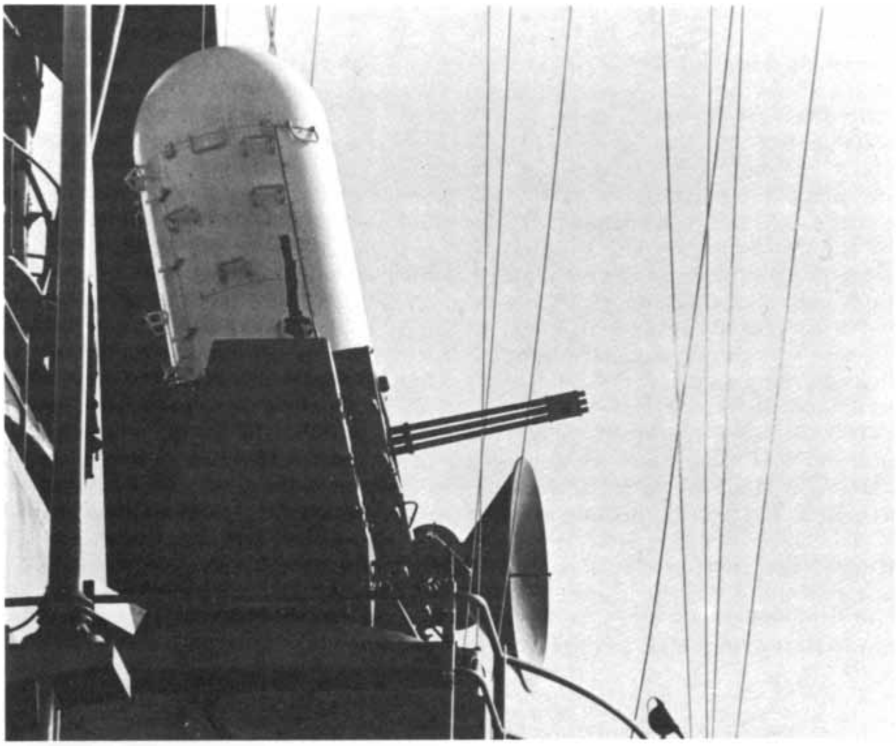
each fighter in turn can be equipped with a maximum of eight air-to-air missiles. This gives a total of 192 shots at incoming aircraft and missiles, which is probably enough in limited engagements but probably insufficient in intense battles. It is for this reason among others that carrier task forces often operate in tandem, thereby doubling the number of available interceptor aircraft.

According to the Navy, the Phoenix missile on board F-14 fighters has an 80 percent kill probability. The General Accounting Office stated in a report to Congress in 1979, however, that the weapon would not be able to maintain that standard against the current generation of Russian aircraft. Another air-to-air missile in the U.S. arsenal, the heat-seeking Sidewinder, proved very reliable in the Falklands conflict, destroying 24 of the 27 Argentine aircraft it was fired at, for an 89 percent kill rate. Nevertheless, some penetration of the outer defense perimeter of a carrier task force seems inevitable. It could be accomplished through an assortment of tactics: jamming of the defensive radars, successful attacks on the E-2C's or the F-14's, the deployment of decoys or simply saturation of the defense by an overwhelming number of attacking aircraft and missiles.

The effectiveness of the middle zone of fleet defense is a function of both the speed and the capacity of the defensive combat-control system. Incoming supersonic aircraft and missiles allow only a few minutes for the defensive system to detect a target, identify it, track it, pass it on to the fire-control radar and fire the defensive missile—all before the attacker moves to within a few miles of the carrier task force.

Short of the Aegis system, the Tartar-D surface-to-air missile system is generally regarded as the most effective U.S. fleet-defense system. Yet it has serious operational limitations; for example, the rotating antennas of its search radars take four seconds to sweep the horizon, thereby setting a lower limit on the system's reaction time, an important factor in identifying and tracking a target. Moreover, the detection, identification and tracking of a target are done by human operators until the system locks onto the target, making it hard to deal with multiple targets flying at supersonic speeds. Indeed, a recent congressional study called the reaction time and firepower of the Tartar system "inadequate," referring specifically to its "considerable degradation" in the face of sophisticated electronic countermeasures. Furthermore, there is evidence of a serious lack of fire-control coordination within a task force. Several Tartar-type systems may lock onto a single incoming missile, allowing other missiles to penetrate the middle defense zone.

The Aegis control system overcomes some of these operational difficulties. The Standard 2 missile gives it greater firepower; its phased-array search radar (which points the beam electronically rather than mechanically) sweeps the horizon in only one second; it is fully automated, and it is better able to coordinate fire control within the task force. Nevertheless, the Aegis system reportedly lacks the ability to intercept sea-skimming missiles such as the Exocet. Furthermore, it is designed primarily to protect itself and any vessels behind it (with respect to the incoming missile). In operational terms, therefore, the Aegis system can handle multiple threats only within a limited sector; if attacking missiles were to arrive from several directions at once, throwing what the Navy calls "crossing targets" at the system, penetration would be virtually certain.



PHALANX ANTIMISSILE GUN is described as the Navy's first all-weather, fully automatic point-defense gun. The six-barrel, 20-millimeter Gatling gun fires high-velocity, uranium-core bullets at a maximum rate of 50 per second. Nevertheless, it has a number of operational limitations: its magazine holds only 989 rounds, allowing it to fire for less than 20 seconds at the maximum rate; furthermore, its barrels wear out after seven minutes of maximum firing, and the entire gun wears out after 50 minutes. The gun can therefore be easily overwhelmed.

That leaves the inner, point-defense zone to pick up and destroy (or deflect) all the remaining aircraft or missiles. The constraints here are very tight: a missile arriving at the speed of sound (Mach 1) requires a reaction time of less than 30 seconds. The incoming missile can be fired at by either an automatic antimissile missile, such as the NATO Sea Sparrow, or an automatic rapid-fire gun, such as the Phalanx. Both have their operational limitations, however. The NATO Sea Sparrow is fired from a deck-mounted canister that contains only eight missiles and can therefore be overwhelmed. The Phalanx has an impressively high rate of fire—3,000 rounds per minute—but its magazine holds only 989 rounds, enough for less than 20 seconds of fire at the maximum rate. It too can be overwhelmed. Finally, passive countermeasures that attempt to "fool" the missile rather than destroy it can be effective in certain circumstances but not in others, as the British experience in the South Atlantic showed.

The absolute last resort of a ship is to absorb a missile hit, as the *Sheffield* did, and to control the damage, as the *Sheffield* could not. The designers of combat ships take great care to provide adequate damage-control systems to enable a ship to survive a first hit and ideally to enable it to continue combat operations. The U.S. Navy maintains that its ships are designed to withstand several hits and still go on fighting. The hulls of most large warships are compartmentalized and are made of heavily armored steel; the superstructures, although they are often constructed of a lighter material such as aluminum, are shielded in some of the newer vessels by a layer of Kevlar, a rigid, resin-based protective armor. The steel armor of the recently reactivated battleship U.S.S. *New Jersey* is more than a foot thick in many places; it is doubtful an Exocet missile could penetrate very far into such armor.

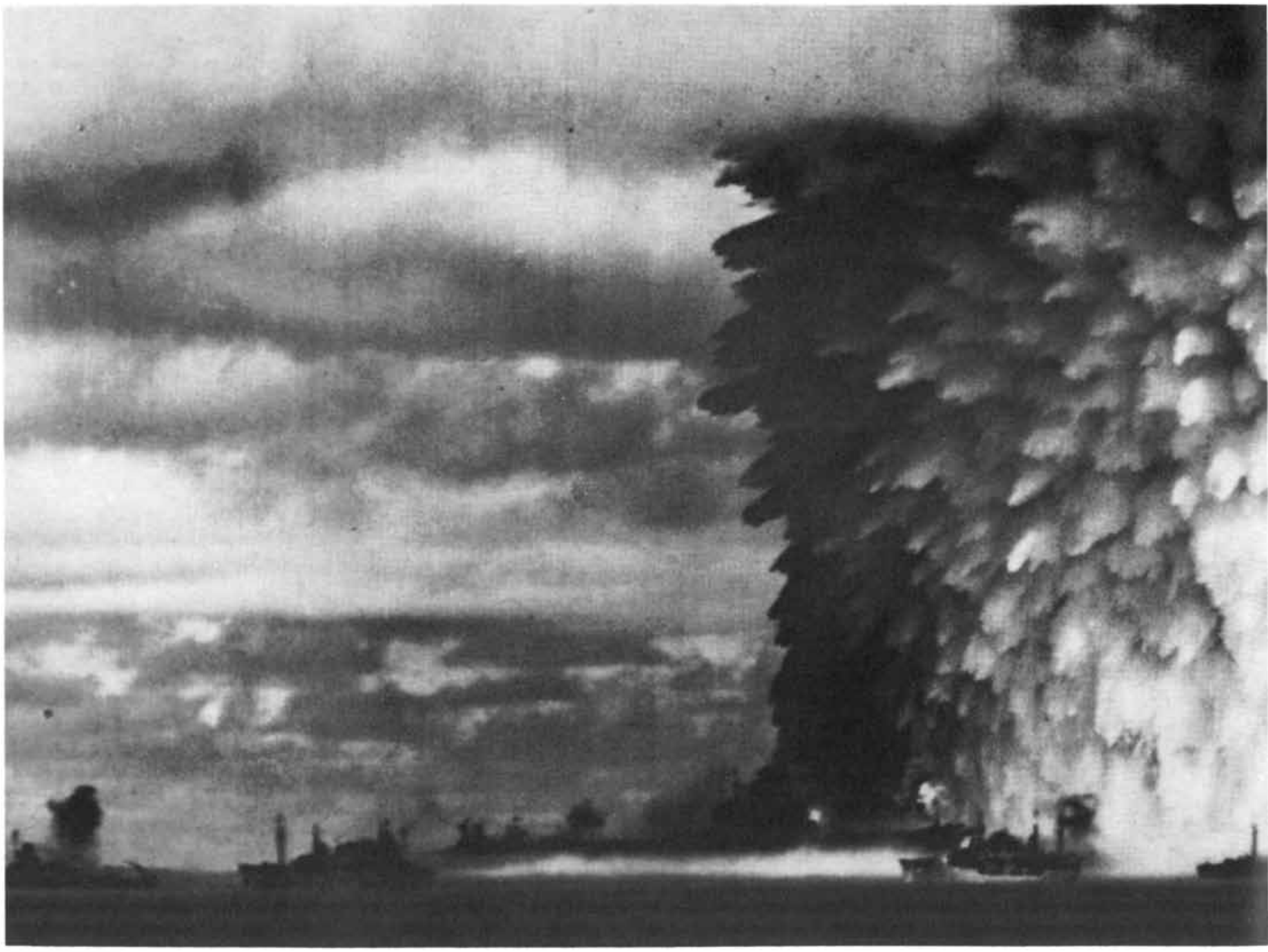
The history of naval warfare, however, is replete with incidents in which ships have been put out of action by seemingly minor hostile actions or shipboard accidents. The German admiral Alfred von Tirpitz once remarked that a warship has three goals: first, to stay afloat; second, to stay afloat; third, to stay afloat. Large modern warships may well be able to do that after being hit by a smart weapon or even by several weapons, but to carry on normal operations may be another matter. Most naval officers readily admit that in modern warfare a large ship will stay out of the line of fire until the enemy is "neutralized," most likely by aircraft.

The vulnerability of the U.S. Navy is most apparent when one considers the maximum potential threat it faces: not a half-dozen Exocet missiles fired by a few Super Étendard aircraft but the Russian fleet of more than 700 surface warships (mostly limited-range frigates, corvettes and coastal patrol vessels), almost 300 submarines (mostly diesel-powered) and more than 1,500 antiship cruise missiles. The missiles, which are based on surface ships, submarines, aircraft and shore batteries, have ranges of up to 350 miles and speeds as high as Mach 4. The Phalanx gun, which is designed to stop a missile in the final kilometer of its flight, would have slightly more than one second to hit an incoming Mach 4 weapon. Although many of these missiles would not be encountered outside the coastal zones of the U.S.S.R., some of the longer-range Russian aircraft, including the Backfire bomber, are capable of projecting force at greater distances.

Another critical factor in any assessment of the vulnerability of the U.S. surface fleet is the existence of nuclear weapons. Most of the antiship missiles in the Russian arsenal can carry either a nuclear warhead or a conventional (non-nuclear) one. No ship, no matter how large or well armored, can survive a nuclear explosion nearby.

In short, the advent of smart missiles, with or without nuclear warheads, is bound to place a premium in naval warfare on stealth, deception, mobility and dispersion rather than on size, armor plate or coordinated defensive firepower. To be sure, the technological advantage will continue to alternate somewhat between offense and defense, as it has in the past. Antiship missiles will continue to become more accurate, more maneuverable, longer in range, faster and stealthier, and so will defensive missiles. At the same time ships will continue to evolve into more effective platforms for improved radar and fire-control systems, rapid-fire guns, antimissile missiles and passive countermeasures. In the long run, however, surface warships must become increasingly expendable.

This conclusion leads one to question the current naval strategy of the U.S. The declared policy is one of mari-



ULTIMATE ANTISHIP WEAPON currently available in the arsenals of both the U.S. and the U.S.S.R. for possible use against the other side's naval forces is a long-range precision-guided nuclear warhead. The effects of such a hypothetical attack are suggested by this photograph of the Baker test explosion conducted by the U.S. at Bikini atoll in 1946. Some 70 unmanned ships of various types were anchored around the site of the shallow underwater burst, which had an explosive yield of about 20 kilotons (small by today's standards). The

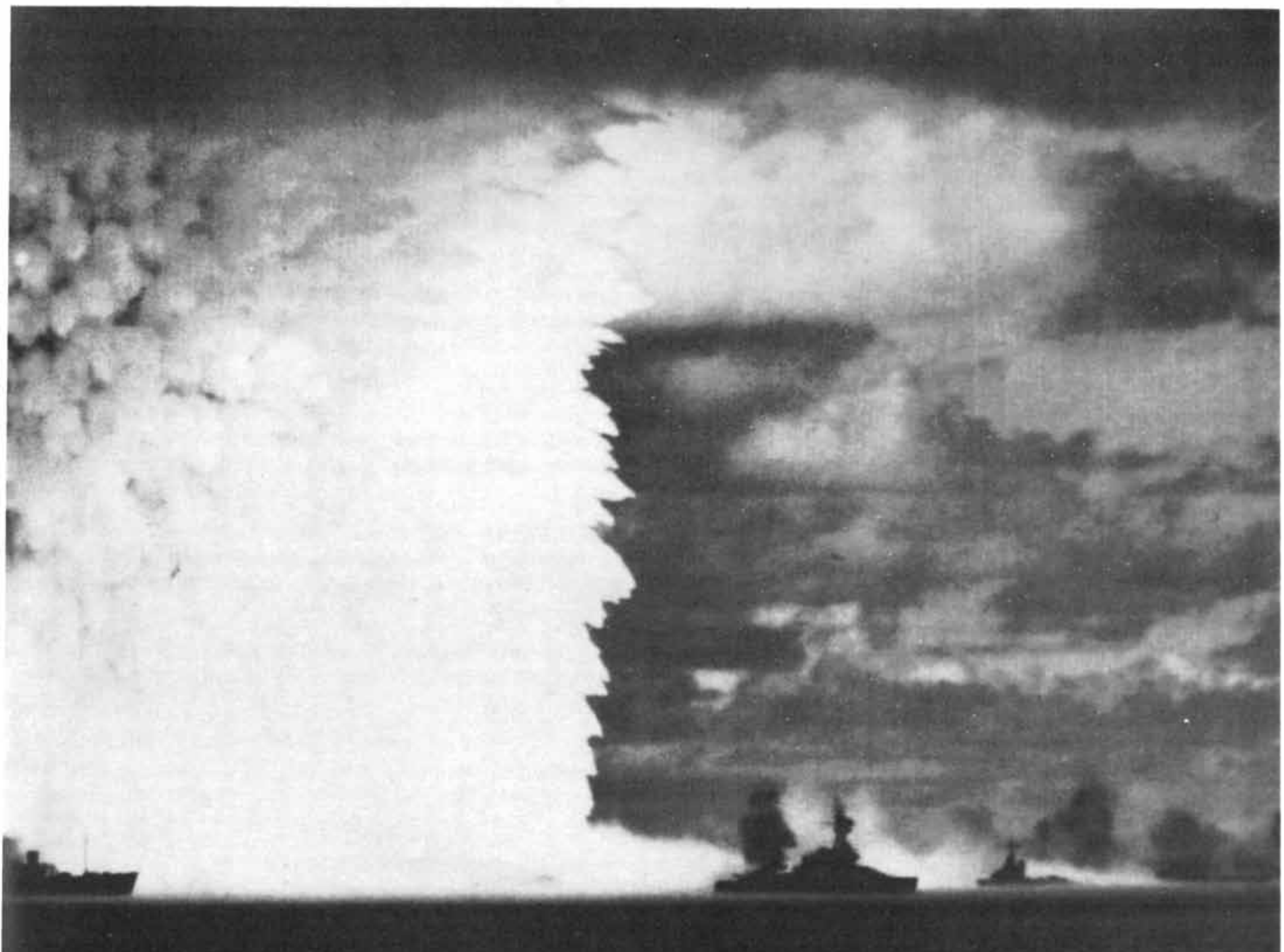
time supremacy based on large-deck aircraft carriers. Secretary of the Navy Lehman explains: "Carriers are the only way to get superiority over seven-tenths of the earth's surface." He and Secretary of Defense Caspar W. Weinberger argue that the Falklands experience only strengthens the case for additional carriers and that it did not teach any "brand-new lessons."

The lessons may not be new, but they are cautionary nonetheless. Capital warships are increasingly at risk, regardless of their armament, to comparatively inexpensive smart missiles and torpedoes. For a carrier task force to be effective against a capable opponent such as the U.S.S.R. one must assume that (1) the fleet will remain hidden, (2) virtually all the attacking missiles and aircraft will be defeated and (3) nuclear warheads will not be employed. Given the increasing effectiveness of satellite and aerial reconnaissance, the permeability of fleet defense and the strong reliance of the Russians on nuclear weapons in their military planning, none of these assumptions appears plausible.

It may be, therefore, that the U.S. Navy considers its fleets expendable. Retired Admiral Hyman G. Rickover, the "father" of the nuclear navy, testified before a congressional committee last year that a modern aircraft carrier would last "about two days" in a major war. The Joint Chiefs of Staff project the need for 25 such carriers for "minimal risk"; that number or more might be needed just to keep pace with attrition in a major war. Yet a new, 90,000-ton aircraft carrier that costs more than $3.5 billion to build (in 1982 dollars) and another $15 billion or so to outfit and maintain—and that has a crew of more than 6,000—is not an obviously expendable item.

In meeting their stated objectives of "controlling the seas" and "projecting power" aircraft-carrier task forces are also intended for foreign intervention. As sophisticated antiship technology proliferates into the Third World, however, even small nations will represent a serious threat to large surface vessels. To cite just one example, Libya is reported to have 13 kinds of smart missiles in its arsenal.

Perhaps the final lesson of the Falklands war is that when it happens again, it will be even costlier. The days of gunboat diplomacy are over, and yet some of the major powers fail to recognize the fact. Military custom dies slowly. "In harm's way," a traditional American naval motto, will eventually give way to the recognition of the growing dominance of the naval conflict by land-based aircraft and missiles, by small, mobile, fast (and expendable) ships armed with long-range precision-guided weapons and by submarines. Aircraft carriers, reactivated battleships and other large, expensive surface vessels will retain one safe role: to show the flag in peaceful waters.



ships suffered damage from a number of causes, including underwater and atmospheric shock waves as well as surface water waves. For example, the aircraft carrier U.S.S. *Saratoga,* which was moored in the lagoon about 400 yards from "surface zero," had the central part of its island structure folded down onto the deck by a 90-foot wave generated by the explosion. The radiation effects on the ships' crews in the event of such an attack would of course extend far beyond the immediate destructive effects on the ships themselves.

Vibrations of the Atomic Nucleus

The nucleus can quiver, ring or even "breathe"; the coordinated motion of the nuclear particles reveals much about the forces between them. Six modes of vibration have been detected so far

by George F. Bertsch

Vibrating systems, from the swinging pendulum to the oscillating electromagnetic field of a light wave, have long had a privileged place in the physical sciences. The vibrational modes and frequencies observed in a system can reveal much about the nature of the forces acting within it. An analysis of vibrational motions was notably important in the understanding of atomic structure in the early years of the 20th century. More recently a richly varied spectrum of vibrations on a far smaller scale has been discovered in the nucleus of the atom. The study of the nuclear vibrations is proving to be a major source of information on the structure of the nucleus and on the forces that hold it together.

The protons and neutrons that make up a nucleus are collectively called nucleons. Quantum mechanics describes the arrangement and the motion of the nucleons by means of a wave function. For the stable nuclei found in nature the wave function does not change with the passage of time. The nucleus can be set in motion by external forces, however, leaving it internally excited. In the simplest excitations, called giant vibrations or giant resonances, all the nucleons oscillate coherently and the motion follows a simple pattern. The difference between coherent and incoherent motion in an excited nucleus is roughly analogous to the difference between the coherent motion of the liquid in a cup of tea that has been bumped and the random thermal motion of the molecules in the hot tea. In a coherent vibration there is a pattern: like liquid sloshing in a cup, the nucleons in a vibrating nucleus pass cyclically from one distribution to another.

The motion of the individual particles in a vibrating body can be coordinated in various ways, giving rise to the distinctive patterns of motion called vibrational modes. The molecules in the cup of tea can oscillate from the side of the cup to the center and back again or from one side of the cup to the other, and the nucleons can also oscillate in several patterns.

So far six giant nuclear vibrational modes have been observed. Most of them have been detected experimentally only in the past few years. The giant vibrational modes fall into two classes. The first class consists of the modes the nucleus shares with other spherical bodies, such as a drop of water or the earth itself. The spherical vibrational modes of the nucleus that have been observed are, in the order of their discovery, the giant dipole, the giant quadrupole, the giant monopole and the giant octupole. The terms were borrowed from the names for electric fields of a given spatial complexity. For example, the protons in a nucleus vibrating in the quadrupole mode give rise to an oscillating electric field that resembles the field generated by four poles, or point charges.

The other class of nuclear vibrations involves the spin orientation of the nucleons. Protons and neutrons, like electrons and many other particles, spin on an internal axis. In the unexcited nucleus the spin axis is fixed with respect to the wave function of the nucleon and the spin orientation of the other nucleons. During a spin vibration some of the nucleon spins are tipped slightly and the spin axes begin to precess, that is, they describe circles around their original orientation. The precession of neutrons with respect to protons can be coordinated in different ways, giving rise to different spin-vibrational modes. For example, the neutron spins can be tipped in the same direction as the proton spins or in the opposite direction. As a result of the coordinated precession of the nucleons the nucleus can acquire a net spin or a net magnetic moment that oscillates at the precession frequency. So far two vibrations of this kind have been observed: they are called the giant Gamow-Teller resonance and the giant magnetic-dipole resonance.

An understanding of nuclear vibrations requires knowledge of the forces acting between nucleons and the mechanical laws governing the motion of the nucleons under the influence of these forces. Given the complexity of the problem in the many-body system of the nucleus, physicists have long resorted to simple models. The nucleons are viewed as moving under the influence of a single generalized force whose effects approximate those of the interactions of the many nucleons. The vibrational modes of the nucleus provide an experimental check on the validity of the models; a theoretical justification requires the more fundamental quantum theory.

A body vibrating in a given mode oscillates at a specific frequency called the resonance frequency. The resonance frequency is determined in part by the internal forces opposing the motion of the oscillating particles. A drop of water and a solid sphere vibrating in the same spherical mode have different resonance frequencies. The solid strongly resists any distortion and has a high vibrational frequency. The liquid drop, on the other hand, returns to a spherical shape only because of forces associated with surface tension. These forces are quite weak and lead to a low vibrational frequency.

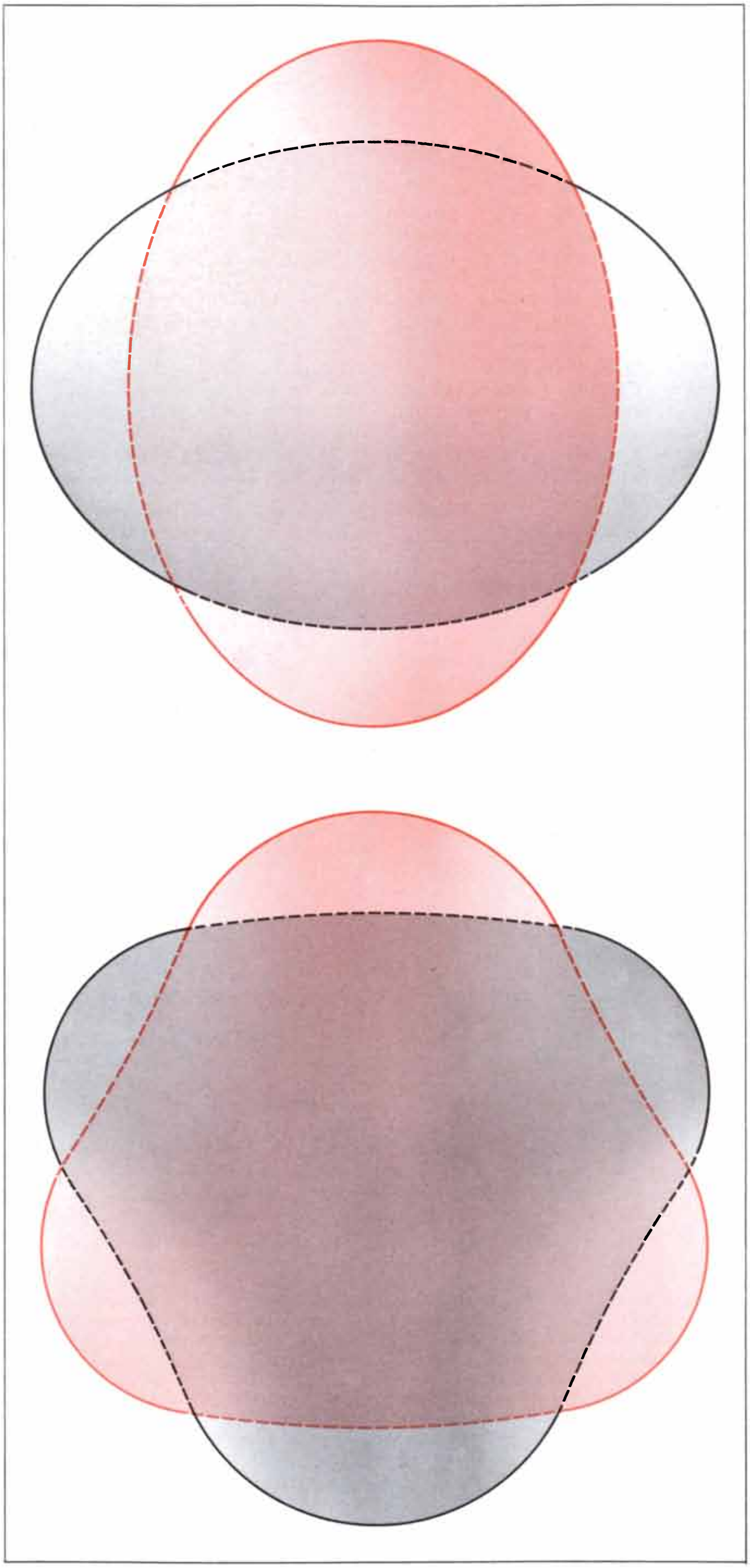
One model of the nucleus, the liquid-drop model, likens the forces acting between the nucleons to those acting between the molecules of a low-viscosity liquid. Another model, based on the shell model of the nucleus, likens the forces to those between the particles in an elastic solid. The resonant frequencies of nuclear vibrational modes were calculated from the liquid-drop model and from other simple models long before the vibrations could be excited in the laboratory and their frequencies measured. No single model accurately predicts the resonant frequencies of all modes of vibration. In some circumstances the nucleus acts like a liquid and in others like an elastic solid. In general its response is rather like that of a class of non-Newtonian fluids, of which Silly Putty is the most familiar example.

These substances respond elastically to sudden forces but flow as liquids over longer periods of time.

The most accurate and sophisticated description of nuclear vibrations is given by the time-dependent mean-field theory, which is based on quantum mechanics. In the mean-field theory the forces acting on the nuclear particles are calculated from the quantum-mechanical motion of the particles themselves, whereas the simpler models are based on specific assumptions about the relation between the forces and the particle motion. At first sight these approaches seem quite divergent, but the motion in the vibrations is simple, and in many cases it can be shown through the quantum theory that the assumptions made in the simple models are justified.

In general, nuclear vibrations are excited by bombarding nuclei with high-energy photons (quanta of electromagnetic radiation) or other particles. The vibrations are detected by observing how the projectile is absorbed or diffracted by the nuclei. In the case of the first vibration to be reported, the giant dipole vibration, both the excitation and the detection of the vibration proved to be relatively straightforward. For the other modes it has proved quite difficult to probe the nucleus in just the way needed to excite a particular vibration. Indeed, nearly 30 years passed between the discovery of the giant dipole vibration and the discovery of another giant vibrational mode.

If a nuclear vibration is to be excited, the first and simplest condition to be met is that the energy imparted to the nucleus must be equal to the energy associated with the vibration. The nucleus vibrates at extremely high frequency; the energy of the vibration is equal to the frequency multiplied by Planck's constant, and so the energy is also comparatively high. A typical vibrational frequency is 5×10^{21} hertz, which corresponds to a vibrational energy of 20 million electron volts (MeV). The energy of the photon or other particle that



VIBRATIONS OF NUCLEAR SHAPE, in which the nucleus is deformed from a spherical to an ellipsoidal or pear shape, are akin to the vibrations of a macroscopic body such as as a drop of water. In both cases the vibration is a coherent oscillation of the particles of which the body is composed and hence gives rise to a pattern of motion of the body as a whole. The vibrations have had an important role in testing models of nuclear structure. Various models differ in their predictions of the frequency and the other characteristics of a given mode of vibration in the nucleus of a given atom. It seems that the nucleus sometimes acts like an elastic solid, sometimes like a liquid and sometimes like Silly Putty.

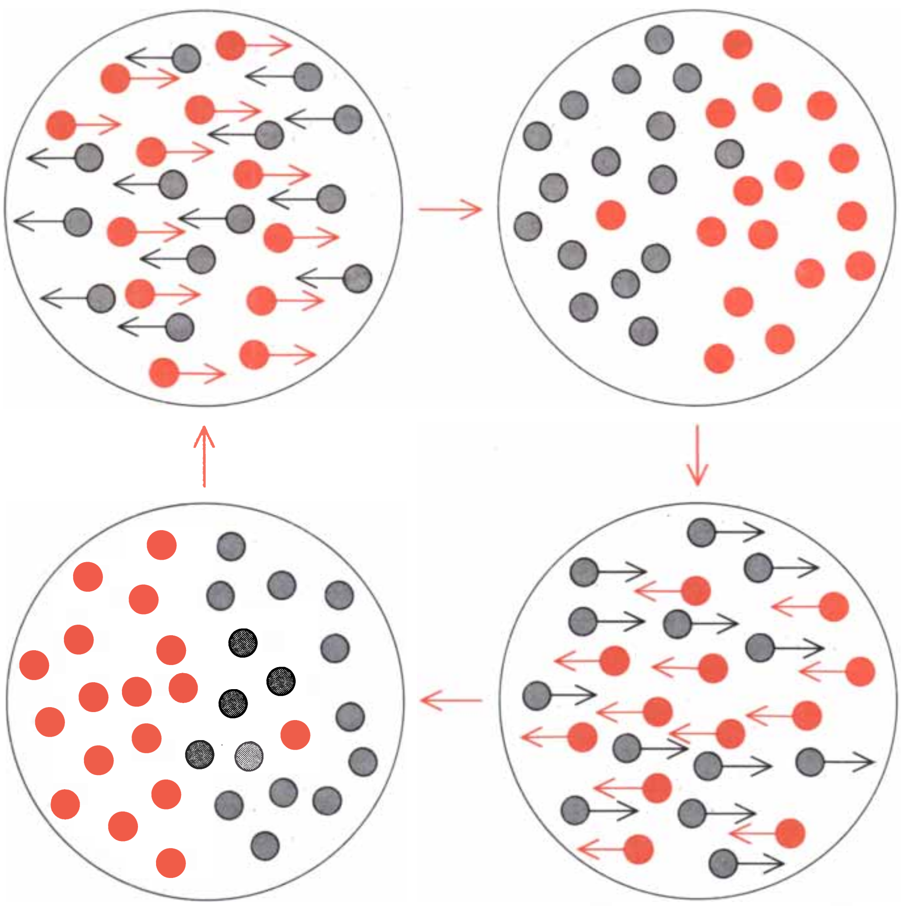
excites a nuclear vibration must be at least as high as the vibrational energy.

In the case of the giant dipole resonance the energy requirement was the only condition that had to be met. Physicists were able to excite the giant dipole vibration simply by bombarding nuclei with photons having an energy equal to the vibrational energy of the mode. The discovery of the giant dipole depended only on the availability of a source of monoenergetic high-energy photons, or gamma rays. The detection of the vibration followed closely on the development of such sources (bremsstrahlung, or braking radiation, from electron accelerators) in the mid-1940's.

It is not difficult to understand how gamma rays excite the dipole vibration. A photon carries with it an oscillating electric field. Although the wavelength of a gamma ray is smaller than that of other forms of electromagnetic radiation, such as visible light, it is large with respect to the diameter of a nucleus. As a result the electric field associated with a passing gamma ray is nearly uniform across the nucleus. The field exerts a force on the positively charged protons, moving them away from the neutrons. The neutrons themselves are electrically neutral, and so the field has no direct influence on them. Because the center of mass of the nucleus remains at rest, however, the neutrons move in the opposite direction. The restoring force of the vibration is the attractive force between protons and neutrons, namely the strong nuclear force responsible for binding the particles together. The strong force is independent of electric charge.

The giant dipole was not only relatively easy to excite but also relatively easy to detect. The photons that excite the vibration are simply absorbed in the nucleus. This phenomenon, which is a form of resonance, arises in any vibrating system excited by an external source. A person singing in a shower provides a familiar example of resonance. When the frequency of the voice matches a natural vibrational frequency of the air mass in the shower, the amplitude of the vibration becomes very large. The sound intensifies; the purity of the frequency is perceived in a resonant tone. In such situations the large amplitude leads to an increased absorption of energy. Similarly, when the frequency of the oscillating electric field associated with the gamma rays matches the resonance frequency of the dipole mode of the nucleus, the gamma rays are absorbed.

The tendency of the nucleus to absorb incident particles is expressed quantitatively as an effective cross section, measured in units of area. The dipole resonance is seen in a sharp increase in the absorption rate of the target at a particular photon energy. This is expressed as an increase in the effective absorption cross section of the nucleus. It is as if the nucleus suddenly grew larger and therefore intercepted more photons, although what actually changes is not the nuclear size but the absorbance. Because the response of a body to an electromagnetic force depends on the electric charge of the body, it is possible to determine from the effective cross section of the vibrating nucleus that all the protons in the nucleus are participating in the vibration; in short, the dipole vibration is in fact a giant vibration.



GIANT DIPOLE VIBRATION, identified in the late 1940's, is excited by bombarding the nuclei in a target with high-energy photons. The protons in a nucleus (*color*) are accelerated in one direction by the electric field associated with a passing photon. The neutrons (*gray*) are unaffected by the field, but they move in the direction opposite to that of the protons so that the center of mass of the nucleus remains stationary and momentum is conserved. The restoring force, which ultimately reverses the motions of the protons and neutrons, is the strong nuclear force responsible for binding them together. The frequency of the vibration can be calculated from a simple model in which only the inertia of the nucleons, the restoring force and the laws of classical mechanics are considered. The inertia of the system is proportional to the number of nucleons and thus to the volume of the nucleus. The restoring force is proportional to both the volume of the nucleus and the displacement of the nucleons. The dependence of the variables on nuclear volume suggests that the frequency of the vibration should vary inversely with the radius of the nucleus; the experimental data on large nuclei bear out this conclusion. Here the amplitude of the vibration has been greatly exaggerated for the purpose of clarity.

The next vibration to be reported was the giant quadrupole vibration. The quadrupole, unlike the dipole, is a shape vibration, in which the shape of the nucleus as well as the distribution of the nucleons changes. A nucleus vibrating in the quadrupole mode is distorted from a spherical shape to an ellipsoidal shape and moves back through a spherical shape to an ellipsoidal shape of another orientation. It is intuitively clear that the vibration could be induced by pushing in on the nuclear surface along one axis and pulling out along a perpendicular axis.

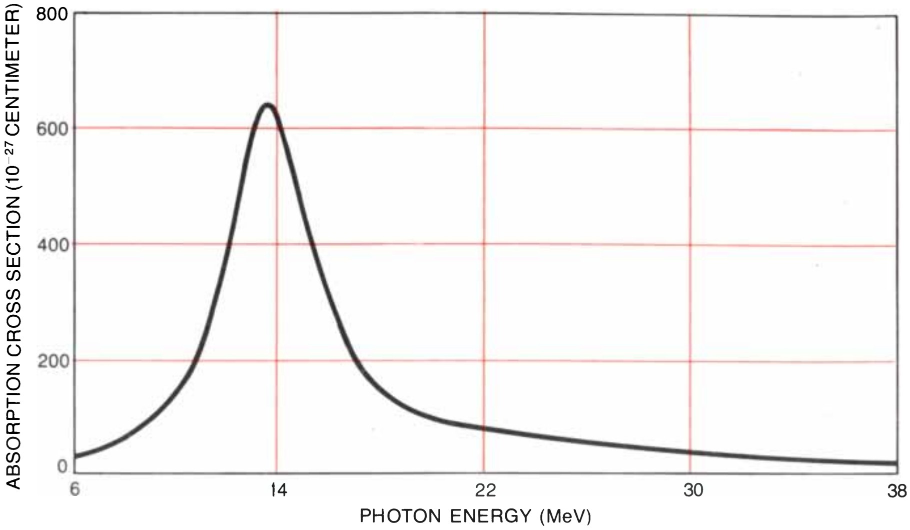
In order to excite the shape vibrations it is necessary not only to impart a given energy to the nucleus but also to distribute the energy in such a way that the nucleons are set in motion in different directions. Even though the vibrational energy of the quadrupole mode is from 10 to 20 MeV, within the energy range of gamma rays, they do not excite the quadrupole mode because they cannot meet the second condition. Gamma rays interact with the nucleus through the electromagnetic force; they can accelerate only the protons and, given the wavelength of the gamma rays, accelerate them only in one direction.

One answer to the problem is to excite the vibration by the inelastic scattering of a particle from the nucleus rather than by a complete absorption process. In classical physics a particle that strikes an extended object and bounces off can cause localized motion and leave the object vibrating. A similar phenomenon can take place in a quantum-mechanical system. The particles that can serve as projectiles for inducing nuclear excitations include electrons, protons and more massive nuclear particles such as the alpha particle (the nucleus of the helium-4 atom).

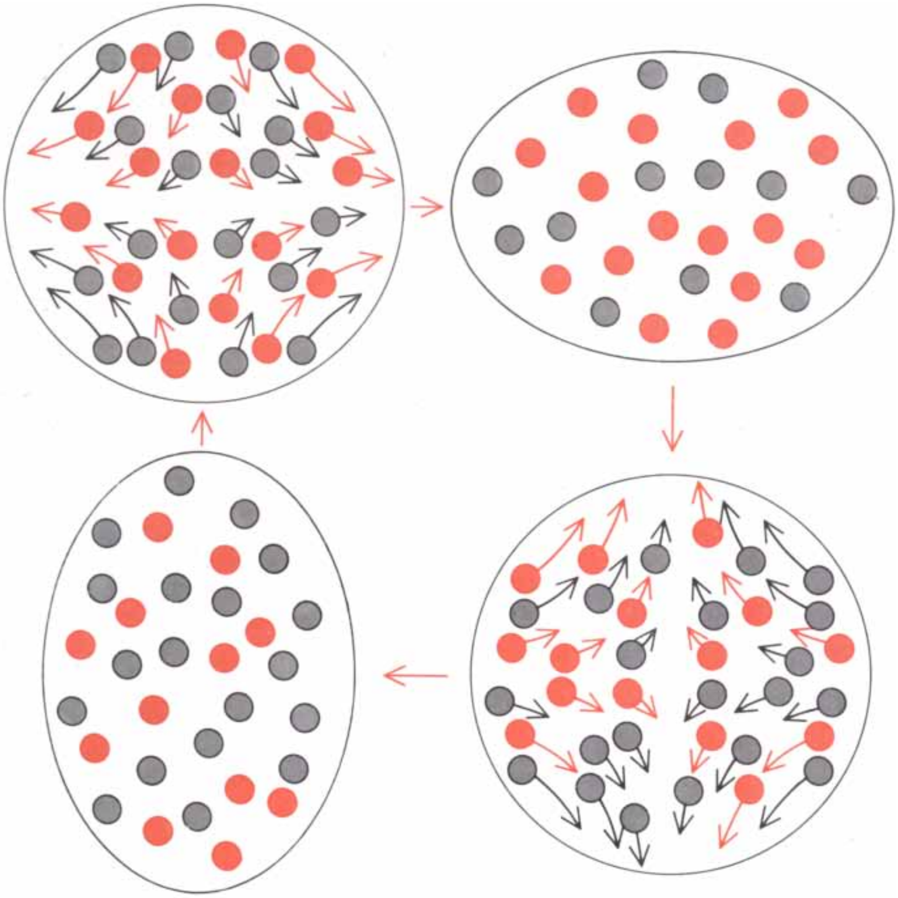
There are several advantages to choosing nuclear projectiles such as protons or alpha particles to excite the nucleus. The principal interaction in such events is mediated by the strong nuclear force, which is charge symmetric, that is, the same for protons and neutrons. The charge symmetry makes it easier to excite vibrations such as the quadrupole in which the protons and neutrons move together. Vibrations in which protons move opposite to neutrons, such as the dipole, are not excited at all by charge-symmetric forces. Thus to observe the quadrupole vibration without interference from the dipole, it is better to employ strongly interacting projectiles such as alpha particles rather than electrons, which interact with the electric field and can give rise to both kinds of vibration.

Another advantage of nuclear projectiles is that the different shape modes can be selectively excited. A fast proton going by a nucleus transmits a fleeting impulse to the nearby target nucleons, causing them to move in the wake of the projectile. The range of the strong nuclear force is small compared with the size of a nucleus, raising the question of how a coherent vibration of the entire nucleus can be set in motion. To understand how particular vibrations can be selected by the scattering process requires one of the basic notions of quantum mechanics, the wave aspect of particles. A proton going by a nucleus can be thought of as a wave enveloping the nucleus; the wave is subject to the laws of diffraction like any other wave.

Depending on how the wave interacts with the nucleus, it is diffracted with a characteristic pattern. For example, a light wave passing over a black sphere is diffracted into the shadow region, forming a pattern of rings in the center of the shadow. In the case of nuclear projectiles and targets the diffraction pattern depends on several factors. In order to produce a clear diffraction pattern at all, the wave must be absorbed in the interior of the nucleus. In physical terms, when the projectile penetrates the nu-



PEAK IN THE ABSORPTION CROSS SECTION of lead nuclei is evidence that they are vibrating in the dipole mode. Cross section is a measure of the fraction of the photons absorbed by nuclei. It increases dramatically when the photon frequency matches a vibrational frequency of the nuclei, a phenomenon called resonant absorption. The resonance in this nucleus is at a photon energy of 14 million electron volts (MeV), equivalent to a frequency of 3×10^{21} hertz.



GIANT QUADRUPOLE VIBRATION can be excited by bombarding nuclei with projectiles that interact with protons and neutrons equally by means of the strong force. The quadrupole vibration therefore differs from the dipole in that the protons and neutrons move together rather than in opposite directions. As in the case of the dipole mode, the vibrational frequency can be calculated from the inertia of the system, the restoring force and the laws of classical mechanics. Again the inertia is proportional to the volume of the nucleus, but it is not immediately clear how to describe the restoring force. If the restoring force is assumed to be analogous to surface tension, as in a vibrating drop of water, the force is proportional to tbe surface area of the nucleus. If the restoring force is assumed to be analogous to stress energy, or resistance to deformation, as in a vibrating elastic solid, the force is proportional to the volume of the nucleus. Experimental data indicate that the frequency of the giant quadrupole vibration varies inversely with the radius of the nucleus, which favors the model based on an elastic solid.

cleus, it must interact so strongly that it does not emerge intact. The position of the rings in the diffraction pattern is determined by the size of the nucleus, the wavelength of the projectile and the type of vibration being excited.

The interaction of the wave with the nucleons on the nuclear surface alters the form of the wave. After the interaction each small area of the nuclear surface acts as the origin of a spreading wavelet. The phase of the wavelet depends on the motion of the originating surface. Wavelets originating from outward-moving surface areas have the same phase, or sign, whereas wavelets from inward-moving surface areas have the opposite phase. As the wavelets propagate outward from the nuclear surface they overlap and interfere. Whether the interference is constructive or destructive at any point depends on the phase of the various wavelets at that point.

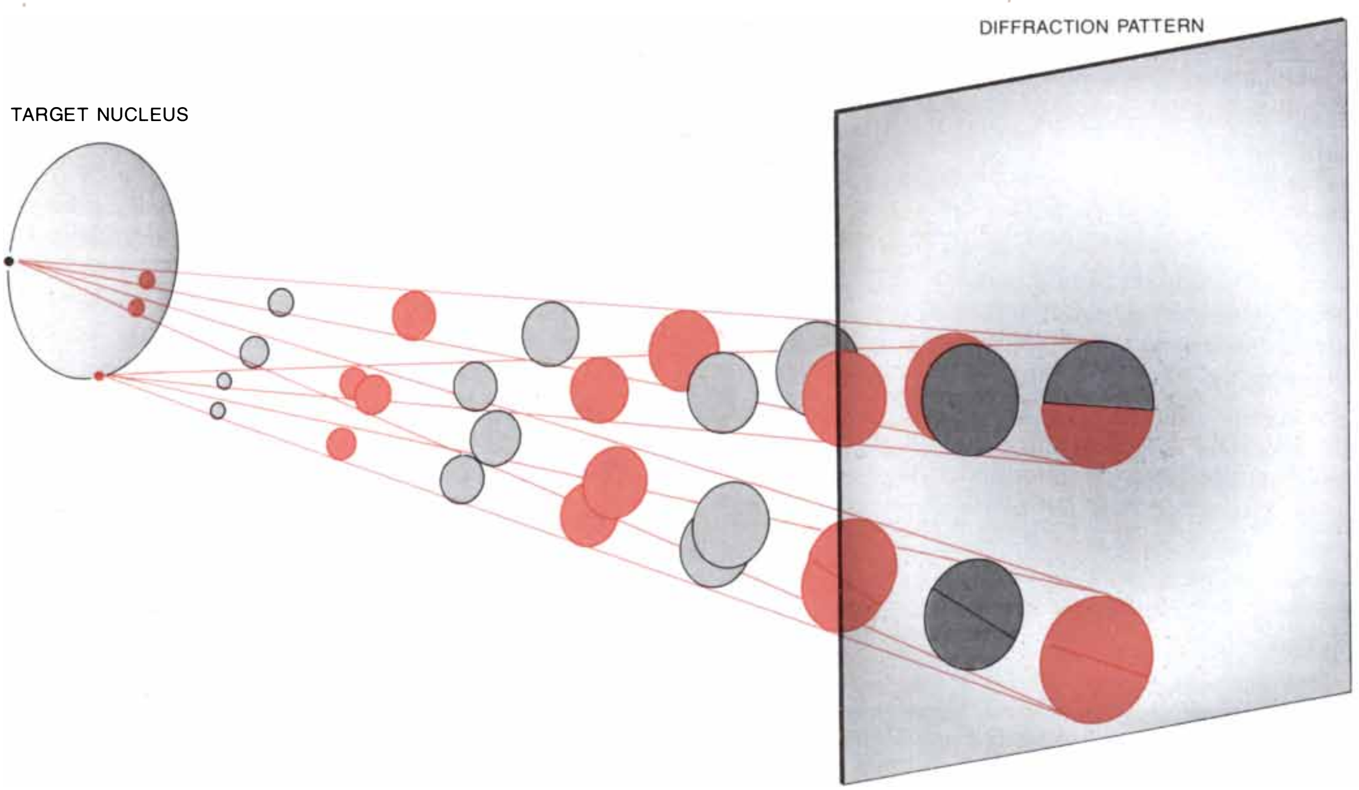
The nature of the diffraction pattern can be understood by considering a plane "downstream" from the target nucleus and perpendicular to the beam axis. If the phase of the wave were not altered by its passage near the nucleus, the phase of the wavelets reaching a given position on the plane would be determined entirely by the distance from the point of diffraction to that position. Of particular importance, wavelets diffracted to the center of the plane along the beam axis in the shadow of the nucleus would all travel the same distance and would arrive in phase. The resulting constructive interference would create a bright spot (a region of large wave amplitude) in the center of the plane.

When the wave stimulates a quadrupole vibration, however, the phases do not remain unaltered. On the contrary, the wavelets originating from the elongating sides of the nucleus are initially 180 degrees out of phase with the wavelets originating from the contracting sides. As a result the interference in the center of the plane is destructive. There are other points in the plane, however, where the distance to an elongating region and the distance to a contracting region differ by exactly half a wavelength. The wavelets arrive at these points in phase, interfering constructively. Because the axes of the inward- and outward-deforming regions of the nucleus can have any orientation, the strongest diffraction is directed into a ring centered on the axis of the beam.

The amplitude of the wave function of a particle at any point in space determines the probability that the particle will be found at that point. Consequently the projectile particles are most likely to be scattered into regions of constructive interference, where the amplitude of the wave function is largest.

Like absorption, inelastic scattering can be measured in terms of a cross-sectional area. Even in the absence of a vibrational resonance a certain fraction of the projectiles bombarding a target are scattered inelastically, and the fraction scattered in any one direction varies smoothly with the angle between the beam axis and the scattering direction. The quadrupole vibration shows up as a sharp increase in the number of particles scattered inelastically in particular directions.

For a quantitative prediction of the cross section as a function of scattering angle the interaction of the wave with the nucleus is described by a phenomenological optical model. The wave equation is then solved numerically to find the precise cross section. The cross section for the quadrupole mode reaches its maximum value at an angle that depends on the size of the nucleus as well as on the wavelength of the projectile. For example, in the case of the lead 208 nucleus bombarded by 100-MeV al-



ANNULAR DIFFRACTION PATTERN formed when projectiles are scattered by nuclei in the target is associated with the quadrupole vibration. The projectile particle can be described quantum-mechanically as a plane wave perturbed by its interaction with the surface of the nucleus. The interaction of the wave with the nucleons is attractive along one axis, so that the nuclear surface is pulled outward, and repulsive along the perpendicular axis, so that the surface is pushed inward. The single plane wave is diffracted into many small circular waves that propagate outward and interfere with one another. The waves originating from the inward-deforming regions of the nucleus are 180 degrees out of phase with those from the outward-deforming regions. (Here the phase differences are represented by color differences.) The pattern created by the interference of the waves is determined by this phase difference. Waves arrive at any point along the beam axis out of phase and cancel. There are points off the axis, however, where the waves arrive in phase and interfere constructively.

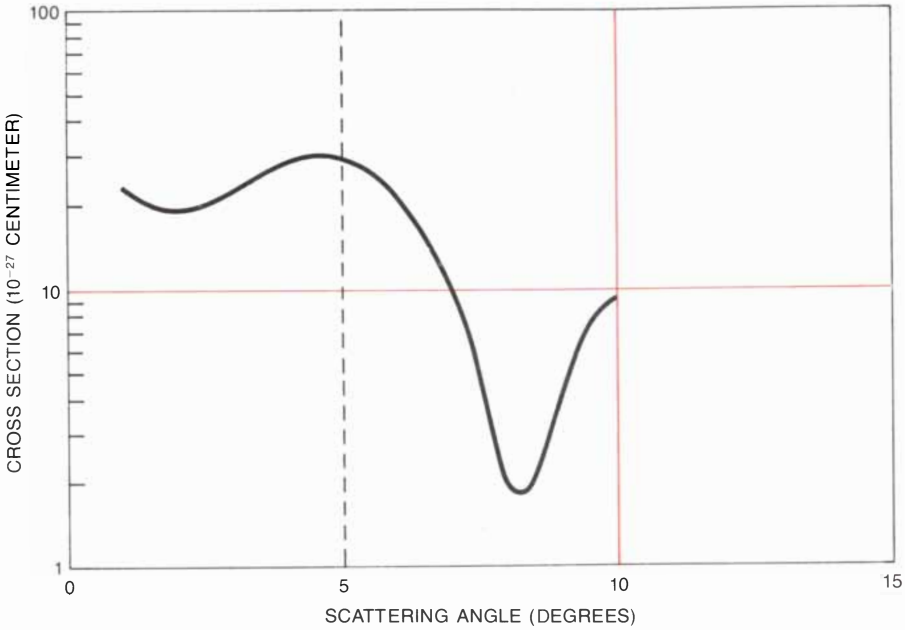
pha particles, the main diffraction peak is at five degrees from the beam axis.

The third nuclear vibrational mode to be observed, the giant monopole vibration, is excited and detected in much the same way as the giant quadrupole. The monopole vibration is a "breathing" mode: the nucleons move inward and outward from the center of the nucleus in phase with one another, so that the nucleus expands and contracts. Intuitively it would seem that the best way to induce this motion would be to push uniformly on the surface of the sphere or, equivalently, to pull outward in a radial direction.

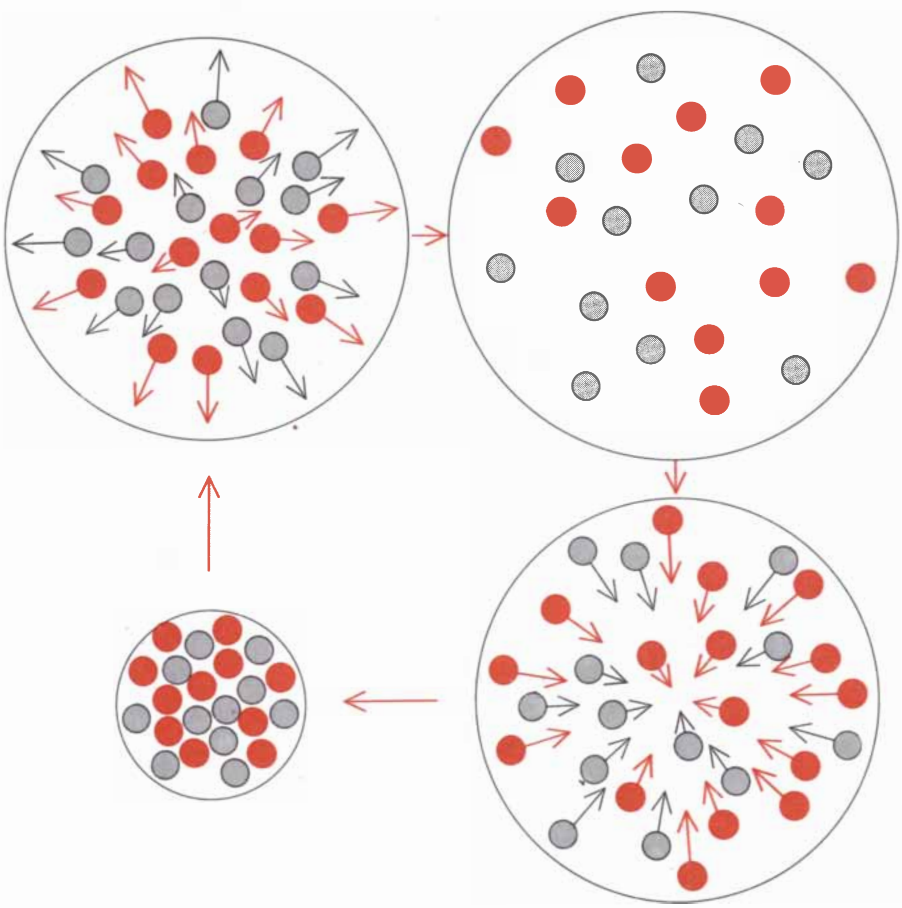
Such perfect conditions for exciting the monopole vibration cannot be achieved with inelastic scattering. If the projectile hits the nucleus, it loses most of its energy and leaves the nucleus in a very highly excited state. Only when the projectile grazes the surface of the nucleus is the interaction gentle enough to excite the simple vibratory motion. The part of the surface the projectile wave acts on is a ring-shaped region circling the beam axis. As in the quadrupole excitation, the interaction gives rise to many wavelets originating from each small area of the nuclear surface. The wavelets interfere as they spread out from the surface of the sphere. Because the monopole excitation is completely symmetric, the wavelets start out from the curved surface with the same phase. Hence along the beam axis, where the distance to all areas on the periphery of the sphere is equal, the wavelets arrive in phase and interfere constructively.

The diffraction pattern formed by the particles exciting the monopole vibration is a prominent spot centered on the beam axis and surrounded by faint rings. In other words, the monopole vibration is characterized by a maximum in the scattering cross section at zero degrees from the beam axis. The pattern is essentially the same as the one formed by particles diffracted by a nonvibrating nucleus. Indeed, it was largely for this reason the monopole vibration was harder to detect than the quadrupole vibration; it proved difficult to separate the particles that are inelastically scattered along the central axis of the beam from the particles that pass through the target without interacting with the nuclei. The only way to discriminate between the inelastically scattered particles and the beam particles is by measuring the energy dependence of the inelastic scattering.

At this point it is possible to return to the question of why it took so long to obtain experimental evidence of quadrupole and monopole vibrations. Rough predictions of the vibrational frequencies of these modes had been



PERCENTAGE OF PARTICLES scattered at a given angle to the beam axis increases sharply when the quadrupole vibration is excited. The scattering cross section varies with angle in a way that reproduces the diffraction pattern associated with the quadrupole mode. The scattered particles have a specific energy, namely the energy of the impinging beam minus the vibrational energy of the quadrupole mode. In this case the projectiles are alpha particles with an energy of 96 MeV striking a target consisting of the isotope samarium 144. The quadrupole mode is preferentially excited with a scattering angle of five degrees to the beam axis.



GIANT MONOPOLE VIBRATION is also called the breathing mode: in it the nucleus expands and contracts. In the model of the giant monopole vibration based on classical mechanics the restoring force is the resistance of nuclear matter to compression. The observed frequency of the monopole vibration gives a value for the compressibility coefficient of nuclear matter.

made from theoretical models of the nucleus years before the vibrations were seen in the laboratory. Thus experimental physicists had known for some time approximately where to look and what they could expect to see. What they lacked were instruments of sufficient power and sensitivity, in combination with accelerators providing projectiles of sufficient energy.

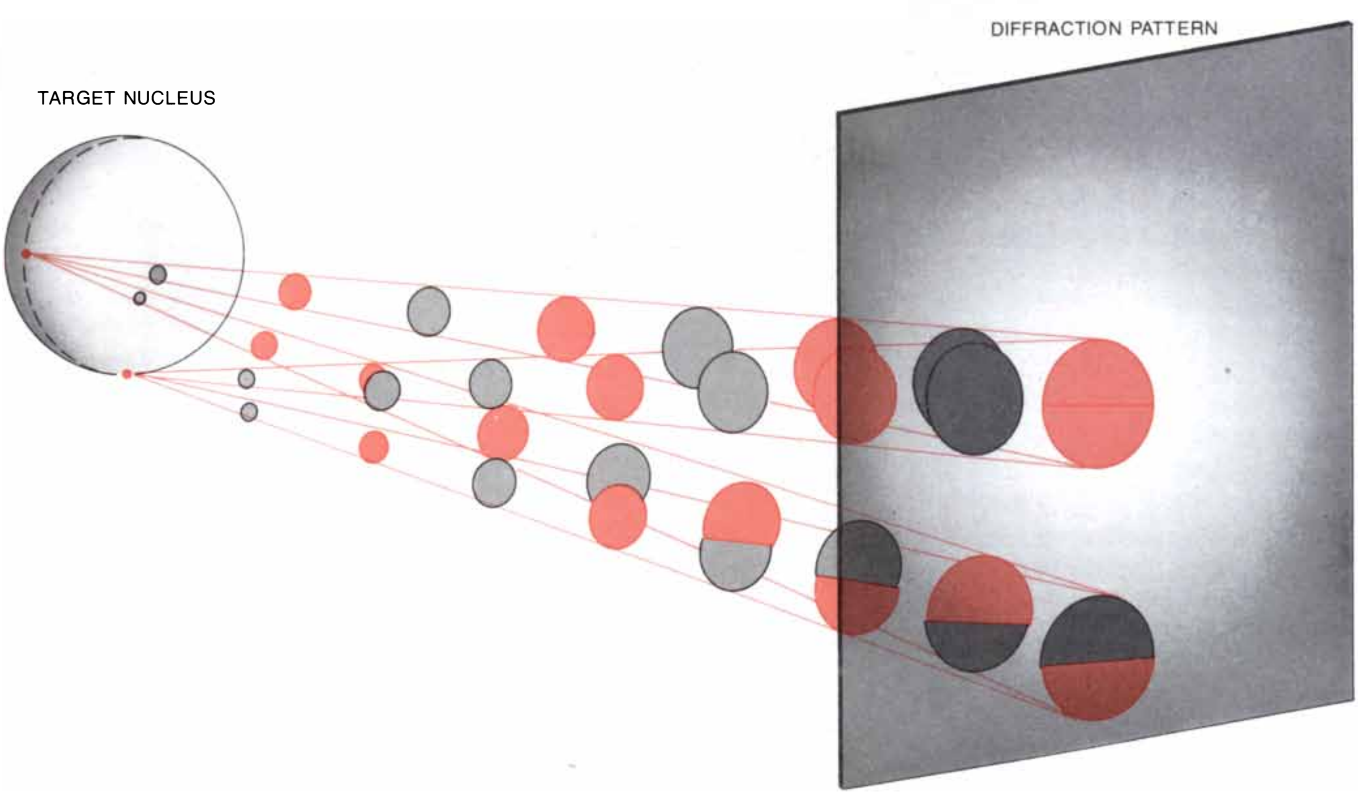
Even though the vibrational energy of the quadrupole and monopole modes lies in the range from 10 to 25 MeV, the projectile energy must be several times higher. The velocity of the projectile is an important consideration in judging its suitability as a vibrational probe. The velocity of a particle of a given mass depends on its energy. High-energy projectiles are needed to excite giant vibrations because lower-energy (and hence lower-velocity) projectiles interact with the target for a relatively long time, allowing more complex excitation processes to take place. A slower projectile might induce a vibration composed of several fundamental modes, or it might induce more complicated motions in the nucleus by exchanging nucleons with the target. These processes do not give rise to a diffraction pattern but contribute to a background cross section that must be subtracted from the data before the vibrational modes can be analyzed.

High-energy projectiles are needed for another reason. Simple diffraction patterns are obtained only if the wavelength, and therefore the energy, of the exciting projectile does not change significantly during the excitation process. A projectile that excites a nuclear vibration gives up an amount of energy equal to the energy of the vibrational mode; the higher the initial energy of the projectile, the less significant the resulting change in wavelength. On the other hand, from the experimental point of view the small change in energy makes it difficult to distinguish the diffracted particles from the much larger number of beam particles that do not interact with the target nuclei.

The discovery of the quadrupole and monopole vibrations in the 1970's was based on the development of two types of scientific instrument: machines capable of accelerating particles of various kinds to energies higher than 50 MeV and sensitive magnetic spectrometers (spectrographs) capable of separating particles of slightly different energies. Once the instruments were available nuclear vibrations began to be reported in rapid succession. The giant quadrupole vibration was first seen in electron-scattering measurements done at Darmstadt in West Germany by Rainer Pitthan and Th. Walcher and in proton-scattering measurements done at the Oak Ridge National Laboratory by F. E. Bertrand and M. B. Lewis. The giant quadrupole vibration has now been excited in practically all species of nuclei; it has a frequency that varies inversely with the radius of the nucleus. The equivalent energy is in the range from 10 to 20 MeV.

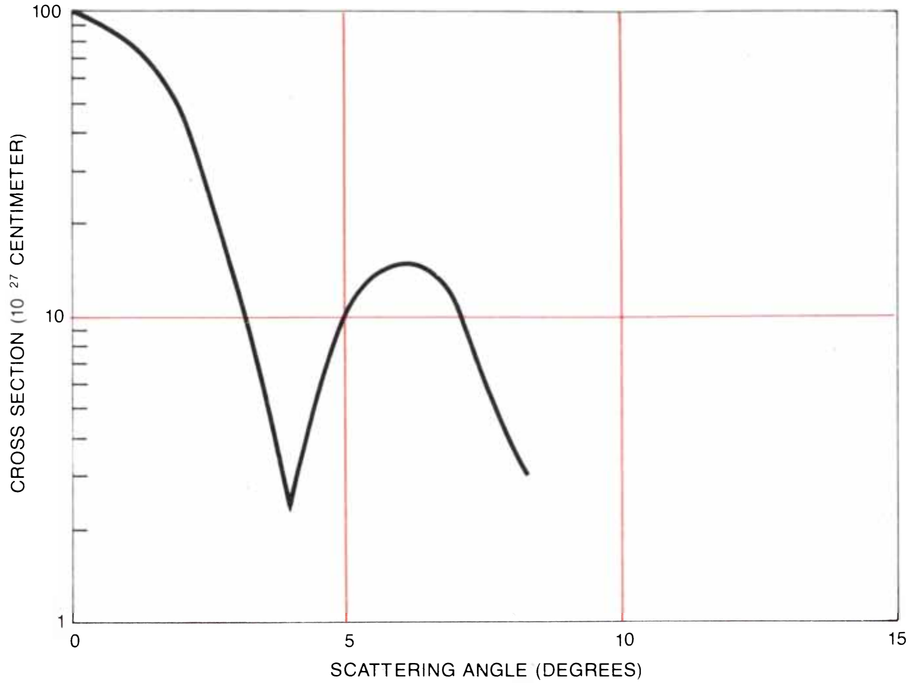
The first indications of the breathing mode were found with deuteron scattering by Nadine Marty and her collaborators at the Institute of Nuclear Physics of the University of Paris (Paris-Sud) at Orsay. Observation of the diffraction peak in the forward direction, providing definitive proof of the monopole mode, was first obtained in 1977 with alpha-particle scattering by a group at Texas A&M University led by Dave H. Youngblood. Like the quadrupole, the breathing mode has a frequency that decreases with increasing size of the nucleus. In energy units the range is from 15 to 25 MeV, slightly higher than the energy of the quadrupole. The giant octupole vibration was first detected in 1980 in scattering experiments done with 800-MeV protons at the Los Alamos Scientific Laboratory. The octupole mode has an energy ranging from 20 MeV in heavy nuclei to 30 MeV in lighter nuclei.

The history of theoretical speculation about nuclear vibrations and the interplay between theory and experiment



DIFFRACTION PATTERN **associated with the monopole vibration is a point on the beam axis. The pattern is known as the Poisson spot; in the 19th century Siméon Denis Poisson proved mathematically that if light consists of waves, there should be a small bright spot in the shadow of a sphere. Poisson put forward his argument as a proof that light cannot have a wave nature, since no such diffraction pattern had been seen at the time. The pattern is formed by a disk or a sphere whose surface absorbs light uniformly. Particles that interact uniformly with the surface of a nucleus, as do those that excite the monopole vibration, are diffracted with the same spot.**

form an equally interesting thread in the story. Theoretical calculations predicted a nuclear phenomenon for which there was no experimental evidence, giving physicists some clues, albeit imperfect, about where to look and what it would look like if it were found. The experimental evidence in turn showed where the simplified theoretical models of vibrating nuclei were in error and helped to lay the groundwork for more powerful models based on more accurate assumptions.



SCATTERING CROSS SECTION of nuclei vibrating in the monopole mode peaks on the beam axis (at zero degrees). The graph shows the angular distribution of alpha particles inelastically scattered from lead nuclei. The monopole vibration was identified seven years later than the quadrupole vibration, largely because of the difficulty of separating particles scattered along the beam axis from the particles passing through the target without being deflected.

The frequency of a vibrating body depends on two properties: the inertia of the constituent particles, which governs how quickly they respond to a force, and the restoring force, which opposes the displacement of the particles. The larger the inertia is, the slower the body vibrates, and the stronger the restoring force, the higher the vibrational frequency. More quantitatively, the square of the vibrational frequency is directly proportional to the strength of the restoring force and inversely proportional to the inertia.

In 1944 the Russian physicist Arkadii B. Migdal predicted the frequency of the dipole vibration by applying those simple mechanical laws to the nucleus. It is easy to construct a plausible model of the inertia of the nucleus vibrating in the dipole mode. Assuming that all the nucleons are moving with the same speed in the dipole vibration (but in different directions for protons and neutrons), the inertia is equal to the total mass of the nucleus. Modeling the restoring force is more difficult, and it is here the simplification in theoretical models of the nucleus generally comes in. In the case of the dipole mode the restoring force is mainly due to the attractive strong force between protons and neutrons. The strength of this force in a vibrating nucleus can be inferred from the binding energy of various nuclei. Among all nuclei those with nearly equal numbers of protons and neutrons have the greatest binding energy, apart from the effects of the protons' electric charge. A model of the proton-neutron interaction is constructed to fit the systematic variation in binding energy with nuclear composition; the same model is then used to calculate the restoring force when the nucleons are displaced in the dipole mode.

The measured frequency of the dipole vibration agrees remarkably well with the frequency Migdal predicted from his model. Nevertheless, we now know that more than the potential energy of separated protons and neutrons needs to be considered in modeling the restoring force of the dipole mode; furthermore, the inertial mass of the system is not simply that of free nucleons. The inertial mass is slightly smaller than Migdal assumed because in a quantum-mechanical description of the nucleus, protons and neutrons are not the only particles present. There are also pi mesons, or pions, the subatomic particles responsible for transmitting much of the strong interaction between nucleons. Pions are much lighter than nucleons and hence reduce the average inertial mass. For the dipole vibration, however, the discrepancies are minor and the predictions based on binding-energy calculations are essentially correct.

Another model of the nucleus, the liquid-drop model, was proposed by Niels Bohr in 1936. Bohr took note of the fact that a nucleon in the interior of a nucleus is surrounded by other nucleons that pull it equally in all directions, so that the net force is zero. A nucleon at the surface, on the other hand, has other nucleons on only one side, and so it is pulled toward the center. The effect on the surface nucleons is analogous to the surface tension of a drop of water; in both cases the force tends to make the system take on a spherical shape.

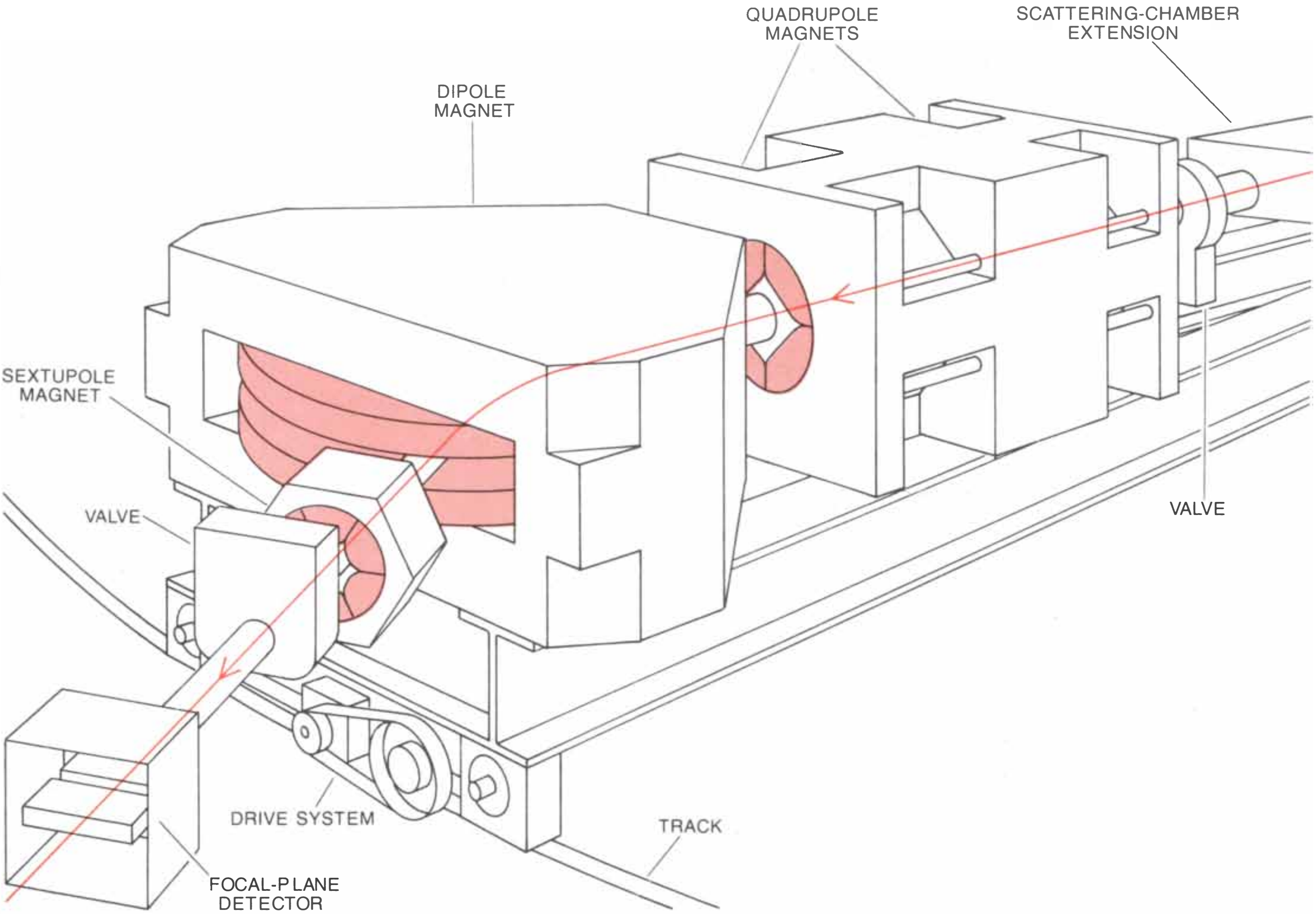
The frequencies of liquid-drop vibrations were worked out by Lord Rayleigh at the end of the 19th century. In 1952 Niels Bohr's son, Aage Bohr, and Ben R. Mottelson suggested that the liquid-drop model might be applied to vibrations such as the quadrupole vibration, in which the nucleus oscillates between spherical and deformed shapes. In these vibrations the dominant restoring force should be the nuclear surface tension. Because the nuclear surface tension is relatively weak, Bohr and Mottelson predicted that the frequency of the quadrupole mode would be low. Quadrupole-shaped motions were found at low frequencies, but only a few of the nucleons, typically fewer than 10 percent, participate in the motions. The giant quadrupole vibration has a frequency much higher than the predicted one.

The problem with the liquid-drop model is that nucleons are not free to move in the nucleus in quite the same way as molecules are free to move in a drop of liquid. Nucleons must obey the Pauli exclusion principle, which states that no two identical nucleons can have exactly the same quantum-mechanical state of motion. As a consequence two protons or two neutrons having the same spin orientation must occupy different orbits in the nucleus. Their motion is thereby constrained to some extent; they must keep out of each other's way.

To include the quantum effects properly requires a much more elaborate theory, which I shall describe below. The theory gives a quite simple and unexpected result for the giant vibrations, namely that the nucleus has a rigidity making it respond more like a solid than a liquid. The restoring force for the quadrupole vibration is governed by the elastic constant of the nuclear medium. The value of the elastic constant can be estimated from the quantum theory; it is proportional to the kinetic energy of the nucleons in their shell-model orbits. Similar physical principles determine the forces between ordinary atoms in a solid. If two atoms are pushed together, there is a repulsive force that can be traced to the electrons' increased kinetic energy, which in turn can be traced to the requirements of the exclusion principle. Nuclear matter is, however, much stiffer than ordinary matter because the nucleons in the nucleus have much higher kinetic energy than the electrons in the atom. Given the elastic constant of nuclear matter from quantum theory, the frequency of the quadrupole vibration can be calculated from a formula for the elastic vibrations of a sphere worked out more than a century ago by the British physicist A. E. H. Love.

The great stiffness of nuclear matter clearly suggests that nuclear quadrupole vibrations should be of high frequency. How, then, is one to understand the low-frequency quadrupolelike motions in which a few percent of the nucleons take part? If the disturbance that excites a vibration is slow, it is relatively easy to move a few of the nucleons into empty orbits of nearly the same energy. The exclusion principle does not forbid such a transition, and so the nucleus acts more like a liquid. When the nucleus is struck by a high-speed projectile, on the other hand, there is no time for internal rearrangement and the giant vibrations are more prominent. Overall the nuclear response to an external force resembles that of Silly Putty, which responds like an elastic solid to sudden forces and like a viscous liquid to slow ones.

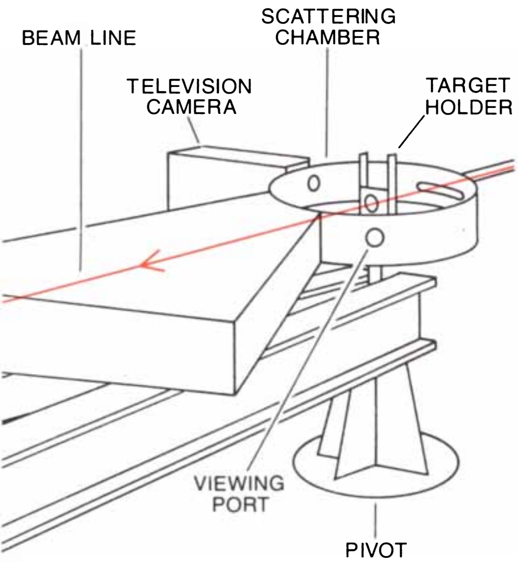
The frequency of the monopole vibration can also be worked out from a simple model of the nucleus in which only the inertia of the nucleons and the dominant restoring force of the vibration are considered. The restoring force in this case is the resistance of nuclear matter to compression. (In the "breathing" motion of a monopole vibration the nucleus is alternately compressed and rarefied.) The breathing model was suggested long before quantum calculations of the monopole vibration were available; later mean-field theory confirmed the



THE MAGNETIC SPECTROGRAPH, a sophisticated version of the mass spectrometer, separates particles that have lost a specific amount of energy in exciting a nuclear vibration from other particles in a beam. The beam of particles from an accelerator is directed onto a thin foil of target material in the scattering chamber of the spectrograph. After interacting with the target the projectiles are sorted according to energy by a dipole magnet that deflects them into a circular path. A particle with a higher energy follows a wider arc in the magnetic field than a particle with a lower energy. Only particles with the selected energy pass through the magnet to the detector. Quadrupole magnets focus the beam; a sextupole magnet corrects focusing aberrations. In the Michigan State University spectrograph shown here the detector is a multiwire gas ionization chamber. The high-energy particles ionize atoms of gas in the chamber. The electrons lib-

validity of a force based on the compressibility of nuclear matter. It has not been possible, however, to calculate the compressibility coefficient reliably or to measure it by other means, so that the observation of the monopole vibration provides the most direct information on the compressibility of nuclear matter. The coefficient derived from the breathing model together with the observed frequency of the vibration has nonetheless received spectacular confirmation from a very different source.

When a massive star comes to the end of its life, the inward-pushing gravitational forces are no longer balanced by the outward pressure of hot gases produced by nuclear reactions and the star begins to collapse. Gravitation compresses the core of the star to the density of nuclear matter, at which point the great resistance of nuclear matter to further compression begins to counteract the gravitational collapse. Depending on the compressibility of nuclear matter (among other things), the collapse may continue to the formation of a black hole or it may be stopped by an outward-moving shock wave that blows off the outer layers of the star in a supernova explosion. A dense neutron star is left at the center of the explosion; the maximum mass such a star can have depends directly on the compressibility coefficient. To date there is good agreement between the observed range of neutron-star masses and the compressibility coefficient deduced from the nuclear monopole vibration.



erated are attracted to the closest point on an array of parallel wires in the detector chamber. Currents induced in the wires are individually detected to determine the position of the particle. The spectrograph can be pivoted around the chamber so that the scattering probability can be measured as a function of angle.

An exact theory of nuclear structure (as opposed to a model) would specify in detail the forces exerted by each nucleon on every other nucleon. In a nucleus of moderate size, for example oxygen 16, there are 120 pairings of the nucleons to be considered. Furthermore, according to quantum mechanics, all the possible configurations of the nucleons have to be considered simultaneously, and each configuration must be assigned some amplitude in the quantum wave equation. The relative amplitudes would be independent of time in a description of the unexcited nucleus, but in the vibratory motion they would of course vary with time. In either case the task of describing the system in this way is mathematically intractable. The shell model offers a simplified description; it is approximate but nonetheless retains most of the quantum physics and is quite accurate in accounting for many of the properties of nuclei. The shell model does not attempt to calculate all the interactions of the individual nucleons. Instead a single potential, or mean field of force, is defined; it represents the collective effect of all the particles on any given particle. The quantum wave equation is solved for one particle at a time in this common potential.

The problem then becomes one of choosing the appropriate potential. The starting point is the distribution of particles in an unexcited nucleus; summing the fields associated with the individual nucleons yields an approximate collective potential. The next step is to go back and alter the wave functions of the individual particles in accordance with the estimated potential. By successive approximation one finds a potential and a set of particle wave functions that are mutually consistent. This method was introduced by William Hartree as a means of describing the electrons in an atom; the technique has been very useful in nuclear physics and is the basis of the nuclear shell model. The usual shell model is a static theory, but one can easily allow the potential field to depend on time in order to describe vibrations.

The time-dependent mean-field theory was first applied to nuclear vibrations in the 1960's. The change in the potential caused by an external force is calculated from the change in the distribution of the nucleons from one point in the vibrational cycle to the next. Again a repetitive procedure is employed to arrive at an accurate description of the field. The potential acting on the nucleons is calculated from their movement. The resulting approximation of the field in turn serves to refine the initial wave functions, which specify a new approximate potential. Ultimately a set of wave functions is found that fits both the distribution of particles and the potential field for the succeeding point in the vibrational cycle. The procedure can be simplified for small-amplitude vibrations.

The mean-field theory, rather than the simpler models of the nucleus, has provided the theoretical basis for vibrational studies for the past 10 years. In 1952 Mottelson predicted the frequency of the quadrupole vibration on the basis of the liquid-drop model; as I have stated, the predicted frequency turned out to be too low. In 1969 Mottelson predicted the frequency of the same vibration on the basis of the mean-field theory. When the vibration was finally detected in 1971, the observed frequency agreed with his second prediction.

The dipole, quadrupole and monopole vibrations are all geometric deformations with clear analogues in the vibrations of ordinary macroscopic bodies. The spin vibrations of the nucleus are quite different. Spin is an intrinsically quantum-mechanical property, and analogous motions are not to be found in macroscopic systems. The spin vibrations differ from the shape vibrations in that the spatial distribution of the nucleons may remain frozen with only the spin orientation varying with time. When there is no spatial motion, the Pauli principle is more restrictive and only a few of the nucleons can participate in the spin vibration. In the unexcited nucleus the spins are nearly all paired; the excitation process tips the spins of some of the nucleons and they precess. There are several ways the spin precession can be induced and detected. The nucleon spin has an associated magnetic moment and can interact through the electromagnetic field. Hence photon absorption and electron scattering are two techniques for studying spin properties of nuclei.

There are other fields as well that interact with nucleon spins. One of these is the field of the pi meson. When a pion is absorbed by a nucleon, it changes the nucleon's spin orientation. An individual nucleon is surrounded by a pion field, so that even nucleon projectiles can induce spin vibrations. Another field that interacts with spin is the weak field, mediated by the recently observed *W* particle. The weak field is responsible for beta decay, one of the major processes in the formation of stable elements. In beta decay a proton is changed into a neutron or vice versa, usually with a reorientation of the spin of the affected particle.

One of the best experimental techniques for studying spin vibrations is the inelastic scattering of high-energy protons from a nuclear target. The pion field of the proton interacts with nucleons in the target, and much of the character of the interaction is due to the pion. If the exchanged pion is a neutral one, the electric charges of the projectile and the target remain the same and the collision is seen as an instance of ordinary inelastic scattering. It is also possible to exchange a charged pion, in which case the charges of the projectile and the target both change. The bombarding proton is turned into a neutron, and in the target one of the neutrons is changed into a proton. In spite of the charge exchange the same diffractive effects determine the angular distribution of the scattered beam. The simplest possible spin vibration, which is uniform over the entire surface of the nucleus, gives rise to a diffraction pattern with a peak at zero degrees, just as the giant monopole vibration does.

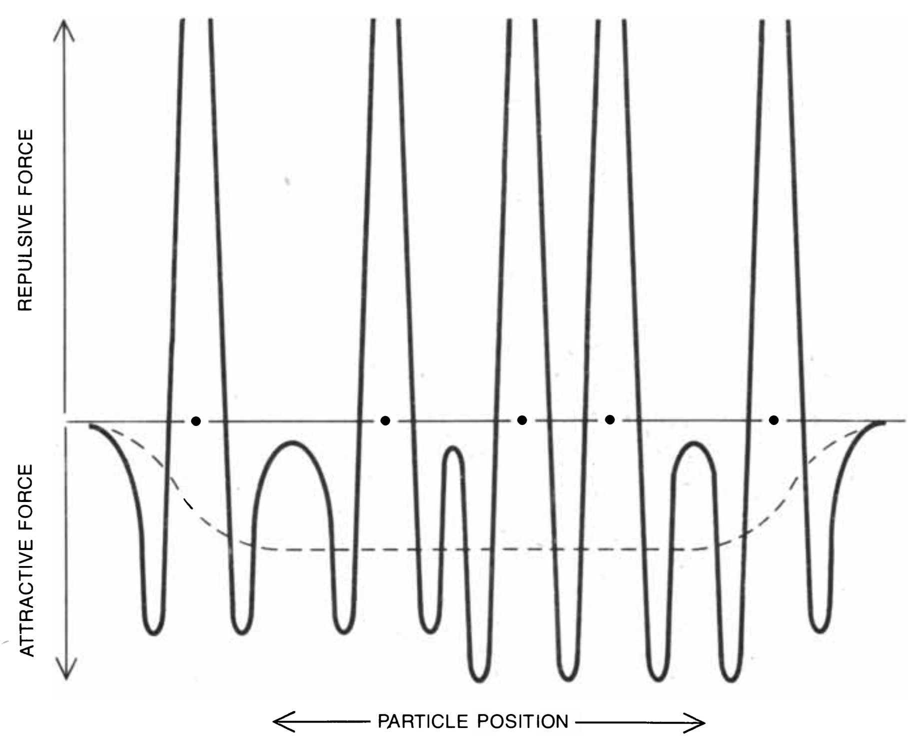
The charge-exchange spin vibration is called the giant Gamow-Teller resonance because of its relation to the spin-flip beta-decay process originally described by George Gamow and Edward Teller. The giant Gamow-Teller resonance was first observed in 1976 by Robert Doering, who was then a graduate student at Michigan State University. The energy of the proton beam was rather low and the forward diffraction peak was just barely distinguishable from the background of neutrons arising from more complex interactions with the target. More recent experiments done at the Indiana University Cyclotron Facility with higher-energy beams have given diffraction patterns in which the background is much reduced. The clearer patterns allow the properties of the vibration to be measured more accurately. The energy of the vibration lies in the range from 10 to 15 MeV.

The measured energies can be compared with theoretical predictions to check our understanding of the spin forces. There are no classical models for the spin vibrations, but the mean-field theory can be applied. The predictions agree well with the measurements, demonstrating that the theoretical description of the spin forces, based largely on the pion fields of the nucleons within the nucleus, is essentially correct. Even apart from theory, the existence of the giant Gamow-Teller resonance had long been suspected from indirect evidence. Beta-decay transition rates tend to be much lower than expected from the shell model, showing that the amount of spin precession at low frequency is small. From this finding one infers that the spin forces cause a higher-frequency precession—the giant vibration.

The last vibration I shall discuss is the magnetic-dipole resonance, which is much like the Gamow-Teller resonance. The main difference is that the number of protons and neutrons is not altered in the magnetic vibration. In the magnetic-dipole resonance the spin of the proton is tipped in the direction opposite to the spin of the neutron. Since the magnetic moments of protons and neutrons have opposite signs, the overall magnetic moment is maximized by this configuration. As the name of the vibration suggests, the magnetic field has a dipolar pattern.

It is relatively easy to study the magnetic-excitation properties of light nuclei because the structure is fairly simple when only a few nucleons are present. Depending on the nucleus and the degree of pairing, a prominent spin vibration may be present. Until recently, however, little was known about the magnetic excitations of heavy nuclei, in spite of a long search for the magnetic-dipole resonance. This situation has changed recently; in 1981 the magnetic-dipole resonance was found in the nucleus of zirconium 90 by using inelastic proton scattering to excite the spins of the target nucleons. The experiment was done at Orsay with 200-MeV protons and a spectrograph capable of identifying inelastic-scattering events at angles as small as two degrees. The excitation energy of the magnetic-dipole resonance ranges from 8 MeV in heavy nuclei to 15 MeV in light nuclei.

One interesting and somewhat puzzling aspect of both the Gamow-Teller and the magnetic-dipole resonances is that only a third of the expected number of nucleons seem to participate in the vibration in heavy nuclei. This is one of the reasons it was difficult to observe the magnetic-dipole resonance. The reduced strength has led nuclear physicists to look for a complete understand-



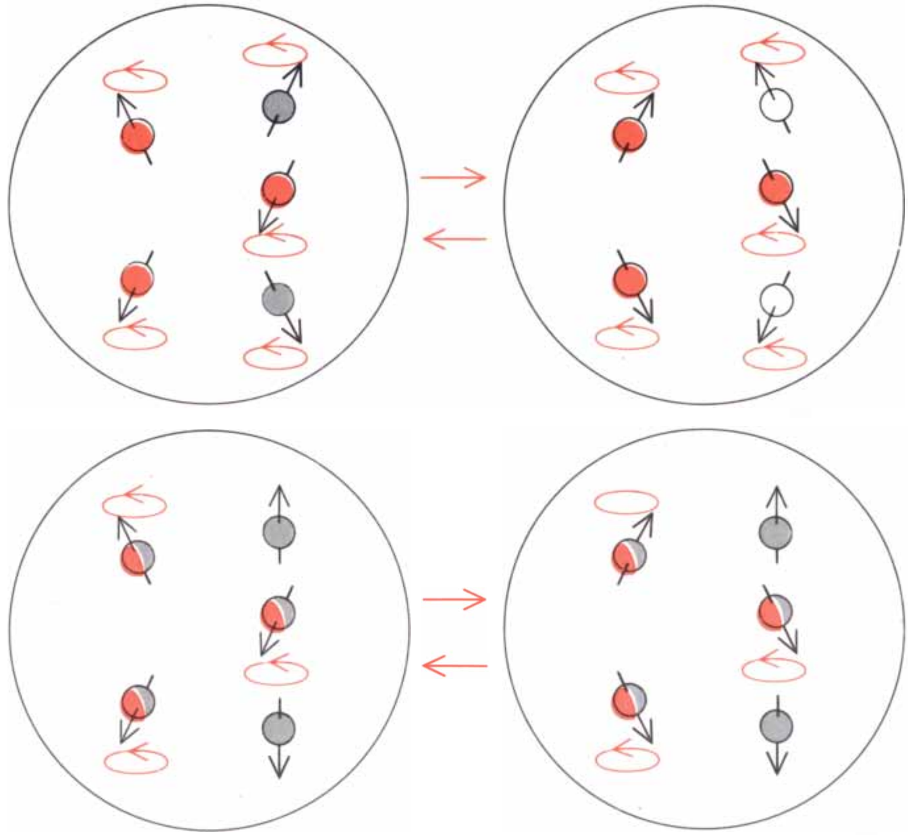
QUANTUM-MECHANICAL MODEL called the mean-field model provides a more satisfactory explanation of nuclear vibrations than models based on classical mechanics do. The force exerted on any one nucleon by the other nucleons in a nucleus fluctuates widely. Calculating the path of a nucleon in this field poses insurmountable difficulties. Not only is the field complex but also it changes rapidly because the nucleons creating the field are themselves in motion. The mean-field model simplifies the problem by postulating a smooth although varying field; it is the average of all the fluctuations of the actual field. An approximate mean field is calculated from the motion of the particles; the motion in turn is then calculated from the estimated field. Iterated calculations yield an approximation of the varying field that corresponds to the varying distribution of the nucleons during the vibration. The field shown here is for a one-dimensional array of nucleons; the actual distribution is a three-dimensional one.

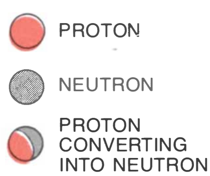
ing of the vibrations on a deeper level. Protons and neutrons are no longer thought to be elementary particles; they are composite objects made up of the more fundamental particles called quarks. Any change in the nucleon's spin necessarily entails some change in the quarks' state of motion. The quark spins interact much more strongly than the nucleon spins and give the system a tendency to precess at a much higher frequency. The apparent number of spins vibrating at low frequency is thereby reduced.

The reduced participation in the Gamow-Teller vibration has consequences for astrophysics, particularly for the theory of supernovas. During the collapse of a star the nuclei in the core undergo inverse beta decay in which a proton, an electron and an antineutrino combine to form a neutron. The amount of energy available to blow off the outer layers of the star turns out to be proportional to the fraction of the nuclei that take part in the process. Because the nuclear species involved do not exist in the laboratory, one must rely on the theory of the spin vibrations in seeking an understanding of how stars explode.

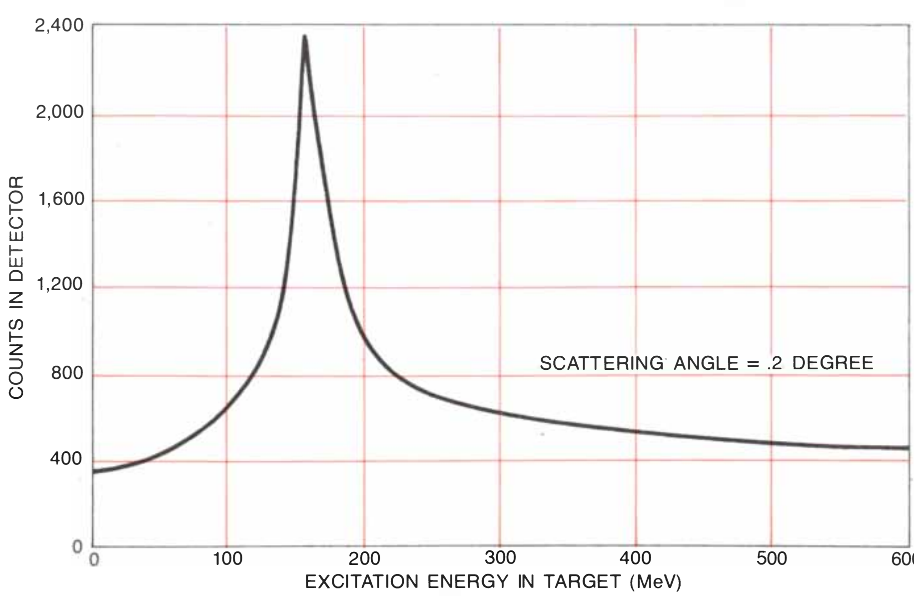
The vibrational models and the mean-field theory have successfully accounted for the frequencies of the vibrations, but they do not explain one important aspect of the motion: the damping of the vibrations, the decrease in the amplitude and the final extinction of the motion with time. Experimentally the damping shows up as a broadening of the resonance peak in measurements of the vibrational frequency. From the width of a resonance, say 4 MeV for a typical dipole or quadrupole vibration, it can be calculated that the nucleus oscillates through about three cycles before the motion is damped out.

The mechanisms that underlie damping are poorly understood. Collisions between nucleons would damp out the motion quickly, but the Pauli exclusion principle severely limits the probability of such collisions. Indeed, the mean-field theory would not work as well as it does if collisions were frequent. Another possibility is that the vibrational motion is coupled to more complex patterns of motion in the nucleus. The mean-field theory, however, is not capable of dealing with complex, uncoordinated motions of the nucleons. These motions are of course present and show up as the background cross sections in most measurements. The question of vibrational damping and the search for a theoretical model that can treat the more complex motions are now among the most active areas of research. No doubt there is much remaining to be learned about vibrations of nuclei, and indeed about the cooperative motion of quantum particles in general.





GIANT SPIN VIBRATIONS represent the coordinated precession of the spin axes of the nucleons. In the giant magnetic-dipole resonance (*top*) a proton and a neutron are tipped in opposite directions; their spin axes precess 180 degrees out of phase with each other. As a result of the coordinated precession of a fraction of the nucleons the nucleus as a whole acquires a net spin and a net magnetic moment. In the giant Gamow-Teller resonance (*bottom*) a proton is converted into a neutron and its spin is tipped from the original orientation.



SCATTERING CROSS SECTION of the giant spin vibration is much like that of the giant monopole vibration, peaking at an angle close to zero degrees from the beam axis. The difference is that the projectile, a proton, exchanges electric charge with a neutron in the target nucleus in the course of exciting the vibration, so that the scattered particles are neutrons rather than protons. In this case protons with an energy of 160 MeV excite the giant spin vibration in lead nuclei and yield a diffracted beam of neutrons with an energy of 146 MeV. Because the neutrons are uncharged, it is easy to separate them from the beam protons and to measure scattering in the direction of the beam. The neutrons are not deflected by a magnetic field, however, and so their energy cannot be measured in a magnetic spectrograph. Instead their velocity is measured by timing their flight to a detector placed 100 meters away from the target.

Microbodies in the Living Cell

A group of subcellular organelles linked by superficial structural similarities actually includes several distinct types, each having different sets of enzymes dedicated to different metabolic tasks

by Christian de Duve

In the early 1950's a Swedish graduate student, Johannes A. G. Rhodin, was exploring mouse kidney cells in the Karolinska Institute laboratory of Fritjof Sjöstrand, one of the pioneers of electron microscopy. He saw a small organelle that had not been described before. It was about half a micrometer in diameter (a micrometer is a thousandth of a millimeter), surrounded by a single membrane and filled with a fine granular matrix. The nondescript appearance of the unknown cell component particularly struck its young discoverer; for want of any characteristic feature to attach a name to, Rhodin (who is now at the University of South Florida College of Medicine) called it a microbody. Similar structures, many of them having a dense, semicrystalline core, were soon detected in liver cells. For several years there was no clue to the identity or function of the mysterious microbody.

Morphology reflects chemistry. By and large, all members of a given subcellular population (a distinctive organelle, for example) have the same enzymatic composition. By and large, each enzyme is found at a single site in the cell (in a particular organelle, for example). The same particle a cell anatomist sees in his micrographs may therefore reveal itself to the biochemist as a set of similar peaks in curves reflecting the distribution of particular enzymes in cell fractions separated by spinning in a high-speed centrifuge. It was therefore predictable that help in characterizing the microbody might come from biochemical studies. It did, but it came in the unpredictable, roundabout way that is common in basic research (to the confusion—would they only learn by it—of those who call for "goal-oriented" programming of scientific investigation).

In the mid-1950's my colleagues and I at the Catholic University of Louvain were in the midst of an exciting series of experiments (themselves prompted by a chance observation) that were just giving us a glimpse of a new intracellular organelle, the lysosome [see "The Lysosome," by Christian de Duve; SCIENTIFIC AMERICAN, May, 1963]. Our main tool was a modified cell-fractionation scheme. It had enabled us to demonstrate that in the liver of the rat the enzyme acid phosphatase is segregated within a special group of particles intermediate in size between mitochondria and "microsomes," the two main cytoplasmic entities recognized in those days. (Mitochondria are the power-generating organelles of the cell; microsomes, once thought to be organelles, are fragments of several internal cell-membrane systems.)

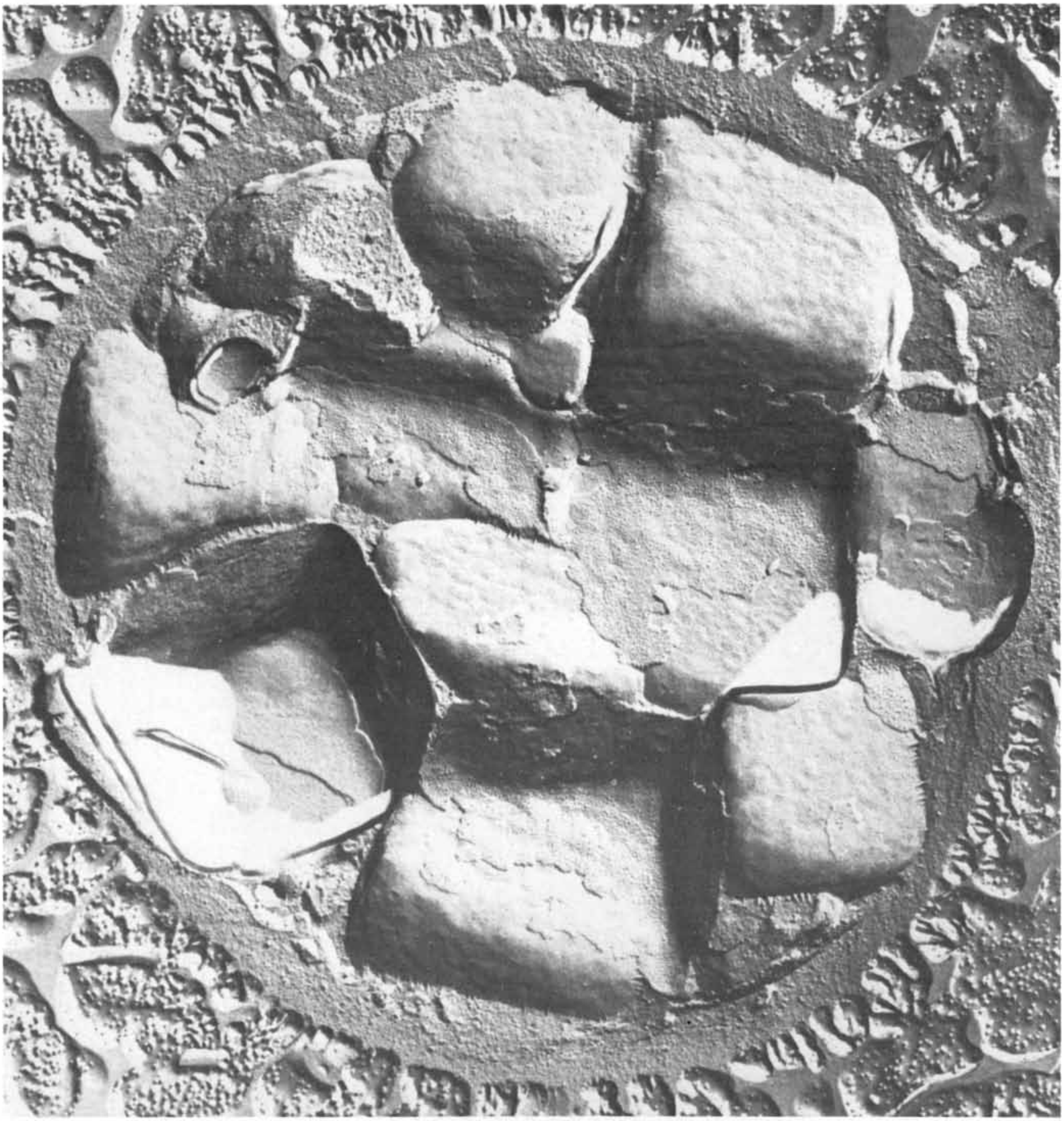
Applying our new fractionation scheme systematically to the enzymes other workers had found straddling the mitochondrial and microsomal fractions, we found that several acid hydrolases (enzymes that split a wide range of biological compounds in an acid medium) ended up along with acid phosphatase in our fractionations. It was the association of these enzymes that led us to the characterization of the lysosome: a saclike particle containing enzymes that digest nutrient molecules, foreign substances and, under certain circumstances, components of the cell itself.

A nondigestive, oxidizing enzyme showed the same distribution pattern, however. It was urate oxidase (or uricase, as it was then sometimes called): an enzyme that oxidizes urate (or uric acid) to allantoin and carbon dioxide. We had included it in our measurements because Alex B. Novikoff, who is now at the Albert Einstein College of Medicine, had reported that its centrifugal behavior was similar to that of acid phosphatase. We found the two enzymes did indeed move together when we fractionated liver tissue by differential centrifugation, which separates entities on the basis of the rate at which they sediment when they are spun in a centrifuge. This did not, however, necessarily mean that urate oxidase and acid phosphatase are associated physically within the same particle. It could instead reflect the cosedimentation of two distinct types of particles that happen to have the same sedimentation coefficient, which is a complex function of size, shape and density.

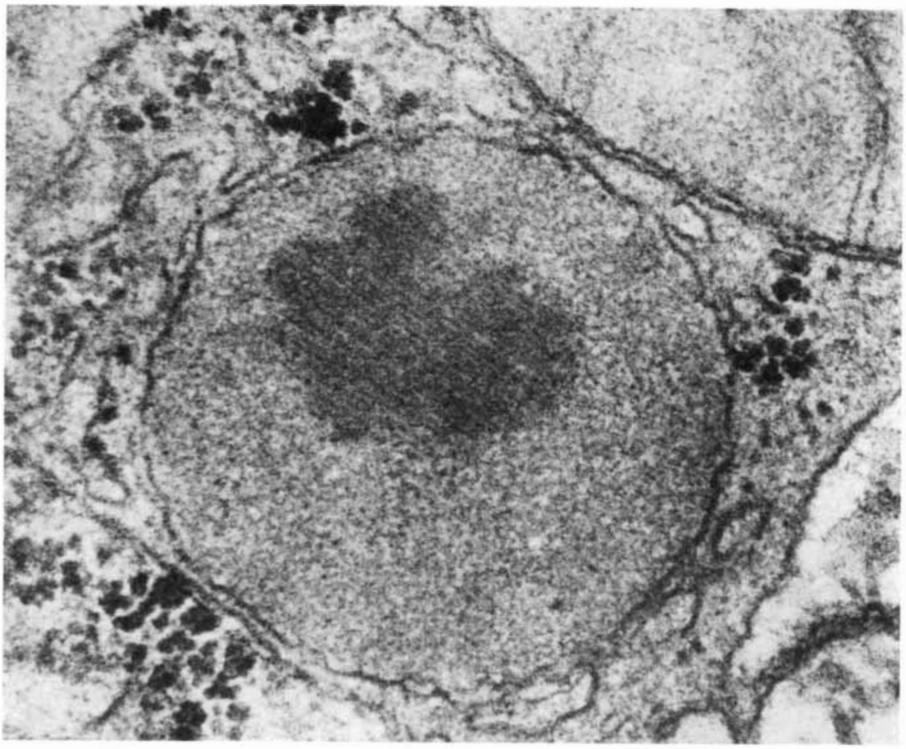
We suspected the second explanation might be true. One of our reasons was that urate oxidase did not (unlike acid phosphatase and the other acid hydrolases of lysosomes) display the property we called structure-linked latency: a seemingly inactive state of certain particle-linked enzymes that is observed when the particles are assayed for enzyme activity under conditions respecting their structural integrity. It was the chance discovery of this phenomenon (as exhibited by rat-liver acid phosphatase) that had attracted us to this field in the first place; we had been able to attribute it to the fact that the "inactive" enzyme is enclosed within a membranous envelope, inaccessible to the compound meant to serve as the enzyme's substrate in the assay. Our suspicion that urate oxidase is not in lysosomes was confirmed when we applied another fractionation procedure, density equilibration in a density gradient, in which each particle sediments only as far as its own density allows. In several systems of this kind acid phosphatase and urate oxidase showed clearly different distribution patterns.

The matter was clinched after my colleague Robert Wattiaux found that when rats are injected with the mild detergent Triton WR-1339, it accumulates in the lysosomes of liver cells and other cells and thereby causes a drastic and selective decrease in the density of those particles. The particles containing urate oxidase (and some other nondigestive enzymes that had been found to sediment with it) failed to show the change in density; they could therefore be separated cleanly from lysosomes, as well as from mitochondria, in a sucrose density gradient. This enabled my colleagues Pierre Baudhuin and Henri Beaufay to identify what we at first referred to as "uricosomes" (a term reflecting our early biochemical information) with the morphological entity Rhodin and others had described as microbodies.

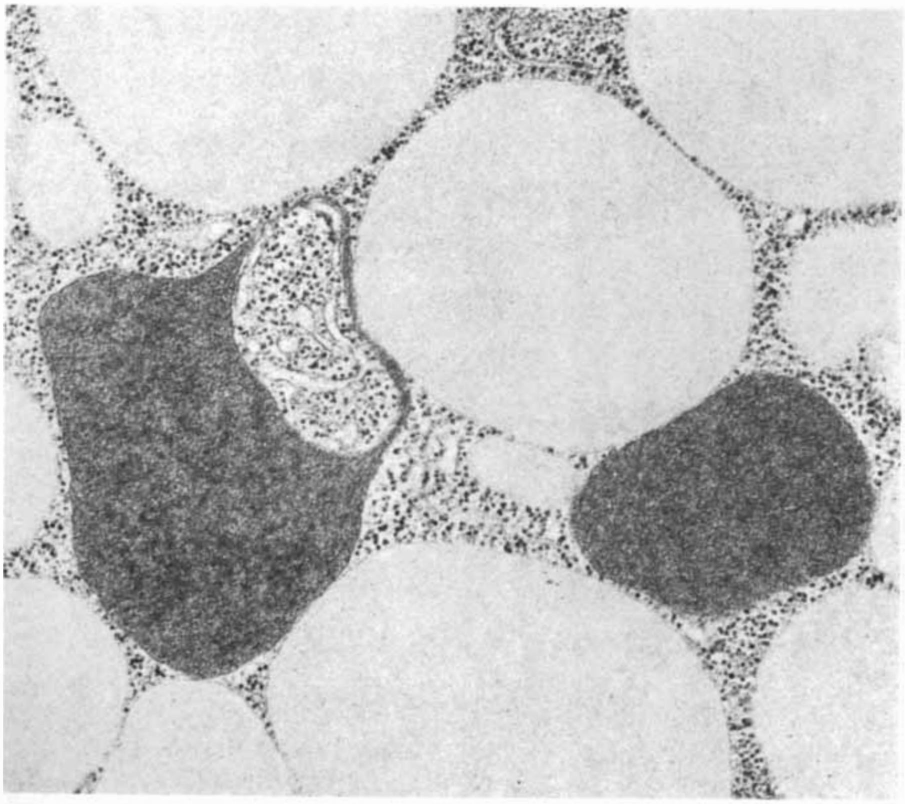
The definition of the microbody kept changing as our search extended to enzymes other than urate oxidase and to biological materials other than rat liver. (The locus of our efforts also expanded. In 1962 I accepted a position at the Rockefeller Institute for Medical Research—now Rockefeller University—in New York. I did not sever my ties with Louvain, however, and have since divided my time between the two institutions. Something of a scientific airlift was soon established between the two laboratories, which continue to exchange investigators, findings and techniques and to collaborate on many projects.) First Baudhuin showed that catalase, an enzyme that breaks down hydrogen peroxide (H_2O_2), seemed to be associated with the rat-liver microbodies contain-



PEROXISOMES, the major type of microbody, are seen crammed into a yeast cell in this freeze-etch electron micrograph made by Marten Veenhuis of the State University at Groningen in the Netherlands. These are highly unusual peroxisomes whose proliferation has been induced by growing the yeast *Hansenula polymorpha* in a medium in which methanol is the source of carbon. The peroxisomes contain the enzymes alcohol oxidase and catalase, which respectively oxidize the methanol, generating hydrogen peroxide, and reduce the hydrogen peroxide to water. The yeast cells were frozen and then fractured. Ice on the fractured surface was sublimed, the surface was shadowed with platinum, and carbon was deposited to make a replica. Here the carbon replica of one cell is enlarged 36,000 diameters.



SINGLE RAT-LIVER PEROXISOME is enlarged 115,000 diameters in this electron micrograph made by Helen Shio in the author's laboratory at Rockefeller University. It is bounded by a single membrane and has a granular matrix with a denser, semicrystalline core. Part of a mitochondrion is visible (*top right*), as are sectioned channels of the endoplasmic reticulum.



TWO GLYOXYSOMES surrounded by lipid (fat) droplets in the cell of a germinating tomato seedling are enlarged 32,000 diameters in an electron micrograph made by Eldon H. Newcomb of the University of Wisconsin at Madison and Peter J. Gruber. One glyoxysome has an inclusion of cytoplasmic material. Glyoxysomes, another type of microbody, are in effect peroxisomes that contain some or all of the enzymes needed to catalyze the glyoxylate cycle, a variant of the Krebs cycle. Glyoxysomes of seedling cells also contain enzymes for what is called the beta oxidation of fatty acids; raw material for that oxidation is present in lipid droplets.

ing urate oxidase. Otto Z. Sellinger, who is now at the University of Michigan Mental Health Research Institute, did the same for D–amino acid oxidase, an enzyme that oxidizes the unusual D–amino acids found in certain bacterial constituents.

On a visit to New York, Baudhuin joined forces with Miklós Müller (who had come to us from Hungary with a major interest in the lysosomes of protozoa) and with a gifted graduate student, the late Brian H. Poole. The results of this collaboration were most instructive. Particles biochemically similar to the liver microbodies (except that they lack urate oxidase) were found in the mammalian kidney (where microbodies had first been detected morphologically) and also in the protozoan *Tetrahymena pyriformis,* a ciliate whose cytoplasm likewise contains microbodylike organelles. In addition to catalase and D–amino acid oxidase the particles were found to contain a hydroxy acid oxidase acting notably on glycolate and lactate and (in kidney but not in the other two materials) an L–amino acid oxidase acting on the usual amino acids that make up proteins.

These findings indicated we were dealing with a cellular organelle likely to be widely distributed in the animal kingdom. They also provided a possible clue to what the microbody does in life. We were particularly impressed with the fact that the various oxidases in the new particles, although they can oxidize a wide range of substrates, have one feature in common: in oxidizing these substrates they reduce oxygen to hydrogen peroxide. Catalase thereupon reduces the hydrogen peroxide to water. To reduce is to add electrons. For catalase the electron donors are such small molecules as methanol, ethanol, nitrite or formic acid or, in the absence of such a donor, hydrogen peroxide itself. Together the oxidases and catalase can therefore be pictured as forming a primitive respiratory chain in which electrons removed from various metabolites are eventually added to oxygen, forming water. Unlike the major respiratory chain of the mitochondria, this one is mediated by hydrogen peroxide and is not coupled to phosphorylations that yield adenosine triphosphate (ATP), the cell's major energy carrier.

On the basis of these considerations I proposed in 1965 that the new particles be named peroxisomes. The term has been widely accepted, although there is still something of a nomenclature problem. Be that as it may, the concept implied by the word peroxisome has provided a useful link connecting a variety of microbodylike particles that are found throughout the eukaryotic kingdoms.

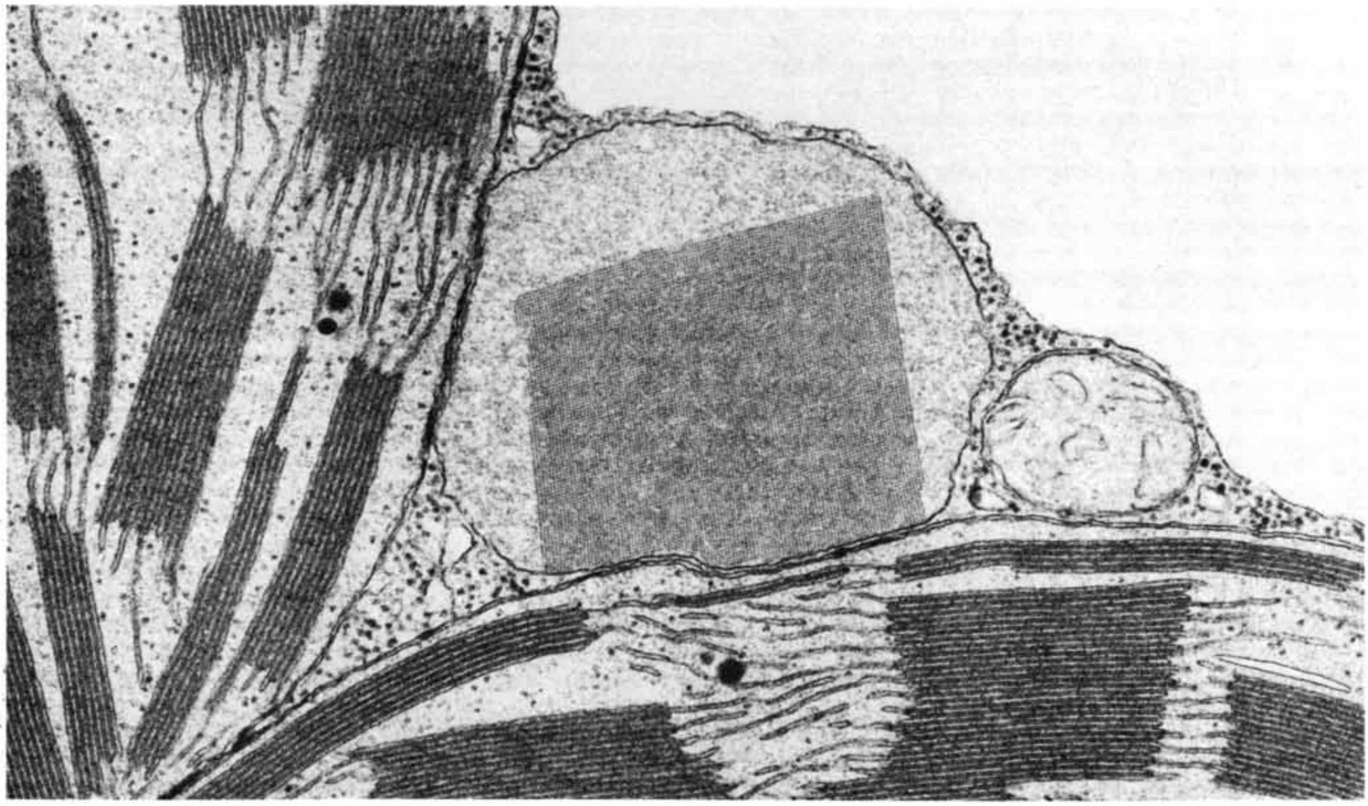
The concept is also physiologically significant, as has been shown most elegantly by Britton Chance and his co-workers at the Johnson Research Foundation in Philadelphia. By means of a sensitive spectrophotometric technique they measured the extent to which catalase is complexed with hydrogen peroxide in intact rat livers maintained in the laboratory. In agreement with the peroxisome concept they found that liver-cell catalase is normally complexed with hydrogen peroxide to be broken down; that the addition of known substrates for the oxidases increases the supply of hydrogen peroxide, and that the addition of substrates for the action of catalase accelerates the compensatory unloading of hydrogen peroxide from the enzyme-peroxide complex.

Implicit in the peroxisome concept is the central importance of catalase as the single agent of hydrogen peroxide disposal, whatever the nature of the oxidase substrates that have supported formation of the peroxide. Indeed, catalase seems to be present in all representatives of the peroxisome family, and often it is a major component of the particles; in rat liver, for instance, it accounts for as much as 15 percent of their total protein. A word of caution is required. Peroxisomes are certainly not the only site of hydrogen peroxide metabolism. There are oxidases that yield hydrogen peroxide in the cytosol (the fluid matrix of the cell), in the outer membrane of the mitochondrion and perhaps elsewhere in the cell. Most cells have at least one peroxidase in addition to catalase that is capable of breaking down hydrogen peroxide. There is also evidence that catalase itself is not always confined to the peroxisomes; there may be significant amounts in the cytosol.

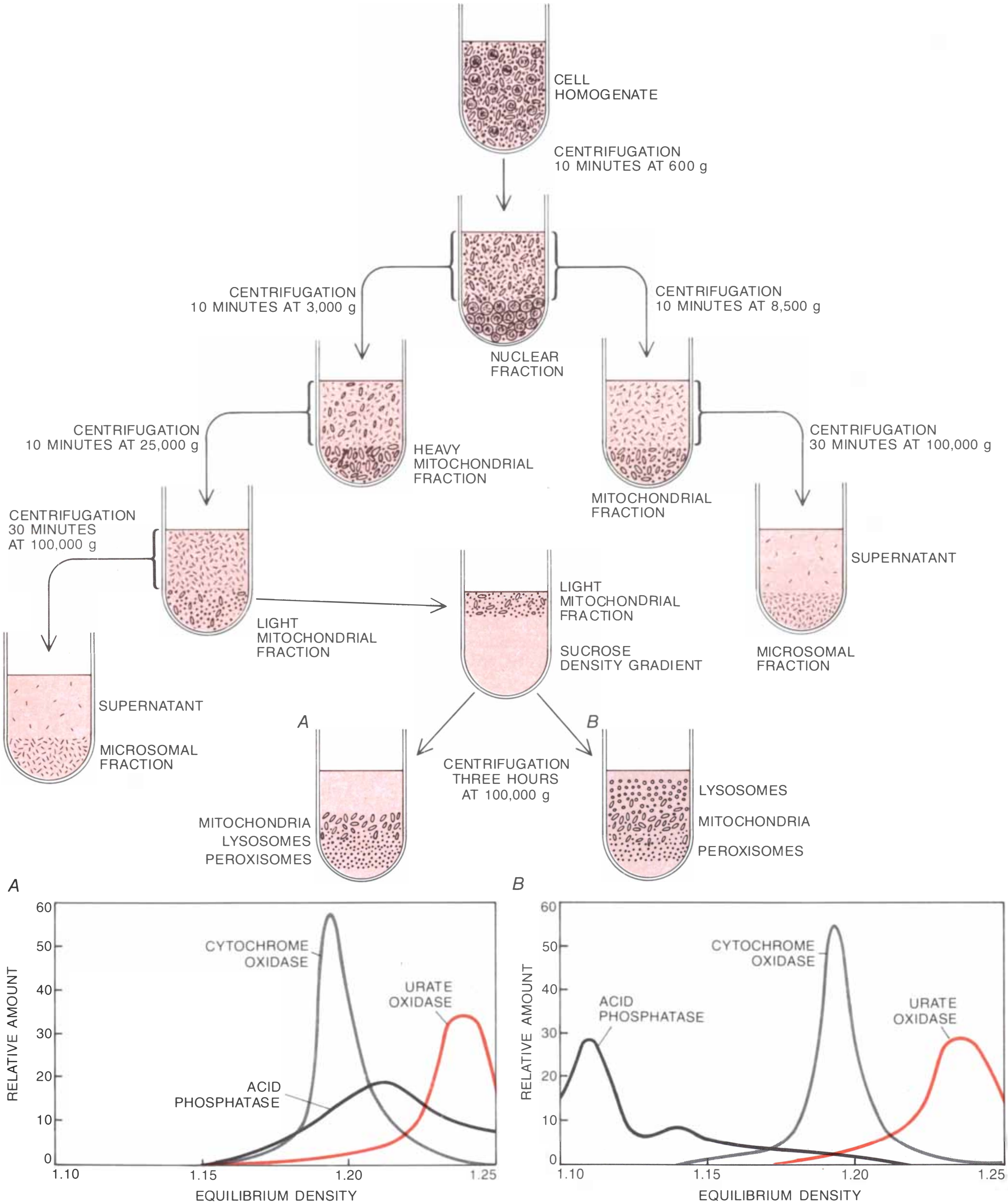
The next installment in the story begins with another transatlantic exchange. At the University of Oxford in 1957, in the laboratory of the late Sir Hans Krebs, his student Hans L. Kornberg discovered the glyoxylate cycle. In the 1930's Krebs had worked out the citric acid (or tricaboxylic acid, or Krebs) cycle, the final common metabolic pathway for the oxidation of fuel molecules. The glyoxylate cycle is a variant of the Krebs cycle. In it the sequence of four reactions whereby isocitrate is converted into malate in the Krebs cycle, with the release of two molecules of carbon dioxide, is bypassed; it is replaced by a two-reaction sequence that consumes an additional acetyl group furnished by acetyl coenzyme A (acetyl CoA) and leads to the formation of succinate as a by-product. Whereas one acetyl group is completely oxidized to carbon dioxide at every turn of the Krebs cycle, in the glyoxylate cycle (named for a key intermediate of the bypass) two acetyl groups are condensed to succinate at every turn.

Now, one of the major sources of acetyl CoA in living cells is what is called the beta oxidation of fatty acids. Moreover, the succinate generated from acetyl CoA by the glyoxylate cycle can be converted into carbohydrate by way of a segment of the Krebs cycle, phosphoenolpyruvate and reverse glycolysis. The glyoxylate cycle can therefore support gluconeogenesis, or the new formation of carbohydrate, from fat. It is the cycle's main biological function, and one of major importance.

Two workers attracted to Oxford by Kornberg's discovery subsequently took their interest to the U.S. One was Harry Beevers, who moved from Britain to Purdue University and set out to investigate the way in which the glyoxylate cycle is put in motion in germinating castor beans. This phenomenon, which is common to all fatty seedlings, serves to provide the young shoots with carbohydrate (at the expense of the seeds' oily stores) until the first green leaves appear and photosynthesis begins. Beevers and his co-worker Rowland W. Breidenbach located the two enzymes catalyzing the two glyoxylate-bypass steps in what ap-

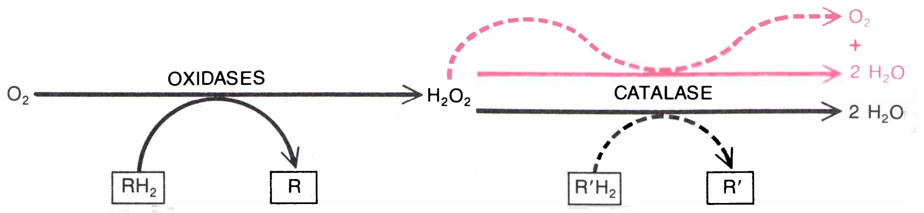


LEAF PEROXISOME, a large microbody housing a remarkably regular crystalline aggregate believed to consist of the enzyme catalase, is enlarged some 58,000 diameters in this micrograph made by Newcomb and Sue Ellen Frederick. The peroxisome is wedged between two chloroplasts and is adjacent to a mitochondrion in a tobacco-leaf cell; the large open area is the cell's central vacuole. The close association of the three kinds of organelle facilitates a circuitous metabolic pathway. Glycolate formed in chloroplasts is oxidized to glyoxylate in peroxisomes; subsequent steps in mitchondria and then back in peroxisomes lead to the formation of carbohydrate in chloroplasts.



CENTRIFUGAL FRACTIONATION separates cell components into fractions, which can be analyzed for their enzyme content to reveal the association of particular enzymes with particular cellular structures. A cell homogenate is spun in a centrifuge at progressively higher speeds. The classic fractionation scheme (*right*) yielded nuclear mitochondrial and microsomal fractions and a largely fluid supernatant. A modified procedure (*left*) developed in the author's laboratory some years ago subdivided the mitochondrial fraction. The resulting light mitochondrial fraction was layered onto a sucrose density gradient and subjected to density equilibration (*center*), separating its components into three layers on the basis of their density. The layers were defined by peaks in curves for three marker enzymes (*bottom*): cytochrome oxidase (a mitochondrial enzyme), acid phosphatase (known to be present in lysosomes) and urate oxidase, which appeared to be associated with a different particle. The peaks (and the fractions they define) were close together when normal rat cells were fractionated (*A*). When rats were injected with a detergent that accumulates in lysosomes, the density of the lysomes was selectively lowered (*B*), and they could be clearly separated from particles containing urate oxidase: the microbodies now known as peroxisomes.

peared to be a novel kind of cytoplasmic particle. The same particles contained the three other glyoxylate-cycle enzymes, which could also be found in the mitochondria as part of a regular Krebs cycle. The new particles could be purified by centrifugation in a sucrose gradient, where they equilibrated at a high density, as our liver peroxisomes had; they were found to have the general morphological characteristics of microbodies. In accordance with the new particles' biochemistry Beevers called them glyoxysomes.



PRIMITIVE RESPIRATORY CHAIN mediated by hydrogen peroxide (H_2O_2) is characteristic of peroxisomes and gives them their name. One of various oxidases in the particle oxidizes a substrate (RH_2), passing its electrons to oxygen to form hydrogen peroxide, which is reduced to water by catalase. Electrons for the reduction come either from one of various small molecules ($R'H_2$) or, if no other donor is available, from hydrogen peroxide itself (*color*).

Plant and animal biochemists are notoriously insulated from each other, and the connection between these findings and ours might long have gone unnoticed but for another former associate of Kornberg's. James F. Hogg of Queens College of the City University of New York had visited Oxford and there had begun investigating the glyoxylate cycle in *Tetrahymena pyriformis*. My colleague Müller, embarking on a search for lysosomes in the same organism, called on Hogg for technical advice and was easily convinced to cast his net wider: to look for glyoxylate-cycle and Krebs-cycle enzymes too. Their collaboration led to the discovery that the two glyoxylate-bypass enzymes are present in the peroxisomes of *T. pyriformis*. In the protozoan they are not, however, accompanied (as they are in castor-bean glyoxysomes) by the three enzymes the glyoxylate cycle has in common with the Krebs cycle; those enzymes are found only in the mitochondria with the other Krebs-cycle enzymes.

In the spring of 1967, at that great annual fair of American biological scientists known as the Federation meetings, Breidenbach announced the first findings on castor-bean glyoxysomes. Müller was there to report on the presence of glyoxylate-cycle enzymes in *Tetrahymena* peroxisomes. After the meeting it did not take the Purdue investigators long to detect in their castor-bean preparations both catalase and some oxidases (including urate oxidase) that yield hydrogen peroxide, thereby identifying glyoxysomes as peroxisomes. The name glyoxysome continues to take precedence over peroxisome whenever glyoxylate-cycle enzymes are recognized in the particle. This creates a certain amount of confusion because it is not exactly clear at what stage a peroxisome becomes a glyoxysome. Does it have to contain a complete glyoxylate cycle, as in fatty seedlings? Or are the two bypass enzymes enough, as in *T. pyriformis*? And what if only one of the two enzymes is found?

Apart from the nomenclature problem, the difference between the castor-bean glyoxysomes and the *Tetrahymena* peroxisomes is illuminating. In the seedlings the glyoxysomes contain a complete glyoxylate cycle and the mitochondria a complete Krebs cycle, so that there is duplication of the three enzymes common to both cycles. Gluconeogenesis and the oxidation are clearly separated, and each has its own supply of acetyl CoA. In *Tetrahymena* the mitochondria hold the whip hand by controlling the supply of isocitrate to the peroxisomes. Oxidation is clearly dominant, and only those isocitrate molecules that are not funneled into the Krebs cycle, and leak out of the mitochondria, are available to start down the gluconeogenesis pathway in the peroxisomes. (The mitochondria are needed for gluconeogenesis even in the fatty seedlings. Only there can succinate be converted into phosphoenolpyruvate to enter the reverse-glycolysis pathway.)

In higher animals the enzymes of the glyoxylate cycle appear to have been casualties of evolution. Mammals, including human beings, are unable to make carbohydrate from the common fatty acids. Recently, however, glyoxylate-cycle enzymes have been detected in cytoplasmic particles in epithelial cells of the bladder of the toad; they are said to be associated with catalase, as they are in the glyoxysomes (peroxisomes) of fatty seedlings and *T. pyriformis*. This finding supports the hypothesis that there is a phylogenetic relation between these particles in both the plant and the animal kingdoms.

In pursuing his glyoxysome studies Beevers, with T. G. Cooper, made the surprising discovery of another metabolic pathway: glyoxysomes in germinating castor-bean seedlings have all the enzymes needed for the beta oxidation of fatty acids to generate acetyl CoA, a capability thought to be limited to mitochondria. The topographical association of beta oxidation with the glyoxylate cycle makes for efficiency in the overall gluconeogenetic process. The glyoxysomal enzymes for beta oxidation catalyze the same reactions as their mitochondrial counterparts, with one interesting exception. The enzyme that initiates each cycle of oxidation, which in mitochondria is a dehydrogenase feeding electrons into the phosphorylating respiratory chain, is a simple hydrogen peroxide–producing oxidase in the glyoxysomes. Shades of peroxisomes!

It is astonishing that the wider significance of this observation failed to be recognized for almost 10 years (even by those who firmly believed plant glyoxysomes and animal peroxisomes are related phylogenetically) in spite of a highly suggestive finding made in the course of studies on clofibrate, a drug given to lower serum lipids in patients with a high blood-fat level. Investigators at the CIBA, Ltd., laboratory in Basel had found as early as 1965 that clofibrate induces a dramatic proliferation of peroxisomes in the rat liver.

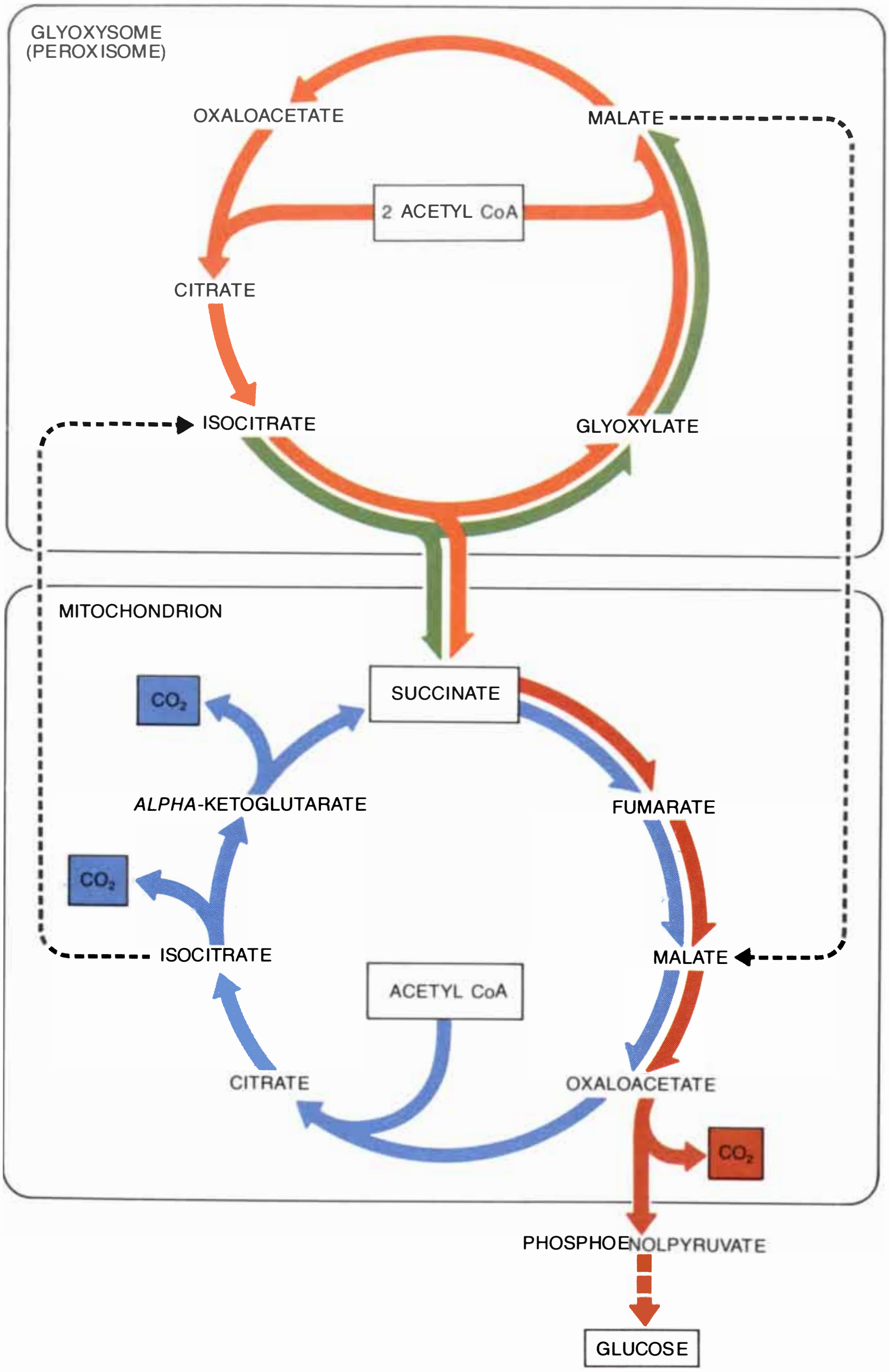
The observation attracted a great deal of interest and was confirmed elsewhere; other drugs that lower blood fats, even some unrelated structurally to clofibrate, were found to have the same effect, and the statement was repeatedly made that peroxisomes must play a role in lipid metabolism. Yet not until 1976 was this statement given experimental corroboration, when my associate at Rockefeller, Paul B. Lazarow, demonstrated that beta-oxidation enzymes are present in rat-liver peroxisomes and are enhanced by a factor of 10 or more after clofibrate treatment.

Lazarow's report elicited follow-up studies in a number of laboratories. The main peroxisomal beta-oxidation enzymes from rat liver were purified and characterized by Takashi Hashimoto of the Shinshu University Faculty of Medicine in Japan, who showed them to be entirely unrelated chemically to the corresponding mitochondrial enzymes. The peroxisomal beta-oxidation system has been induced experimentally with many different fat-lowering drugs; it can also be induced by high-fat diets or by exposure to cold, to both of which enhanced fatty-acid oxidation would seem to be an appropriate physiological response.

Imitating in reverse the role played by castor-bean glyoxysomes, liver peroxisomes mediated an important transdisciplinary fertilization process, this time with plant biochemistry as recipient and animal biochemistry as donor.

The result of the exchange was new information about photorespiration, a paradoxical phenomenon that has puzzled plant biochemists for decades.

As the term indicates, photorespiration is a process whereby light induces an increase in the amount of oxygen consumed by the tissues of green plants. It is associated with the production of glycolate by chloroplasts, the photosynthetic organelles. In the chloroplast ribulose-bisphosphate carboxydismutase, the key enzyme of carbon assimilation, has a dual function. In the absence of oxygen the enzyme acts exclusively to fix carbon dioxide, converting ribulose bisphosphate into two molecules of 3-phosphoglycerate. When oxygen is present, however, it competes with carbon dioxide; part of the ribulose bisphosphate is diverted into an alternative oxidative pathway that yields one molecule of phosphoglycolate and one of phosphoglycerate instead of two phosphoglycerates.



GLYOXYLATE CYCLE (*orange arrows*) is a variant of the Krebs cycle (*blue*), the major metabolic pathway for oxidation of fuel molecules. In the glyoxylate cycle four Krebs-cycle steps are bypassed by a two-reaction sequence (*green*), which supplies succinate as a by-product. The succinate can be converted into glucose and other carbohydrates (*red*). Seedling-cell glyoxysomes have enzymes for a complete glyoxylate cycle; the Krebs-cycle enzymes are in the mitochondria, and so the enzymes catalyzing the three steps common to the two cycles are duplicated. In the protozoan *Tetrahymena pyriformis* the peroxisomes contain only the two bypass enzymes and lack the other three glyoxylate-cycle enzymes; the mitochondria supply isocitrate to the peroxisomes, which return malate to the mitochondria (*broken black arrows*).

The subsequent breakdown of phosphoglycolate by a phosphatase yields glycolate, which is a major product of isolated chloroplasts and of certain unicellular algae when they are illuminated in the presence of oxygen. The larger the photorespiratory response, the smaller the net photosynthetic yield of a plant, and so the magnitude of photorespiration can have economic significance. The magnitude depends essentially on the ratio of oxygen to carbon dioxide, so that it varies inversely with a plant's ability to concentrate carbon dioxide. Tropical grasses such as corn, sugarcane and sorghum are particularly efficient in this respect.

The glycolate formed photooxidatively in chloroplasts was known to be further oxidized to glyoxylate, with the consumption of additional oxygen. Where and how this further oxidation takes place remained uncertain until a hydroxy acid oxidase acting on glycolate was found in association with catalase in the peroxisomes of rat liver. That finding stimulated N. Edward Tolbert of Michigan State University to overcome various difficulties besetting the homogenization and fractionation of green-leaf tissue. He thereupon discovered that peroxisomes are the site of glycolate oxidation. Like other representatives of this versatile family, leaf peroxisomes equilibrate at a high density in a sucrose gradient, and they display the morphological characteristics of microbodies.

These green-leaf microbodies, which have been studied particularly by Eldon H. Newcomb and his co-workers at the University of Wisconsin at Madison, are objects of great beauty. They may measure as much as 1.5 micrometers in diameter and often contain dense nucleoids, or crystalline inclusions, that are believed to consist of catalase. According to Tolbert, the glyoxylate made by oxidation of glycolate is converted into the amino acid glycine by an enzyme (a transaminase) in the peroxisomes. The scene then shifts to the mitochondria, where two glycine molecules are transformed oxidatively to the amino acid serine and carbon dioxide. The serine goes back to the peroxisomes, where it is converted into glyceric acid, which, after phosphorylation, can contribute to the formation of carbohydrate in the chloroplasts. This circuitous journey is reflected vividly in the morphological disposition of leaf peroxisomes: they are often seen wedged tightly between chloroplasts and mitochondria.

Microbodies containing peroxisomal and glyoxysomal marker enzymes were detected in the yeast *Saccharomyces cerevisiae* in 1968 by Charlotte J. Avers of Rutgers University. The study of yeast strains that can grow on unusual substrates suggests that their adaptability is due to certain remarkable inducible metabolic properties of their peroxisomes.

Saburo Fukui of Kyoto University and his colleagues first noted a marked proliferation of catalase-containing microbodies in strains of the yeast *Candida* that are grown on the long-chain hydrocarbons called alkanes. The particles contain catalase, several oxidases, a complete beta-oxidation system, the two glyoxylate-bypass enzymes (but not the three other enzymes of the glyoxylate cycle) and enzymes that remove hydrogen from long-chain alcohols and aldehydes. The mitochondria of the same organisms house the Krebs-cycle enzymes and alcohol dehydrogenases but no beta-oxidation system. These findings suggest that the peroxisomes (with the assistance of mitochondria to fill in the glyoxylate cycle) play a major role in making carbohydrates from long-chain alcohols that are derived from the alkanes.

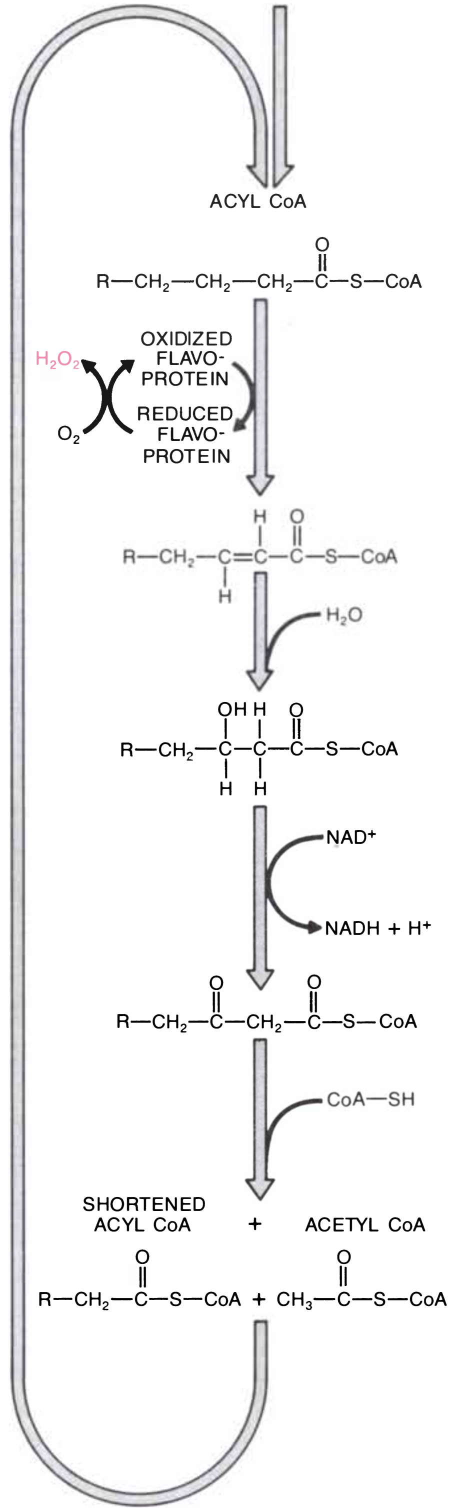
A most dramatic metabolic adaptation is seen when the yeast *Hansenula polymorpha* is grown on a medium in which the only source of carbon is methanol. Several groups of investigators in the Netherlands, Germany and Japan find such yeast cells are filled with large microbodies packed so tightly that they become almost cubic in shape. Much of the volume of these microbodies is occupied by a large crystalline structure composed of alcohol oxidase. The enzyme oxidizes the methanol to formaldehyde, generating hydrogen peroxide that is disposed of by catalase; the formaldehyde is subsequently converted into carbohydrate by a special pathway in the cytosol. There is evidence that yeasts may develop other peroxisome-mediated metabolic adaptations when they are grown on certain nitrogenous substrates such as D-amino acids, urate, choline or simple amines.

In mammals typical microbodies are seen only in a few organs, notably the liver and the kidney; these organs also have much higher catalase levels than other tissues. Small amounts of catalase are found, however, in every type of cell. When it is not in typical peroxisomes, the enzyme appears to be associated with smaller membrane-bounded particles, some .1 to .2 micrometer in diameter, which are often elongated and sometimes are even snakelike in form. These particles have been named microperoxisomes by Alex and Phyllis M. Novikoff and their collaborators. Increasing evidence for the presence of oxidases in these small structures suggests that they may indeed be related functionally to peroxisomes, even though they are different from microbodies morphologically.

How do peroxisomes (or glyoxysomes) originate? Do they resemble mitochondria, which are present in the germ cell and give rise to more mitochondria by growth and division? Or are they more like primary lysosomes, which arise from a pinching off of bits of the Golgi apparatus, which in turn is derived from the cell's major network of membranous channels, the endoplasmic reticulum? Many morphologists have reported seeing continuities between the membranes of plant or animal peroxisomes (or glyoxysomes) and the endoplasmic reticulum. These observations, together with some early reports that liver catalase is synthesized on ribosomes bound to endoplasmic reticulum (the "rough" endoplasmic reticulum), inspired the hypothesis that peroxisomes arise from the reticulum by budding. Peroxisomal proteins, it was proposed, are synthesized on membrane-bounded ribosomes, delivered into the lumen of the endoplasmic reticulum and thence transferred to peroxisomes by way of connecting channels.

In our Rockefeller laboratory first Poole and then Lazarow could find no evidence for this model. Lazarow and his collaborators find instead that proteins of the peroxisomal matrix are synthesized on free ribosomes in the cytosol. In experiments in which animals are fed amino acids labeled with a radioactive isotope, newly made (labeled) peroxisomal proteins are recovered entirely in the soluble fraction of homogenized liver cells. After a brief sojourn (an average of less than 20 minutes for catalase) in what one assumes (until there is proof to the contrary) to be the cytosol, the proteins are transferred to the peroxisomes by an unknown mechanism. Other investigators working on rat liver have reported similar results, and the main features of this scheme seem to apply also to the biogenesis of glyoxysomes in several different kinds of seeds.

The mechanism of translocation of the newly made proteins across the peroxisomal membrane does raise a puzzling problem. Proteins made by ribosomes on the rough endoplasmic reticulum are in effect injected into the membrane system by the ribosomes. In the case of proteins introduced from the cytosol into mitochondria, a "preprotein" chain is shortened as it passes through the mitochondrial membrane, perhaps providing energy for the transfer. In the absence of either active ri-



FATTY ACIDS are converted in seedling glyoxysomes to yield acetyl coenzyme A, the fuel for the glyoxylate cycle and hence for gluconeogenesis (the new production of glucose and other carbohydrates). The beta-oxidation pathway that accomplishes this conversion begins with an acyl coenzyme A, an activated fatty acid: a chain of CH_2 groups (some represented by *R*, for radical) linked by a CO group to a sulfur atom of coenzyme A. Each round of beta oxidation removes two CH_2 groups and yields a molecule of acetyl coenzyme A. In glyoxysomes (but not in mitochondria, the other site of beta oxidation) the first enzyme of cycle generates hydrogen peroxide, hallmark of peroxisomal metabolism.

bosomes to "push" large protein molecules such as enzymes into a membranous structure or proteolytic clipping to "pull" it in, where does the needed energy come from? One is tempted to think in terms of a self-assembly process within the peroxisome, with proteins that become attached to the particle membrane by a receptor-binding mechanism being somehow pulled through it and trapped inside the particle in multiprotein complexes. It is suggestive in this respect that dense aggregates, sometimes crystalline in nature, are so often observed inside peroxisomes.

Whatever the nature of the translocation mechanism, there is evidence that it can be reproduced in the test tube. When peroxisomal (or glyoxysomal) proteins, newly synthesized in the laboratory and labeled with a radioactive isotope, are mixed with peroxisomes (or glyoxysomes), some of the protein can be recovered from within the particles. This has been done with cucumber-seedling glyoxysomes by Wayne M. Becker of the University of Wisconsin and with rat-liver peroxisomes by Yukio Fujiki in Lazarow's laboratory.

The finding that proteins in the peroxisome's matrix are translocated from the cytosol, does not in itself rule out the possibility that the particle's membrane may originate from the endoplasmic reticulum and may even remain attached to it. It is equally possible, as Lazarow has pointed out, that peroxisomal-membrane proteins too are synthesized on free ribosomes, and that they are then inserted into the membrane of preexisting peroxisomes. Mitochondrial-membrane proteins and even some proteins of the endoplasmic reticulum originate in this way.

The fact that the Novikoffs and others have described direct continuities between peroxisomes and the endoplasmic reticulum would seem to argue in favor of the first possibility, were it not for the difficulty of identifying endoplasmic reticulum unequivocally on purely morphological grounds. Lazarow and Helen Shio have looked for continuities by staining for specific products of peroxisomes and of the endoplasmic reticulum. In no case were such products seen en route from one structure to the other; there is no luminal communication between the two. Lazarow has also shown, with Fujiki and others, that the membranes of the two structures have a very different chemical composition.

In view of these facts one should consider the possibility that what looks like a system of connections to the endoplasmic reticulum may conceivably be a system linking peroxisomes into some kind of special "peroxisomal reticulum" (Lazarow's phrase) or "microbodial compartment" (a term used by K. Gorgas of the University of Heidelberg). Evidence for such a compartment in at least certain cell types has come from workers in Germany and Portugal who have combined staining with serial sectioning. Such reconstruction has not yet been done for liver cells, but the characteristic clustering of liver peroxisomes and their striking biochemical homogeneity have long suggested to us that there too microbodies may be interconnected. It remains for future work to clarify this important issue.

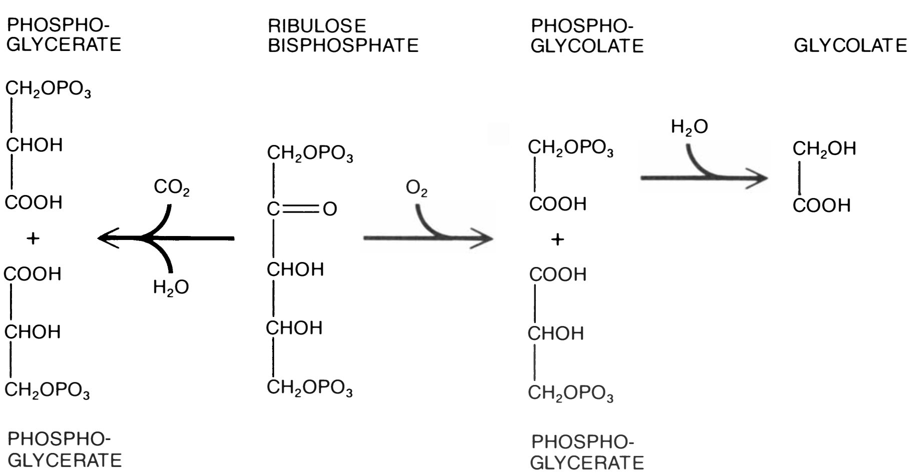
Particles identified biochemically as peroxisomes or glyoxysomes have been found in a wide variety of cells in all kinds of organisms above the bacteria. Whether these particles belong to plants, fungi, protozoa, invertebrates or higher animals, they show so many similarities and overlaps of all kinds that one can consider them members of a single family. For the sake of simplicity I shall refer to them henceforth as peroxisomes. My reason is that glyoxysomes, whether complete or incomplete, share the important properties of peroxisomes, whereas the reverse is not always true. To call them microbodies, on the other hand, could be misleading. As will become clear, not all microbodies are peroxisomes.

Peroxisomes in different organisms and tissues harbor various overlapping sets of enzymes from a growing list that already includes more than 40 distinct enzymes. So far no peroxisome has been found that has all the enzymes on the list; whether this is owing to the absence of certain genes or merely to their repression in a particular species is a question worth raising, given the striking degree of inducibility manifested by many peroxisomal enzymes.

For all of the functional diversity of peroxisomes, two metabolic features emerge as constants. One is the association of hydrogen peroxide–producing oxidases with catalase, the characteristic to which peroxisomes owe their name. The substrates of this oxidative activity are numerous and varied. They include amino acids (both D and L), hydroxy acids, fatty acids, alcohols, amines and a purine, but not the major electron carrier reduced nicotinamide adenine dinucleotide (NADH). Apparently any electrons delivered to NAD^+ (the oxidized form of the carrier) inside the peroxisomes have to be shuttled to the mitochondria for oxidative disposal. Alternatively, they could be handed over in the cytosol to a potential peroxisomal substrate such as lactate or ethanol and so be returned to the peroxisomes for final disposal. No energy-retrieval mechanism has yet been found to be coupled to peroxisomal oxidations.

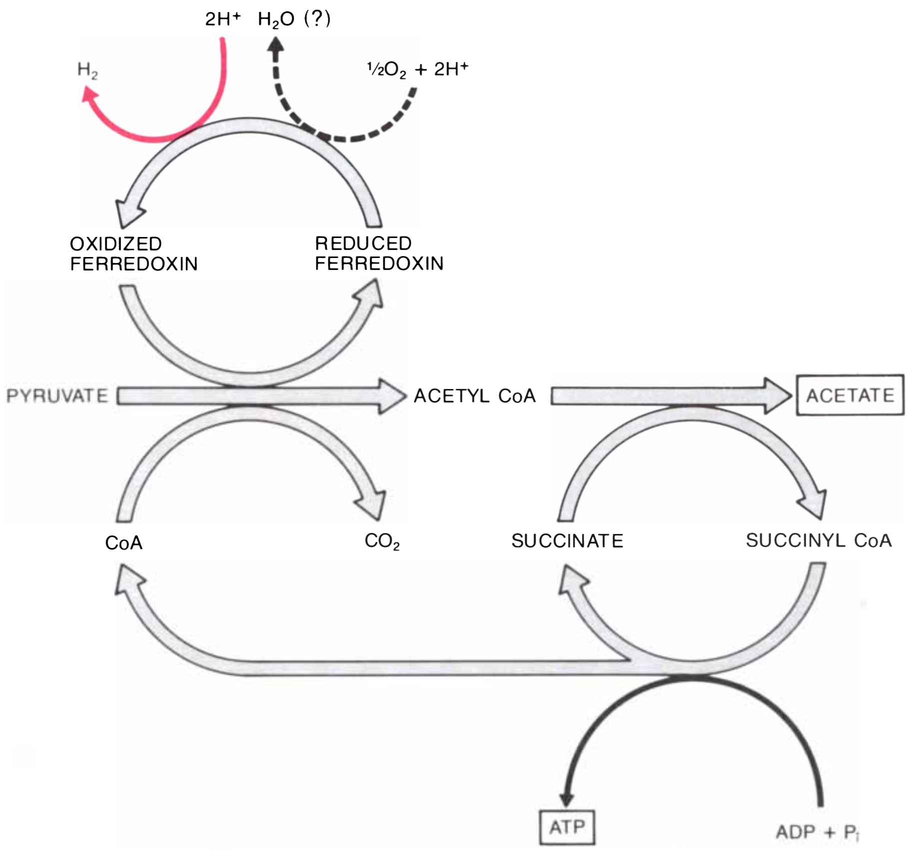
The second function that comes through as some kind of metabolic leitmotif of peroxisomes is gluconeogenesis. I have mentioned its occurrence in germinating fatty seedlings, in *Tetrahymena,* in the leaves of photorespiring plants and in yeasts growing on alkanes or methanol. Even in mammals one is struck, as we were many years ago, by the fact that peroxisomes are seen mainly in the two major sites of gluconeogenesis, the liver and the kidney, but so far there is no evidence that they participate in the new formation of carbohydrate in these tissues.

The wide distribution of peroxisomes in all major groups of eukaryotes (or-



PHOTORESPIRATION competes with photosynthesis in the chloroplast. In photosynthesis carbon dioxide is fixed by an enzyme that converts ribulose bisphosphate into two molecules of phosphoglycerate (*arrow to left*). In the presence of oxygen the same enzyme acts as an oxygenase, diverting some of the ribulose bisphosphate into a pathway yielding only one molecule of phosphoglycerate along with a molecule of phosphoglycolate (*arrow to right*). The hydrolysis of phosphoglycolate produces glycolate. The glycolate is exported to peroxisomes and oxidized to glyoxylate, which is converted into glycine. Two glycine molecules are transformed in the mitochondria to make serine and carbon dioxide. Back again in peroxisomes, the serine is converted into glyceric acid, which contributes to carbohydrate formation in chloroplasts.

ganisms with nucleated cells) indicates that the organelles existed in the primitive unicellular organism thought to be the common ancestor of all eukaryotes. The rudimentary character and energetic wastefulness of the peroxisome's respiratory apparatus suggest that peroxisomes preceded mitochondria in the ancestral organism.



UNUSUAL MICROBODY, the hydrogenosome, was discovered in *Tritrichomonas foetus*, a protozoan parasite that lacks mitochondria. Hydrogenosomes house enzymes that oxidize pyruvate, in the process generating adenosine triphosphate (ATP), the cellular energy carrier. Electrons removed from pyruvate are transferred to the electron-transport protein ferredoxin and finally to oxygen if it is available (*broken black arrow*), possibly forming water. In absence of oxygen electrons are transferred to protons (*H^+*), generating molecular hydrogen (*colored arrow*).

If peroxisomes supported oxidative metabolism in some primitive cell devoid of mitochondria, the possibility exists that the original cell line may have survived to this day. This idea prompted Müller to look for protozoa that have no mitochondria but do have microbodies. He chose the trichomonatids, a group of flagellates that, in addition to being endowed with beautiful microbodies and no mitochondria, have great economic and medical importance as parasites of the genital tract in animals and human beings and as agents of sexually transmitted diseases.

With Donald G. Lindmark, who is now at Cornell University, Müller succeeded in isolating and characterizing the trichomonad microbodies. He did not find what he was looking for, however. Instead of descendants of an ancestral peroxisome he found hydrogenosomes: organelles capable of oxidizing pyruvate to acetate and carbon dioxide. The oxidation is coupled to the synthesis of ATP. The electrons released by the oxidation are transferred to oxygen if it is available; under anaerobic conditions the electrons are transferred instead to protons (H^+) to form hydrogen (H_2).

Hydrogenosomes owe their ability to produce hydrogen to their having both the appropriate hydrogenase and a special electron-transport protein (a ferredoxin with a low oxidation-reduction potential) that serves as the electron acceptor for the enzyme pyruvate dehydrogenase. The ferredoxin also enables hydrogenosomes to reduce certain low-potential nitroimidazole derivatives and thereby convert them into compounds that are highly toxic to the cell. Such a property can be turned against its owner, and nitroimidazole drugs have become the standard treatment for trichomonad infections. Natural selection has not failed to wage its counteroffensive, and Müller is now tracking down pathogenic strains that have become resistant to the available drugs.

Quite aside from its considerable biological interest, the discovery of hydrogenosomes thus contributes to the war against what has become the most widespread venereal disease in the U.S. And once again I must point out to the funders and planners of scientific research that Müller was not proposing to enroll in this public-health cause—or even trying to understand hydrogen production in trichomonads. He was angling for an ancient peroxisome and came up instead with a hydrogenosome at the end of his line. So far no biochemical kinship has been found between the two types of microbodylike particles. Presumably they are not related.

The latest (but perhaps not the last) of the microbodies was discovered by Fred R. Opperdoes in the laboratory of Piet Borst at the University of Amsterdam. Behind its perfectly legitimate morphological credentials this microbody turned out to be hiding a large segment of the glycolytic chain, the major pathway for the breakdown of glucose into smaller carbon compounds. A number of protozoan blood parasites (including the agents of African sleeping sickness, of the dreaded Chagas' disease of Latin America and of leishmaniasis in many parts of the world) have been found to harbor this unusual organelle. It has been named the glycosome.

In all other animal cells so far investigated in this respect, the enzymes of the glycolytic chain are free in the cytosol; to find much of this system sequestered within a membrane-limited particle is very peculiar indeed. Aerobically glycosomes collaborate with mitochondria in supporting the energy needs of the cells in which they are found. They can also provide energy on their own in the absence of oxygen or if the mitochondrial oxidative pathway is blocked by the drug salicylhydroxamic acid. Their enzymes have physicochemical properties quite unlike those of the usual glycolytic enzymes, even though they catalyze most of the same reactions. Glycosomes may therefore be a vulnerable target for chemicals that interfere selectively with their special enzymes; such chemicals could serve as drugs to combat the deadly diseases caused by glycosome-carrying parasites. (Again basic research has practical application. Opperdoes now heads the tropical-disease unit at the International Institute of Cellular and Molecular Pathology, which we have established in Brussels specifically to facilitate this kind of fertilization.) Glycosomes, for all of their morphological similarity to the microbodies of animal and plant cells, are very different biochemically from peroxisomes and are probably not related to them.

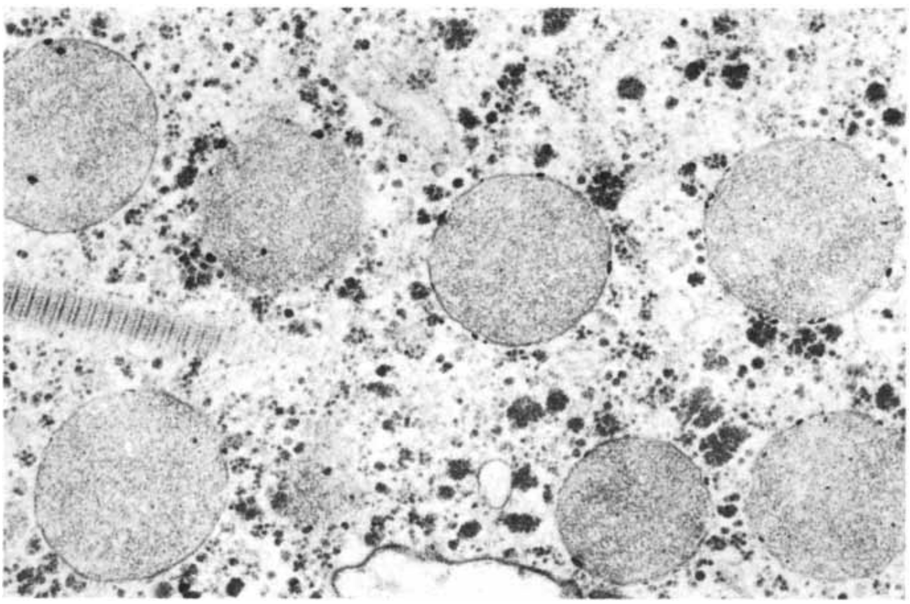
A seemingly simple and still widely accepted explanation for the evolutionary origin of microbodies is to view them as offshoots of the main cytomembrane system. As I have indicated, much of what is known about the biogenesis of peroxisomes and glyoxysomes (there is no information on that of hydrogenosomes or glycosomes) does not support such an origin. What one sees in the case of peroxisomes—the posttranslational transfer of proteins through the cytosol—resembles instead the biogenetic pathway of many proteins that are encoded in the nucleus, synthesized in the cytoplasm and then introduced into mitochondria and chloroplasts. These are the two membrane-bounded organelles many biologists believe are derived from endosymbionts. The idea is that small bacterial cells, having respectively the ability to catalyze oxidative phosphorylation and to photosynthesize, were engulfed by a primitive ancestor of today's eukaryotic cells and stayed on to develop a symbiotic relation with the larger cells and eventually to become organelles. And so the question arises: Could microbodies too have originated as endosymbionts?

Now, the main evidence for such an origin in the case of mitochondria and chloroplasts is that both of those organelles contain their own DNA, specifying some of their own proteins, and the genetic machinery to replicate it, transcribe it into RNA and translate the RNA into protein. A few years ago there was a report that what looked like DNA could be seen in hydrogenosomes, but Müller has not been able to confirm the claim biochemically; there is no evidence that any of the microbodies has its own genetic system. A secondary argument for a symbiont origin of mitochondria and chloroplasts is that both of those organelles are bounded by a double membrane, the outer one of which may be a descendant of a vacuolar membrane that once housed the original bacterial symbiont. The microbodies (except for hydrogenosomes) are bounded by a single membrane.

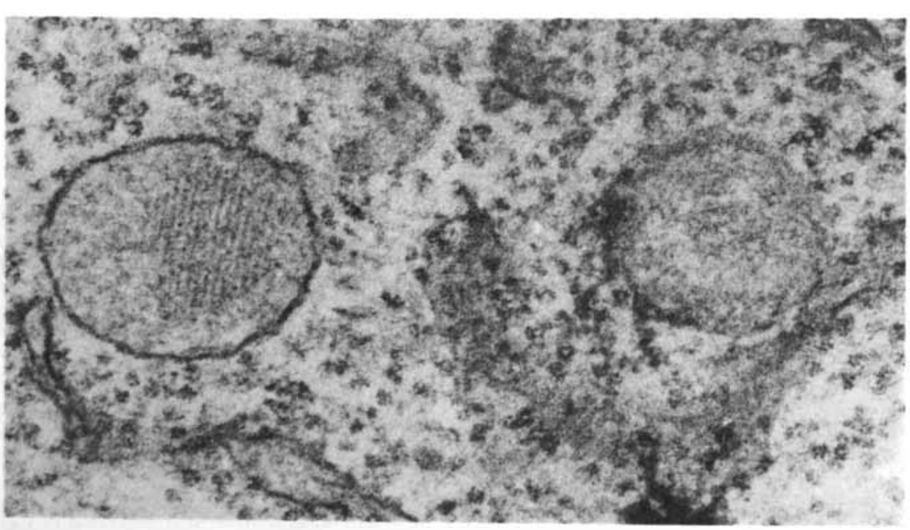
These negative arguments do not make a watertight case against an endosymbiont origin for microbodies. One could point out that mitochondria and chloroplasts have lost to the nucleus more than 90 percent of their genetic information. Might the loss not be 100 percent in the case of organelles that may have been adopted as endosymbionts earlier? As for the two-membrane argument, not all endoparasites are enclosed in a vacuole. Some are naked in the cytoplasm.

What makes the endosymbiont hypothesis attractive, besides the biogenetic argument already mentioned, is the primitive—one is tempted to say "bacterial"—character of the metabolic functions subserved by microbodies. The glycolytic chain is thought to be among the oldest enzyme systems, dating back to before photosynthesis, when there was virtually no oxygen in the atmosphere; it could survive in a particularly ancient form in the glycosome. The ability to reduce protons to molecular hydrogen, characteristic of the hydrogenosome, may also belong to early anaerobic days; it has been preserved in a similar form in *Clostridium*, a group of anaerobic bacteria that includes the agents of gas gangrene. As for the peroxisome, it could embody an early protective adaptation to the appearance of oxygen. A simple, nonphosphorylating peroxisomal type of respiration could have preceded the more advanced phosphorylating form found in some bacteria and in mitochondria. A final peroxisomal quality that is also characteristic of bacteria is ready adaptation to the availability of unusual nutrients.

These are flimsy arguments, to be sure. When it comes to events that happened more than a billion years ago, facts are scarce and imagination has an open field. The possibility that microbodies originated as symbionts would hardly be worth mentioning if it were not for the fact that it can be tested. One can determine the amino acid sequence of microbody proteins (or the nucleotide sequence of the genes encoding them) for comparison with the sequence of homologous molecules in various eukaryotic and bacterial cells, and the degree of relatedness may reveal something about the ancestry of microbodies. Such work is now in progress.



HYDROGENOSOMES of *Trichomonas vaginalis* are enlarged some 40,000 diameters in this micrograph made by Shio. The large dense objects are rosettes of glycogen, a storage form of glucose. The striated object at the left in the micrograph is a riblike structure, the costa.



GLYCOSOMES, which are generally smaller than other microbodies, are enlarged 105,000 diameters in a micrograph made by Pierre Baudhuin. The particles are seen in *Trypanosoma brucei*, a protozoan parasite of domestic animals in Africa that is transmitted by the tsetse fly.

Museum piece, circa 1987.

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

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

The Datakode magnetic control surface is a thin, transparent layer of iron oxide particles (approximately 9 billion per square inch) coated across the entire back of the film during manufacture. Less than 8 microns thick, this transparent layer provides the means to record machine-readable information (100 bits of binary data in a single track on each frame of film), and makes possible a uniform frame-indexing code that can be used with both film and videotape. Best of all, it accomplishes this interface without altering the quality or characteristics of the final image.

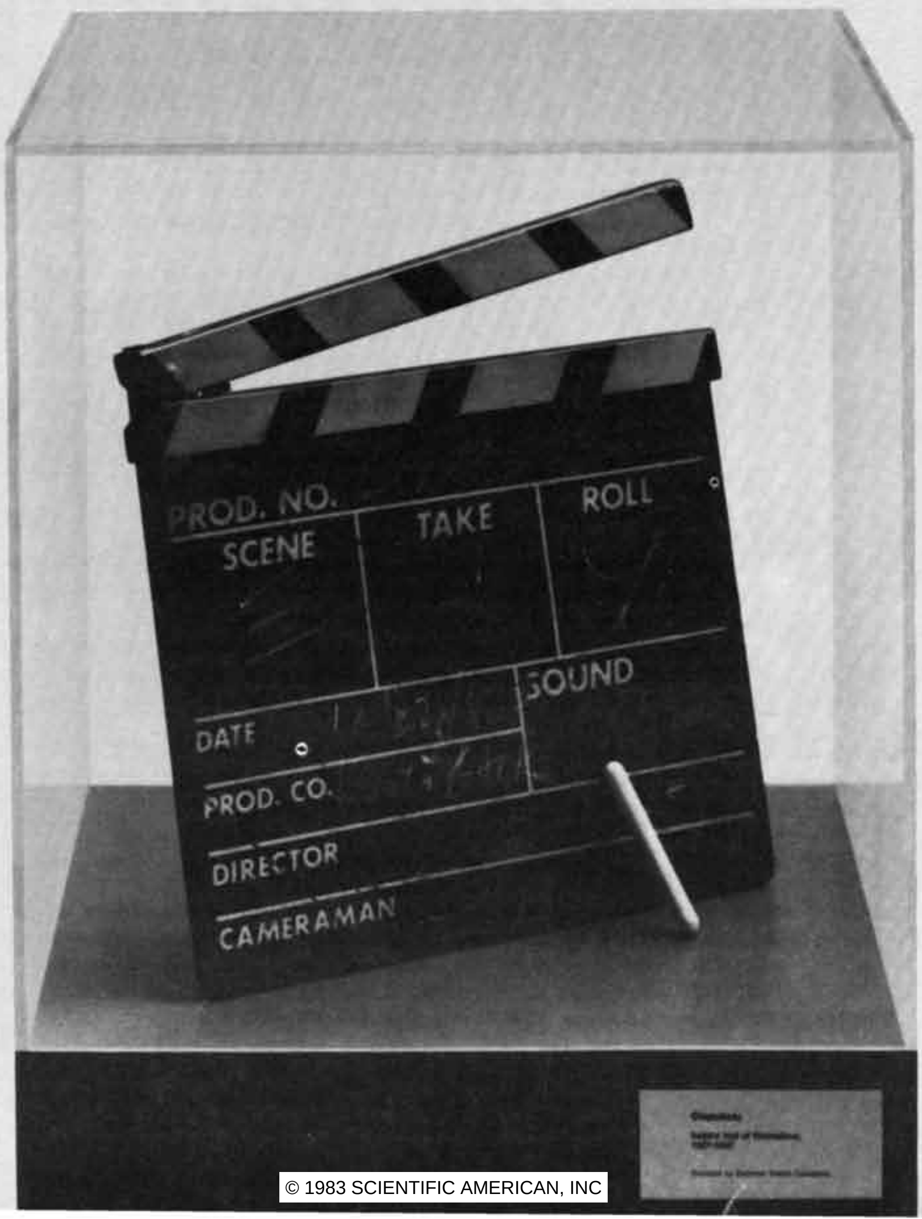
12 34 30 00

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

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

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

Kodak. Where technology anticipates need.



The new Honda Prelude began as a clean piece of paper.

Honda engineers sketched a superior aerodynamic shape. A low wedge with a wide stance and increased overall length.

Once the design was conceived, they set about perfecting it. They made the windshield and door handles flush with the body. New retractable halogen headlights raise in such a way to cause less turbulence.

Underneath is a new front suspension. It's a unique double wishbone, low-profile concept inherited from Honda Formula racing car experience. Independent MacPherson struts support the rear wheels. A low center of gravity contributes to handling and stability.

A new engine improves performance. It has 12 main valves, 2 intake and 1 exhaust for each cylinder's main combustion chamber. The result is better breathing and efficiency. And the dual carburetors make highway passing power more than passable.

Engine displacement is upped to 1829cc. With the standard 5-speed shift, there's an increase of 33 percent in horsepower. There's also a new 4-speed automatic transmission available. Variable-assist power steering comes only with the automatic.

Particular attention was given to the new interior. There's more room. Front bucket seats hold you firmly and comfortably for extended touring. Push-button controls are all within easy reach of the driver. Your hands easily grip the thicker 3-spoke steering wheel.

And when you aren't carrying passengers in the rear seats, the rear seatback unlocks with a key and folds down for added luggage space. It also gives you access to the trunk. And trunk size has been increased by 20 percent.

A power-operated Moonroof is included in the long list of Honda standard equipment. Along with our best AM/FM stereo radio with digital electronic tuning and an autoreverse cassette and four speakers.

When you first drive the new Prelude, you'll think you are in something else. You'll be right.

HONDA

The new Prelude.

SCIENCE AND THE CITIZEN

The Trouble with Trident

The decision by the Reagan Administration to accelerate the development of the Trident II sea-launched ballistic missile (SLBM) reflects among other things a growing appreciation of certain advantages inherent in deploying strategic nuclear weapons at sea rather than on land. Unlike land-based intercontinental ballistic missiles (ICBM's) such as the Minuteman and its proposed replacement, the MX, SLBM's on submerged submarines are difficult to locate accurately; hence they are essentially invulnerable to a preemptive "counterforce" attack and seem certain to remain so for the foreseeable future. Moreover, U.S. ballistic-missile submarines, by virtue of their range and mobility, are capable of launching a prompt retaliatory strike on the U.S.S.R. from many directions, thereby complicating any attempt to thwart a counterattack by means of an anti-ballistic-missile (ABM) system. In short, SLBM's come close to being an ideal deterrent force against a nuclear attack.

The mobility of ballistic-missile submarines, combined with the need for them to go undetected, has until now made SLBM's considerably less accurate than ICBM's and accordingly less suitable as counterforce weapons against "hard" targets such as missile silos and command, control and communications facilities. The Trident II, however, promises to overcome this limitation. Technological advances in missile-guidance systems are expected to give the Trident II an accuracy comparable to that projected for the MX. As a result the new generation of SLBM's should in principle be capable of destroying the entire range of potential targets in the U.S.S.R.

From the viewpoint of arms control the accelerated Trident II program must be considered a mixed blessing. On the one hand, it can be argued that greater reliance on submarine-based missiles would contribute to strategic stability; because submarines are apparently invulnerable, no attack could destroy the entire retaliatory force. Even if the Russians were to launch an attack, at least SLBM's would not draw fire to the continental U.S.

On the other hand, any analysis of counterforce-capable SLBM's ought to consider how the Russians are likely to react to the deployment of such missiles by the U.S. The U.S.S.R. now has most of its strategic nuclear weapons in the form of fixed land-based missiles. In addition it is generally agreed that the U.S.S.R. is far behind the U.S. in the development of an antisubmarine-warfare capability. Consequently Russian leaders might well perceive the Trident II program as a potentially dangerous first-strike threat to their ICBM force. One "quick fix" response could be the deployment of mobile land-based missiles. Another could be the adoption of a "launch on warning" policy, an outcome fraught with ominous possibilities, particularly given the reduced warning time available in the case of an actual or apparent SLBM attack.

The Administration's proposed military budgets for the fiscal years 1984 and 1985 allocate respectively $1.5 and $2.2 billion for development of the Trident II, up sharply from the current fiscal year ($369.7 million). If Congress approves the acceleration of the Trident II program, the first of the new missiles could become operational as early as 1988. As matters now stand, the full Trident II program calls for the deployment of 20 Trident submarines over the next decade, each one bearing 24 Trident II missiles; each missile in turn is reported to have a range of 6,000 miles and to be capable of carrying between 10 and 15 nuclear warheads with an explosive yield of 475 kilotons each. (The current generation of Trident I missiles are less accurate, have a range of 4,800 miles and carry between eight and 10 warheads of 100-kiloton yield.) The projected total cost of the Trident II program, including submarines, missiles and bases, is said to be roughly $60 billion.

Betting the Farm

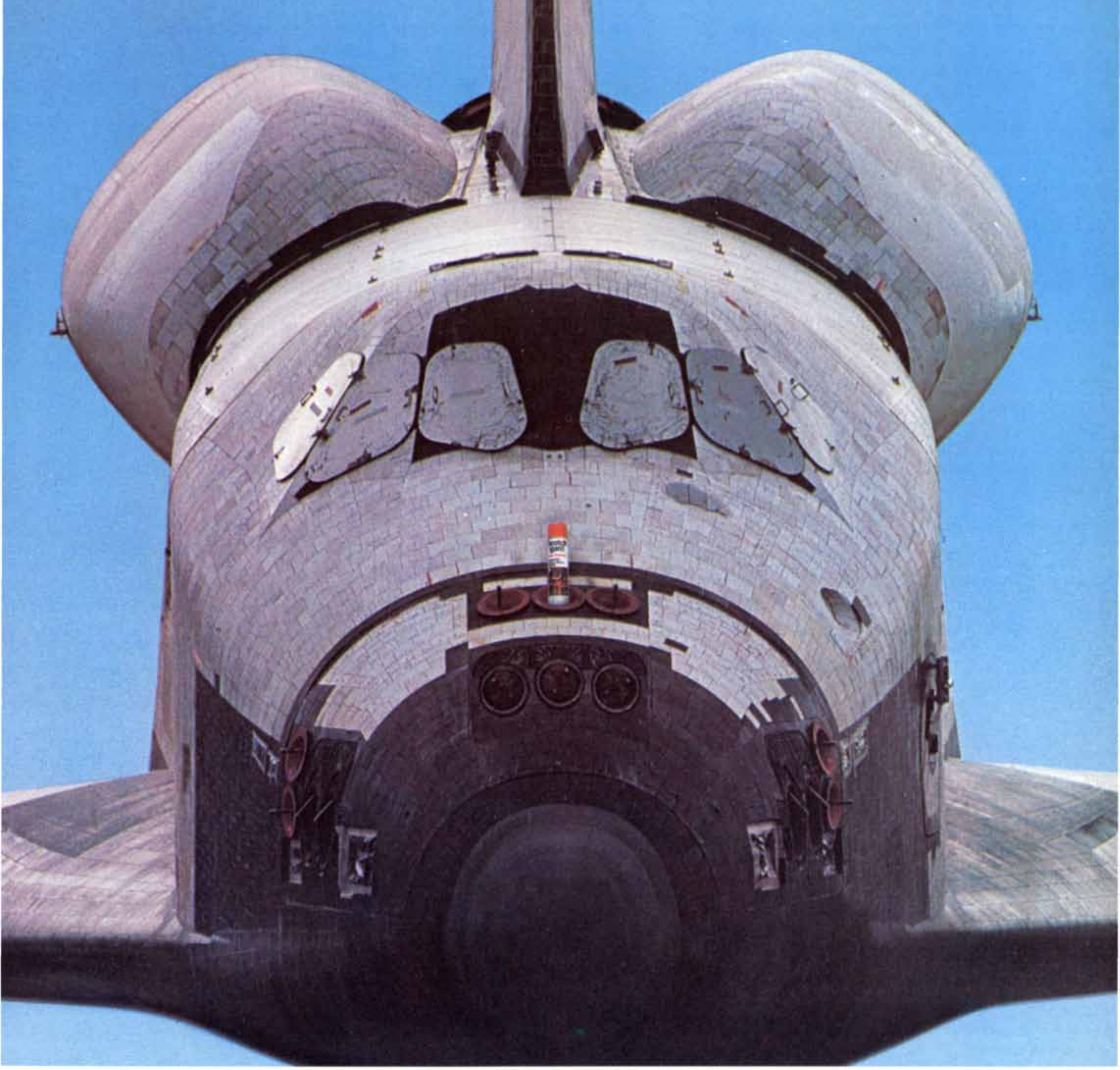
The largest program of agricultural research in the U.S. is sponsored by the Department of Agriculture. Some of the work is done by the department's staff of scientific workers and some is done by outside agencies with funds provided by the department. Most of the in-house work is done by the Agricultural Research Service, which has a research budget of more than $400 million per year. In recent months the ARS has been the subject of considerable criticism both from within and from outside the Federal Government. Partly in response to the criticism, the agency has proposed a new research agenda that would increase the amount of work it does in basic science, particularly in areas related to new biological technology. The agenda would reduce the amount of work directed toward the improvement of plant and animal productivity, which has long been the largest single project carried out by the ARS.

Some of the most pointed criticism of the ARS came at a conference held last summer at Winrock, Ark. The conference was organized by Denis J. Prager, assistant director of the Office of Science and Technology Policy, and it included representatives of government, universities, foundations and private industry. *Science for Agriculture,* the report of the conference, concludes that the quality of research done by the Department of Agriculture has been declining steadily for several decades. According to the report, until the 1930's workers in the department did much of the fundamental work in basic science related to agriculture. Furthermore, because of its scientific preeminence, the staff of the department was able to provide strong formal and informal leadership in establishing priorities for all agricultural research done in the U.S.

In the past 50 years, however, the work of the ARS has become progressively more parochial, more fragmented and less fundamental in a scientific sense. Most projects are now aimed at solving specific local problems in animal and plant husbandry, such as how to eliminate a particular parasite in a small geographic area. As a result the department no longer exercises effective leadership in the setting of research priorities.

At least two outside influences have apparently contributed to a decline in the quality of ARS research. The first influence is that of Congress. Particularly since the 1950's, members of Congress, motivated by a desire to have Federal research facilities in their state, have been instrumental in the proliferation of ARS stations. There are now 148 ARS experimental stations in the U.S., ranging from the Agricultural Research Station at Beltsville, Md., with 450 scientific workers, to stations with only two or three workers. It is generally agreed that the cumbersome and inefficient institutional structure of the agency has contributed to its scientific decline, but attempts to close small facilities have met with much political resistance.

The second influence is a sharp decrease in the proportion of financial support for agricultural research done by state agencies that is provided by the Federal Government. Under the Hatch Act of 1887 the Government distributes funds for agricultural research to state governments. The amount depends on the size of the state's rural and farm population. Most of the money goes to support the state agricultural experiment stations. The states are required to match the amount provided by the Department of Agriculture, which is now about $140 million. In recent years, however, the contributions of the states have far exceeded those of the Depart-



We're keeping the feathers dry.

Before the last four flights of America's most famous bird, the Columbia space shuttle, our "Scotchgard" Brand Fabric Protector was sprayed right from the can onto the shuttle's surface. It's the same product you buy to protect clothes and furniture.

On the Columbia it kept rain from seeping into the surface tiles.

That solved a problem from the Columbia's first flight, when moisture in the tiles froze in the cold of space, expanded, cracked the tiles and endangered a safe re-entry.

When people ask us to solve tough problems, we often have the solution right on the shelf.

Then too, by listening to people's needs, we've pioneered over 200 high-performance consumer products. We make everything from sandpaper and skincare products to home video and computer tape.

And it all began by listening.

3M hears you...

For your free 3M Consumer Products Brochure, write: Department 070705/3M, P.O. Box 4039, St. Paul, MN 55104.

Name__________

Address__________

City______ State____ Zip____

Or call toll-free: **1-800-323-1718,** Operator 367. (Illinois residents call 1-800-942-8881.)

3M

ment of Agriculture: their contribution is now about $700 million.

The effect of both congressional and state influence has been to shift research in the direction of applying well-known technologies and scientific results to the solution of local agricultural problems. Obtaining new fundamental scientific results and developing new technologies has largely been left to universities and private industry. The new ARS research agenda, which covers the years from 1984 through 1990, represents an attempt to reverse these trends. The proposal assumes that the current ARS research budget of $413 million will remain unchanged through 1990, after the effects of inflation have been taken into account. Under the proposal no major work now done by the ARS will be eliminated, but funds will be redistributed among the existing categories of research. Work on the productivity of crop plants and farm animals, which currently constitutes 60 percent of the ARS budget, would decrease to 52 percent by 1990. Most of the savings would be distributed among projects concerned with human nutrition, soil and water conservation, and commodity conversion and delivery.

Funds would also be redistributed within the major categories. For example, under the heading of plant productivity, work on insects, diseases and nematodes in field horticulture would decrease from $27 to $14 million. In the same category work on maintaining the genetic diversity of crop plants would increase from $10 to $12 million. A total of $75 million would be redistributed among existing projects.

Among those who attended the Winrock conference there are sharp differences of opinion on the new direction at the ARS. John A. Pino, director for agricultural sciences of the Rockefeller Foundation, described the agenda as "a move in the right direction [that] provides a stronger focus on some of the basic aspects of agricultural science." Pino added that in his opinion the ARS had a responsibility to do work that would lead to "an understanding of some of the fundamental processes... involved in the interaction of plants and animals." He said of the current proposal: "I hope it's more than a cosmetic change." He noted, however, that the agenda merely redistributes funds without eliminating any of the work currently being done. To make a fundamental change in the direction of research, substantial cuts will have to be made, he said. "Some surgery is going to be needed. It remains to be seen whether the surgery can be done and upheld."

James T. Bonnen, professor of agricultural economics at Michigan State University, on the other hand, questioned the notion that the solution to the problems of the ARS is an increased commitment to basic science. According to Bonnen, "some critics of the agency talk as if the ARS ought to do only basic science. If that is so, how do you get the very large national agenda of applied work done? If Federal agencies won't do it, it won't get done." The primary role of the ARS, he said, should be to work on national problems in applied science and technology. "The people who are critical of the ARS come from one end of the spectrum: basic science. They apparently have no appreciation for the intervening institutions between basic science and [agricultural] productivity."

Living Color

In recent years the most important tool for the study of how the brain sees has been the microelectrode. Inserted into the brain of experimental animals, it has revealed that nerve cells in the retina, in the cerebral cortex and in the lateral geniculate nucleus (a group of cells interposed in the visual pathway that leads from the retina to the cerebral cortex) respond preferentially to particular patterns of lightness and darkness in the visual field. The neural mechanism underlying the perception of color, however, has remained elusive. Two recent discoveries promise to change that. One of them was made with microelectrodes in the brain of the monkey; the other is based on investigations of how colors look to a human observer under special circumstances. The new discoveries support the growing consensus that the brain performs several simultaneous analyses of the visual world, processing visual information in a number of more or less independent ways before reintegrating the data to produce the unified conscious sensation of sight.

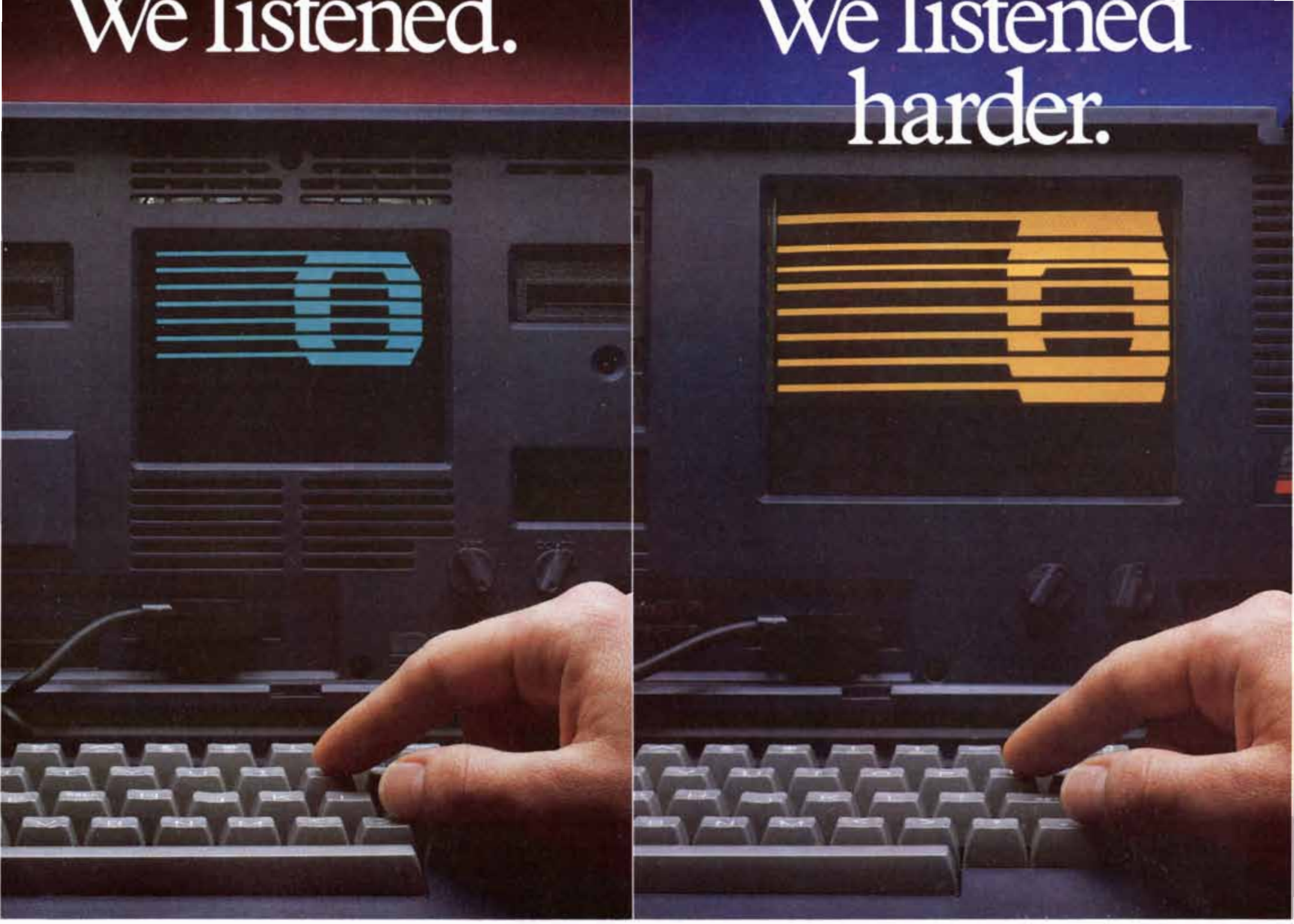
Some of the simultaneous analyses the brain brings to bear on vision became apparent in the 1950's and 1960's as David H. Hubel and Torsten N. Wiesel of the Harvard Medical School described the physiological properties of nerve cells in the visual cortex, the specific part of the cerebral cortex receiving visual data from the retina by way of the lateral geniculate nucleus. They worked mostly with cats. The cells with the simplest properties were found to be most responsive (that is, they changed their pattern of electrical activity most markedly) when the investigators presented a small bar of light at a particular orientation in a particular part of the cat's visual field. If the orientation was altered or the bar was enlarged, the cell became less responsive. Indeed, its electrical activity could be inhibited. Evidently the cell reacted to "on" regions and "off" regions in the visual field. In many cases the regions were arrayed so that the resulting "receptive field" of the cell (the locus in the visual field to which the cell is sensitive) had an "on" center and an "off" surround.

Other cells in the visual cortex have more complex receptive fields. They respond best to discontinuous bars of light at particular places and orientations. An overall pattern emerged, in which nerve fibers carrying visual data from the lateral geniculate nucleus pass their signals to cells in a certain layer of the visual cortex; it is designated layer 4*c*. The cells in this layer monitor specific parts of the visual field but have no specificity for the orientation of the stimulus. They communicate, however, with cortical cells in layers 2 and 3, and the communications are "wired" in ways that make the latter cells quite sensitive to stimulus orientation. The cells in layers 2 and 3 communicate in turn with cells in layers 5 and 6 in ways that give many of the cells in these deepest cortical layers still more complex sensitivities. The pattern offered tantalizing hints of how the brain might analyze images for their forms, or at least for features such as edges and corners. It offered no hint of how the brain sees colors.

While Hubel and Wiesel were making their discoveries, Edwin H. Land and his colleagues at the Polaroid Corporation were formulating the retinex theory of color vision. Land began by noting that the colors of a scene look much the same to an observer no matter how the scene is illuminated (in sunlight, for example, and then in deep shade). Moreover, a scene illuminated with artificial light in a quite narrow range of wavelengths often turned out to retain a surprising wealth of perceived colors. Land concluded that the colors seen by an observer have no simple relation to the wavelengths of the light entering the eyes. The colors result instead from computations performed by the brain.

The retina is known to have three distinct types of cone cells, which are the light-receptor cells sensitive to color. One type is most sensitive to the shorter wavelengths of visible light, the second to the middle wavelengths and the third to the longer wavelengths. Land therefore proposed that each type of cone cell supplies information enabling the brain to compute ratios of brightness in the corresponding range of wavelengths. The ratios would be calculated at the boundaries between objects of differing color encountered along arbitrary paths across the visual field. The process amounts to the assignment of three brightness values (short-, middle- and long-wavelength "lightness") to any given part of the visual field. The triplets of brightness values are scaled by conditions that apply across the entire field; thus the ratios are not much affected by how the field happens to be illuminated. As a result the triplets can yield more or less invariant color sensations.

"Since we are uncertain of the loca-



Introducing the new Osborne™ Executive™ Series.

We listened when you told us you wanted a personal business computer.

We delivered the Osborne 1.

And in the two years since its introduction, the Osborne 1 has become the largest selling portable business computer on earth.

We listened harder.

The result is the new Osborne Executive Series. Consider these new advantages:

☐ A big, 80-column, eye-easy amber display lets you see more of the information you're using, with less effort. ☐ Memory is expanded, for larger spreadsheets. ☐ The new Executive Series now includes a database system—Personal Pearl.™ (That's in addition to an already formidable list of standard Osborne software, including word processing and electronic spreadsheets.) ☐ Terminal Emulation provides access to information stored on other computers. ☐ IBM Personal Computer™ compatibility (on the Executive 2) lets you use software packages designed for the IBM system.

And there's much more.

Add up the value of all the hardware features and software systems included with either the Executive or Executive 2 and you will see that the sticker price is one of the incredible bargains available in personal computing.

Visit your local authorized Osborne dealer. Once again, Osborne has set the standard for portable business computers.

For the name of your nearest Osborne dealer, dial (800) 772-3545 ext. 905 (in California); (800) 227-1617 ext. 905 (outside California).

OSBORNE™
COMPUTER CORPORATION

Trademarks: Osborne 1, Executive, Executive 2: Osborne Computer Corporation. Personal Pearl: Relational Software. IBM Personal Computer: International Business Machines.

tion of the mechanisms that mediate these processes," Land wrote in 1977, "I have coined the term retinex (a combination of retina and cortex) to describe [them]." Land, in collaboration with Hubel and with Margaret Livingstone of the Harvard Medical School, now reports the results of an experiment that enables the investigators to specify a location in the brain: it is the cerebral cortex. In each trial of the experiment subjects were asked to call out a succession of small letters flashed rapidly at the same place (a "fixation point") on a screen. Well to the left of the fixation point lay a collage of colored paper. Attending to the flashing letters ensured that the collage lay entirely on the left side of each subject's visual field.

The collage could be illuminated by light shone through various filters. Each filter, however, had a hole cut in it, so that the illumination of a small test spot surrounding the fixation point never changed. If the perception of color depended on only the wavelengths of light, the test spot would always look the same. In fact it did not. "With the filters in place," Land and his colleagues write in a paper to be published in *Nature,* "the test spot was a chalky white, whereas with no filters it was a deep purple—although the flux from the test spot had not changed at all." In accord with the retinex theory, the altered illumination of the collage had affected the perception of color throughout the visual field.

Next Land and his colleagues enlisted the assistance of a 29-year-old man, identified as J.W., who had undergone neurosurgery two years earlier in a successful effort to reduce the number and severity of his epileptic seizures. The surgeon had cut the corpus callosum: a massive plate of nerve fibers that connect the cortex of the left and right cerebral hemispheres. Like other split-brain patients, J.W. showed no obvious subsequent impairment. His color vision was normal. Tests determined that the parts of the cerebral cortex specialized for language were on the left side of his brain. Thus J.W. could give verbal descriptions of things in the right half of his visual field. (Information from the right half of the field passes from the eyes to the left side of the brain.) Things in the left half of the field could not be described; information from the left visual field continued to pass from the eyes to the right side of the brain, but since the corpus callosum had been severed, the brain no longer had a path by which information could cross from the right side to the language centers on the left.

When J.W. was tested with Land's apparatus, the results were remarkable. At first the collage was in the left half of the visual field. The test spot spanned the vertical midline, which divides the halves of the field. J.W. "consistently reported that the test spot was 'white' regardless of the illumination." Then a mirror was interposed, so that the collage was in the right half of the visual field. J.W.'s first comment was that "he saw 'all colors....' It turned out that he was describing the entire [collage], something he had apparently not been able to do when it was presented to his nonverbal hemisphere." After that he described the test spot as purple or as white, "in complete agreement with the appearances to a normal observer."

If the retinex calculations were done by neurons in the retina, whose grasp of the visual world extends well across the vertical midline, the severing of the corpus callosum would have no effect on the perception of color. Each side of the brain would get the same color sensations based on identical retinex calculations, and J.W.'s responses would always duplicate those of a normal observer. J.W.'s responses did not. They show that the retinex calculations need the corpus callosum to provide a path from one side of the brain to the other. "The calculations," Land and his colleagues conclude, "must [occur] in the cerebral cortex."

Without the second recent discovery

Land's finding would have been vexing, because the cells in the cerebral cortex known to be sensitive to color are incapable of performing retinex calculations. Opponent-color cells, for example, had been known since the late 1950's. Gunnar Svaetichin of the Venezuelan Institute of Neurology and Brain Research and Edward F. MacNichol, Jr., of Johns Hopkins University first described them in the retina of various shallow-water fishes, and they have since been found by Hubel and Wiesel in the visual cortex of the monkey. Like other cells in the visual cortex, they respond in opposite ways to "on" and "off" regions in their receptive field. In the case of the opponent-color cells, however, the response depends on the wavelength of light. An opponent-color cell might be excited by a small red spot and inhibited by a large field of green. ("Red" and "green" refer here to wavelengths of light, not to perceived colors.) Presumably the input to the cell derives ultimately from at least two types of cone cells. The problem is that a red spot on a green background would leave the cell unaffected. The excitation and the inhibition would cancel each other. Since the cell could make no response to a boundary between contrasting colors, it could make no contribution to retinex calculations.

What is needed for the calculations is a network based on what are termed double-opponent-color cells. A cell of this type might be excited by, say, a small red spot in a particular part of the visual field. It would be inhibited by a small green spot in the same position. The colors surrounding the position would have the opposite effect: red would be inhibitory and green would be excitatory. The cell would thus make no response to white light of any pattern, nor to diffuse light of any wavelength. It would make no response to a large expanse of red or of green; the center excitation and the surround inhibition would cancel. Its greatest excitation would be in response to a red spot on a green background, its greatest inhibition in response to a green spot on a red background. Hence it is specialized to respond most markedly to contrasting colors.

In the 1960's Nigel W. Daw, who is now at the Washington University School of Medicine in St. Louis, found double-opponent-color cells in the retina of the goldfish. Soon a number of investigators began to find them in the visual cortex of the monkey. In all accounts, however, the cells were disconcertingly few and gave no sign of being concentrated in any particular cortical locus. Then in 1978 Margaret T. T. Wong-Riley, who is now at the Medical College of Wisconsin, stained the visual cortex of the monkey by a technique that reveals the distribution of cytochrome oxidase, an enzyme found in mitochondria. Mitochondria are the organelles that produce energy in living cells; thus the stain is thought to mark sites of great activity in a tissue. Wong-Riley found dense concentrations of cytochrome oxidase in flask-shaped blobs about 200 micrometers in width spaced at .5-millimeter intervals across the visual cortex. In 1980 it emerged that the blobs form a polka-dot array in layers 2 and 3. It has since emerged that the blobs receive nerve fibers from the lateral geniculate nucleus.

Hubel and Livingstone employed a microelectrode to probe the cells in the blobs. None of them was sensitive to the orientation of a bar of light. Apparently they take no part in the analysis of edges and corners. On the other hand, as many as 90 percent of them are sensitive to color. Most are double-opponent-color cells.

Vaccine Gene

Two of the powerful new tools of molecular biology, monoclonal antibodies and the molecular cloning of recombinant DNA, have been exploited


GM

to make the development of a vaccine against malaria feasible in principle. The manufacture and testing of such a vaccine is probably some years away, however, in part because of a disagreement between the biotechnology company that would manufacture the vaccine and an international agency that has helped to support the research.

After World War II there was reason to hope that a combined attack on *Anopheles* mosquitoes with DDT and on *Plasmodium* parasites with new specific drugs might soon bring malaria under control. For a time there was progress. Now the incidence of malaria is on the rise, however, and so is the death rate. The logistics of spraying and of drug delivery have been daunting in underdeveloped regions and particularly in tropical Africa, where malaria is endemic and a million children die of the disease every year.

In this situation a malaria vaccine would be an ideal weapon, but the plasmodium is a difficult target for immunologists. Four species cause malaria in human beings and each one carries on its outer surface a different set of antigens, the proteins that provoke (and become the targets of) an immune response by antibody-secreting cells. The antigens may also vary in different strains of a single species. Indeed, the plasmodium displays different sets of antigens successively in the several morphological stages of its complex life cycle in the mosquito and in the human victim. The mosquito's bite injects sporozoites, which infect liver cells and divide to form small merozoites. Each merozoite invades a red blood cell, grows through several stages and divides into from 12 to 24 new merozoites that burst out of the cell and invade other cells. A few merozoites develop into gametocytes, whose ingestion by a biting mosquito initiates a new cycle and spreads the disease.

The complexity of the various parasite forms and their antigenic variability have made it clear that intact parasites, whether attenuated or killed, will not serve as the active agent of a vaccine. The effort recently has been to develop a vaccine by finding and isolating a particular antigen, preferably one whose role in infection is so important that it precludes much variability from strain to strain. For some years Ruth and Victor Nussenzweig of the New York University School of Medicine and various collaborators have taken that approach. They have isolated from the surface of the sporozoites that cause a number of animal and human malarias a protein whose composition is similar in all species and that has a role in the process of infection. By crossing specific antibody-producing cells with cells of the tumor called a myeloma they have made hybridoma cells that secrete monoclonal antibodies directed specifically against the sporozoite surface protein. Injected into chimpanzees, the antibodies give a measure of passive immunity against malaria, suggesting that the surface protein should serve as an effective antigen.

To make the large quantities of the protein needed even for testing a vaccine one might dissect millions of mosquito salivary glands and isolate the sporozoites and then the protein. An alternative is to find the gene encoding the protein, introduce it into *Escherichia coli* and get the bacteria to manufacture the protein. That is what G. Nigel Godson of N.Y.U., with his students Joan Ellis and Luiz Ozaki, has done.

Godson's method, which is reported in detail in *Nature*, relied heavily on the existence of the Nussenzweigs' monoclonal antibodies. He extracted messenger RNA (the nucleic acid into which the DNA of genes is transcribed and that in turn is translated to make protein) from mosquitoes infected with *P. knowlesi*, a plasmodium that causes malaria in monkeys. He made DNA from the RNA, broke up the DNA into thousands of short fragments and inserted each fragment into a plasmid: a small circular piece of bacterial DNA that serves as a vehicle for cloning. Each plasmid was inserted into bacteria. To find any bacteria that were synthesizing the *P. knowlesi* surface protein, and must therefore carry on a plasmid the gene encoding the protein, he tested each bacterial colony with the monoclonal antibody to the protein. When he found such bacteria, he grew their plasmid in quantity and showed by various tests that it did indeed carry the gene for the designated protein. The next step will be to show that when the protein made in bacteria is injected into monkeys, it confers immunity to malaria.

With the gene for the *P. knowlesi* protein in hand, it should not be hard to find the gene (which must be quite similar) for the analogous protein of *P. falciparum*, the agent of the lethal form of human malaria. That effort is under way at N.Y.U. Meanwhile the institution has filed for a patent on the process and has been negotiating an agreement licensing Genentech, Inc., a San Francisco company that specializes in recombinant-DNA technology, to develop, test, manufacture and market a vaccine. Any income to N.Y.U. would be devoted to research in tropical medicine. The licensing agreement requires the approval of public agencies that have helped to fund the work.

One of the agencies, the World Health Organization (WHO), objected to a provision giving Genentech exclusive marketing rights, without which the company is reluctant to develop the product. WHO is presumably worried less by the financial aspects of the agreement than by the possibility that Genentech, having tied up rights to the vaccine, might never market it. Agency officials have often complained that drugs having critical importance for the underdeveloped world are sometimes not made available because there is no profitable market for them in affluent countries.

Prime Time

When a computer that can do a million operations per second is responding to keystrokes made at intervals of some tenths of a second, it can have a lot of spare time. Paul A. Pritchard of Cornell University decided to put a computer to work in those idle moments. He instructed the computer, whenever it was not busy doing something else, to search for a series of at least 18 prime numbers that advance in an arithmetic progression. (A prime number is divisible only by itself and 1; in an arithmetic progression the difference between each number and the next is the same).

Pritchard wanted 18 such prime numbers because the previous record, set in 1977, was 17. Nipping in for a microsecond here and a millisecond there on a computer that is devoted mainly to word processing and runs 24 hours per day, he was able to find a series of 18 arithmetic-progression primes in less than a month. The computer was actually working on the problem for an average of about 10 hours per day. The series begins with 107,928,278,317 and rises by an increment of 9,922,782,870 to 276,615,587,107. The numbers before the first in the series and after the last (taking into account the same increment) are not primes.

The strategy for the search was based on the fact that in any arithmetic progression of n primes the common difference must be divisible by all the primes less than or equal to n, except in the rare case of a series whose first number is n. In Pritchard's search n was 18, and so the common difference had to be divisible by 2, 3, 5, 7, 11, 13 and 17. Pritchard also included 19 in case the computer hit on a series of more than 18 arithmetic-progression primes. The common difference must be a multiple of the product of these eight primes, namely 9,699,690. Pritchard therefore had the computer choose a prime in some designated range of numbers, construct an arithmetic series by repeatedly adding a multiple of the product and test the resulting numbers for primality.

Pritchard had his in-and-out program going on two computers, each a fairly standard machine that is not notably fast. One of them is still searching after thousands of hours. Pritchard is continuing the search with both machines. He expects to find more 18-term primes in arithmetic progression and hopes to find one with 19 terms.

Modern Pork Production

Pork currently provides some 25 percent of the energy and 9 percent of the protein human beings get from animal sources. The interaction of economics and biology has given rise to a new kind of pig farming

by Wilson G. Pond

The pig lingers in the memory of many people as a farmstead animal that roots in the ground for various edibles, wallows in mud and is fed slop and swill, meaning the liquid and semiliquid leavings from the farmer's kitchen and the operation of the farm. That is not the pig of modern agribusiness. Such a pig lives indoors for its entire brief life: born and suckled in a farrowing unit and raised to slaughter weight in a nursery and later in a growing-feeding unit. It is fed a computer-formulated diet based on cornmeal and soybean meal with supplements of protein, minerals and vitamins. Unless it is destined for breeding, it is sent to market at five or six months of age, having reached the slaughter weight of 220 pounds or more from its birth weight of about two pounds. In the U.S., which is one of the leading countries in the production of pork, some 97 million pigs with a value of $8.9 billion went to market in 1980.

It is best to be precise about terminology. All pigs, wild and domestic, are swine, members of the family Suidae, which in turn is a member of the mammalian order Artiodactyla (even-toed ungulates). In the U.K. all domestic swine are called pigs, but in the U.S. the term applies technically to young swine weighing less than about 120 pounds; the others are called hogs. One hears also of boars (uncastrated males), barrows (castrated males), gilts (virgin females), sows (females that have bred) and shoats (newly weaned piglets).

The domestic pigs of today are descended from the European wild pig (*Sus scrofa*), which is also known as the European wild boar. Pigs living in the wild are still found in many regions of the world. Except for old males, which are solitary, wild pigs live in groups. They are nocturnal and omnivorous, digging for edible roots and tubers, gathering fruits and nuts on the ground and also eating a large variety of insects, reptiles, birds and small mammals. The wild pig stands about 90 centimeters (35 inches) tall at the shoulder and has a thick coat of bristly hair. It is a fast runner and a good swimmer.

The time of domestication of the pig is lost in history. One study has it that the earliest known domestication of pigs was in what is now Iraq in about 6750 B.C. A Chinese scholar has asserted that domesticated pigs were to be found in his country by 2900 B.C. Probably the attractiveness of the pig as a domestic animal originally was that it is an efficient scavenger and will eat a wide variety of foods. That is still its role in many developing countries.

In modern agribusiness the pig is attractive because of its efficiency in converting feed into food. Today's pig gains a pound of weight for each three or 3.5 pounds of food it eats. Fifty years ago it took more than four pounds of food to achieve the same result. In terms of the pig's yield of edible energy per calorie consumed it outdoes cattle, lambs and poultry. Indeed, the production of pork, like that of beef, lamb and poultry, is based on the economic advantage for the farmer of selling his crops through animals at a higher profit than he would realize by selling the crops directly for human food.

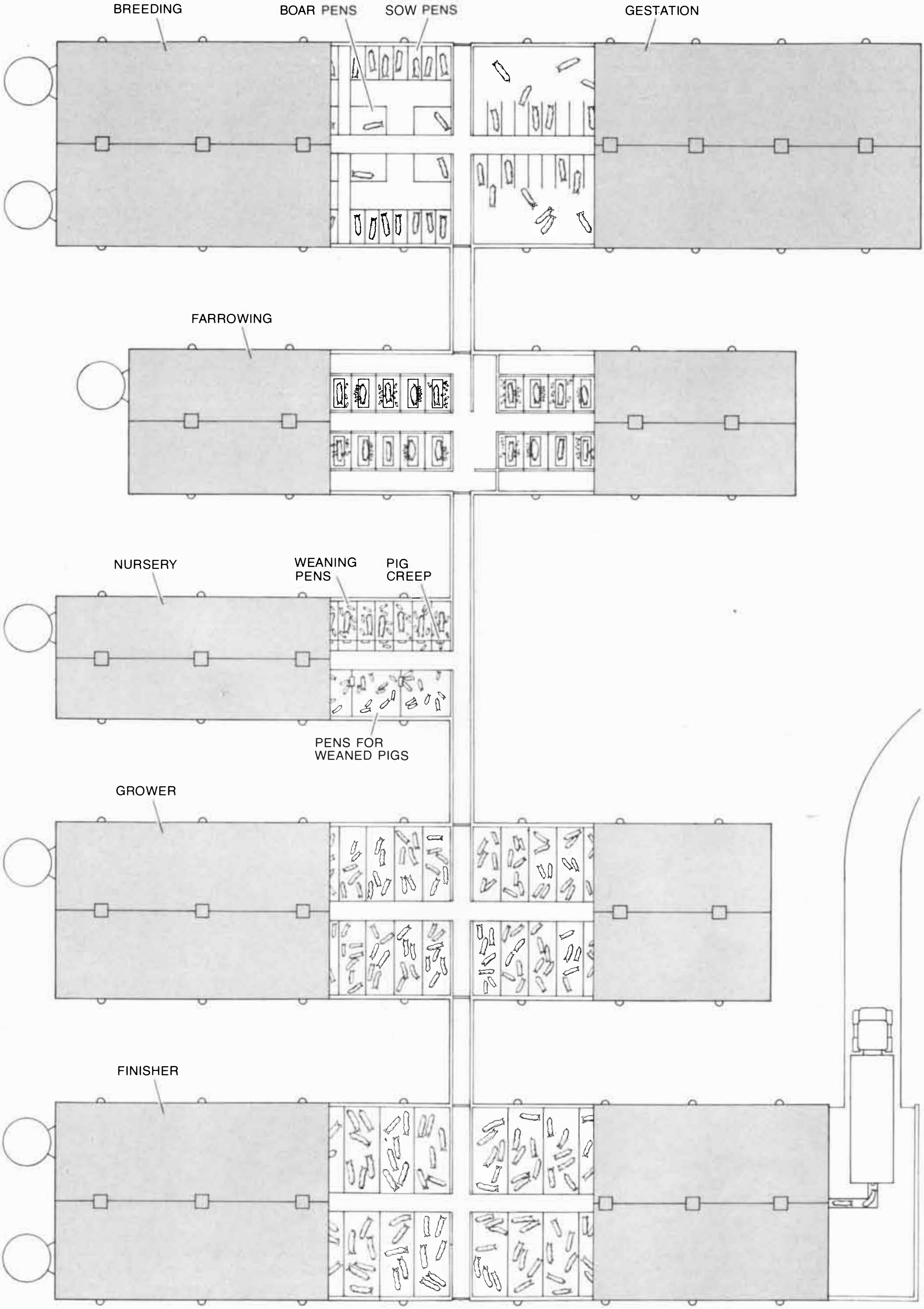
From the consumer's point of view the meat of the pig has high nutritional values. A four-ounce serving of lean pork supplies about three-fourths of the adult's daily need for thiamin and iron, from a fourth to a third of the need for niacin, vitamin B-6 and vitamin B-12, half of the need for protein and most of the requirement for trace minerals. In North America and Europe the annual consumption of pork is about 60 pounds per capita; the worldwide average, however, is less than 20 pounds per capita. Since the demand for animal protein in the diet increases linearly with average per capita income in a country, the geographic disparities in the consumption of pork largely reflect the huge disparity in per capita income among countries.

Modern Practices

The trend that took the pig out of the pigpen and into a carefully controlled environment resulted from high labor costs, the rise in land values, the fact that it became more profitable to raise crops on productive land than to keep animals on it, the ability to keep the pig's diet under closer control indoors and the possibility for more effective control of diseases and parasites. On a farm that raises pigs from birth to marketing age, which is a strong modern trend, a typical physical arrangement includes a unit housing pregnant sows, a farrowing unit where the sows give birth, a nursery and a growing-finishing unit. They all have slatted floors for easier removal of wastes. In Temperate Zone farms each unit (except sometimes the growing-finishing unit) has a heating system.

The main task of husbandry is feeding the pigs, an operation that accounts for from 55 to 85 percent of the total cost of the commercial production of pork. The variations result from the relative costs of feed, labor and housing between one farm and another or from one season to the next. The economics of feeding swine depend on the local availability and price of feedstuff, competition for a particular foodstuff for consumption by people or by other domestic animals and the prices of the protein, mineral and vitamin supplements.

TOTAL-CONFINEMENT UNIT depicted on the opposite page keeps pigs indoors from birth until they are sent to market. The perspective is straight down on a five-building complex, with sections of the roofs removed to show the activity inside. In the farrowing unit pigs are born and suckled while the sow remains in a metal crate that prevents her from rolling onto her piglets. In the nursery, the grower and the finisher the pigs are fed carefully formulated diets designed to bring them to the slaughter weight of 220 pounds or more at the age of five or six months. Many farms that raise pigs in this way have the grower and finisher as one building.


BREEDING
BOAR PENS
SOW PENS
GESTATION
FARROWING
NURSERY
WEANING PENS
PIG CREEP
PENS FOR WEANED PIGS
GROWER
FINISHER

The pig's digestive system and the human one are similar, so that the pig farmer is in direct competition with the human population for food supplies. The degree of competition is related to cultural differences in food preference. For example, in the U.S. wheat and potatoes are usually not fed to pigs because the human demand for these crops holds the price too high, whereas in other parts of the world pigs often get wheat and potatoes in their diet.

The farmer must plan on using up about 800 pounds of a typical concentrated diet based on cornmeal and soybean meal for each pig marketed at 220 pounds. The total includes feed that goes to the breeding herd and also takes into account the fact that even in a well-run operation a fair number of pigs die before reaching marketable age. Of the 800 pounds of feed from 100 to 200 pounds consist of the protein, mineral and vitamin supplements needed to make up the nutritional deficits in corn and cereal grain. The nutritional requirements of pigs at various stages of production have been so precisely determined that commercial feed manufacturers now formulate diets by computer. The number of farmers working the diets out on their own computers is increasing.

Although unprocessed corn and cereal grains make up a major part of most diets for pigs, the computer-formulated least-cost rations contain progressively higher proportions of by-product foodstuffs, including wheat and corn bran from mills, grains and solubles from distilleries and blood meal, meat meal and bone meal from slaughterhouses. Such by-products would have little value if they were not sold as feed for livestock; if they were simply discarded, the cost of disposing of them would significantly increase the market price of the final product.

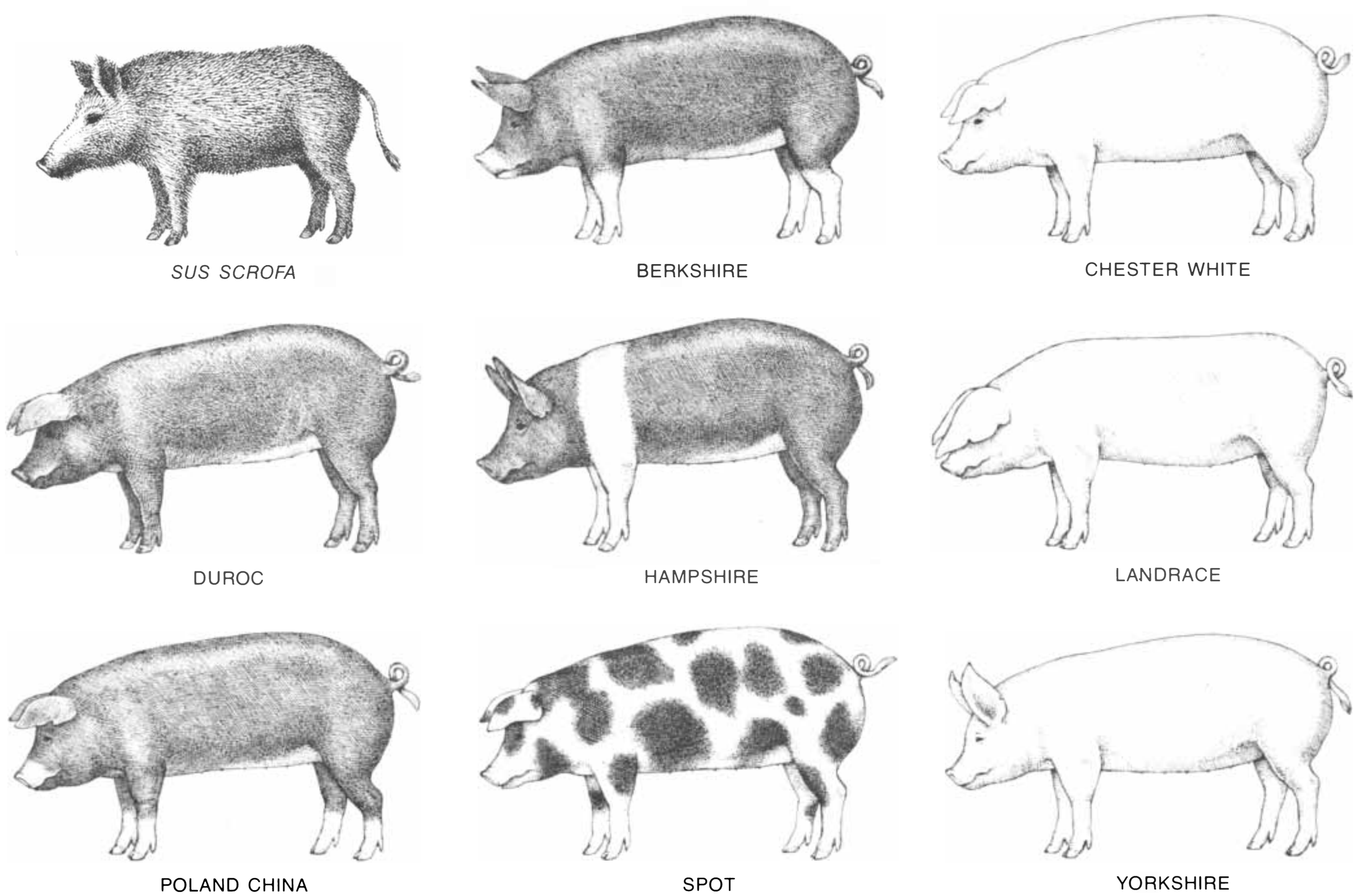
It has also been found that forage crops such as alfalfa can be incorporated in the diet of pigs, particularly pregnant females. The reproductive performance of such females has been shown in recent tests to be normal when high-quality alfalfa meal constitutes 96 percent of the diet, with mineral and vitamin supplements making up the other 4 percent.

Biology of the Pig

The yearly production calendar of the pig farm must be geared to the biological timetable of the pig. Although a pig reaches puberty at five or six months of age, it is common for mating to be delayed until about eight months so that the animal will be larger and physiologically more mature. The gestation period is about 114 days, and the lactation period of the female continues for from six to eight weeks. The estrous cycle takes up about 21 days and repeats throughout the year. Pregnancy halts the cycle, which is also inhibited in the pig (as it is not in the cow) by lactation. The cycle resumes in from three to 10 days after weaning, which normally is done when the pigs are from four to six weeks old.

The pig's natural course of development dictates that the female can complete her first pregnancy at about one year of age and that she can complete no more than 2.7 cycles of reproduction and lactation per year thereafter. In practice the average reproductive rate of a herd is closer to 1.8 litters per year after adjustments are made for barren sows, failure to conceive on the first mating and other factors. A litter consists typically of nine or 10 piglets, of which 7.2 on the average survive to maturity.

Although a newborn pig weighs only about 2.2 pounds, it grows at a phenomenal rate, normally doubling its weight in the first week and tripling it by the end of the second week. A weight of 15 pounds after three weeks is not uncom-



PUREBRED BREEDS provide the stock for the crossbred herds found on most commercial pig farms. Also portrayed is the ancestral pig, the European wild boar (*Sus scrofa*). The Berkshire and the other domesticated purebreds shown here constitute the seed stock for most of the crossbreeding in the U.S. A few countries raise mainly purebreds, such as the Landrace herds predominating in Denmark.

mon. A market weight of from 220 to 240 pounds is attained by about five months of age, but individual pigs are known to have reached 200 pounds in less than 120 days having eaten less than two pounds of feed per pound of gain in weight.

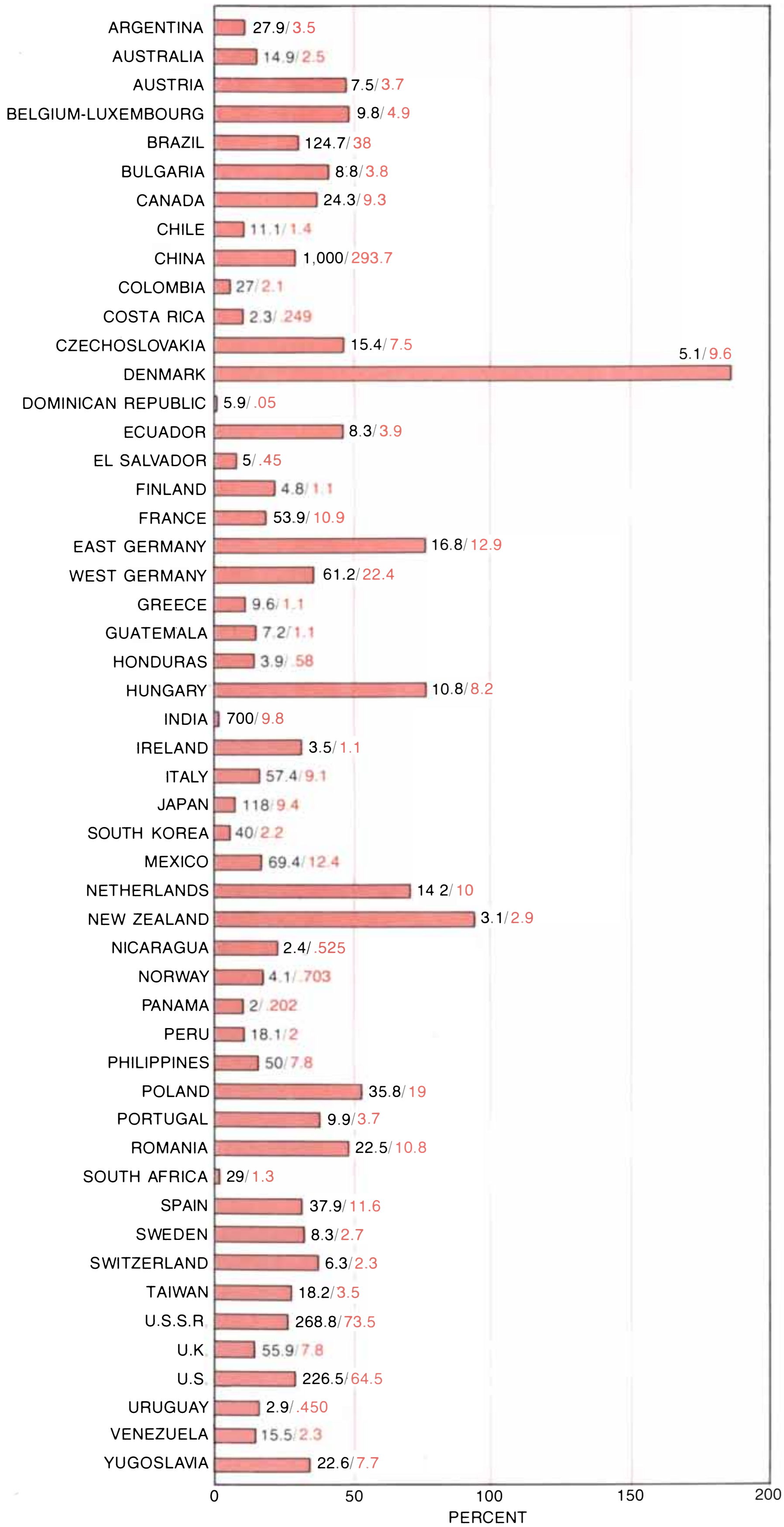
Male pigs are normally castrated during the first few days or weeks of life to avoid the marketing of pork with the objectionable taste associated with the meat of mature boars. The surgery can be done much later, however, with the same effect, and it is noteworthy that intact males grow faster and are leaner than castrates. As a result research is under way in several countries to develop mature boars that yield meat lacking the objectionable taste. Farmers often delay the castration of potential breeding boars in order to have more time to choose the ones that seem likely to make the best sires.

Role of Crossbreeding

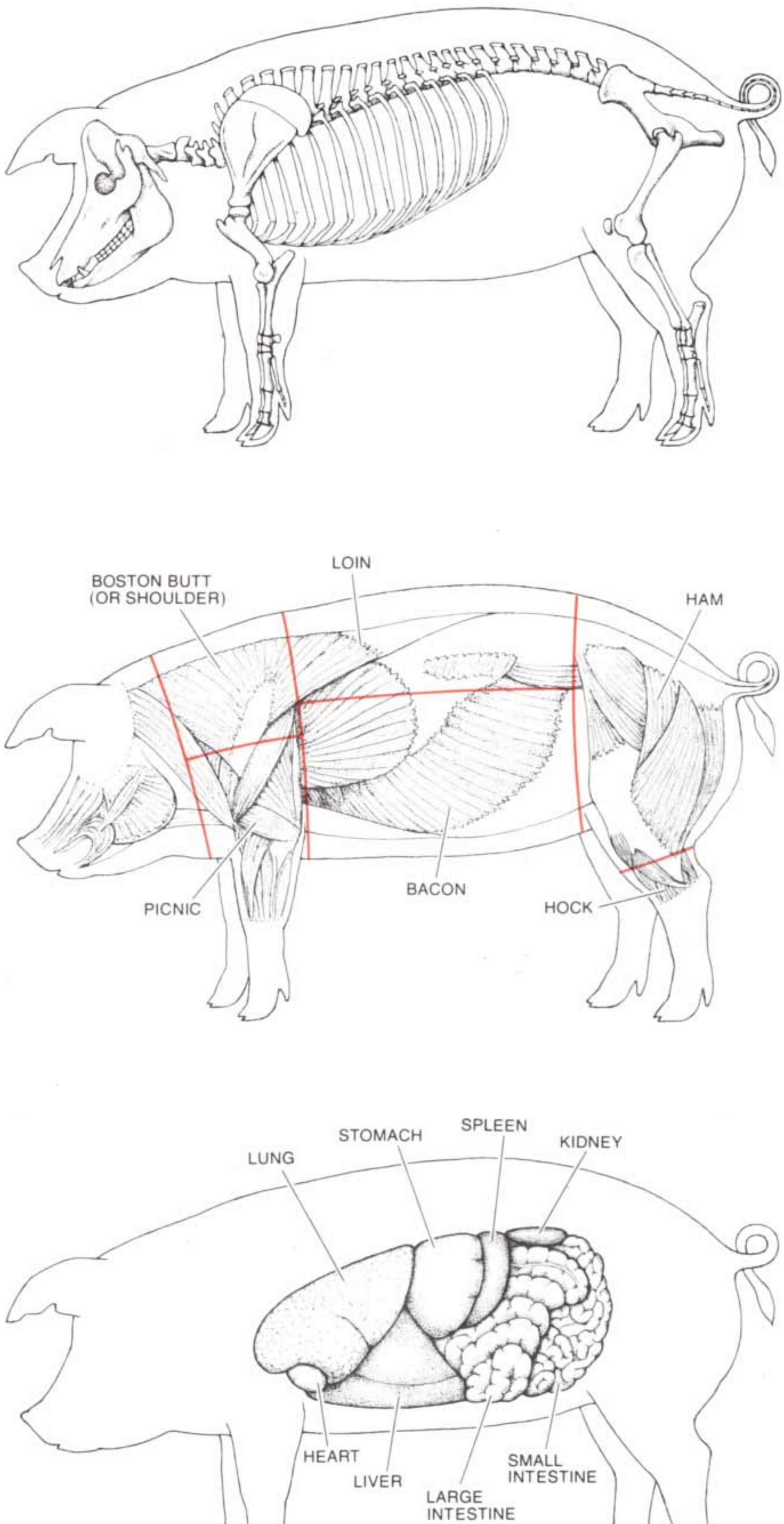
The genetic base for improving the breed in economically important ways is a group of purebred populations that constitute less than 10 percent of the total population of swine in most countries. There are exceptions, as in Denmark, where most of the pigs produced are purebred Landrace. In most other developed countries a few major breeds provide the seed stock for a commercial pork-production industry based on crossbreeding. In the U.S., for example, some eight breeds fulfill that role. In addition to the Landrace they include the Berkshire, the Chester White, the Duroc, the Hampshire, the Poland China, the Spot and the Yorkshire. Most of the swine breeds that figure in commercial production throughout the world originated in the U.S. or the U.K. The Yorkshire of the U.S. is similar to the Large White of Europe, and the American Landrace is similar to the Landrace of the Scandinavian countries.

Crossbreeding, involving from two to four breeds in a typical systematic breeding program, accounts for more than 90 percent of the total annual supply of pork. Crossbred pigs are superior to their purebred parents in such economically important traits as rate of survival, rate of growth, efficiency in utilizing feed, production of milk and number of offspring. Crossbreeding is a valid method of production only if it continues to result in heterosis ("hybrid vigor"): the improved average performance of the crossbred progeny over that of their purebred parents. Continued progress in increasing the efficiency of swine production by crossbreeding depends on the intensive selection of purebred stock for excellence in the appropriate heritable traits.

Some of the pig farms in the U.S. are "farrow to finish" enterprises, in which



EXTENT OF PIG PRODUCTION in various countries is suggested by this comparison of human and pig populations. For each country the black numeral is the human population, the colored numeral the pig population, both in millions. The bar shows the number of pigs as a percentage of the human population. The comparison is somewhat inexact because of variations in the "extraction rate," the ratio of the number of pigs sent to market per year to the number on hand at any one time. For example, in the U.S. the rate is about 1.5 and in Brazil about .4.



ANATOMY OF THE PIG is depicted in a presentation that shows the skeleton, the musculature and the internal organs. A widely employed butcher's terminology for cuts of the meat is also shown. Variations of and additions to the terms (chitterlings, for example) are often seen.

the pig spends its entire life on one farm. Others are "feeder pig" operations that raise piglets to the age of from eight to 10 weeks (shortly after weaning) and then sell them to "finishing pig" enterprises that feed them to marketable weight. The two types of operation that handle piglets do certain things routinely: clip the "needle teeth," or incisors, which might otherwise injure the sow when the piglet suckles; inject iron; disinfect the navel; notch ears for purposes of identification, and dock the tail because pigs have a tendency to chew the tails of other pigs.

Pig farms also maintain vaccination programs to control infectious diseases, spraying and drenching programs to control internal and external parasites, mating and culling programs to maintain reproductive efficiency and to optimize the use of the farm's facilities, and marketing programs to maximize the net income. A farm with purebred animals must in addition keep detailed and accurate records of the genetic identity of its pigs. It must also develop effective advertising and marketing programs so that other pig farmers will be aware of what the farm has to offer for crossbreeding. A finishing enterprise has fewer critical requirements in husbandry and is less labor-intensive, but it has to be aware of the condition of the market so that the farm does not have too many or too few pigs at a given time.

A major task in managing a pig farm is looking after the health of the animals. Pigs are vulnerable to a large number of infectious, metabolic and nutritional diseases. Modern advances in the prevention of infectious disease and in the control of external and internal parasites have reduced the veterinary bills the pig farmer once paid for the treatment of acute illness among his animals. Such advances have brought about increased emphasis on the health of the herd and on disease-prevention programs involving vaccination against infectious diseases, routine control of parasites and close restrictions on the movement of pigs from one farm to another. Many commercial pork-production enterprises maintain herds that are relatively free of disease by measures that include limiting human and animal traffic and quarantining new breeding stock brought into the herd; in the jargon of the industry such a herd is SPF, or specific-pathogen-free.

Certain metabolic diseases still cause significant financial losses even in well-managed herds. One is porcine stress syndrome, which is associated with specific breeds and genetic groups. Another is the metritis-mastitis-agalactia complex, a multifactor disease syndrome resulting in a high rate of death among newborn pigs.

Nutritional diseases are less common now than they once were, but a few still appear even with today's sophisticated techniques for formulating diets. For example, an acute deficiency of selenium and vitamin E causes excessive deaths among growing pigs in certain regions of the U.S. and elsewhere that have a low content of selenium in the soil, resulting in a deficiency of the element in the plants grown there. As such deficiency syndromes are recognized the feed industry responds by providing appropriate supplements. Sometimes a supplement cannot be added to feed in the U.S. without the approval of the Food and Drug Administration. Selenium is an example, since it was known for its toxic properties before it was recognized as an essential nutrient; the FDA authorizes the addition of inorganic selenium to the diet of young pigs at levels of up to .3 part per million.

Prospective Changes

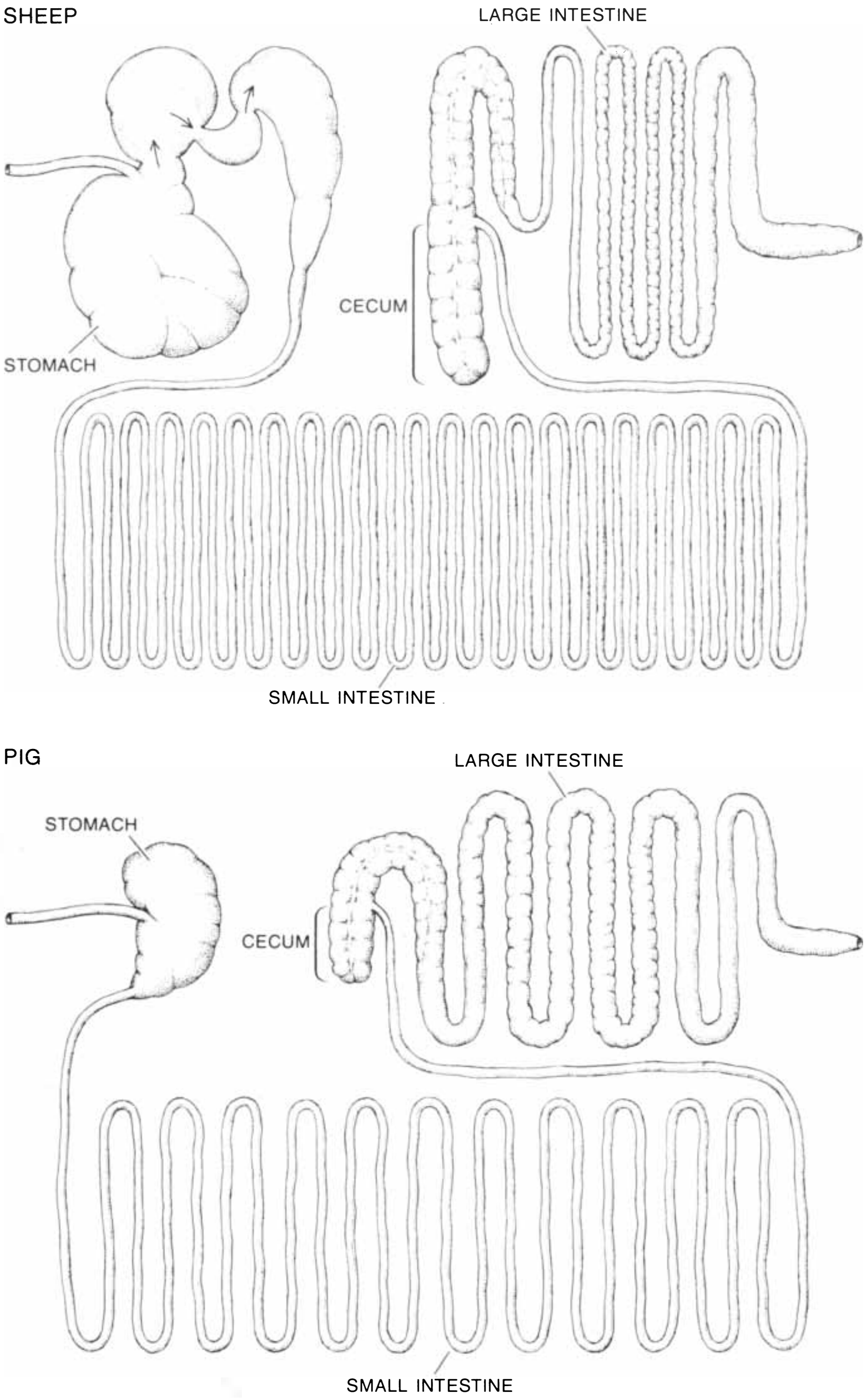
Although the pig industry has been greatly transformed over the past 30 years by advances in agricultural technology, biological knowledge and veterinary medicine, it seems probable that biological research now in progress will result in even further changes. They can be considered under the headings of genetics, nutrition and reproduction.

In genetics the farmer seeking better heritable traits in his crossbreeding activities is being aided by the work of population geneticists, who not only capitalize on the sophisticated modeling techniques made possible by computer technology but also introduce genetically diverse breeding stock into crossbreeding programs. Much research by geneticists in the U.S. Department of Agriculture and private organizations is directed at improving recognized traits that are beneficial in the production of pork and at finding new traits that might be beneficial.

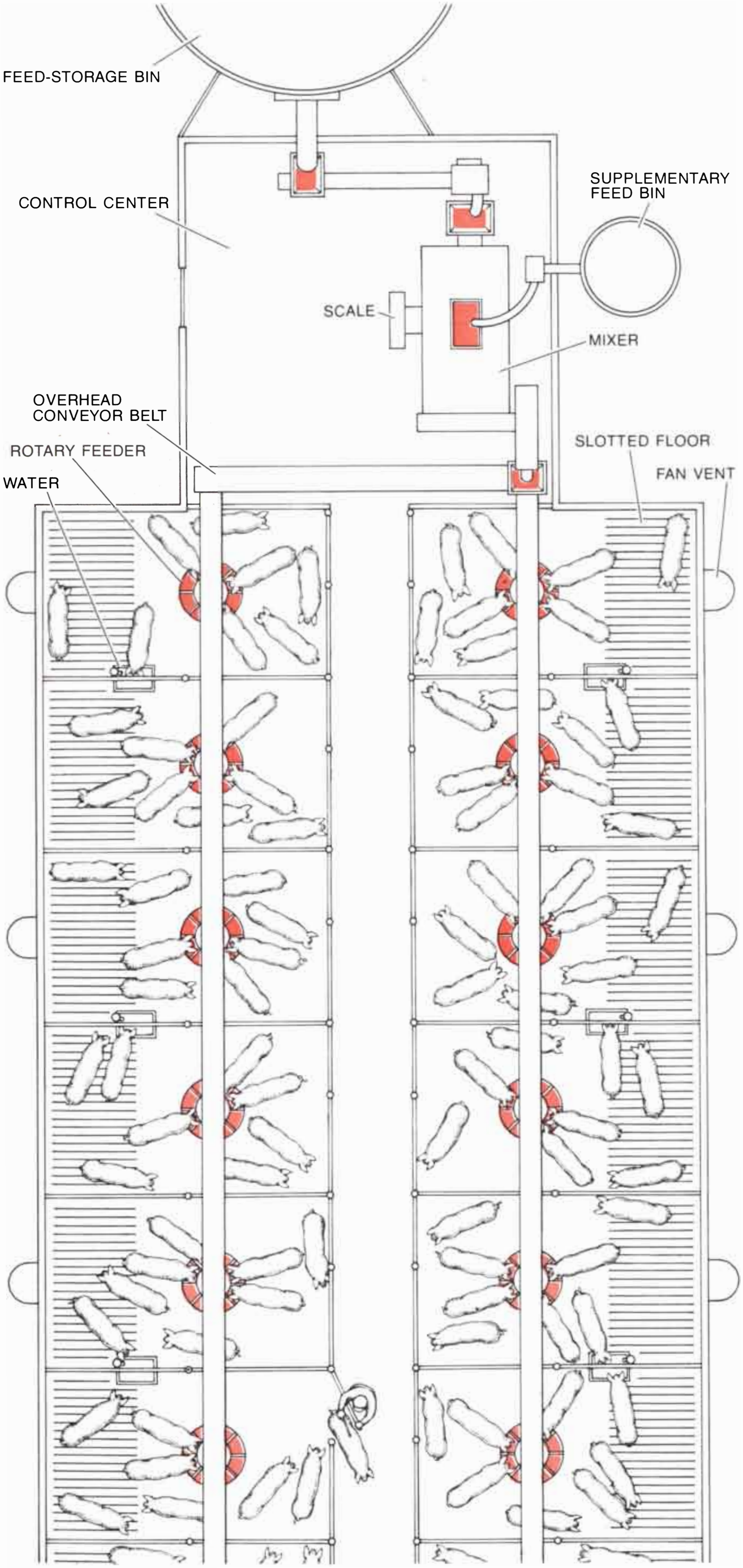
Techniques for estimating accurately the body composition of live animals, such as the measurement of back fat by ultrasonic devices, and for predicting the subsequent body composition of an animal by appropriate measurements done early in life offer possibilities for more effective selection of desired traits such as leanness. Genetic engineering by recombinant-DNA techniques ("gene splicing") shows promise of opening the way to the large-scale production of pig hormones by bacteria. With growth hormone made in this way the farmer might get his pigs to market weight sooner. Synthetic growth hormone is also being investigated as a modulator of milk secretion and the growth of lean tissue in a number of farm animals, including the pig.

Genetic variation in economically important traits such as growth and lactation is related to hormones, enzymes and various intracellular processes. The genetic bases for these relations are not well understood. Laboratory work on the growth of muscle cells in culture, together with other in vitro efforts to understand how protein and fat are synthesized and broken down, offers the possibility of fuller knowledge of the metabolic processes that control the buildup of muscle tissue in pigs and other animals.

In nutrition recombinant-DNA techniques may soon be applied to change the activity of the microflora in the pig's gastrointestinal tract in order to make



DIGESTIVE SYSTEMS of the sheep and the pig are compared. The sheep is a ruminant, with a large stomach consisting of four parts: the rumen, the reticulum, the omasum and the abomasum. The system evolved to digest cellulose. Grass eaten by the sheep passes first into the rumen; it is later regurgitated and chewed further. After it is swallowed again it passes successively into the other three parts of the stomach. The pig is an omnivore, but it cannot digest much cellulose. Research under way aims to exploit recombinant-DNA techniques ("gene splicing") to change the activity of the microflora in the pig's gastrointestinal tract to deal with preparations derived from grass and other fibrous plants, greatly expanding the sources of pig feed.



cellulase, the enzyme that cleaves cellulose, the most abundant constituent of plants, into its glucose monomers. Such a development would change radically the constraints on pig production imposed by the inability of the animal to digest cellulose and other complex carbohydrates. Foods the pig cannot now eat or fully digest would then be available to pig husbandry.

Recombinant-DNA techniques applied in another important way may soon make amino acids, the monomers of protein, available to animal husbandry. Recombinant-DNA strains of bacteria that are efficient producers of amino acids can be grown on a pilot scale. If amino acids made in this way were available at prices competitive with the cost of standard feed, more than four billion pounds of feed protein could be saved annually. The metabolic needs of the farm animal could be met with a low-protein feed combined with a well-balanced mixture of amino acids.

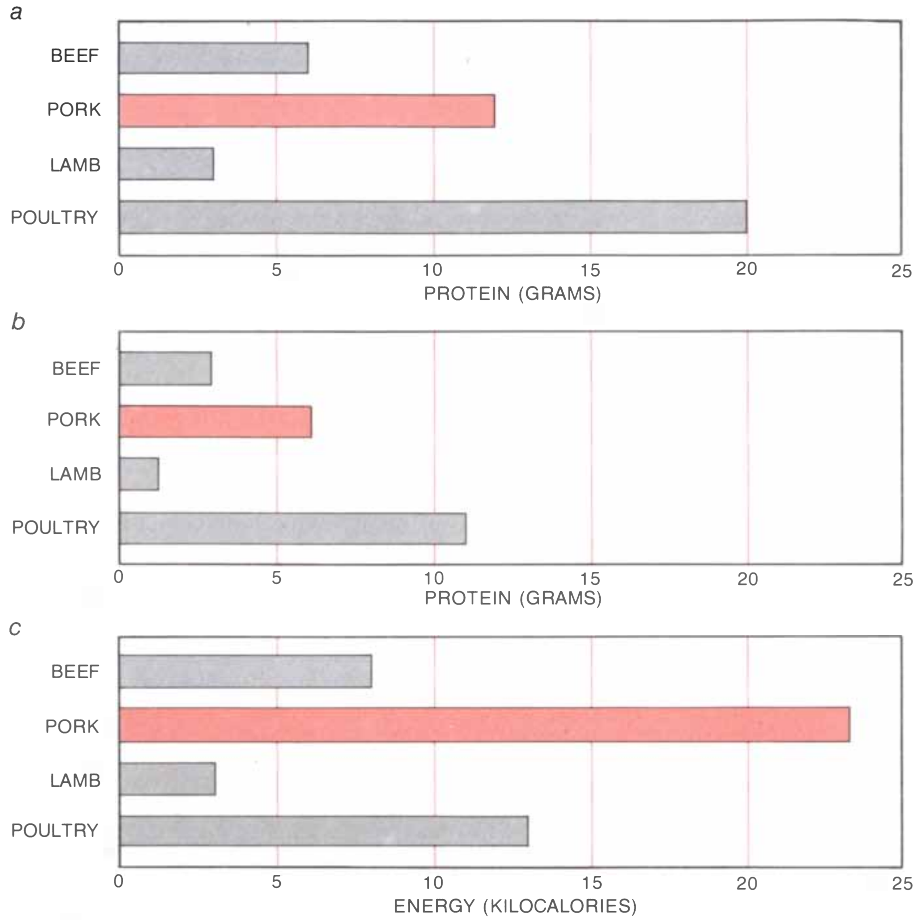
New types of feed will also play an important role in the future of pork production. Plant breeders continue to develop new varieties of corn, barley and oats with a high content of lysine, an amino acid essential to animal nutrition that is deficient in older varieties of these grains. Animals fed with the high-lysine varieties would need less supplemental protein. In developing countries pigs are being fed local products that were not traditionally a part of the pig's diet; examples include sweet potatoes and fish meal in New Guinea and bananas in Ecuador. Finally, several microbial sources of protein, including dried bacteria, yeast and algae, offer a potential supply of feed for pig farmers.

The interest in improving reproduction stems from the high death rates among pigs before birth and soon afterward. In the U.S. prenatal mortality is normally from 20 to 40 percent. A sow releases about 17 ova per period of estrus. Although about 95 percent of the eggs are fertilized, only 9.4 pigs are born on the average and 7.2 survive to marketable age. Even in Denmark, where emphasis is laid on efficiency in pork production, 20 percent of the pigs in a litter die before being weaned at the age of from six to eight weeks.

AUTOMATIC FEEDING ARRAY for pigs is designed to be part of a total-confinement unit. The feed from the storage bin is corn or cereal grains, moved from the bin to the mixer by an electrically operated auger. Protein, mineral and vitamin supplements are added at a rate of about one pound for each four to eight pounds of grain. The mixture is moved to the feeders by overhead conveyor belts or augers. The slotted floor facilitates waste removal. The feeder portrayed on the cover of this issue is a unit in such a feeding system.

Research on the genetic, endocrine and nutritional factors associated with high prenatal mortality shows the promise of making it feasible to induce superovulation—the release of a larger number of eggs—by administering suitable hormones. It has been possible for some time to synchronize ovulation in a group of virgin females by administering hormones, but the practice has not been widely followed because techniques for freezing the semen of boars have not been available. Synchronized ovulation or superovulation coupled with artificial insemination with fresh or frozen semen from genetically superior boars may make it possible to achieve higher rates of live births per pregnancy.

An improvement in reproductive efficiency must be accompanied by improved methods of caring for piglets in the days following birth. The sow's colostrum, the milk she secretes in the first few days after she gives birth, contains antibodies against pathogens that may infect the piglet. It has been found that after the piglets have received the colostrum for two or three days they can be reared successfully on liquid or dry diets. Indeed, dry diets may be associated with higher survival rates and a lower incidence of diarrhea. Combining the technology that would raise the average of viable pigs per litter to 15 or 20 with the early weaning of part of the litter after two or three days could result in a significant increase in the number of pigs marketed annually per breeding female from the present norm of from 12 to 14 to an average of 24 or more.



CONVERSION EFFICIENCY OF FOUR MEAT ANIMALS is charted. Shown are the amount of edible protein for each 100 grams of food consumed by the animal (*a*), the amount of edible protein per 100 kilocalories of metabolizable energy the animal eats (*b*) and the kilocalories of edible energy per 100 kilocalories of metabolizable energy in the animal's food (*c*).

Attainable Goals

The prospective advances of pig farming imply it is possible to envision certain goals that should be attainable within the next 20 to 30 years. This has been done by workers at universities and by my colleagues at the Meat Animal Research Center of the Department of Agriculture for eight factors that are economically important in the production of pigs. The number of litters borne per sow annually should rise from 1.8 to 2.2 and the number of pigs marketed per sow annually from 13 to 24. Today's high losses should decline; expressed one way, as the percentage of loss of pigs between birth and weaning, the change should be from the current 15 to 30 percent to 5 to 10 percent, and expressed in terms of the number of pigs raised per litter the change should be from 7.2 to 11. The number of pounds of feed per pound of gain in body weight from birth to market is now 3.5 and can be expected to be 2.5 or less, and the months of age from birth to market should decline from the five or six of the present to four or five. Because leanness of meat is increasingly sought by consumers, one can expect the percentage of fat in the edible parts of the carcass to decline from 41 percent to 32 percent or less and the average thickness of back fat at the time of marketing to decline from 1.6 inches to less than .8 inch.

In 1978 Henry A. Fitzhugh and some of his colleagues at the Winrock International Livestock Research and Training Center projected the proportions of food energy and protein that human beings will obtain from pigs, ruminants and poultry by the year 2000. They foresaw an increase of 23 percent in the number of pigs, 23 percent for poultry (chickens, ducks and turkeys) and 29 percent for ruminants (cattle, sheep and goats) between 1970 and 2000. In addition they predicted that food energy from pigs will be 215 billion megacalories in 2000 compared with 144 billion in 1970 and that protein from pigs will be three million metric tons compared with two million in 1970. Pork will provide 23 percent of the energy and 8 percent of the protein from animal sources, including milk, compared with 25 and 9 percent respectively in 1970. In other words, although the production of pork is predicted to increase substantially, the overall contribution of pork to human nutrition is predicted to decline slightly.

Other Roles of the Pig

Besides its importance to agriculture the pig is valuable as a model animal for many physiological and nutritional studies related to human health. Pigs are similar to human beings in the anatomy and physiology of several organs and organ systems, including the teeth, the eyes, the kidneys, the cardiovascular system, the skin and the digestive system. Studies of the pig have helped in elucidating a variety of human disorders such as cardiovascular disease, diabetes, gastric ulcer, muscular dystrophy, alcoholism and obesity.

In clinical medicine the heart valves of the pig have often been inserted as a replacement for damaged human heart valves. Several vital substances extracted from the tissues of the pig are widely employed in medicine. They include insulin from the pig pancreas and heparin (an anticoagulant) from the pig lung. Moreover, the skin of the pig can be grafted successfully onto the human body to repair burns.

New Inorganic Materials

Synthetic organic materials have come to play a key role in modern civilization, but unusual materials made out of sand, clay and other minerals have the advantage of requiring a smaller input of energy

by J. D. Birchall and Anthony Kelly

Look around you and you are likely to see things made of a wealth of diverse materials: metals, plastics, ceramics, glass and concrete. They are the stuff of a technological society. It takes energy to make them. In fact, the worldwide production of materials for construction and manufacturing annually calls for the equivalent in energy of 10^9 tons of oil, or 15 percent of the total worldwide expenditure of energy. The amount is equal to the total worldwide expenditure of natural gas.

The amount of energy required to make a material varies greatly from one material to another. It is useful to take portland cement as a standard, because it is made from chalk and clay, two readily available constituents of the crust of the earth. Moreover, portland cement is produced on a huge scale: almost 10^9 tons per year worldwide. It takes about 3×10^{10} joules of energy to make a cubic meter of portland cement. It takes six times as much to make a cubic meter of polystyrene plastic and 29 times as much to make a cubic meter of stainless steel.

Plainly portland cement, an inorganic material, has a considerable advantage in energy over plastics, which are made from organic material, and metals, which are smelted by processes that require very high temperature. This advantage, together with the wide and ready availability of the earth chemicals (the oxides of silicon and aluminum, the carbonates of calcium and magnesium and so on), suggests that technology should reexamine inorganic materials as replacements for energy-expensive materials, particularly at a time when energy itself is expensive and an end can be seen to the ready availability of hydrocarbons in the form of oil. Research we shall describe in this article is demonstrating that inorganic materials can in many applications indeed supersede the energy-expensive materials of present-day technology.

The first materials employed by human beings were flint and other kinds of stone, along with materials of animal or vegetable origin such as skin, bone and wood. Together they enabled human beings to excel at hunting and food gathering. Flint is hard, brittle and amorphous in its microscopic structure; thus it fractures conchoidally and can be fashioned into a sharp cutting edge. The Stone Age use of flint grew to such a scale that flint had to be mined rather than searched for in surface deposits.

The next material was bronze, which must have arisen from the chance juxtaposition of easily smeltable rocks (ores of copper and zinc) with sources of heat and with carbon. At much the same time it was discovered that wet clay would harden when it was baked and that before the baking it could be molded into a useful shape. Somewhat later came the reduction of iron ore into metallic iron. The earliest synthetic-material technologies, then, were metallurgy and ceramics.

A much later advance was the discovery, by the Romans, that a mixture of volcanic ash and lime reacts with water to form a hard, solid mass: a cement. This was the first low-temperature inorganic technology. In the reaction the silica (SiO_2) in the ash combines with the lime ($Ca(OH)_2$) to form a calcium silicate that has no regular chemical composition. In 1824 Joseph Aspdin improved on the technology by heating lime with a material such as clay, which contains silica. The result, after cooling, is portland cement clinker, which includes the silicates $2CaO \cdot SiO_2$ and $3CaO \cdot SiO_2$. When the clinker is mixed with water, it sets to a hard, firm, rocklike solid. Instead of shaping rock by chipping or cutting, civilization could now cast or mold it into the familiar form of concrete (cement mixed with sand and stone).

Metals, ceramics, rock and cement continued to dominate technology until the early years of this century. Cement, however, served only in low-grade applications. Then came the age of plastics, which can be said to have begun in 1907, when Leo H. Baekeland produced the first synthetic resin, Bakelite. Even before that, some natural polymers had been modified into materials with novel and useful properties. Natural rubber, for example, stretches when it is pulled. In 1839 Charles Goodyear found that the addition of sulfur stiffens it. The sulfur atoms form bridges between the hydrocarbon chains of which the rubber is composed. Nitrocellulose, developed by Christian Friedrich Schonbein in 1845, is brittle. In 1864 Alexander Parke discovered that the addition of camphor, a gummy substance from the camphor tree, makes it pliable. The camphor molecules act as a lubricant between hydrocarbon chains. The plasticized nitrocellulose served as the first motion-picture film.

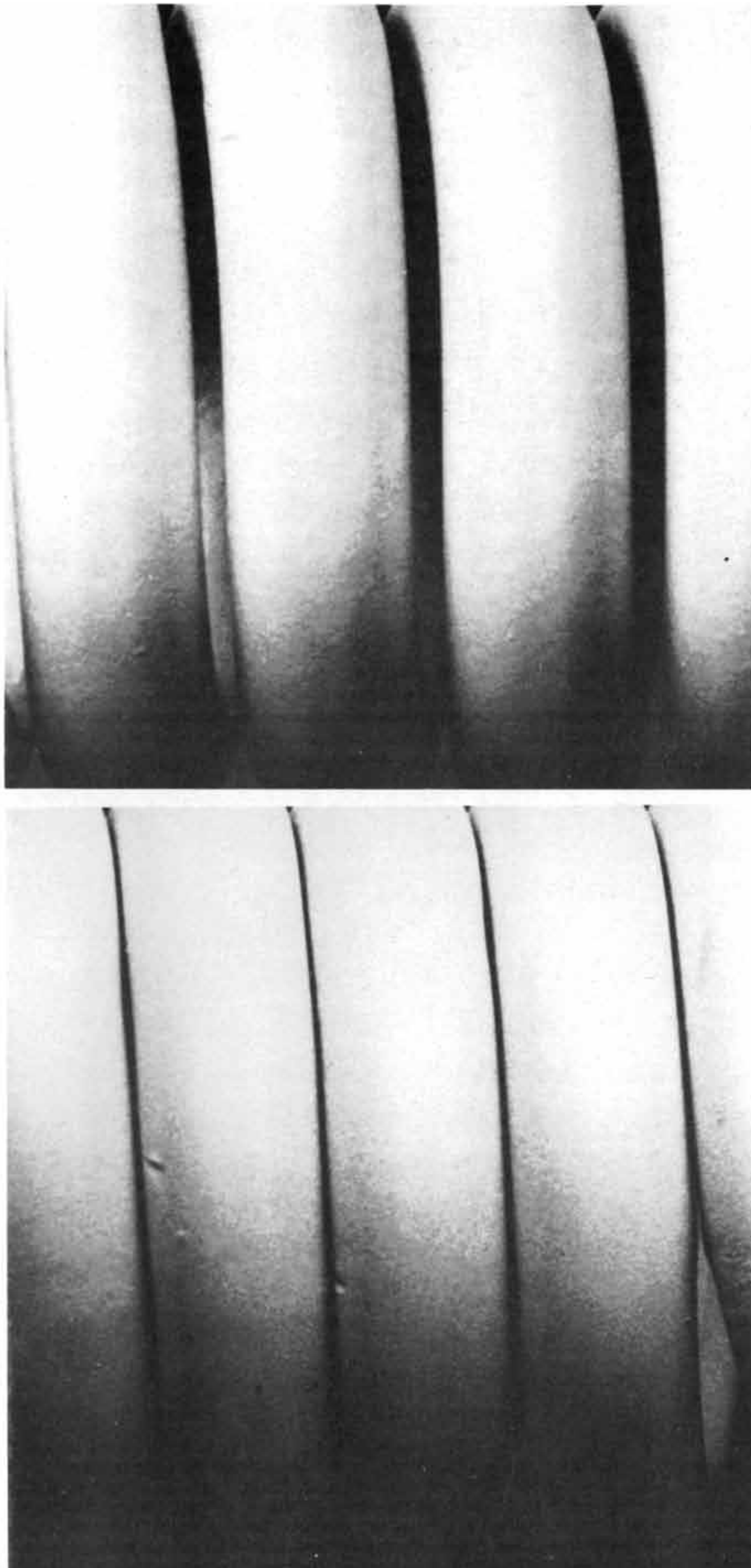
In the wake of Bakelite the age of plastics began in earnest. Synthetic rubber appeared in 1910, polystyrene in 1925 and nylon, polyethylene and polyesters in the 1930's. By 1940 the synthetic-polymer industry was well established. In that year the raw material for some 95 percent of all synthetic organic chemicals was coal. Today, four decades later, when the production of synthetic organic chemicals has increased more than a hundredfold, the raw material for some 97 percent is petroleum. Although the price of petroleum is currently decreasing somewhat, over the past eight years it has increased some fifteenfold.

What is the future for materials technology? One prospect is to derive synthetic polymers from coal once again. Surely that will happen. Another prospect is that the earth chemicals will again become prominent as raw materials. Apart from their availability, inorganic solids are intrinsically much stiffer than organic polymers. Moreover, concern is growing about the fire hazard of organic materials in homes and public buildings. Inorganic materials do not burn.

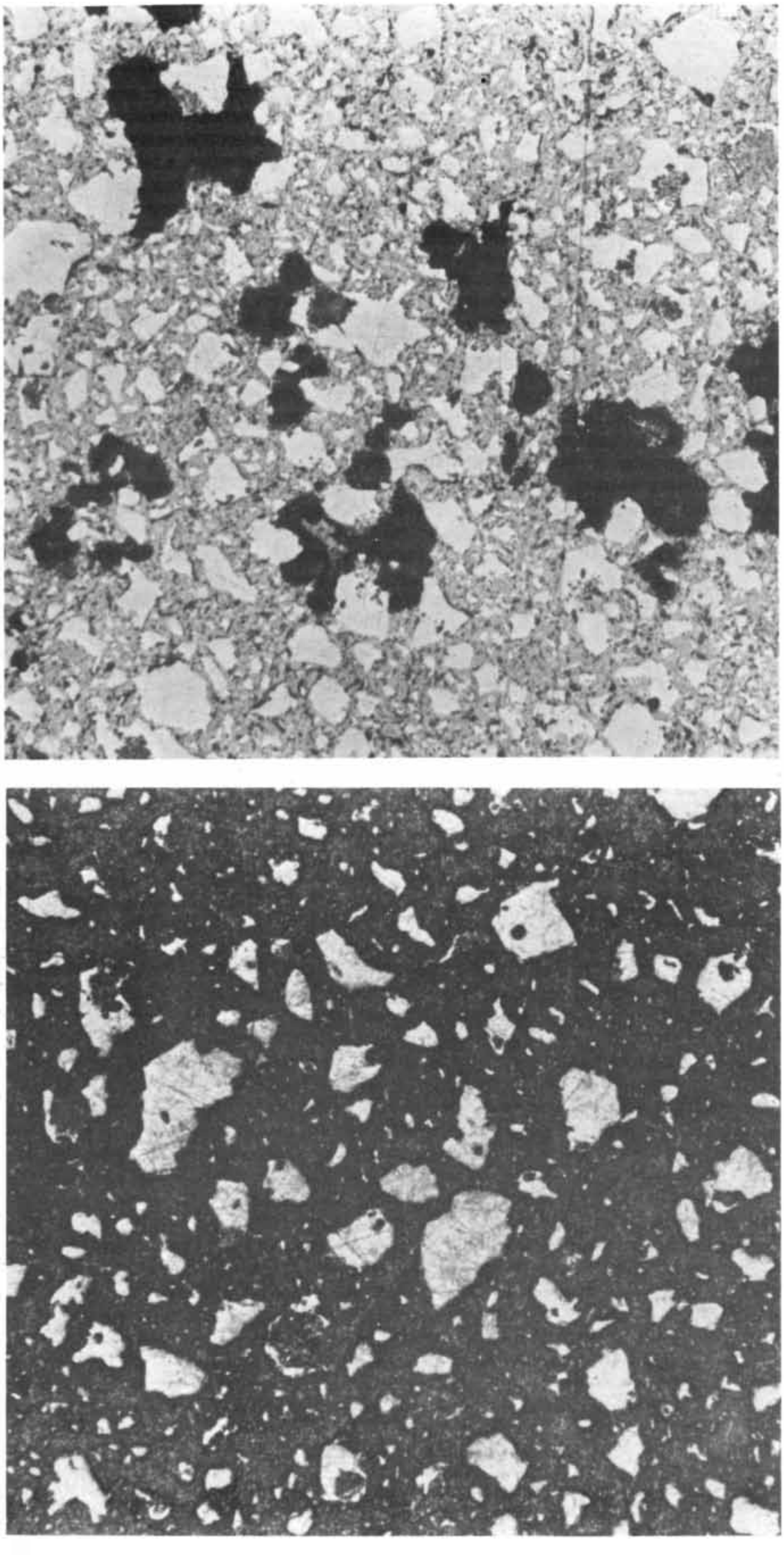
In the 1960's it was hoped that inorganic polymers would be synthesized in great variety and would find commercial application. Each inorganic polymer would be analogous to the organic polymers in being a long chain of repeating monomers, but the backbone of the chain would consist of atoms other than carbon. To date the only successful synthetic inorganic polymer that has emerged is silicone. In silicone the backbone consists of silicon atoms and oxygen atoms in alternation. Hence synthetic inorganic polymers have largely remained a dream. On the other hand, inorganic polymers taken fairly directly from nature are very widely used indeed; an example is glass. Moreover, the silicate rocks of the earth's crust consist of both chains and rings of silicon atoms and oxygen atoms in alternation. The problem is that the production of materials such as glass, as it is currently conceived, calls for high temperature, and high temperature defeats the objective of turning to inorganic materials in order to conserve energy.

There is a second problem. Consider the basic properties of materials that determine how the engineer incorporates them into a design. Three such properties are stiffness (the resistance of a material to bending), tensile strength (resistance to pulling) and fracture toughness (resistance to impact). They are properties any do-it-yourselfer would judge informally as he handled a material and decided whether or not to use it. Organic polymers and organic materials such as wood are not as stiff as inorganic materials such as ceramics and glass (one reason many synthetic organic materials are reinforced by incorporating into them fibers of inorganic material). The inorganic materials, however, lack toughness. If inorganic materials are to stand any chance of replacing plastics (and metals) in some applications, their toughness will have to be improved.

Can these two problems be solved? That is, can sufficiently tough inorganic materials be made at low temperature? The study of living organisms suggests they can. Living organisms exploit calcium carbonate, calcium phosphate and silica with remarkable control and sophistication. The abalone shell, for example, is 99 percent calcium carbonate; it is almost entirely chalk. Yet its tensile strength is more than 100 megapascals (14,500 pounds per square inch) and its



SPRING MADE OF CEMENT is shown under 300 pounds of tension (*top*) and under no tension (*bottom*). The cement is specially prepared to have no pores or other internal flaws larger than a few micrometers across. As a result the cement, which is called macro-defect-free, or MDF, cement, is highly resistant to fracture. Indeed, it is as strong as aluminum.



ABSENCE OF LARGE PORES in MDF cement is displayed in micrographs of a polished section of ordinary cement (*top*) and MDF cement (*bottom*). The pores in ordinary cement (*black patches*) can compose as much as 30 percent of the volume of the material. They are responsible for its lack of strength. The micrographs were made at an enlargement of 50 diameters.

toughness is more than 1,000 joules per square meter of new surface generated by fracturing, which makes it comparable in that respect to polystyrene and Plexiglas.

An examination of the shell's microscopic structure by means of electron microscopy shows how these properties are achieved. Flat plates of calcium carbonate, each about .2 micrometer thick, are stacked with impressive uniformity and are separated from each other by a thin layer of protein. A crack in the shell is thus forced to traverse a tortuous path and so the shell's fracture toughness is great. In a way that is not yet understood the protein serves the abalone as a template for the deposition of an ordered pattern of inorganic material at the temperature of the environment. The same is true of the growth of bone.

Let us examine in detail how a man-made inorganic material, say the ceramic porcelain, is created. The contrast with how the abalone fashions its shell is instructive. The basic process of making porcelain was discovered in China and was brought to the West in the early 18th century by Father F. X. d'Entrecolles, in what was probably the first recorded instance of industrial espionage. d'Entrecolles had spent time in Jingdezhen, a center of porcelain making in Jiangxi Province, and had learned Chinese in order to determine the role in porcelain making of the kind of clay the Chinese called kaolin. (To this day kaolin is one of the few Chinese words in the English language.)

High-quality ceramics such as porcelain are made by firing, that is, by heating to high temperature a mixture of quartz, feldspar and clays (usually kaolin and ball clay, a fine-grained clay that has some organic matter in it). To begin with, each of the powdered raw materials is mixed with water. The resulting pastes are pressed in a cloth to remove the excess water; then the pastes are combined and extruded through a screw under vacuum to remove most of the air in the mixture. The mixture is dried and shaped, much as a potter would shape it, and the resulting object is put into a gypsum mold to remove still more of the water in it. Next the object is coated with a glaze. The glaze is usually a clay combined with an additional flux: a material that readily fuses.

After further drying the object is fired. At temperatures of some 400 degrees Celsius a number of chemical reactions begin. Then at 600 degrees the material becomes glassy and crystals form in it, so that the material shrinks. The temperature is only about a third of the material's melting point, but that is deliberate: melting the material would give rise to large, irregular crystals rather than tiny, regular ones and in any case would re-

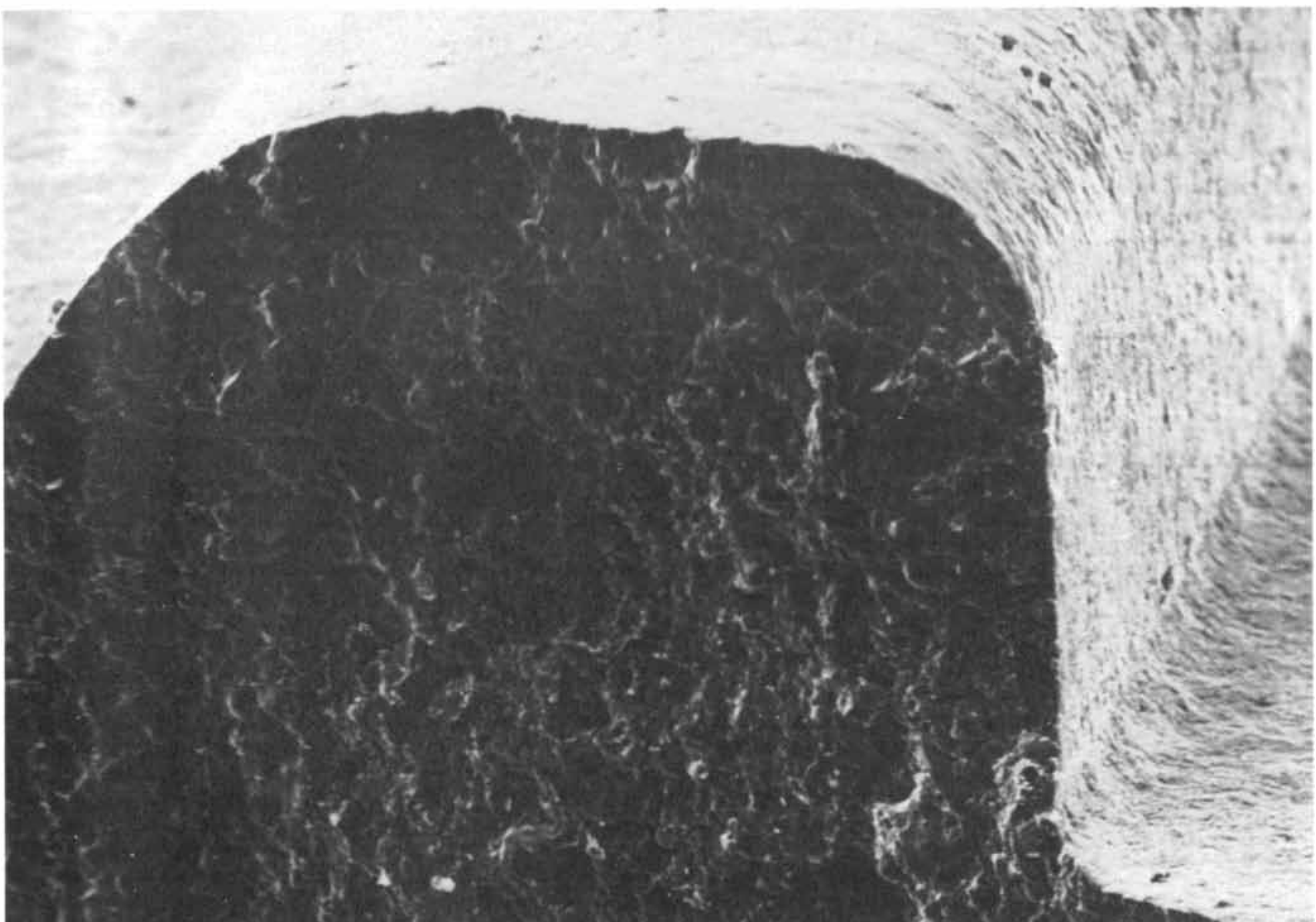
quire a large and costly expenditure of energy. On the other hand, atoms diffuse slowly in an unmelted material, so that when water evaporates, pores are left behind. Moreover, holes remain unfilled where clay particles fail to pack tightly. Further still, the individual crystals in the material are of varying composition and so are quite different in their mechanical properties. Notably they differ in their coefficient of thermal expansion: the degree to which heating makes them expand and cooling makes them contract. When the ceramic is cooled, therefore, some of the pores and holes may fill up but cracks and other blemishes develop. When the ceramic returns to room temperature, it is brittle and weak.

That is why all useful ceramic articles are coated with a glaze. The glaze is chosen to have a coefficient of expansion lower than that of the ceramic itself. Hence the cooling that follows the firing of the ceramic puts the glaze into compression. In effect the glaze holds the ceramic in a kind of vise, making it difficult for cracks to enter the ceramic from its surface and thereby lead to breakage.

Paradoxically, then, the high-temperature process of firing, which is intended to give the ceramic strength by allowing diffusion and shrinkage to fill up the gaps in the material, succeeds only at the expense of introducing cracks. High temperature, however, is not the only way to make atoms move so that they fill up gaps. Another way is to make water serve as a solvent for ions and as a medium for their diffusion. That is what happens in the hydraulic cements: solids that set and harden irreversibly in the presence of water. Gypsum is one example; portland cement is another.

When the calcium silicates in portland cement are mixed with water, some of the water takes part in the production of hydrated calcium silicates. Even today, 150 years after the invention of portland cement, the reactions are not well understood. The rest of the water is lost by evaporation, leaving pores in the cement. The pores vary in diameter from a few tens of angstrom units to more than a millimeter; together they can occupy 25 to 30 percent of the volume of the solid.

It has been known for many years that the strength of a cement, like the strength of a fired ceramic, depends on this porosity. The exact nature of the dependence has been elusive. First it was found that the strength of a cement depends on the ratio of water to solid when the cement is made. The strength increases as the proportion of water is reduced. Then it was proposed that the strength depends on the volume of the



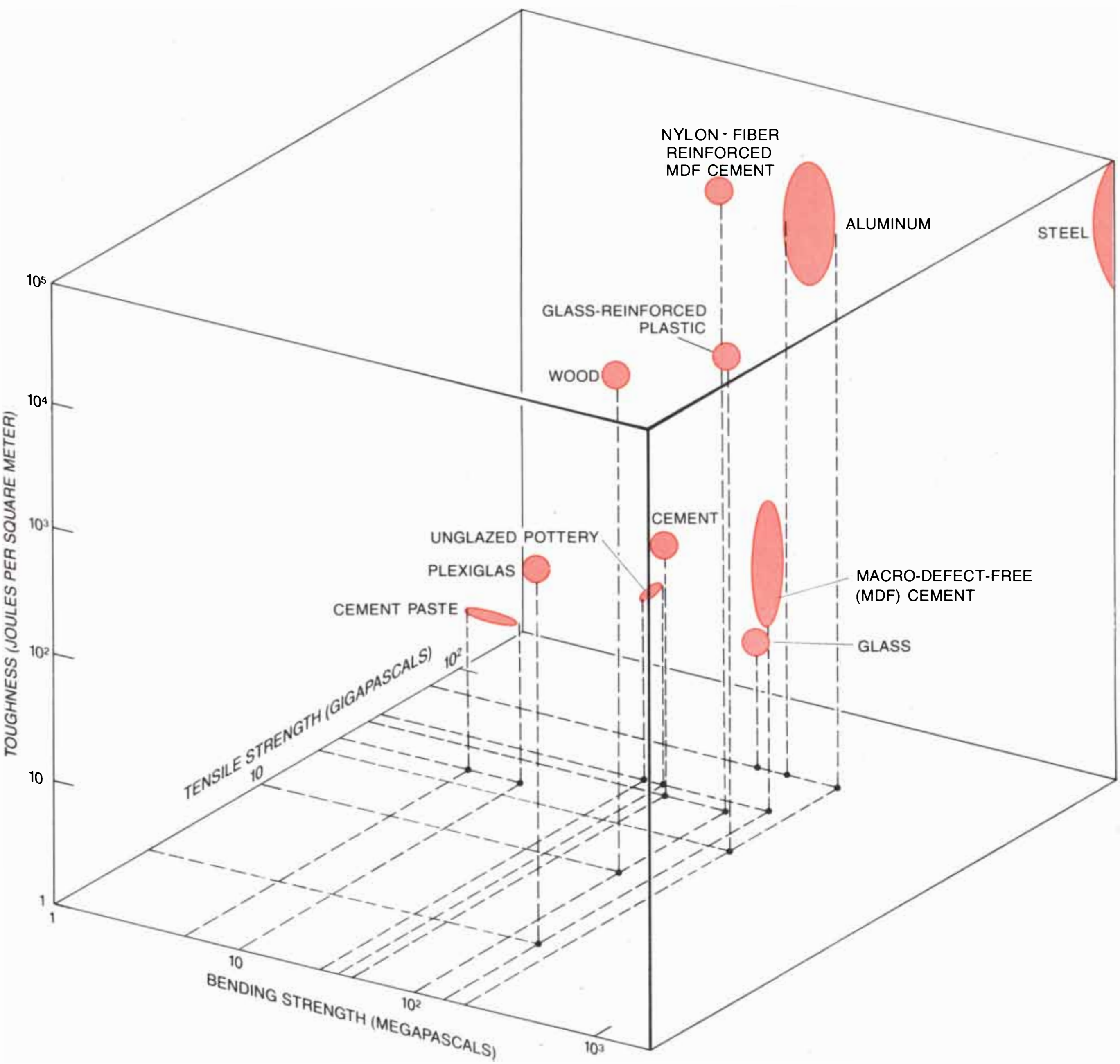
CEMENT BOTTLE CAP is an instance in which MDF cement replaces plastic. The cap has been fractured, and the fracture is shown crossing the thread of the cap in a scanning electron micrograph made at an enlargement of 75 diameters. The fracture is the dark area.

pores remaining in the cement after it has set. The strength increases as the porosity is reduced. The strength of a typical cement is actually rather low: its flexural strength is usually less than five megapascals and its tensile strength is only seven times higher. That is why cement and concrete are employed in construction only under compressive loads and why steel reinforcement is added when the material must bear a tensile load.

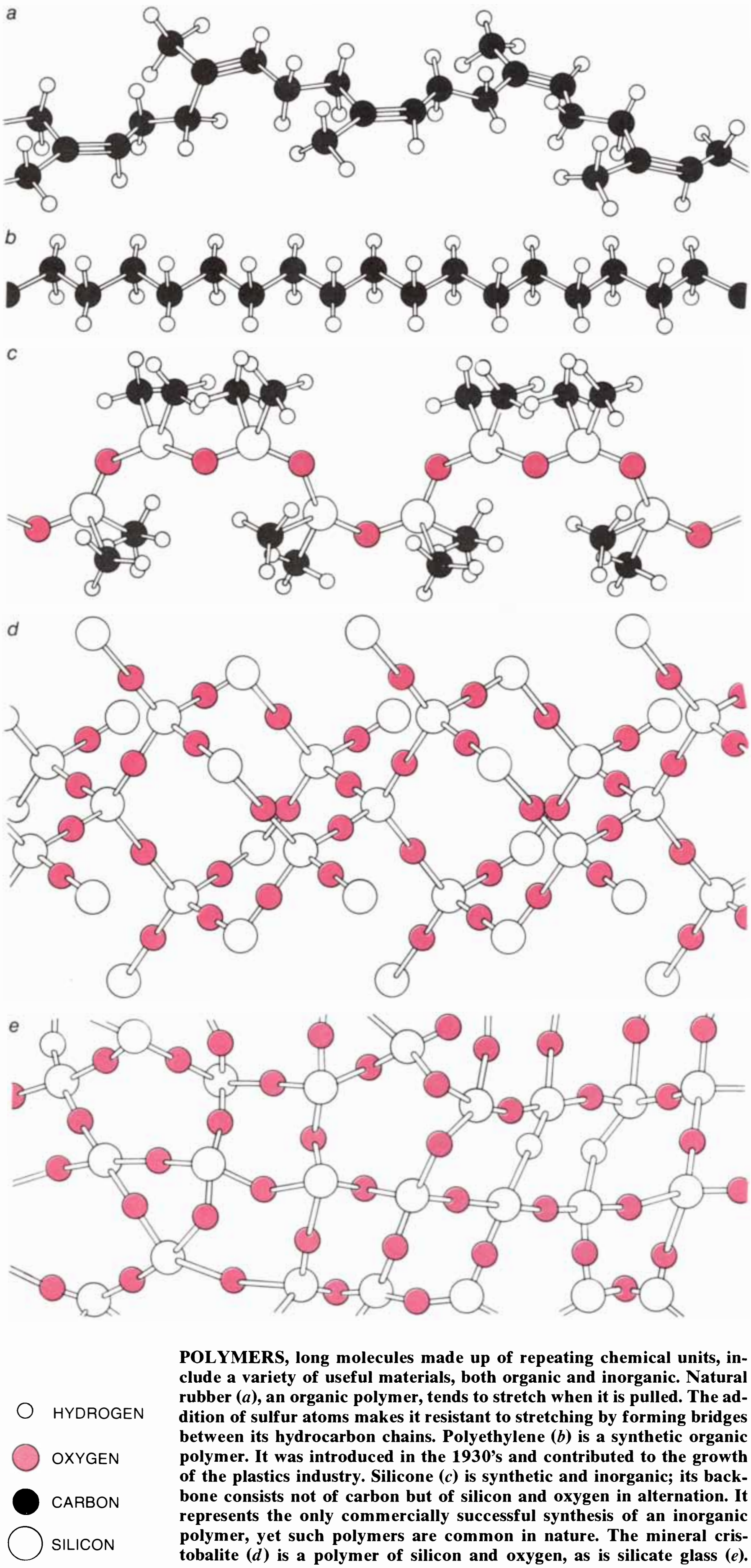
Suppose the relation between porosity and strength is extrapolated to predict the strength of a cement that has no pores. The result is a flexural strength of about 20 megapascals. The flexural strength of aluminum alloys is some five times as great, and the flexural strength of steel is 50 to 100 times as great. Even if the cement were reinforced with asbestos or glass fibers, one would predict a flexural strength of no more than 40 megapascals.

The extrapolation to zero porosity, however, is now known to be invalid. The simplest theory to explain the fracture of a brittle solid was developed in 1920 by Alan A. Griffith, who was then at the Royal Aircraft Establishment at Farnborough in England. He proposed that the tensile strength of a brittle solid is determined by the size of the largest flaw in the material. If the theory is correct, the strength of a cement should increase dramatically as the diameter of its largest pores is reduced. Conversely, the strength of the cement should depend only weakly on the number of pores or their volume.

That turns out to be the case. Anthony James Howard, Kevin Kendall and one of us (Birchall), working at the Imperial Chemical Industries (ICI) Research



THREE PROPERTIES of a material are its stiffness (resistance to bending), its tensile strength (resistance to pulling) and its fracture toughness (resistance to impact). Here the properties are plotted for various organic and inorganic materials. The scales are logarithmic. Organic materials such as wood rate high in toughness but low in stiffness. Inorganic materials such as glass are essentially the opposite. MDF cement rates almost as high as aluminum in all three properties. Stiffness and tensile strength are in pascals; one megapascal equals 145 pounds per square inch. Fracture toughness is the energy (in joules) that opens fractures adding up to one square meter of surface.



POLYMERS, long molecules made up of repeating chemical units, include a variety of useful materials, both organic and inorganic. Natural rubber (*a*), an organic polymer, tends to stretch when it is pulled. The addition of sulfur atoms makes it resistant to stretching by forming bridges between its hydrocarbon chains. Polyethylene (*b*) is a synthetic organic polymer. It was introduced in the 1930's and contributed to the growth of the plastics industry. Silicone (*c*) is synthetic and inorganic; its backbone consists not of carbon but of silicon and oxygen in alternation. It represents the only commercially successful synthesis of an inorganic polymer, yet such polymers are common in nature. The mineral cristobalite (*d*) is a polymer of silicon and oxygen, as is silicate glass (*e*).

Laboratories at Runcorn in England, have now shown that cement obeys Griffith's theory. Moreover, we have developed ways in which the maximum diameter of the pores in cement can be reduced from the usual millimeter or so to no more than a few micrometers. In essence the pores are kneaded out. In order to make the kneading thorough the cement grains are graded in size and a small quantity of a polymer is added as an aid to flow.

The new material has become known as macro-defect-free (MDF) cement. Tested in bending it shows a strength of more than 150 megapascals and in that respect is comparable to aluminum. Indeed, it is entirely possible to use cement to fabricate a spring.

The fracture toughness of MDF cement is also notable. Most common manmade inorganic materials have a fracture toughness of less than 10^2 joules per square meter of new surface generated by fracturing. In contrast the fracture toughness of wood is 10^4 joules per square meter across the grain. (It drops to 10^2 joules per square meter with the grain.) The mineral jade has a fracture toughness of 10^3 joules per square meter; that is why it can be carved delicately to shape. The fracture toughness of MDF cement is also about 10^3; on a conventional lathe, therefore, a block of it can be hollowed into a tube, a remarkable performance for an inorganic material.

Still, no cement is highly resistant to impact until it is reinforced with fibers. The technology of doing so has recently matured. In fact, strips of ordinary cement can now be made pliable. They can be bent like strips of metal. The bending does make cracks in the cement, but the cracks are quite narrow and the reinforcing fibers hold the cement together. Since the cement is made at low temperature, the fibers can be inexpensive organic ones that have a low melting point.

The new high-strength cement has the appearance of a ceramic; indeed, it would make good cups and saucers. It also has many of the characteristics of a ceramic. It is strong, hard, brittle and stiff, and it is formed largely by covalent chemical bonding. It is not, however, a refractory: a material that keeps its integrity at high temperature. The reason is that it contains water in chemical combination with its calcium silicates. At high temperature the water is driven off, leaving behind a much weaker solid. It would obviously be of value to devise methods of manufacturing refractory materials at low temperature.

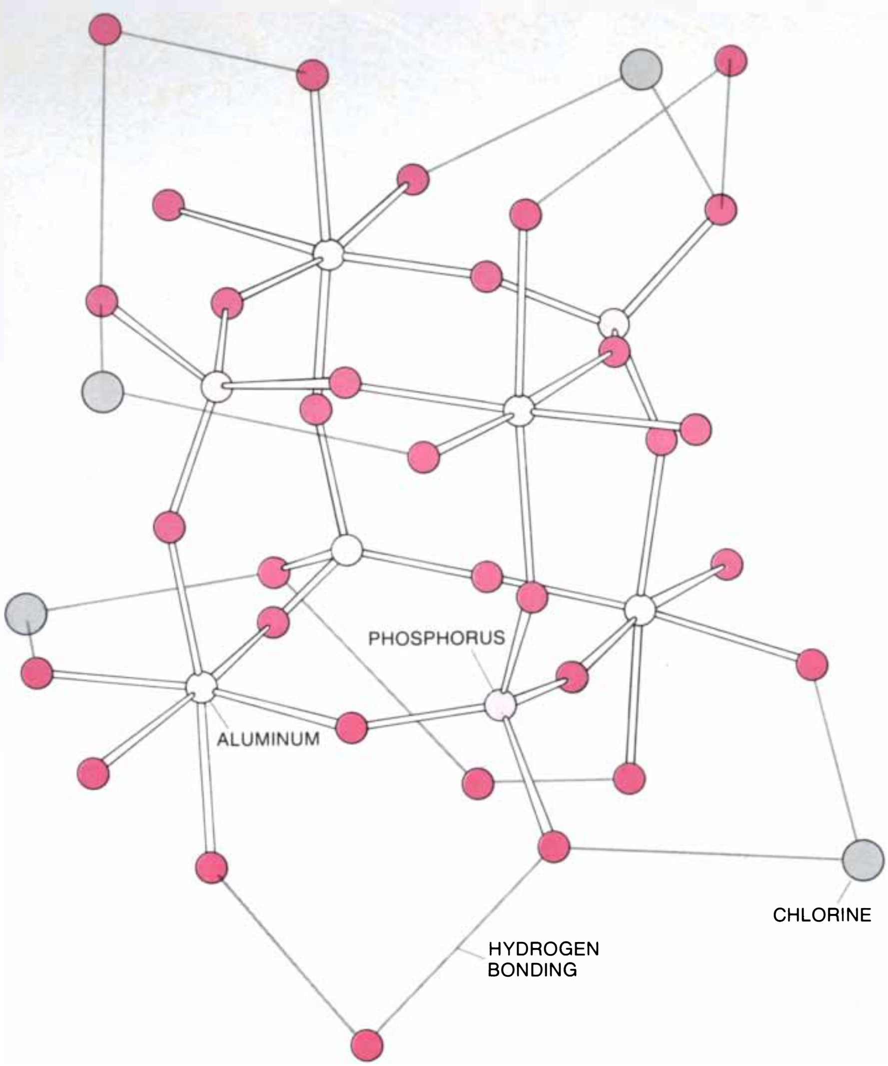
Two approaches are promising. We noted above that the silicate rocks and minerals in the earth's crust are polymers in which alternating oxygen atoms


Split fold-down rear seat-backs give you maximum flexibility in carrying people or cargo.
SL models offer 4-way adjustable seats with adjustable headrests for exceptional riding comfort.
Power steering, standard on all 200-SX models, makes the handling sure and easy.
The unique NAPS-Z engine meets your various driving demands with the smooth power and economy of fuel-injection.
Power-assisted disc brakes on four wheels provide fade-resistant stopping power.
DATSUN 200-SX
IT'S GOT PERFORMANCE DOWN TO A SCIENCE.
The sleek Datsun 200-SX is a scientific paradox. Although it burns gas "faster" than conventionally powered cars, the 200-SX gets even better gas-mileage ratings this year than last. And last year's ratings were already remarkably high for a car that won two dozen races in just nine months.
How? The 2.2-liter Nissan NAPS-Z engine. A microprocessor continuously adjusts the fuel/air mixture injected into the engine's crossflow hemihead combustion chambers and selects which of each cylinder's twin sparkplugs to fire, a unique concept in engine design.
Combustion in this "fast-burn" system is completed within 90° of crankshaft rotation. Ordinary engines don't complete combustion until 170°, wasting both power and fuel. The Nissan NAPS-Z saves both by burning gas "faster."
Most important, you can have this scientific advance without compromising looks or luxury features. This unique power plant delivers its high performance wrapped in a sculptured body that turns science into art. See and test drive the 200-SX at your Datsun dealer now.
38
EPA EST HWY
27
EPA EST MPG
Use EPA estimated mpg for comparison, with standard 5-speed. Actual mpg may differ depending on speed, trip length and weather. Highway mpg will probably be less.
NISSAN
WE ARE DRIVEN
DATSUN
NISSAN
© 1983 NISSAN MOTOR CORPORATION IN USA

and silicon atoms form chains and rings. In some of these minerals the chains and rings lie in silicate sheets some tens of angstroms thick. The sheets are separated from each other by layers of cations (positive ions) such as magnesium. One such mineral is vermiculite. George F. Walker of the Division of Applied Mineralogy of the Australian Commonwealth Scientific and Industrial Research Organization has discovered that the silicate sheets can be separated by removing or replacing the cations and then applying a shearing force. The result is a multitude of shards (fragments of the silicate sheets) thickly suspended in water. The suspension can be dried to make a strong, flexible, translucent and refractory film. Alternatively, the suspension can be whisked and then allowed to dry to make a refractory inorganic foam much like the organic polymer polystyrene. Here, then, is the complete re-creation of a natural mineral.

The second approach to the manufacture of refractory inorganic materials at low temperature is emerging from the study of solids such as glassy aluminum phosphate ($AlPO_4$). Aluminum phosphate is closely akin to silica, the archetypical inorganic refractory ceramic, and in fact it displays similar crystal forms. It is virtually impossible to produce glassy aluminum phosphate merely by melting the compound itself. The compound decomposes before it melts. On the other hand, a crystalline precipitate with the composition $AlPO_4 \cdot HCl \cdot (C_2H_5OH)_4$ can be prepared by reacting aluminum trichloride ($AlCl_3$) and phosphoric acid (H_3PO_4) in ethyl alcohol (C_2H_5OH) at temperatures lower than zero degrees C.

The structure of the precipitate is curious. It consists of aluminum, oxygen and phosphorus atoms in a cubic arrangement, but the cubes are prevented from linking by intervening molecules of ethyl alcohol. On gentle heating to about 100 degrees C. the alcohol is lost and the cubes do link. They form a three-dimensional lattice of aluminum, oxygen and phosphorus that is glassy and is refractory up to 1,600 degrees C., the temperature at which aluminum phosphate begins to decompose. The glass is inert and insoluble, but the precipitate from which the glass emerges is soluble both in water and in organic solvents. The precipitate can thus be dissolved and warmed. The process yields glassy coatings. Moreover, the precipitate can be used to bind together particles of compounds such as alumina (Al_2O_3) and so can serve the creation of a ceramic at a temperature of about 100 degrees C.

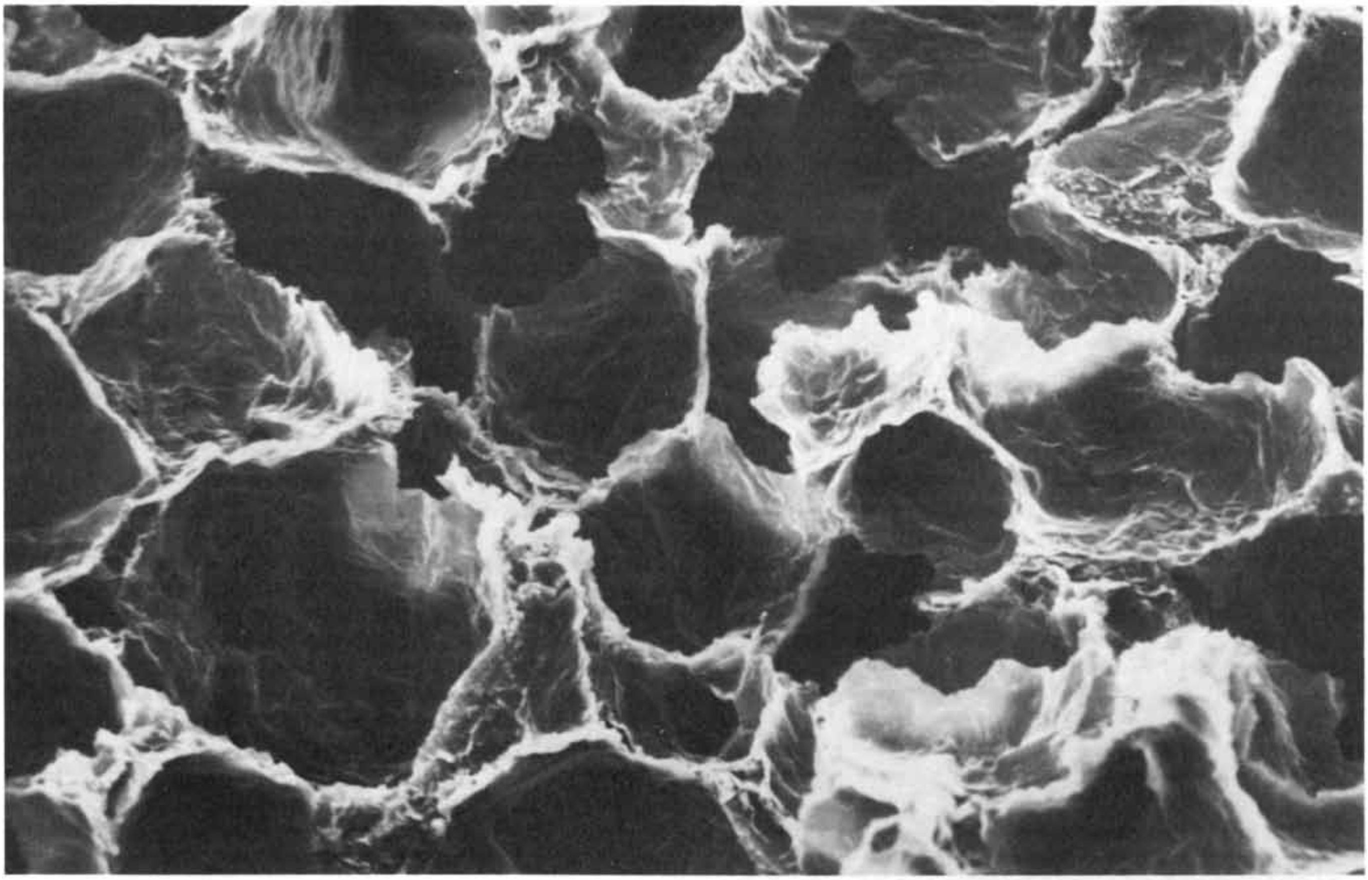


GLASSY ALUMINUM PHOSPHATE ($AlPO_4$) emerges from the reaction of aluminum trichloride ($AlCl_3$) with phosphoric acid (H_3PO_4) in ethyl alcohol (C_2H_5OH). The reaction itself yields the intermediate material shown here. Aluminum and phosphorus atoms form the vertexes of an approximate cube; oxygen atoms lie along its edges. The cubes are separated from one another by ethyl alcohol molecules. At a temperature of approximately 100 degrees Celsius the alcohol is lost and the cubes coalesce into a glass. The structure of the intermediate material was determined by John E. Cassidy and his colleagues at Imperial Chemical Industries.

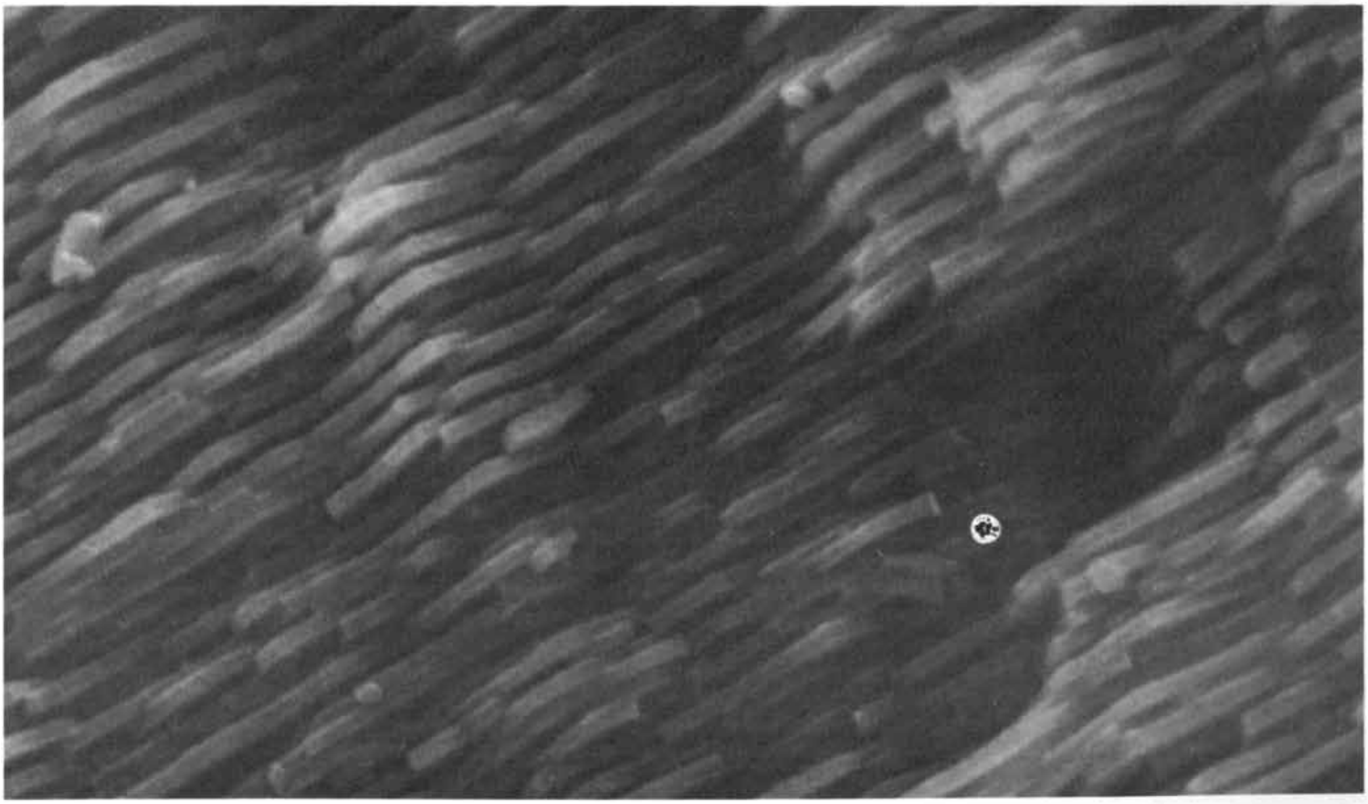
In related techniques developed by B. E. Yoldas of the Westinghouse Research Laboratories and by Masayuki Yamane and Tesuo Sakaino of the Tokyo Institute of Technology glasses are made at low temperature from alkoxides: compounds that emerge from the reaction between an alcohol and an inorganic acid. Methyl alcohol (CH_3OH), for example, reacts with silicic acid ($Si(OH)_4$) to make silicon alkoxide ($Si(OCH_3)_4$) plus water. If the alkoxide is reacted with water, it is transformed into an oxide and an alcohol. The alcohol can be removed by gentle heating, leaving a glassy oxide behind. In principle the glass could be reinforced with organic fibers that have a low melting point, a strategy that is not possible in glass made conventionally at high temperature.

Like the cement described above, the glass produced at low temperature is porous. Here again, however, the pores are small, and so they have little effect on many physical properties. In making glass one often aims to control not only the mechanical properties of the material but also the optical properties. In particular the transparency of a glass requires that no pores be as large as a wavelength of the light the glass is intended to transmit. Many of the glasses produced by Yoldas are splendidly transparent; they contain no pores larger in diameter than 100 angstroms.

In the ways we have described, inorganic chemistry is preparing to contribute to the conservation of energy and of hydrocarbons by manipulating inorganic chemical compounds at low temperature. This manipulation is being aided by advances in the understanding of material properties such as porosity. Interestingly, the importance of the control of porosity had not emerged from all the previous work on high-temperature methods of fabrication. It may be too early to speak of a new Neolithic age, but its beginnings are clearly with us.



FOAM MADE FROM VERMICULITE is a newly devised inorganic material that can supersede organic polymers such as polystyrene plastic. In vermiculite, silicate sheets are interleaved by layers of magnesium ions. The ions are removed and the sheets are sheared. The resulting slurry is whisked and dried to produce the foam. This scanning electron micrograph was made at an enlargement of 75 diameters.



ABALONE SHELL, seen in a scanning electron micrograph, is almost entirely chalk (calcium carbonate, or $CaCO_3$), yet it is stiffer than aluminum and as tough as Plexiglas. It confirms that inorganic materials can be designed to have new and useful properties. The shell consists of calcium carbonate plates that are .2 micrometer thick and are separated from each other by a thin layer of protein.

Computer-Intensive Methods in Statistics

They replace standard assumptions about data with massive calculations. One method, the "bootstrap," has revised many previous estimates of the reliability of scientific inferences

by Persi Diaconis and Bradley Efron

Most statistical methods in common use today were developed between 1800 and 1930, when computation was slow and expensive. Now computation is fast and cheap; the difference is measured in multiples of a million. In the past few years there has been a surge in the development of new statistical theories and methods that take advantage of the high-speed digital computer. The new methods are fantastic computational spendthrifts; they can easily expend a million arithmetic operations on the analysis of 15 data points. The payoff for such intensive computation is freedom from two limiting factors that have dominated statistical theory since its beginnings: the assumption that the data conform to a bell-shaped curve and the need to focus on statistical measures whose theoretical properties can be analyzed mathematically.

These developments have profound implications throughout science, because statistical theory addresses a grand question: How is one to learn what is true? Suppose 15 measurements of some quantity yield 15 moderately different values. What is the best estimate of the true value? The methods of statistics can answer such a question and can even give a quantitative indication of the estimate's reliability. Because empirical observations are almost always prone to error, conclusions in the sciences (and in many other fields) must often be based on statistical measures of truth. As a result any development that makes statistical inferences more accurate or more versatile can be expected to have broad consequences.

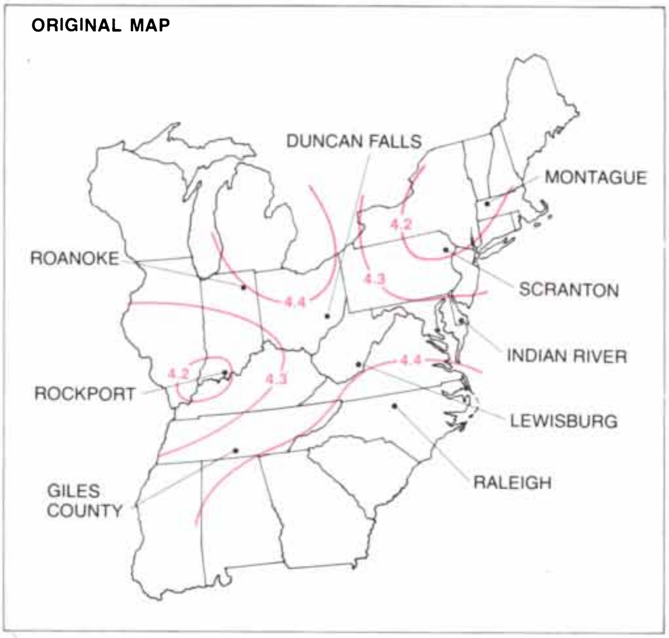
The two advantages of the new methods are best appreciated by comparing them with older ones. First, in older methods it was generally necessary to make certain unverifiable assumptions about the data before statistical analysis could proceed. The assumptions often involved the bell-shaped curve, which is also called the normal or Gaussian distribution, after the German mathematician Carl Friedrich Gauss. When the Gaussian distribution is employed, it is assumed that random fluctuations, or errors, in the experimentally observed values of some quantity are scattered symmetrically about the true value of the quantity. Moreover, it is assumed that the greater the error between the experimental value and the true value is, the less likely it is that the experimental value will be observed. Experience has shown that Gaussian theory works quite well even when the Gaussian distribution is only roughly approximated by the data, which is why statisticians can give reliable predictions even without computers. For sets of data that do not satisfy the Gaussian assumptions, however, the results of statistical methods based on such assumptions are obviously less reliable. Computer-intensive methods can solve most problems without assuming that the data have a Gaussian distribution.

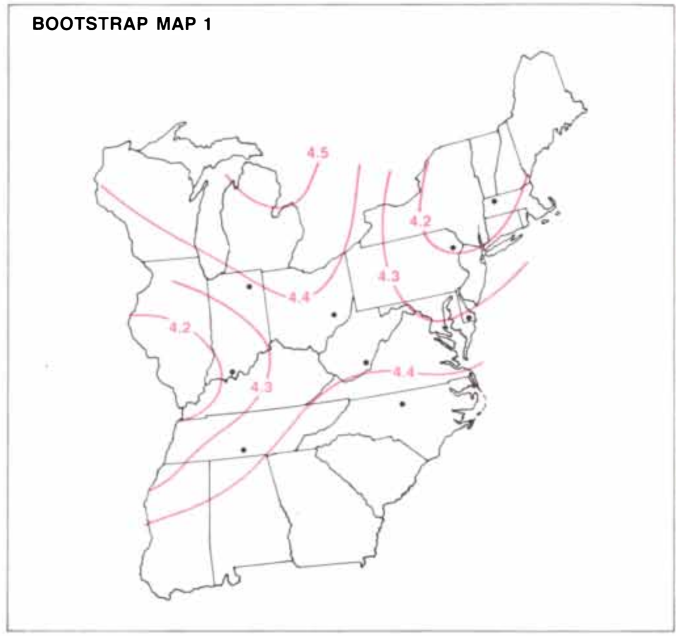
Freedom from the reliance on Gaussian assumptions is a signal development in statistics, but the second advantage of the new techniques is even more liberating. In older practice the arithmetic operations associated with statistical analysis had to be done by hand or with the aid of a desk calculator. Such calculations can often be simplified immensely if the formulas on which the calculations are based have a concise analytical form. Hence statistical theory tended to focus on only a few properties of a statistical sample, such as the mean, the standard deviation and the correlation coefficient, that can easily be manipulated analytically. Many other properties of a sample, however, are of interest to the statistician but are beyond the reach of exact mathematical analysis. The new computer-based methods make it possible to explore such properties numerically, even though their exact analysis is currently impossible. Thus the new methods free the statistician to attack more complicated problems, exploiting a wider array of statistical tools.

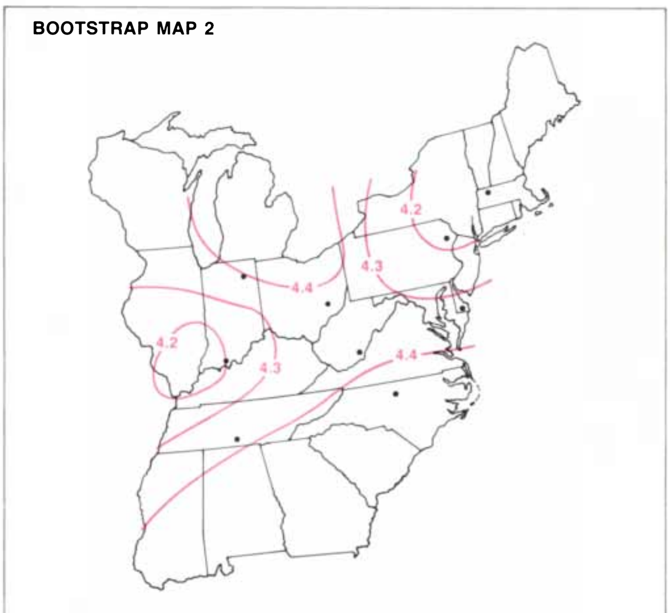
To illustrate how the computer has been applied to statistical inference we have chosen a problem in which only 15 data points appear. We shall apply a method called the bootstrap, invented by one of us (Efron) in 1977, which is quite simple to describe but is so dependent on the computer that it would have been unthinkable 30 years ago.

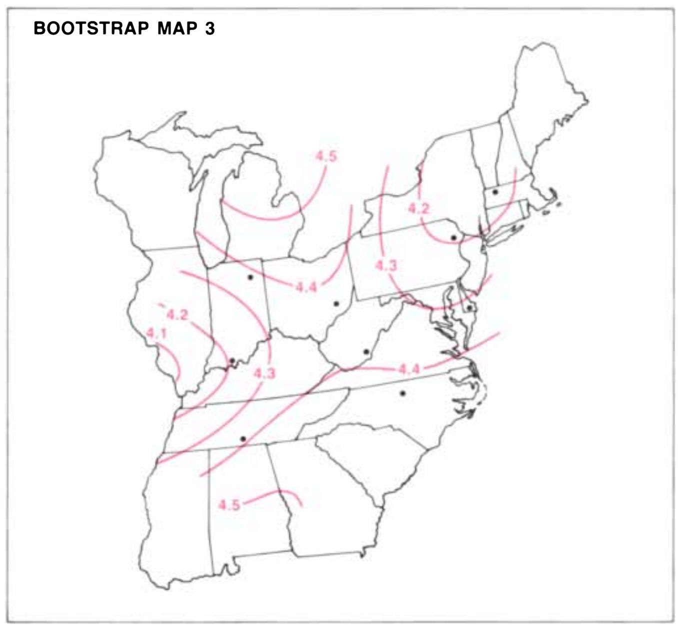
Consider a group of 15 law schools for which two overall characteristics of each entering freshman class are measured: the average undergraduate grade-point average (GPA) and the average

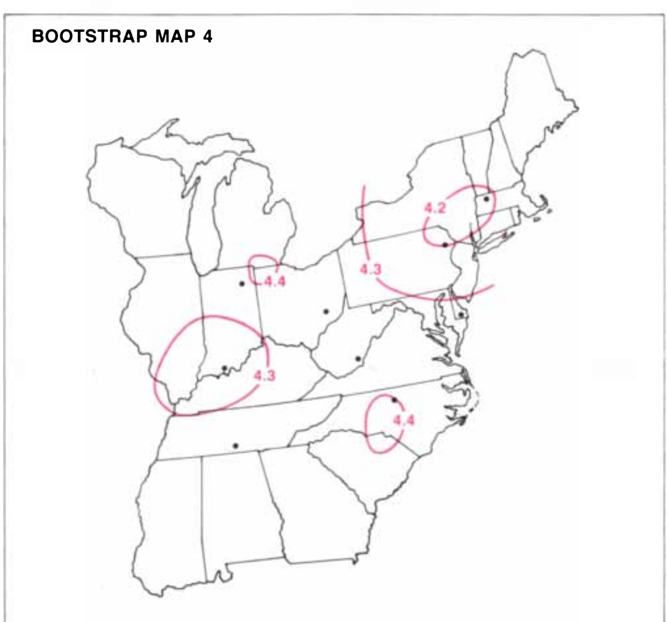
LARGE VARIABILITY of contour lines on a map is revealed by the statistical method called the bootstrap; the method requires so many numerical calculations that it is feasible only with the aid of a computer. The map at the upper left was constructed from 2,000 measurements of the acidity, or *p*H value, of every rainfall recorded at nine weather stations over a period of two years. (The lower the value of the *p*H, the greater the acidity.) The contours were drawn according to a procedure that can be proved optimal under certain conditions. Nevertheless, the 2,000 data points are subject to considerable random variability: contours based on another sample of 2,000 measurements for the same region and time period might have looked quite different. The bootstrap, which was invented by one of the authors (Efron), can estimate from the single set of 2,000 data points the amount of variability the contours would show if many sets of 2,000 data points could be compared. The results of five bootstrap calculations, done with the aid of a computer by Barry P. Eynon and Paul Switzer of Stanford University, are shown in the other five maps. The variability of the contours shows that the original map must be interpreted cautiously: corridors of low acidity on the original map can become islands on subsequent maps.

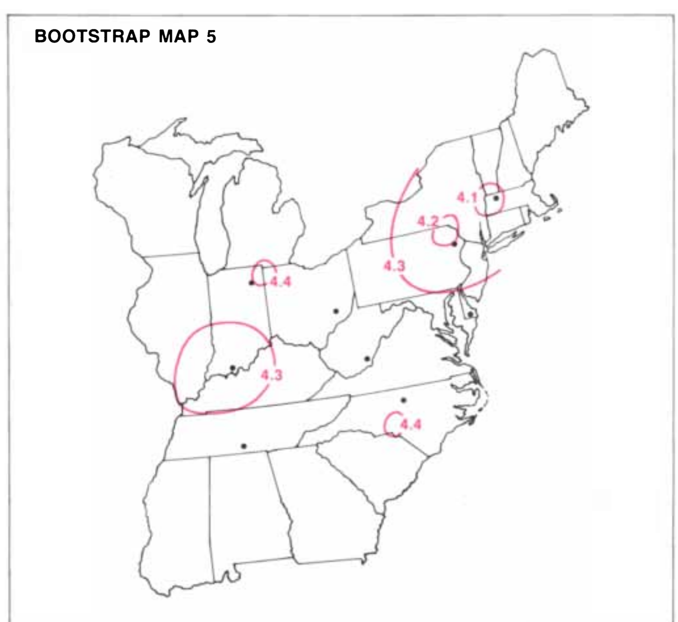

ORIGINAL MAP
DUNCAN FALLS
MONTAGUE
ROANOKE
SCRANTON
INDIAN RIVER
ROCKPORT
LEWISBURG
GILES
COUNTY
RALEIGH
4.2
4.3
4.4
4.2
4.3
4.4

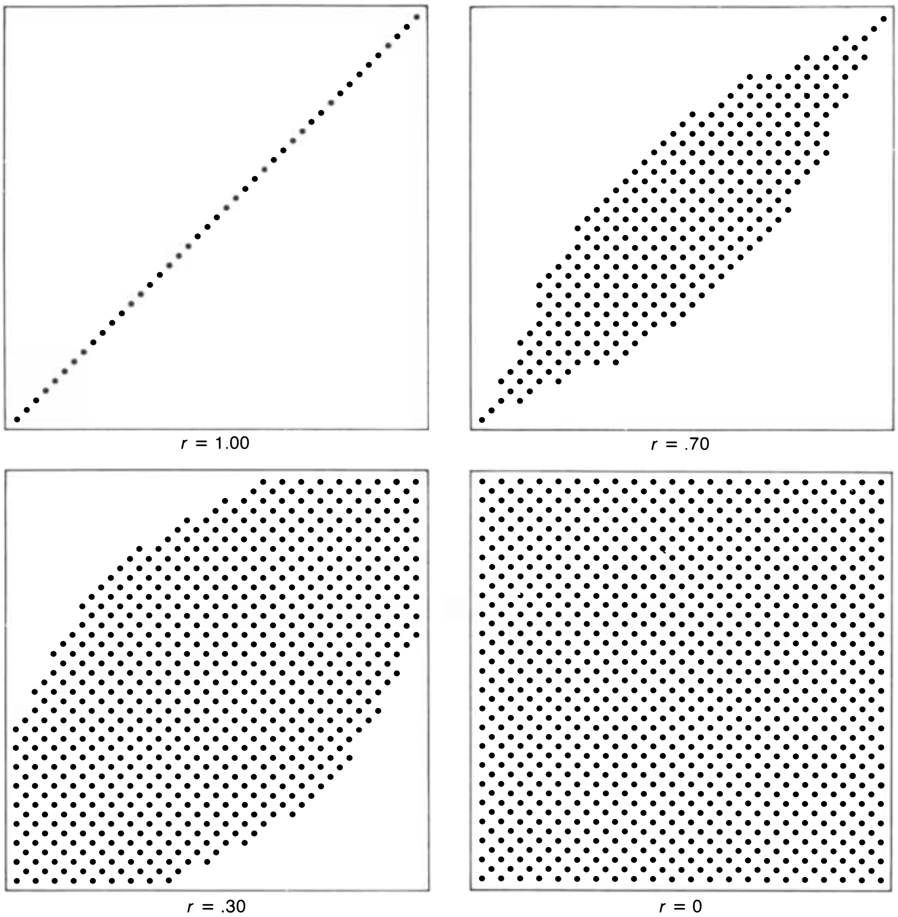

BOOTSTRAP MAP 1
4.5
4.2
4.3
4.4
4.2
4.3
4.4


BOOTSTRAP MAP 2
4.2
4.3
4.4
4.2
4.3
4.4

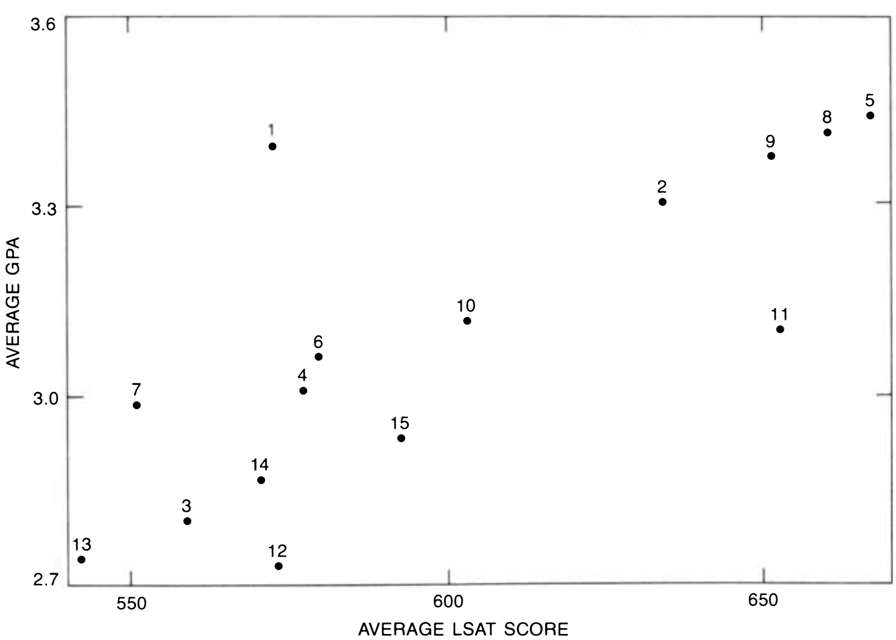

BOOTSTRAP MAP 3
4.5
4.2
4.3
4.4
4.2
4.1
4.3
4.4
4.5


BOOTSTRAP MAP 4
4.2
4.3
4.4
4.3
4.4


BOOTSTRAP MAP 5
4.1
4.2
4.3
4.4
4.3
4.4



CORRELATION COEFFICIENT is a measure of the tendency of data points plotted on a graph to cluster about a line. The coefficient is usually designated by the letter *r* and can have any value between 1 and −1. The more linear the clustering, the greater the absolute value of *r*.



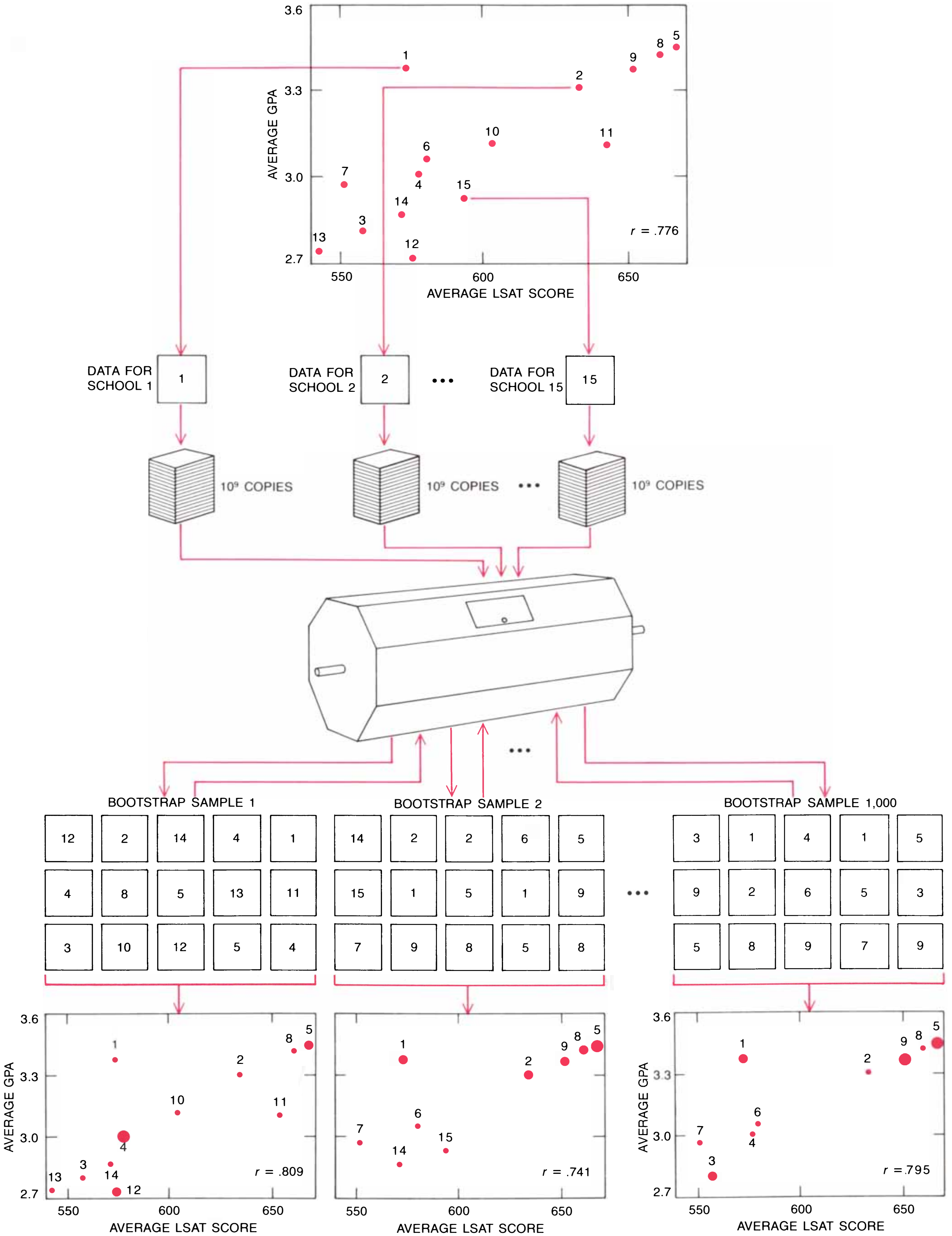
TWO MEASURES for the potential academic pertormance of the students in the entering classes of 15 American law schools are graphed for each school. Each point on the graph represents the undergraduate grade-point averages (GPA) and the scores on the Law School Admission Test (LSAT), averaged over all the students in one entering class. The graph shows that for the sample of 15 law schools the two measures tend to be proportional: their correlation coefficient *r* is .776. One would like to know how accurately .776 approximates the true value of *r* for all American law schools. That is, one would like to know how much, on the average, the observed value of *r* for a random sample of 15 law schools differs from the true value of *r*.

score on the Law School Admission Test (LSAT). It seems reasonable to suppose the two measures ares roughly proportional to each other: the entering classes with a high average GPA tend to have a high average score on the LSAT. It is unlikely, however, that the proportionality is exact: the entering classes of one or two law schools may show a high average GPA but a low average LSAT, whereas a few other schools may have a low average GPA but a high average LSAT. The statistician wants to know first of all how close the relation between the two measures is to proportionality. Moreover, the statistician must try to estimate the degree to which the available data justify the extrapolation of the first result to all other law schools. In short, how confident can one be that the sample of 15 law schools gives an accurate picture of the population of law schools as a whole?

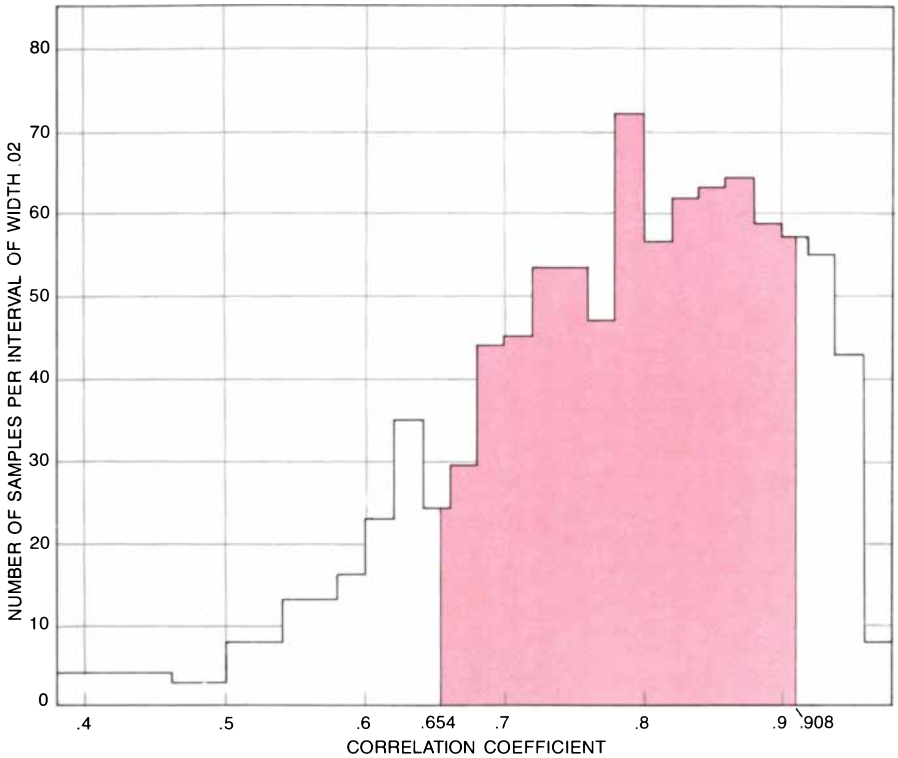
The standard measure of the tendency toward proportionality between two variables such as GPA and LSAT is the correlation coefficient; it is usually designated by the letter *r*. Suppose the data for the law schools are plotted on a graph where the vertical axis represents GPA and the horizontal axis represents LSAT. The correlation coefficient is a measure of the degree to which the points on such a graph tend to cluster along a line. The value of *r* is 0 if the points are scattered at random and gets increasingly close to 1 or −1 as the points tend to cluster along a line of positive or negative slope. (The slope of a line is positive if the line slopes up and to the right, and the slope is negative if the line slopes down and to the right.) The correlation between degrees Fahrenheit and degrees Celsius, for example, is 1 because the two variables are directly proportional to each other. The correlation between the height of fathers and the height of their sons is about .5. Tall fathers tend to have tall sons, but the correspondence is not exact. The correlation between daily consumption of cigarettes and life expectancy has been shown to be negative; that is, the greater the daily consumption of cigarettes, the shorter the life expectancy.

The observed correlation between GPA and LSAT for the 1973 entering classes of the 15 law schools is .776. In other words, there is a strong correlation observed between the two variables and a strong tendency for the points defined by the coordinates of each school to cluster along a line of positive slope. A straightforward mathematical procedure, which takes about five minutes with a desk calculator, was applied to determine the value of *r*: the details of the calculation are not important except that they give a well-defined value of *r* for any collection of data points.

What grounds does one have, how-



BOOTSTRAP METHOD is applied to the sample of 15 law schools shown in the illustration at the bottom of the opposite page in order to assess the accuracy of the correlation coefficient calculated for the sample. The data for each law school are copied perhaps a billion times and all 15 billion copies are thoroughly shuffled. Artificial samples of 15 law schools, called bootstrap samples, are created by selecting sets of 15 data points at random from the 15 billion copies. The value of r is then calculated for each of the bootstrap samples. Although it is simple in concept, the application of the bootstrap requires so many calculations that it is not feasible without a computer.



FREQUENCY DISTRIBUTION of the correlation coefficient *r* is plotted for 1,000 bootstrap samples. A widely accepted measure of the accuracy of a statistical estimator such as *r* is the width of the strip under the central part of its frequency distribution whose area is 68 percent of the area under the entire distribution. The central strip for the bootstrap distribution is shaded in color; its width is .254. Half of the width of the interval, .127, is a good estimate of the average amount by which the observed value of *r* for a sample differs from the true value of *r*.

ever, for believing the true value of *r* is close to .776 for all law schools? After all, the sample could be highly atypical of law schools in general. The law of large numbers guarantees that in large samples the statistical estimate of *r* calculated for the sample is very likely to approach the true value of *r* for the entire population. A sample of only 15 law schools, however, is not a large sample. Hence some measure is needed that can assess the statistical accuracy of the value of *r* given by the sample, namely .776. The bootstrap is intended to provide such a measure.

To understand what statistical accuracy means for an estimate such as *r*, suppose data were available for additional sets of 15 law schools, different from the sets already sampled. For each set of 15 law schools the value of *r* could be calculated, and the amount of variation in the values of *r* for many samples could thereby be described. For example, if 99 percent of the values of *r* calculated for the hypothetical samples were between .775 and .777, one would assign high accuracy to the estimate .776. On the other hand, if the values of *r* were spread out evenly from −1 to 1, the estimate of *r* given by the original sample would have no accuracy and would therefore be useless. In other words, the statistical accuracy of an estimated value of *r* depends on the width of the interval bracketing the estimated value that is associated with a certain percentage of all the samples. Unfortunately the data needed to calculate the value of *r* for many different samples are generally lacking. Thus, because the law school example is intended to reflect the conditions of real statistical practice, we shall assume for the moment that the only data available are those for the original sample of 15 law schools. Indeed, if more data were available, they could be employed to give a better estimate for the value of *r* than .776.

The bootstrap procedure is a means of estimating the statistical accuracy of *r* from the data in a single sample. The idea is to mimic the process of selecting many samples of size 15 in order to find the probability that the values of their correlation coefficients fall within various intervals. The samples are generated from the data in the original sample. The name bootstrap, which is derived from the old saying about pulling yourself up by your own bootstraps, reflects the fact that the one available sample gives rise to many others.

In effect, the bootstrap samples are generated as follows. The data for the first school are copied an enormous number of times, say a billion, and the data for each of the other 14 schools are copied an equal number of times. The resulting 15 billion copies are thoroughly mixed. Samples of size 15 are then selected at random and the correlation coefficient is calculated for each sample. On a computer the steps of copying, mixing and selecting new sets of data are all carried out by a procedure that is much faster but mathematically equivalent: the computer assigns a number to each law school and then generates the samples by matching a string of random numbers to the numbers that correspond to the law schools.

The samples generated in this way are called bootstrap samples. The distribution of the correlation coefficients for the bootstrap samples can be treated as if it were a distribution constructed from real samples: it gives an estimate of the statistical accuracy of the value of *r* that was calculated for the original sample. We generated 1,000 bootstrap samples from the data for the 15 law schools in our sample. Of the 1,000 samples there were 680, or 68 percent, whose correlation coefficients were between .654 and .908. The width of this interval, .254, is the bootstrap measure of accuracy of the value of *r* for the sample. Half of the width of the interval, .127, can be interpreted as the bootstrap estimate of the average amount by which the observed value of *r* for a random sample of size 15 differs from the true value of *r*.

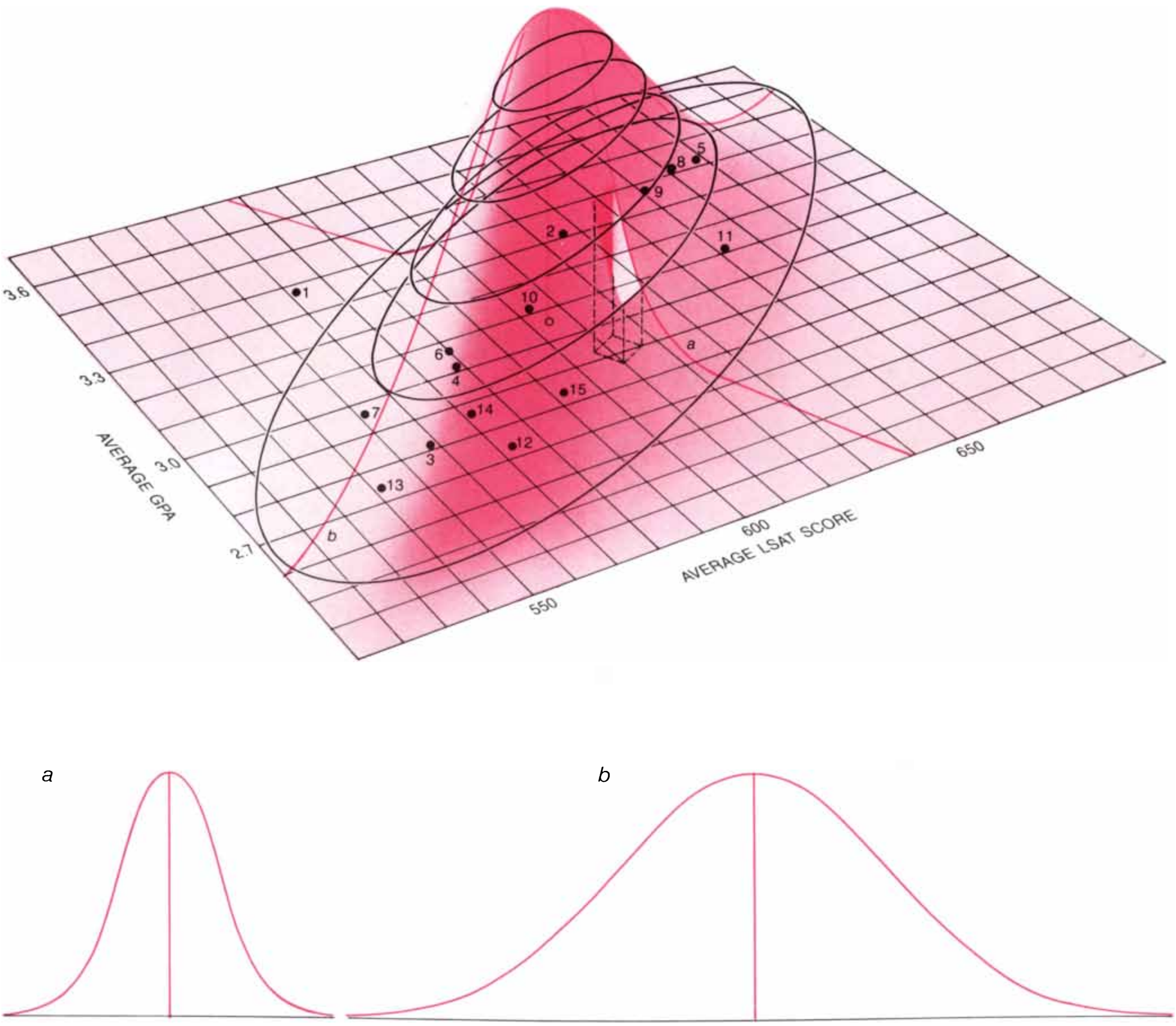
It is worth noting that the statistical accuracy cannot be defined simply as the accuracy of an individual estimate such as .776, that is, the difference between the estimate and the true value of *r*. In a real problem this difference can never be known; if it were known, the problem would vanish, because one could subtract the difference from the estimate and so obtain the true value exactly. Instead the statistical accuracy refers, as we have indicated, to the average magnitude of the deviation of the estimate from the true value.

If the results of the bootstrap distribution can be taken as a measure of the statistical accuracy of the original estimate, then .776 is a rough estimate but not entirely worthless. The true correlation coefficient, that is, the value of *r* for the population of law schools as a whole, could well be .6 or .9, but it is almost certainly not zero. Our theoretical work shows that the bootstrap measure of statistical accuracy is dependable in a wide variety of situations.

We can now abandon our self-imposed ignorance because in the law school example the accuracy of the estimated correlation coefficient can be tested directly. Indeed, we chose the example because all the data for average GPA and average LSAT scores of American law school students in 1973 have already been gathered. There were 82 American law schools in 1973, and the correlation between GPA and LSAT for all the schools was .761. (Hence .761 is the true value of r we referred to above, a quantity that would not be known in most situations.) More important, the true statistical accuracy of the estimate given by the original sample can be calculated, because the distribution of the values of r for many real samples of size 15 can be determined. Samples of size 15 can be chosen at random from the 82 law schools in 82^{15}, or about 5×10^{28}, equally likely ways. In principle the value of r could be calculated for each sample, and so the number of samples for which r lies within a small interval could be plotted for intervals of equal size. The resulting graph is called a frequency distribution.

In practice the frequency distribution for samples of size 15 selected from the 82 law schools can only be approximated; a computer set to work at the beginning of the big bang calculating r for each of the 82^{15} samples at a rate of a billion a second would still not have finished the task. Instead r is calculated for a large but manageable number of randomly selected samples of size 15, say a million samples.

We found that 68 percent of the correlation coefficients for a million samples were grouped between .606 and .876, an interval whose width is .270. In other words, if a sample of 15 law schools is selected at random, the probability that its correlation coefficient lies between .606 and .876 is .68. Note that the width of the interval is in good agreement with that defined by 68 percent of



BELL-SHAPED SURFACE was employed in 1915 by Sir Ronald Fisher in his method for estimating from a single sample how much the correlation coefficient varies from sample to sample. In order to make such an estimate by Fisher's method it is necessary to assume that all the data points in the sample are selected according to probabilities given by the bell-shaped surface. The surface is constructed to fit the data in the sample. In the law school example the highest point of the surface must lie directly over the point on the plane where the GPA and the LSAT points both have their overall average values (*open circle*). The slope and orientation of the surface with respect to the plane of the graph depend on how the data points are scattered. The contours of equal height on the surface are elliptical, and the cross sections are bell-shaped curves of varying width; two cross sections are shown in the lower part of the illustration. Fisher's method can be interpreted as choosing bootstrap samples from among all the points on the plane of the graph. The probability of choosing a point from within a given region on the graph is equal to the volume that lies between that region and the bell-shaped surface (*volume of "hole"*) divided by the entire volume that lies between the surface and the graph. By carrying out the bootstrap sampling with a computer only for the discrete points in the original sample the probability distribution that is given by the bell-shaped surface need not be assumed.

the bootstrap samples, even though the endpoints of the intervals coincide only roughly.

It turns out that the agreement is no accident. Theoretical investigations done by Rudolph J. Beran, Peter J. Bickel and David A. Freedman of the University of California at Berkeley, by Kesar Singh of Rutgers University and by us at Stanford University show that for the correlation coefficient and for a wide variety of other statistics the interval associated with the bootstrap distribution and the interval associated with the distribution of the real samples usually have nearly the same width. (Intervals that include 68 percent of the samples are commonly cited for comparison because for a bell-shaped curve 68 percent of the samples lie within one standard deviation of the peak of the bell.)

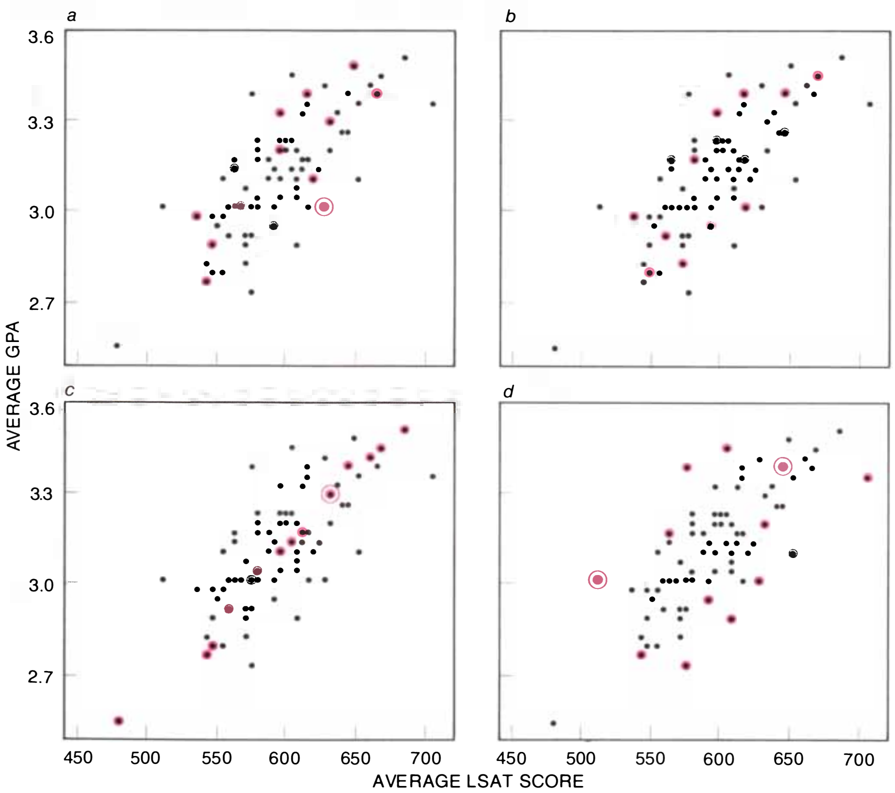
At first glance this theoretical result seems paradoxical: it suggests that from the information in each sample one can derive a good approximation to the frequency distribution of the correlation coefficient for all real samples of the same size. It is as if statisticians had discovered the statistical analogue of the hologram, a pattern of light waves that is preserved on a surface. The scene from which the light waves are emitted can be reconstructed in great detail from the whole surface of a hologram, but if pieces of the surface are broken off the entire scene can still be reconstructed from each piece. Not every sample is like a broken piece of a hologram, however; the good properties of the bootstrap are good average properties. Like any other statistical procedure, the bootstrap will give misleading answers for a small percentage of the possible samples.

Suppose the correlation coefficient for the sample of 15 law schools had been nearly 1. That is, suppose all the data points in the sample lay almost perfectly along a straight line. The circumstance is extremely unlikely, given the real data for the 82 law schools, but it could happen. It would then follow that every sample generated by the bootstrap procedure would also lie along the same straight line and so every bootstrap value of r would be nearly equal to 1. The width of the interval associated with 68 percent of the bootstrap samples would therefore be approximately zero. According to the bootstrap procedure, the statistical accuracy of the estimated value of r would be almost perfect, which is incorrect.

The bootstrap does not always guarantee a true picture of the statistical accuracy of a sample estimate. What has been proved is that the bootstrap gives a good picture of the accuracy of the estimate most of the time. There are always a few samples for which the bootstrap does not work, and one cannot know in advance which they are. The limitation is not so much a failure of the bootstrap procedure as it is a restatement of the conditions of uncertainty under which all statistical analyses must proceed.

What are the advantages of applying the bootstrap? In order to appreciate them it is useful to describe how the accuracy of the correlation coefficient (and of most other statistics) was calculated before the computer became widely available. The earlier procedure can be described in terms of the bootstrap, although it goes without saying that before the invention of the computer statisticians did not characterize their methods in such terms. In 1915 the British statistician Sir Ronald Fisher calculated the accuracy of r theoretically. Fisher had to assume that the data for the two variables, average GPA and average LSAT in our example, were drawn at random from a normal probability distribution, represented by a bell-shaped surface. The surface is a two-dimensional analogue of the one-dimensional bell-shaped curve. There is a family of such surfaces whose shape and orientation can be chosen to fit the available set of data. The surface is fitted to the data in the law school sample by placing the top of the bell directly over the point on the graph where both the GPA and the LSAT points have their overall average values. The surface slopes downward to the graph at a rate that depends on how widely the data points are scattered [*see illustration on preceding page*].

The bell-shaped surface is interpreted as a probability distribution in the same way the graph of values of r for law school samples is a frequency distribution. The probability of selecting a point on the graph of GPA and LSAT scores from within a certain region is equal to the volume that lies under the bell-shaped surface and directly above the region, divided by the entire volume of the space that lies between the surface and the graph. Fisher was then able to



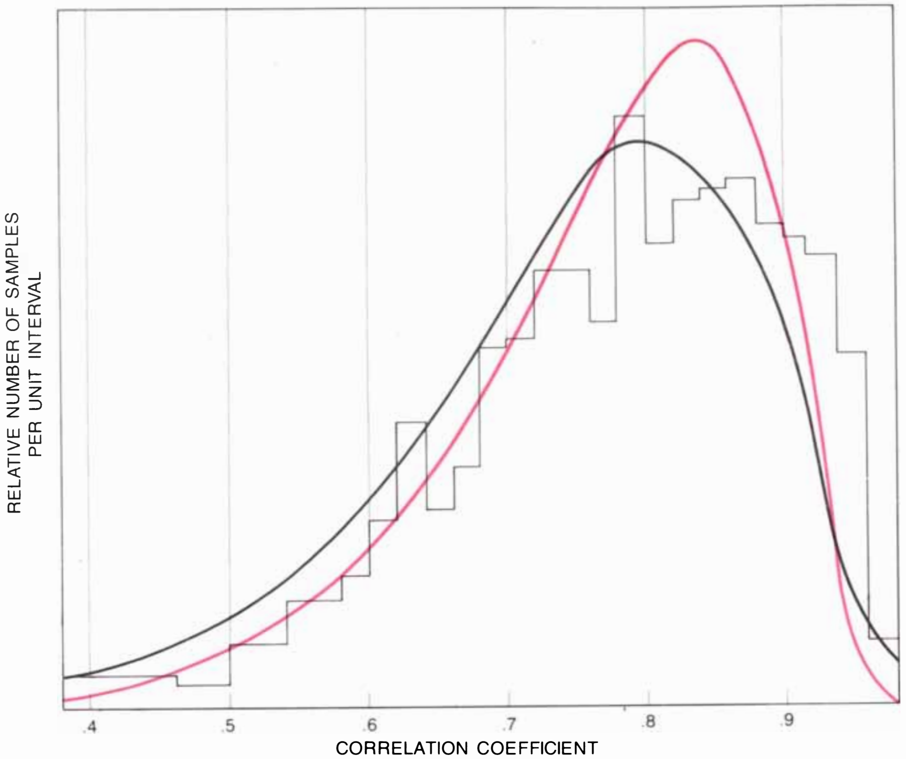
STATISTICAL ACCURACY of the observed value of r for a random sample can be known precisely only if it is known how r varies for a large number of samples. The 15 law schools for which the value of r has been calculated were selected at random from the total population of 82 American law schools. The data points in each of the four graphs represent the average GPA and average LSAT score for each of the 82 law schools. There are 82^{15} ways to choose samples of 15 law schools from the total population; four such samples have been selected by circling the points in color. (It is possible to select a school more than once in a given sample; such schools have been circled more than once.) The observed values of r for samples *a* and *b* are roughly equal to the true correlation coefficient for all 82 schools. The value of r for sample *c*, however, is much too high and the value of r for sample *d* is much too low. The true variability in the value of r for samples of 15 law schools can be determined by finding its value for many such samples because data for many more than 15 law schools (in fact, for all 82 of them) are available. Additional data, however, are often impossible to obtain. The bootstrap can estimate the amount of variability that would be shown by all the samples on the basis of one sample.

generate a distribution for the values of *r* by bootstrapping from the bell-shaped probability distribution. In effect, many samples of 15 data points are selected from the graph according to the probability given by their position under the bell-shaped surface. The value of *r* is calculated for each sample and a frequency distribution for the values of *r* is plotted. According to Fisher's method, the width of the interval that includes 68 percent of the values of *r* is .226, in good agreement with the true value of .270 but not as close in this case as the bootstrap estimate of .254.

The bulk of Fisher's calculation can be done analytically because of the assumption that the data in the sample are selected from a normal probability distribution. This assumption is a disadvantage of the method, however, because it might not be true. It is certainly not true in the law school example. Moreover, even if it is true, there is no easy way to check it; in most situations a much larger sample, with perhaps several hundred data points, would be needed to verify the shape of the surface.

The calculations involved in the bootstrap, in which there are no simplifying assumptions about the probability distribution, would have been quite impractical 30 years ago. As we have mentioned, the calculation of a single correlation coefficient takes about five minutes with a desk calculator, and one must carry out from 50 to 1,000 such calculations before a reasonably accurate frequency distribution for the bootstrap samples can be determined.

Today the calculation of a single value of *r* takes a ten-thousandth of a second with a medium-size computer; at such speed the bootstrap becomes feasible for routine application. If 1,000 bootstrap samples are generated, all the calculations necessary to estimate the width of the interval that includes 68 percent of the samples take less than a second and cost less than a dollar. The cost estimate is based on performing about 100,000 arithmetic operations. More ambitious bootstrap analyses, which give more detailed information about the accuracy of *r*, require about a million arithmetic operations.



BOOTSTRAP DISTRIBUTION of the correlation coefficient *r* (*"skyline" in black*) closely approximates the true distribution of *r* (*smooth curve in black*). The true distribution is actually plotted for a million samples of size 15, chosen at random from the 82^{15} such samples that can be selected from the 82 law schools; differences between the distribution graphed here and the distribution that could in principle be plotted for all 82^{15} samples are not discernible. The shape of the bootstrap distribution also approximates the shape of the distribution that can be estimated according to the probabilities given by a bell-shaped surface (*smooth curve in color*). The agreement suggests the bootstrap can be employed as a measure of the accuracy with which the correlation coefficient of the sample predicts the correlation coefficient of the population. The rather close agreement among the peaks of the distributions is an artifact of the sample.

The bootstrap is not limited to the analysis of the variability of statistics, such as the correlation coefficient, that are mathematically simple. It has been applied to many problems for which the variability of a statistic cannot be assessed analytically. Consider a family of statistics called principal components, which were introduced by Harold Hotelling of Columbia University in 1933. Principal components were devised to solve problems such as the following one, given in a textbook by Kantilal V. Mardia and John T. Kent of the University of Leeds and John M. Bibby of the Open University.

Eighty-eight college students each take two closed-book tests and three open-book tests. Suppose, for the purpose of grading the students, one wants to find the weighted average of the five scores that generates the greatest differences among the students. (In order to make the ratios and not merely the differences of the overall scores as variable as possible, the weights must be scaled so that the sum of their squares is equal to 1.) One set of weights arises if only the final test score is considered; the weights assigned would then be 0, 0, 0, 0 and 1. If all the students had high scores on the final test, however, the summary score generated by this set of weights would not be effective for differentiating the students. Another summary score arises if each test is given equal weight; the weights would then all be equal to $1/\sqrt{5}$, or about .45. The set of weights for the five tests that gives the greatest differences among the students is called the first principal component.

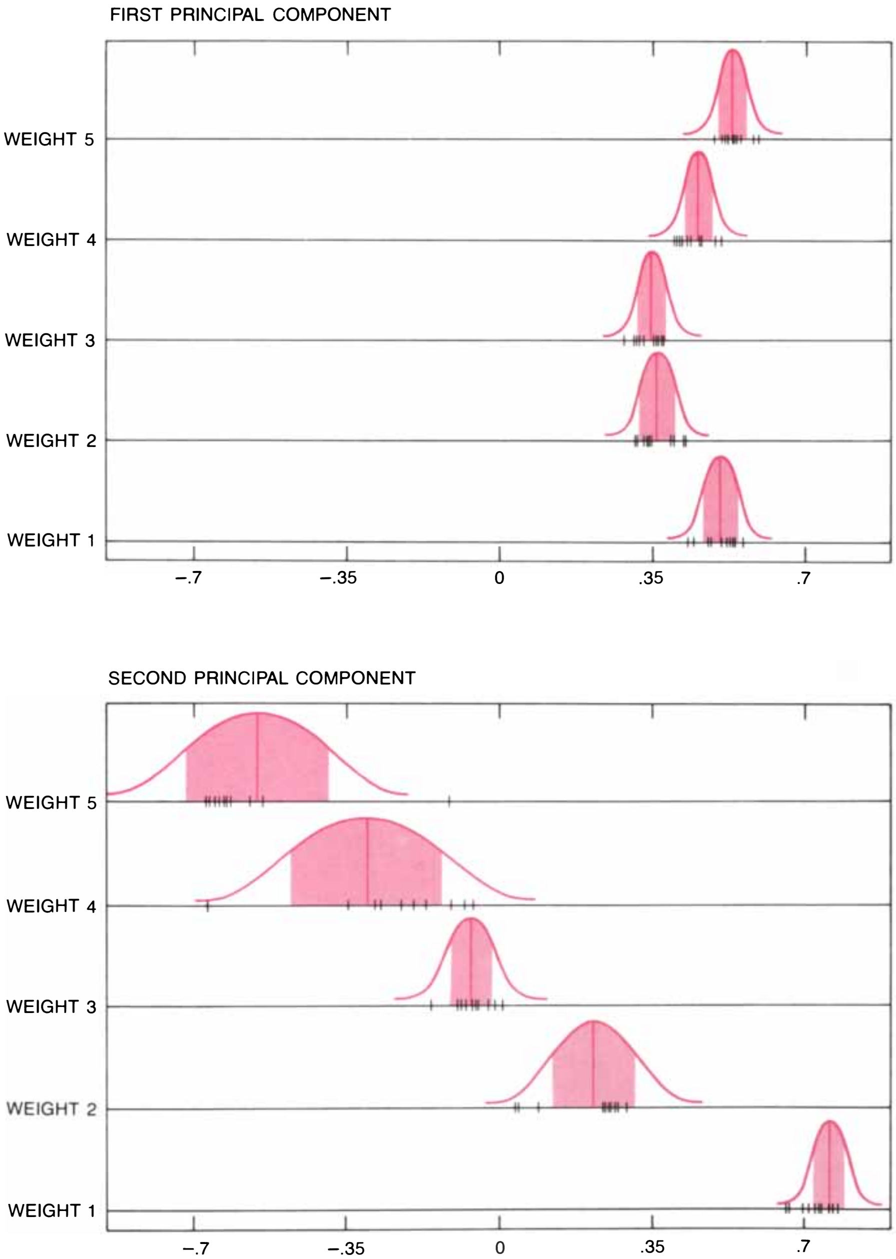
The first principal component is impossible to describe in a mathematically closed form; it must be computed numerically. When the calculation is done for the 88 students, the weights of the first principal component turn out to be roughly equal to one another. The greatest distinctions can therefore be made among the students by finding the average of the five scores.

The second principal component is the set of weights, subject to a mathematical constraint of independence, that gives the second-greatest differences among the students. When the second principal component is calculated for the 88 students, the weights turn out to give the difference between an average of the open-book scores and an average of the closed-book scores. The principal components suggest useful and unexpected interpretations for the averages of the student scores. How reliable are the interpretations? If they are to be trusted, one must try to determine how much variation there is in the values of the two principal components for samples of 88 students selected at random.

The problem of quantifying the variability of principal components for samples of a given size has preoccupied many statisticians for the past 50 years. If the appropriate normal distribution is assumed, partial solutions to questions concerning the frequency distribution for the first principal component can be given; little is known, however, about the second component and higher ones. By applying the bootstrap method a computer can quickly give an estimate of variability for any principal component without assuming that the data have a normal distribution.

In principle the bootstrap analysis is carried out just as it is for the correlation coefficient. Each student's set of five test scores is copied many times (that is, all five scores are written on the same piece of paper) and the copies are thoroughly shuffled. A new sample of size 88 is drawn at random and the principal components are calculated for it. The sampling is repeated many times and a frequency distribution is plotted for each principal component.

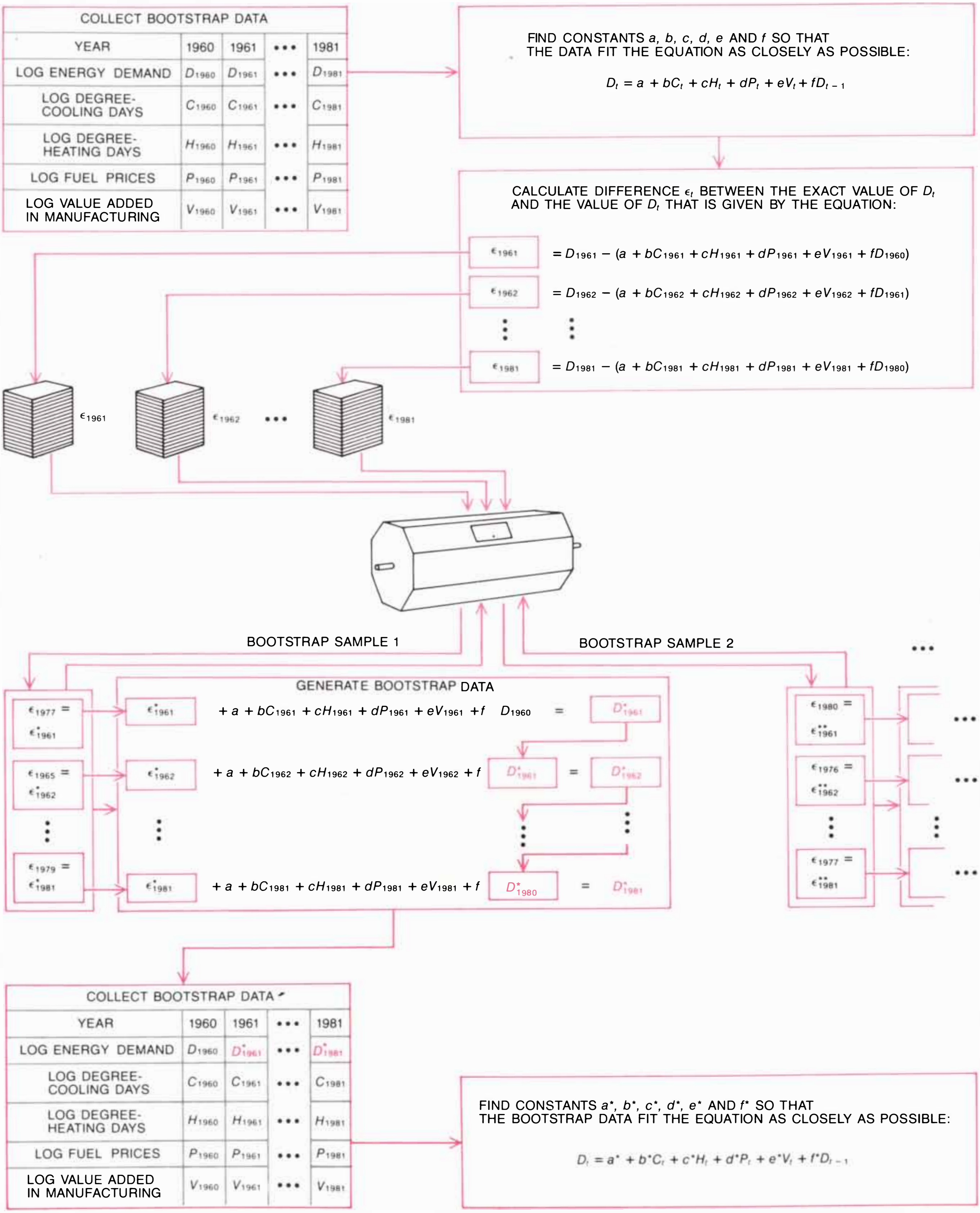
The results suggest that the weights associated with the first principal component are quite stable: they vary only in the second decimal place. The weights associated with the second principal component are less stable, but in a structured way. Remember that the second principal component was interpreted as the difference between an average of the open-book tests and an average of the closed-book tests. The interpretation is confirmed by the bootstrap analysis, but the weights given to the open-book tests are quite variable. The distribution for the principal components generated by the bootstrap is a good estimate of the true distribution of the principal components for samples of size 88. It takes only about two seconds for a large computer to do 100 bootstrap replications.



PRINCIPAL COMPONENTS are statistical estimators that have been widely applied for calculating summary scores on standardized tests. Suppose 88 students take five tests each and suppose, in order to assign a summary score, one wants to find the weighted average of the five test scores that generates the greatest differences among the students. The first principal component is the set of weights that solves the problem. The second principal component is the set of weights, subject to a mathematical constraint of independence, that generates the second most variable combination of the test scores. To assess the variability of the two principal components for many additional samples of 88 students the bootstrap was applied to the single sample. Each student's score for the five tests was written on a piece of paper, and each set of five scores was copied many times. All the copies were then shuffled and bootstrap samples of size 88 were selected at random. The principal components were calculated for each bootstrap sample. The variation in the weights for the first 10 bootstrap samples is shown by the black ticks on each graph; the red vertical lines indicate the observed values of the weights. The width of the central strip under the small bell-shaped curves indicates the variability of the weights. The fourth and fifth weights of the second principal component are particularly unstable.

Not every statistical estimator is a number. Nine weather stations in the eastern and midwestern U.S. recorded the *p*H level, or acidity, of every rainfall from September, 1978, through August, 1980. (A *p*H value of less than 7 is acidic, and the lower the *p*H value, the greater the acidity.) During the two years 2,000 *p*H values were measured. To represent the data Barry P. Eynon and Paul Switzer of Stanford prepared a *p*H-contour map of the region; the *p*H values are constant along a contour line. Such a map can be generated from the data by a well-defined mathematical procedure called Kriging, after the South African mining engineer H. G. Krige. Although the contour map is strictly determined by the data, it represents an extrapolation from the data collected at nine stations to many points in space and time (in fact, to an infinite number) that are not included in the original sample. One can therefore ask how variable the contours on the map would be if random variations yielded different samples of 2,000 *p*H values.

In this example neither the true contour map nor the true variability of all contour maps generated by samples of 2,000 *p*H values can be known. Both estimates must be made from the original data alone if they are to be made at all. By bootstrapping the original sample of 2,000 *p*H values in a way that preserves the geographic relations among the weather stations, Eynon and Switzer generated the maps shown in the illustration on page 117. There is no generally accepted measure of the variability of contour lines on a map, analogous to the width of an interval of a frequency distribution. Intuitively, however, the variability is readily perceived. It shows that the original contour map must be inter-



MODEL OF ENERGY DEMAND called RDFOR (Regional Demand Forecasting Model) was employed by the U.S. Department of Energy to analyze and forecast energy demand in 10 regions of the U.S. For each region the data are fitted as closely as possible to a mathematical model called a regression equation. The demand for energy in any given year is assumed to depend on the demand for energy the year before as well as on several other measures. Each error term ϵ_t is the difference between the predicted value of the energy demand in a given year and the observed value. Bootstrap samples of the error terms are selected at random and artificial data for energy demand are generated by the method shown in the diagram. The bootstrapped data are then fitted to a new regression equation, and the variability of the regression equations generated by the bootstrap gives an estimate of the expected accuracy of the model in predicting energy demand. A bootstrap analysis done by David A. Freedman of the University of California at Berkeley and Stephen C. Peters of Stanford has shown that the variability of the regression equations is from two to three times greater than was previously thought.

preted cautiously. Corridors of relatively low or relatively high acidity on the original map can shrink to become islands on a bootstrap map, depending on the effects of random "noise."

Statistical estimation is often carried out by making the available data conform as closely as possible to some predetermined form, or model. The simplest models are the line, the plane and the higher-dimensional analogues of the plane. Consider the graph of the 15 data points that represent the 15 law schools. Intuitively there are many lines that could be drawn to represent the trend of the data points, and so it is reasonable to agree in advance on a precise method for fitting the points to a line. Probably the most widely used estimator in statistics is a method for fitting points to a line called the least-squares method. The method was invented by Gauss and by Joseph Louis Lagrange in the early 19th century in order to make astronomical predictions.

The least-squares line is the line that minimizes the sum of the squares of the vertical distances between the data points and the line. A straightforward calculation gives the equation of the least-squares line from the data points. If the bootstrap is applied to the data, fake data sets can be generated, and the least-squares method can be applied to each fake set of data points in order to fit them to a new line. The fluctuation of the lines generated by the bootstrap shows the variability of the least-squares method as a statistical estimator for this set of data points.

The least-squares method and its generalizations are particularly useful for complex problems in which an investigator must bring large amounts of diverse information to bear on a single question. The U.S. Department of Energy, for example, has developed a model called the Regional Demand Forecasting Model (RDFOR), which attempts to forecast the demand for energy in 10 regions of the U.S. In the model it is assumed that the energy demand for each region in a given year depends in a simple way on five variables: the amount of variation above 75 degrees F. in summer, the variation below 65 degrees in winter, the price of fuel, the value added in manufacturing (a measure of the economic conditions in the region) and the energy demand during the previous year.

	PATIENT NUMBER						
	149	150	151	152	153	154	155
AGE	20	36	46	44	61	53	43
SEX	M	M	M	M	M	F	M
PRESENCE OF STEROID?	YES	NO	NO	NO	YES	YES	NO
ANTIVIRAL ADMINISTERED?	NO	NO	NO	NO	NO	NO	NO
FATIGUE?	NO	NO	YES	YES	YES	YES	YES
MALAISE?	NO	NO	YES	NO	YES	NO	NO
ANOREXIA?	NO	NO	YES	NO	NO	NO	NO
LARGE LIVER?	NO	NO	NO	NO	YES	NO	NO
FIRM LIVER?	*	NO	NO	YES	YES	NO	NO
PALPABLE SPLEEN?	NO	NO	NO	NO	NO	YES	YES
PRESENCE OF SPIDERS?	NO	NO	YES	NO	YES	YES	YES
PRESENCE OF ASCITES?	NO	NO	YES	NO	NO	NO	YES
PRESENCE OF VARICES?	NO	NO	YES	NO	NO	YES	NO
CONCENTRATION OF BILIRUBIN	.9	.6	7.6	.9	.8	1.5	1.2
CONCENTRATION OF ALKALINE PHOSPHATASE	89	120	*	126	95	84	100
CONCENTRATION OF SERUM GLUTAMIC-OXALOACETIC TRANSAMINASE (SGOT)	152	30	242	142	20	19	19
CONCENTRATION OF ALBUMIN	4.0	4.0	3.3	4.3	4.1	4.1	3.1
CONCENTRATION OF PROTEIN	*	*	50	*	*	48	42
PHYSICIAN'S PROGNOSIS	LIVE	LIVE	*	LIVE	LIVE	*	LIVE
OUTCOME	LIVE	LIVE	DIE	LIVE	LIVE	LIVE	DIE

MEDICAL DATA for seven out of 155 people with acute or chronic hepatitis give the values of 19 variables for each person that, taken together, could predict whether a patient will die or recover from the disease. (An asterisk indicates that information is missing.) It is common practice in statistics to inspect such data before a formal model is constructed; the aim of the inspection is to rule out all but four or five of the most important variables. Peter B. Gregory of the Stanford University School of Medicine eliminated all the variables except the patient's malaise, the presence of ascites (fluid in the abdominal cavity), the concentration of bilirubin and the physician's prognosis. Gregory developed a model based on the four variables that correctly predicted whether or not the patient would die from the disease in 84 percent of the cases.

The five variables can be thought of as if they were plotted on a five-dimensional graph, which is exactly analogous to a two-dimensional graph; every point on a five-dimensional graph corresponds to a possible combination of the five variables. The energy demand in a given year associated with a known combination of variables can then be represented by the height of a point in a six-dimensional space above the corresponding point on the five-dimensional graph. The representation of the data in a six-dimensional space is analogous to representing the dependence of some quantity on two other variables as the height of a point in three-dimensional space above a two-dimensional graph. Thus the energy data determine a set of points at various heights in the six-dimensional space.

The least-squares method specifies a way of drawing a five-dimensional analogue of a plane (called a hyperplane) as close as possible to all the points. Because of the dependence of the energy demand on the demand in previous years, the variables must be fitted to the hyperplane by a generalized version of the least-squares method. The generalized method calls for minimizing a weighted sum of errors after the weights have been estimated from the data. In recent years an elaborate method of estimating the accuracy of the procedure and the accuracy of its forecasts has been developed.

Freedman and Stephen C. Peters of Stanford examined the conventional estimates of the accuracy of the procedure by applying the bootstrap. In their approach it is assumed that the data lie close to an appropriate hyperplane, but it is not assumed that the errors between the data points and points that lie on the hyperplane are independent of one another. Instead the relation of the errors from point to point is allowed to have a complicated structure. The bootstrapping of the data was done in a way that

preserves the evolution of the energy demand from year to year. The variability of the hyperplanes generated by the bootstrap showed that the standard error previously assumed for the energy model is usually too small by a factor of two or three. The predictions of energy demand made by this aspect of the RDFOR model are therefore much less reliable than was once thought.

The examples we have presented so far have involved clearly defined statistical properties of samples. In practice the data can be inspected, sorted, graphed and preanalyzed in several ways before they are formally analyzed. Estimates of variability that do not take such informal practices into account may not give an accurate picture of statistical variability.

Consider a group of 155 people with acute and chronic hepatitis, initially studied by Peter B. Gregory of the Stanford University School of Medicine. Of the 155 patients 33 died and 122 survived, and for each patient there were data for 19 variables, such as age, sex and the results of standard biochemical measurements. Gregory's aim was to discover whether the data could be combined in a model that could predict a patient's chance of survival.

The analysis of the data was done in several steps. First, all but four of the most important variables were eliminated, because statistical experience suggests it is unwise to fit a model that depends on 19 variables with only 155 data points available. The elimination of the variables was done in two stages: each variable was inspected separately, whereupon six variables that appeared unrelated to the patients' survival were eliminated. A standard statistical procedure was then carried out on the remaining 13 variables, which further reduced the number to four. The variables that remained were the patient's malaise, ascites (the presence of a fluid in the abdomen), the concentration of bilirubin in the liver and the physician's prognosis for the patient. The variables were then fitted to a curve that predicts how the proportion of surviving patients depends on the values of the variables.

Such analysis is typical of scientific practice. In order to estimate its overall variability Gail Gong of Carnegie-Mellon University carried out the entire procedure from preliminary screening through the final curve fitting on bootstrapped samples of the original 155 data points. Her results were surprising and informative. The set of "important variables" generated during the initial stages of the analysis was itself quite erratic. For some bootstrap samples only the prognosis of the physician was found to be important, whereas for others such variables as sex, age, level of fatigue, concentration of albumin and concentration of protein were retained. No single variable emerged as significant in as many as 60 percent of the bootstrap samples.

Although the fitted curve is intended to predict whether or not a patient will survive, it misclassifies 16 percent of the original 155 patients. The estimate of 16 percent, however, is too small because the data on which it is based were also employed to generate the curve. The analysis generated by the bootstrap suggests a better estimate for the probability that the fitted curve will misclassify a given patient is .20.

The prospect of bootstrapping the entire process of data analysis offers hope that an extremely difficult problem will begin to yield, namely the connection between the mathematical theory that underlies statistics and actual statistical practice. The effects of preliminary "data snooping" on the final results are usually ignored, for no better reason than that it is impossible to analyze them mathematically. It now appears that the bootstrap, applied with the aid of the computer, can begin to estimate such effects.

The bootstrap is by no means the only statistical method that relies on the power of the computer. Several other methods such as the jackknife, cross-validation and balanced repeated replications are similar in spirit to the bootstrap but quite different in detail. Each of these procedures generates fake data sets from the original data and assesses the actual variability of a statistic from its variability over all the sets of fake data. The methods differ from the bootstrap and from one another in the way the fake data sets are generated.

The first such method was the jackknife, invented in 1949 by Maurice Quenouille and developed in the 1950's by John W. Tukey of Princeton University and the Bell Laboratories; it has been extensively investigated by Colin L. Mallows of Bell Laboratories, Louis Jaeckel of Berkeley, David V. Hinkley of the University of Texas at Austin, Rupert G. Miller of Stanford, William R. Schucany of Southern Methodist University and many others. The name jackknife was coined by Tukey to suggest that the method is an all-purpose statistical tool.

The jackknife proceeds by removing one observation at a time from the original data and recalculating the statistic of interest for each of the resulting truncated data sets. The variability of the statistic across all the truncated data sets can then be described. For the data from the 15 law schools the jackknife assesses the statistical accuracy of the value of r by making 15 recalculations of r, one for every possible subsample of size 14. The jackknife calls for fewer calculations than the bootstrap but it also seems less flexible and at times less dependable.

Cross-validation is an elaboration of a simple idea. The data are split in half and the second half is set aside; curves are fitted to the first half and then tested one by one for the best fit to the second half. The final testing is the cross-validation; it gives a reliable indication of how well the fitted curve would predict the values of new data. There is nothing special about half splits; for example, the data can be split in the ratio 90 to 10 as

BOOTSTRAP SAMPLE NUMBER	VARIABLES SELECTED
491	ALBUMIN, PROGNOSIS, SEX
492	ASCITES, BILIRUBIN, PROGNOSIS
493	BILIRUBIN, ASCITES
494	BILIRUBIN, PROGNOSIS, MALAISE
495	ASCITES
496	BILIRUBIN
497	ASCITES, VARICES
498	SPIDERS, PROGNOSIS, ALBUMIN
499	AGE, PROGNOSIS, BILIRUBIN, MALAISE, PROTEIN, SPIDERS
500	ASCITES, PROGNOSIS, BILIRUBIN, PROTEIN

VARIABLES DESIGNATED IMPORTANT by informal analysis prior to the construction of a formal statistical model can show wide variation. In a bootstrap study that simulated both the formal and the informal aspects of the statistical analysis, Gail Gong of Carnegie-Mellon University programmed a computer to copy the set of 19 variables associated with each patient many times. The sets of data were thoroughly shuffled and bootstrap samples of 155 sets of data were drawn at random from the collection. Formal and informal techniques of data analysis were then applied to each bootstrap sample, just as they had been for the original sample. The variables chosen as important are shown for 10 of the 500 bootstrap samples generated. Of the four variables originally chosen not one was selected in as many as 60 percent of the samples. Hence the variables identified in the original analysis cannot be taken very seriously.

Exxon has the information processor...

that really can be the start of something big.

At Exxon Office Systems, we're bringing the high-tech office down to earth, by designing office automation to grow the way you grow.

You can start an office automation system simply with just one of our remarkable EXXON 500 Series Information Processors.

Then as you grow, we can grow with you, right into our shared resource office automation system ... The EXXON 8400 Series System.

As a fully functioned system its capability is awesome. Operators can create, edit, reformat, file, share and retrieve documents, all with the touch of a key.

There's a dictionary, an electronic mailbox, a program for keeping calendars and scheduling meetings, a tickler file and more.

Start automating your office now with one of our EXXON 500 Series Information Processors. And step into the future without the shock of unnecessary costs from expanding inefficiently.

For more information on the EXXON 500 or our new office automation system, just send in the coupon below. Or call **800-327-6666**.

EXXON OFFICE SYSTEMS

The future...without the shock.

Exxon Office Systems SAM 05 83
P.O. Box 10184, Stamford, CT 06904
I'd like to know more about
☐ the EXXON 8400 Series System
☐ the EXXON 500 Series Information Processor
Please have your representative call.

Name ________ Title ________
Company ________
Address ________
City ________ State ________ Zip ________
Telephone ________

800-327-6666.
IN CONNECTICUT, 800-942-2525.

well. Moreover, there is no reason to carry out the cross-validation only once. The data can be randomly split many times in many ways.

Cross-validation has been widely applied to situations in which a curve-fitting procedure is well defined except in some crucial respect. For example, one might be willing to fit a polynomial to the data by the least-squares method, but the degree, or highest power, of the polynomial to be fitted might still be in doubt. (The higher the degree of the polynomial, the less smooth the fitted curve.) Given that half of the data have been fitted by polynomials of various degree, cross-validation can choose the degree of the polynomial that best fits the second half of the data. Seymour Geisser of the University of Minnesota, Mervyn Stone of the University of London and Grace G. Wahba of the University of Wisconsin at Madison have been pioneers in this development.

Instead of splitting the data in half at random a more systematic system of splits can be employed. The splits can be chosen in such a way that the results are optimal for certain simple situations that allow full theoretical analysis. The balanced repeated-replication method, developed by Philip J. McCarthy of Cornell University, makes splits in the data systematically in order to assess the variability of surveys and census samples. Random subsampling, a related method developed by John A. Hartigan of Yale University, is designed to yield dependable confidence intervals in certain situations.

There are close theoretical connections among the methods. One line of thinking develops them all, as well as several others, from the bootstrap. Hence one must ask what assurance can be given that the bootstrap will work most of the time, and how much it can be generalized. To the first question the answer is simple. The bootstrap has been tried on a large number of problems such as the law school problem for which the correct answer is known. The estimate it gives is a good one for such problems, and it can be mathematically proved to work for similar problems.

We have suggested the answer to the second question through the diversity of complex problems to which the bootstrap has already been applied. What is needed for many of them, however, is independent theoretical justification that the bootstrap estimate of accuracy remains as valid as it is for simpler problems. Current theoretical work seeks to provide such justification and to give more precise statements of accuracy based on the bootstrap. Fisher was able to provide a statistical theory that took full advantage of the computational facilities of the 1920's. The goal now is to do the same for the 1980's.

The Social Influence of the Motte-and-Bailey Castle

The motte was a medieval mound fort; the bailey was an associated enclosure. Their appearance in the second half of the 10th century diminished central authority in Europe and gave rise to chivalry

by Michel Bur

The first project of military engineering ordered by William the Conqueror after his landing in England was the hasty construction of a motte, a fortification of a kind unknown in the British Isles. Designed to provide the Normans at Hastings with a strongpoint, the motte was primarily a mound of earth surrounded by a ditch. Embedded in the mound were heavy timbers that supported a tower put together out of prefabricated members. In the years that followed the Norman victory the occupiers built similar mottes throughout England, thus securing a firm hold over the vanquished.

The Normans brought their knowledge of mottes with them from France, where the building of such simple fortifications had enabled nobles, great and small, to defy the central authority of Charlemagne's successors and to establish their own domains. It is not too much to say that this novel defensive weapon revolutionized medieval Europe. Politically it was the instrument of the feudalism that replaced rule from a single center. Socially its effects were perhaps even more important: the motte served as a kind of schoolroom where the medieval seigneurs and their retinue learned what was needed to create the form of civilized behavior known as chivalry.

Until a few years ago little was known about the mottes of France. The only archaeology concerned with the Middle Ages was the study of major medieval buildings, a subdiscipline of history concerned with describing and analyzing the technical aspects of masonry construction. The examination of other remains of medieval life was left to geographers; archaeologists did not bother to probe them. Recently, however, scholars at universities and other institutions have begun investigating the medieval countryside, and their work has begun to transform long-established patterns of historical thought.

Among the areas where progress in such research is being made is the study of fortified rural habitations. The motte has proved to be the most significant of these primitive seigneurial strongholds of the late 10th century through the 12th. Thanks to archaeological studies outside France, particularly those in Britain, where the details of the Norman Conquest continue to be investigated, French defensive earthworks of the period are well known.

A typical motte, appearing in profile either as a truncated cone or as part of a sphere, could be up to 100 meters in diameter at its base and as much as 20 meters high. Adjacent to most mottes was a large earth-walled enclosure, known as a bailey, surrounded by a ditch and topped with a wood palisade. Hence this type of dual defensive earthwork is called a motte-and-bailey castle. Another type of earthwork consists of a miniature bailey without a motte: a circular or oval bank 30 to 100 meters in diameter, surrounded by a ditch and topped with a palisade. Some of these enclosures served only as pens for livestock. Others, particularly those flanked by separate stock pens, were actually earthen castles of a lesser type.

Since 1961 a systematic inventory of these primitive seigneurial habitations has been under way in France, beginning in the Auvergne. Normandy was added to the survey in 1968 and Champagne in 1972. Since then the work has been extended both eastward and westward to include French Flanders, Dauphiné, Gascony and Saintonge. Its objective is to find, describe, measure and map (on a scale of 1 : 1,000) every vestige of these medieval earthworks that remains before modern agricultural machinery has erased them. The field work is being coordinated with archival research aimed at determining the date of construction and of abandonment or razing, the name given to each motte (such as "stronghold" or "castle") and the role in society of its inhabitants, such as "count" (*comes*), "knight" (*miles*) or "seigneur" (*dominus*). The standardization of such inquiry should result in a uniformity of collected data and so provide, if need be with the help of computer programs, a solid quantitative base for historical speculation.

A mere sampling of sites would not be acceptable. The survey must, like oil poured on water, spread outward from one community to the next until the census is complete. In brief, the study, both methodical and exhaustive, is simultaneously examining in depth the medieval archives and what remains of the medieval landscape.

The motte-and-bailey castle was known as a *château,* a word that like "castle" is from the Latin *castellum.* It was an establishment that served as a permanent abode for a seigneur and his men-at-arms. The *château* differed in size and function from similar but much larger defensive enclosures that had in the past served as places of refuge for the general population. The distinction is of particular importance because the motte, from the military point of view, was the most significant technological innovation of the latter part of the 10th century.

Mottes appeared first in low-lying marshy areas in the region between the Loire and the Rhine and next on the steep hills of the same region. From this "homeland" they spread across Europe from the Atlantic to the Vistula and from Scandinavia to the Mediterranean. With their wood tower—in some instances simply a watchtower, in others a larger residential structure—these earth mounds were, with one significant difference, equivalent to the nearly contemporaneous great stone donjon keeps of the Loire and Normandy. The difference was that in western Europe the medieval mottes were much more numerous than the stone castles were then or at any time later.

The time when the mottes quickly spread across Europe was an eventful one. Even before A.D. 1000 parts of western Europe (at first France and Italy and later Germany) had undergone a profound political and social transformation. The structure of Charlemagne's empire was in ruins and smaller political entities were forming everywhere. In place of the Carolingian system of state control from the center, men grouped themselves around feudal lords. The locally powerful who were sought out as protectors by lesser men now intended to exercise for their own benefit the authority they had previously drawn from the king. It was a revolution of the landed aristocracy, the very group from whose ranks the sovereign had traditionally recruited his administrators and officers. In France the landed aristocracy eventually dethroned the descendants of Charlemagne and in A.D. 987 gave the crown to one of their own, Hugh Capet. In the years that followed the interests of the landed aristocracy stifled those principles of peace and common weal that had formerly had the support of the central authority.

Such an upheaval could not take place without fierce strife: the great nobles against the monarch, the same nobles among themselves and so on down to second- and third-tier seigneurs against one another. The 10th century was a time of particularly violent competition for land and power. If a seigneur was to dominate, he needed abundant resources: the means to maintain troops and reward the fidelity of the rank and file. In the long run, however, even the most determined adventurer was always at risk of chance defeat, perhaps only a temporary setback, perhaps death in a skirmish. No territorial takeover could have been lasting without the motte, a weapon that enabled those who controlled it to defy their adversaries and to transform that defiance into irreversible victory.

Just as the waterwheel mill and the draft-animal harness revolutionized the economics of agriculture, so was the motte an instrument of political and social revolution. An unskilled peasant labor force could easily erect the earthworks in a short time. Moreover, except for its wood superstructure the motte was virtually indestructible. When in the 11th century provincial authority (itself a product of the breakdown of central authority) was challenged in such strongholds as Normandy and Champagne, what followed? Mottes sprang up all over the two provinces; without that weapon those up in arms against

LARGE MOTTE in the Ardennes of northeastern France is seen to the left of the church in this aerial photograph of Cornay, a village near the canal town of Vouziers. Although eroded for more than seven centuries, this earthwork is still 18 meters high, 60 meters in diameter at the base and 35 meters in diameter at the top. The summit, which covers an area of 110 square meters, now serves as a cemetery.

the provincial authorities could neither have initially held their ground nor have eventually sunk their roots.

Motte tactics were not unlike those of a huge but simplified game of chess. The primary objective was to advance one's own pieces without having them captured. One pushed forward and built a motte, strongly barricading one's forces within it to resist the siege that one's rival, hoping to regain lost lands, was sure to begin immediately. In general such sieges were failures, and after the seigneur and his garrison had survived a few such efforts the motte, by now upgraded to the status of a *château,* was guaranteed to endure at least until treason opened the gates or the seigneur himself was captured.

Contemporaneous accounts show that the early seigneurs lived in constant fear of treason, never being entirely sure of their followers' fidelity. They had to avoid capture at all costs, a consideration that explains why, when things went adversely in formal open battles, the seigneur in command of the losing side was the first to leave the field. The reason was that once the loser was in the hands of the winner he was exposed to intolerable pressure, as were his kin and others left behind to guard his domain. Many a captive seigneur was brought to the door of his *château,* to be hanged there if his wife did not open the gates. Many a noble who held as hostages the children of a traitorous seigneur threatened to hang them or blind them if their father did not give himself up. Less dramatic than such episodes but fully as effective in putting pressure on a captive was one involving the comte de Blois in 1044. After his capture by the comte d'Anjou a mere three days of imprisonment convinced him to yield the city of Tours and its surrounding countryside in exchange for his freedom.

Two schools of medieval historians disagree on how the mottes came to proliferate so rapidly. One school takes the view that most such *châteaux* were legally built by those endowed with authority. The other, echoing a phrase coined by Norman chroniclers, sees almost every motte as a *château adultérin,* that is, an unlawful construction. In effect anyone who declares some action to be a revolution necessarily defines as unlawful the means by which the revolution triumphs. Actually whether most of the *châteaux* were built by properly authorized persons or more informally by

MOTTE AT HASTINGS, the site of William the Conqueror's battle with Harold, appears in a detail from the Bayeux tapestry. The corner posts of the tower have been put up and are being covered with fresh earth by two men with pickaxes and three with shovels. For major operations such as this one and perhaps even for lesser operations the motte builders carried with them timbers cut to size.

adventurous landed proprietors matters little since what counts is not so much the source of the power as the purpose to which the power was put. This argument leads to a complementary one. If the act of revolution renders unlawful all other actions against the challenged authority (in this instance a violation of the monopoly on fortifications held by the crown or by lesser nobility), the revolt, if it is successful, soon creates its own legality. One is reminded of a 16th-century epigram of John Harington: "Treason doth never prosper: what's the reason? / For if it prosper, none dare call it treason."

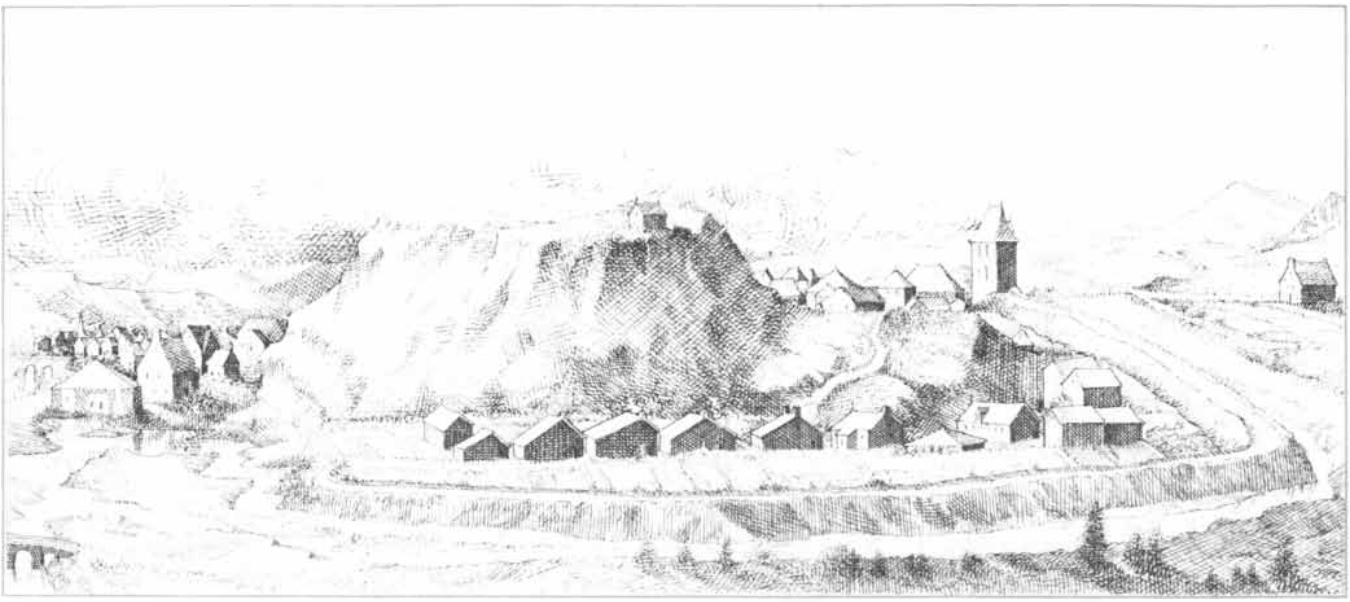
In a military aristocracy social relations were based on the vassal's homage and oath of allegiance to his superior. In other words, the weak became dependent on the strong in return for a reward in the form of land: the fief. This principle, occupying the middle ground between strength and weakness, gradually became the criterion of a new kind of legality that applied equally to kings and to great nobles: to be accepted as a legal occupant the master of a *château* had to hold the fortification in fief from a sovereign or a nobleman. In the give-and-take that naturally put the holders of existing fortifications in opposition to those who sought to build new ones, an equilibrium began to develop. This situation owed nothing to a policy of direct confrontation, an always violent and often sterile course of action. It came about because of recourse to the legal formula of feudalism. For the more politically skilled nobles the building of a stable establishment called not for destroying (or trying to destroy) the *châteaux adultérins* but for bringing them into vassalage. Thus the mottes on the fringe of any particular jurisdiction were progressively transformed into "feudal" mottes, with their existence no longer contested.

The network of the first French mottes had no connection with existing population centers; the formidable mass of the palisaded fortress was raised in total disregard of the local population. Keeping close watch on their rivals, each seigneur's forces camped in the flat countryside like an army of occupation, their ties with the countryfolk limited to levies and labor drafts. Above all, the *château* sought isolation. After that other strategic factors could be weighed. The chosen site might be at a point controlling travel by road, at a central location on a private domain, on property stolen from the Church or even in a town or close to the wall of a fortified city. Beyond any such long-term considerations, however, the seigneur selected a setting that would enable him and his forces both to survive siege and to go over to the offensive effectively.

Under these circumstances it is not

MOTTE AT DOL in Brittany is attacked in a second detail. The light bridge at the left spans the motte's moat. A horse ridden by one of William's men is crossing the bridge to probe the motte's defenses. They are apparently weak: at the center Duke Conan of Brittany escapes by sliding down a rope to the far side of the moat. He fled to another motte at Dinan.

MOTTE AT DINAN surrenders in a third detail. The garrison continues to resist, but two of William's men have crossed the moat and are setting fire to the motte's timbers with torches. At the right Conan is seen extending his lance, passing the keys to William as a token of his submission. Notwithstanding this episode, only a few mottes were ever successfully stormed.



FORTRESS BOURG OF DAMPIERRE-LE-CHÂTEAU in the department of the Marne, depicted in this drawing after a 15th-century engraving, had grown from a motte-and-bailey castle between the 12th century and the time when the engraving was made. Only one structure remained on the summit of the motte. The associated bailey was surrounded both by a rampart and by a moat. Some of the bailey outbuildings, seen in the foreground, are on low ground. Others stand on higher ground with a guard tower at the right. Beyond the rampart and moat at the left are commercial, residential and religious structures, their occupants attracted by the stability of the bourg.

PLAN VIEW OF MOTTE AND BAILEY in the illustration at the top of the page shows its substantial dimensions: 120 by 60 meters at the base and 75 by 35 meters at the top. The motte earthwork itself was 15 meters high. The relation between the motte and the bailey is shown in a section along the line *A–B* that appears at the top left. It shows why that part of the bailey was known as the upper town.

surprising that a good many of the earliest *châteaux,* erected in unsatisfactory locations, were later abandoned by their builders. A number of anonymous vestiges still bear witness to the precariousness and instability of the first seigneurial habitations. The process of site selection, however, soon came to favor locations that were easy to defend, had a good view of the countryside and commanded terrain over which it was possible to venture in all directions.

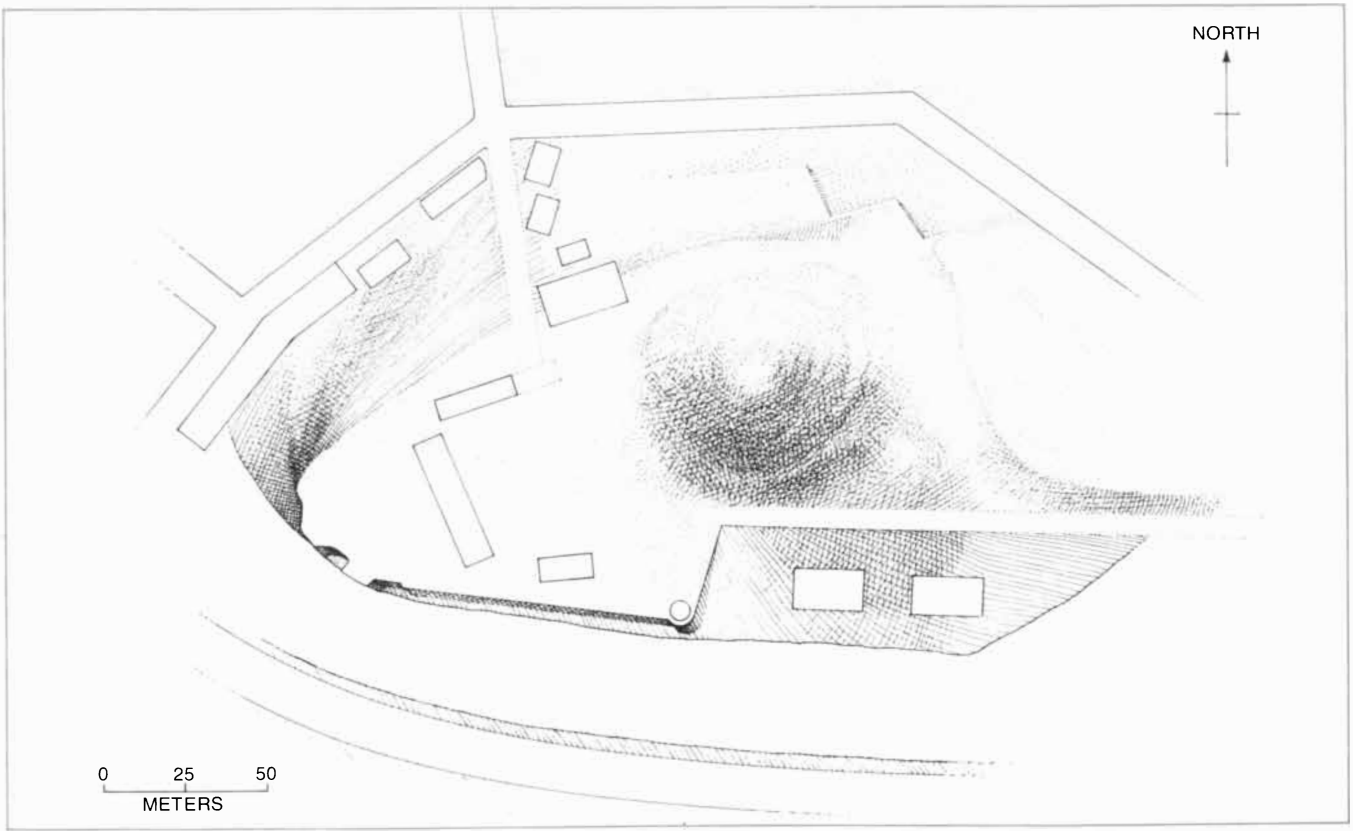
What was *château* life like? Here is a 12th-century description of the dwelling tower of Ardres in French Flanders. The tower of this sizable *château* had three levels. At the ground level were kept the stores of grain and wine. On the next level was the seigneurial chamber, adjoined by smaller rooms occupied by the master's younger children, his table attendants (bearers of wine and bread) and some other servants. On the top level, which was also the level of the guard walk, were communal sleeping quarters for the master's older children, one for the sons and another for the daughters. Adjoining the dwelling tower was another three-story structure. There at the ground level meat was stored; on the next level were the kitchens and quarters for the cooks, and on the top level was the chapel, facing east. Actually few if any of the early *châteaux* were as spacious and comfortable as Ardres; a sensible fear of fire must have made most *château* builders place the kitchens farther away from the dwelling tower.

Nevertheless, at least at Ardres, the seigneur, his family, their table attendants and other servants all lived in the dwelling tower. Their cooks lived in the adjacent building and perhaps so did the chaplain and the sentries of the watch. The rest of the establishment—the knights of the garrison, the administrators of the seigneurial domain, artisans such as the blacksmith and others needed for the upkeep of arms and equipment, and lesser personnel such as stable boys—occupied buildings within the adjacent palisaded bailey. There stables, barns, ovens, workshops and similar outbuildings formed a small, self-contained community that existed only for the seigneur's service and under his protection. The small population of the typical early *château* scarcely interacted at all with the much more numerous peasants who inhabited the villages of the countryside.

This aloofness did not continue for long. By about A.D. 1050 clusters of outsiders began to establish themselves near the *châteaux.* The root causes of this shift appear to have been population growth and a resulting social instability. Carolingian Europe had been covered with deep forest, and in the 11th and 12th centuries forests were actively cleared to allow agricultural expansion. In such newly cleared areas the interests of two quite different groups converged. The first group consisted of surplus labor hungry for land. The second was entrepreneurial: the younger sons of noble families, eager to set up their own domain. The convergence of interests led to the construction of a new kind of fortified site, where the seigneur's *château* and his followers' village all stood inside a single defensive earth wall. Examples are found in southern Aquitaine and Champagne.

Next, beginning in about A.D. 1070, local towns arose, particularly in the west of France. Churches and marketplaces were built in the immediate vicinity of one *château* or another. The population attracted to such amenities swelled the numbers in the growing set-

FORTRESS BOURG OF RETHEL in the Ardennes, depicted in another drawing after a 15th-century engraving, had also grown from a motte-and-bailey castle. By then both the motte and the bailey had been faced with blocks of chalk and surrounded with a stone wall that had towers placed at regular intervals. The same wall had been extended to protect the community that grew up around the priory of Notre Dame, at the center. As at Dampierre-le-Château, the security offered by the original motte had transformed it into a major town.



PLAN VIEW OF MOTTE shows how it took advantage of irregular terrain. The motte itself was on a rocky spur that extended eastward with an almost vertical drop on its south side and an only slightly shallower slope to the east and the north. The motte was 70 meters in diameter at the base and 10 meters in diameter at the summit. Probably built before A.D. 1026, it rose 21 meters above its base.

No matter what you do, we can help you do it better.

All kinds of professionals are finding more than computers at ComputerLand. They're discovering new solutions to old problems and opportunities they hadn't imagined. Better ways to manage a practice, or a research project; to process a brief, or control a production process; to analyze an engine's performance, or a dancer's.

Because ComputerLand doesn't merely offer the widest selection of personal computers on earth. We have everything you need to use a personal computer to work more productively. Know-how to help you choose the right one in the first place. Plus software, service, accessories and professional advice to make the most of it.

Selection, support and expertise. Only ComputerLand provides them all—at over 400 stores around the world.

When it comes to personal computers, nobody does it better.

Why not come in and see what we can help *you* do better.

ComputerLand®

We know small computers. Let us introduce you.™

Over 400 stores worldwide.

For locations call 800-227-1617x118 (in California 800-772-3545x118; in Hawaii or Canada call 415-930-0777 collect).

tlement and soon the seigneur of the *château* found himself master of the adjacent bourg, or borough, economically and administratively the center of a sizable district.

Such bourgs were variants of other settlements, both in the open country and close to the few cities, that were multiplying everywhere at this time. They benefited from the same social dynamics and often enjoyed the same legal privileges. In any event villages without *châteaux* remained more numerous than villages with them. This suggests that a seigneur's effort to recruit a cadre of peasants in order to profit from the economic growth of the newly established village did not always conclude with a company of happy householders settled in the shadow of his tower.

When the fortuitous union of *château* and village gave birth to a bourg, the presence of the seigneur, his entourage and the other clientele of the same social class was certainly crucial to making the enterprise flourish. Yet for the bourg to become firmly rooted other factors favorable to growth had to exist. One such factor was the fertility of the local soil. Another was a location on a main road. Without such advantages there could be little probability of success. Negative factors also played a part; many a bourg withered away when its adjacent *château* was destroyed by chance or by malice.

These observations, concerned with the period of seigneurial consolidation from the middle of the 11th century into the 12th, have taken us a long way from the initial feudal period, a time of struggle when ordinary methods of defending habitations were shaky and when the new weapon, the motte, first came into service as a means of territorial domination. Let us now return to that earlier period in order to see the role the motte played in the evolution of a new kind of man and the molding of his behavior. In effect one cannot separate the feudal warrior from his environment.

To begin with, consider the contradictions of the times: on the one hand freedom and on the other a subjugated and exploited peasantry; on the one hand war and on the other a clergy preaching peace as the ideal. A seigneur inhabited a world of isolation and danger where—in order to protect himself and his family, to sleep in safety, to raise children to be knowledgeable about that harsh and hazardous life—he needed to hide behind an earth rampart and barricade himself on a motte. The only way to soften those harsh conditions was to depend on the links of kinship or, better, on the parent-surrogate links of vassalage and an oath of fealty.

Even with those assurances constant vigilance was necessary. At home the seigneur lived in fear of treason. In battle the knowledge that the stakes were high sharpened cunning and incited treachery. Even a task as simple as finding a wife became enmeshed in haggling or led to matchmaking that left little room for personal choice.

Then, little by little, the violence and insecurity of these early feudal times gave rise, as it were, to their own antibodies. The fear of treason served in the end to reinforce the fidelity of the vassal and to exalt the loyalty of the warrior. The art of war, learned through continuous training from an early age, was not sidetracked into becoming the art of ambush. Instead it evolved into straightforward prowess, so that a victor faced by a valiant but unlucky adversary—virtually a mirror of himself—was careful to exercise fair play. By the same token the young warrior's lust, stimulated by the presence of both unmarried and married women within the *château,* was transformed into civility.

This kind of courtesy was not limited to the men's attitude toward women. It defined the deportment of the men in the seigneur's retinue. Within the confines of the *château* the choice was to live either in fellowship or in antagonism. Clearly the best course was to adhere to certain rules that made life as pleasant as possible or, at the least, to refrain from violence. Thus it was that in the confined world behind the palisade there slowly developed the moral standards that, when feudal society entered its second age at the end of the 11th century, inspired the troubadours. They sang of chivalry and love, but what the songs actually celebrated was the dual achievement of stabilization and colonization. Those foremost figures of chivalry, the seigneur's retinue of knights, began as simple warriors who had attained a higher estate through the practice of cavalry warfare. At the same time their garrison duties imposed on them specific values of courtesy and honor: the rules of the game.

One can conclude that the motte of the 10th and 11th centuries has a historical significance that has not attracted enough notice. As an instrument of domination par excellence it should be evaluated in terms of technological efficiency. The waterwheel mill and the draft-animal harness played a comparable technological role in the economic revolution of the Middle Ages. Consider, however, another factor. The authority of Louis XIV, who ruled France for more than half of the 17th century and on into the 18th, is regarded as having been nothing less than absolute. The fact remains that the Sun King in all his glory was not as well obeyed as any minister in the French republic of today. His orders reached the farthest boundary of his realm, but they got there slowly, and the functionaries on the spot were left a certain margin for interpretation. It is not so today. For example, it is enough for the Minister of the Interior to book a conference call, to lift his telephone and have some 100 departmental prefects standing at attention at the other end of the line. Moreover, any ambiguity in his orders can be eliminated as he goes along. As a means of exercising authority in government the telephone not only is faster than the written order but also can carry more weight.

Viewed in proportion and at a somewhat different level the motte played a role similar to the telephone's but in reverse. It diminished the king's prerogatives, affording the seigneurs a means of weakening central authority and giving birth to as many seigneuries as there were *châteaux.* Unable to exercise sovereignty, the king could only act the part of a feudal overlord, hoping for obedience no longer from subjects but rather from vassals. The conclusion is self-evident: however powerful the ruler and however determined the unruly, the outcome for both remained moot as long as neither had an effective technological advantage.

A second conclusion is that the motte certainly affected the mass of the rural population. From having been more or less self-sufficient they found themselves tied to the *châteaux* and their seigneurs and forced to pay taxes to these new overlords. The effect on the behavior of the overlords, however, was even more profound. Accustomed to a life in large open dwellings in a countryside populated with peasant freemen, the seigneurs in a matter of decades found themselves huddling behind ramparts, armed to the teeth in order first to grasp power and then to keep it. Clinging to this technological instrument of terror, the motte, the seigneurs and their retinues began to evolve a culture both primitive and original. On the one hand the culture was two-faced and dedicated to naked power; on the other it was a culture of courage and faithfulness to the given word. Certain of the popular medieval ballads, the chansons de geste, captured this evolving culture as if in a living mirror.

Little by little, as the generations passed, a certain equilibrium arose. A set of new values acknowledged seigneurial mutuality and made it possible to live less at sword's point. The *châteaux* became open to friends and neighbors; tournaments took the place of warfare; shields, once anonymous, carried symbols indicating the identity of the bearer. Where there had reigned ruse and brutality there now flourished virtues such as prowess, generosity, courtesy and honor. Such is the legacy the period bequeathed us from its second stage of development, a legacy, bred in the environment of the motte, that deserves to be called the civilization of the *châteaux.*

BREAKTHROUGH: A VALLEY OF MIRRORS THAT FOLLOW THE SUN.

At the experimental solar power plant near Barstow, California, our design imitates nature. Just as sunflowers turn to follow the daily progress of the sun across the sky, 1,800 giant mirrored panels (called heliostats) are programmed to tilt and turn to the ideal reflective angle for every minute of the day. They focus the sun's rays on a "power-tower" boiler, creating steam to generate electricity.

Today, Solar One produces enough electricity for Southern California Edison to meet the requirements of a town of 6,000. Solar 100, a new plant ten times larger, is being planned. A plant of this size would save hundreds of thousands of barrels of oil every year.

McDonnell Douglas design innovations would give Solar 100 heliostats 50% more reflective area (and increased power-generating capability) compared with those at Solar One.

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

We're McDonnell Douglas.

Sudden Cardiac Death: A Problem in Topology

Many sudden deaths are the result of fibrillation: a disruption of the coordinated contraction of heart muscle fibers. The cause may lie in a state of affairs described by a mathematical theorem

by Arthur T. Winfree

On the campus of the McGill University School of Medicine in Montreal the morning of November 7, 1914, was bright and chilly. It was a Saturday, and there were few professors or students around to notice George Ralph Mines enter his laboratory to begin some weekend experimental work. Mines, a 28-year-old physiologist, had been doing work on fibrillation, a radical disorganization of the heart's pattern of contraction. Ordinarily the heart contracts as a unified mass, but in fibrillation each small region of cardiac muscle contracts rapidly and with no apparent coordination with the neighboring tissue.

Mines had been trying to determine whether relatively small, brief electrical stimuli can cause fibrillation. For this work he had constructed a device to deliver electrical impulses to the heart with a magnitude and timing that could be precisely controlled. The device had been employed in preliminary work with animals. When Mines decided it was time to begin work with human beings, he chose the most readily available experimental subject: himself. At about six o'clock that evening a janitor, thinking it was unusually quiet in the laboratory, entered the room. Mines was lying under the laboratory bench surrounded by twisted electrical equipment. A broken mechanism was attached to his chest over the heart and a piece of apparatus nearby was still recording the faltering heartbeat. He died without recovering consciousness.

The death of George Mines might have remained a personal tragedy of little scientific interest if it had not been for the subject of his last work. Fibrillation is one of the major causes of the syndrome known as sudden cardiac death, which kills several hundred thousand people per year in the U.S. Moreover, of the several causes of sudden cardiac death fibrillation is the least understood. It can arise with little warning in apparently healthy people. In many instances when an autopsy is done, there is no perceptible clinical reason for the drastic interruption of the normal coordination of the heart. Mines's work shed light on part of the problem. He was the first to demonstrate that fibrillation can develop after a relatively small electrical impulse if the impulse is applied to the heart at the right time. Although much work has been done on sudden cardiac death and fibrillation since 1914, it is still not known conclusively how a small stimulus can elicit such a drastic response in the heart.

Intriguingly, the branch of mathematics known as topology can help to elucidate the problem. Like certain other physiological systems, cardiac tissue is capable of a rhythmic electrical discharge. If an external electrical impulse is delivered to a piece of tissue from the heart, the most frequent effect is to shift the usual rhythm ahead or back, without changing the interval between impulses after the stimulus. By means of a topological theorem, however, it can be shown that there must be a relatively small electrical stimulus that does not yield such a predictable shift if it is applied at the right moment in the cycle of beats. Indeed, the experimental results indicate that after such a singular stimulus the regular beating can stop.

The cessation of contractions in a small region of heart muscle is not fibrillation. It can be shown topologically, however, that around a patch of tissue where the beat has stopped the conditions are favorable for the creation of a circulating wave of electrical impulses. Such a wave could circulate through the heart, usurping the function of the normal pacemaking apparatus. In passing, the circulating wave could give rise to many smaller waves, causing small areas of the heart to contract rapidly in an uncoordinated way. The best current models suggest that a pattern of circulating waves is present in fibrillation.

The arguments I shall present here describing how a singular stimulus can cause fibrillation have by no means been proved. They are primarily mathematical arguments and their confirmation will require much practical work in physiology. Even if they are confirmed, their clinical implications are far from clear. Nevertheless, topology has provided a significant new insight into the problem of sudden cardiac death.

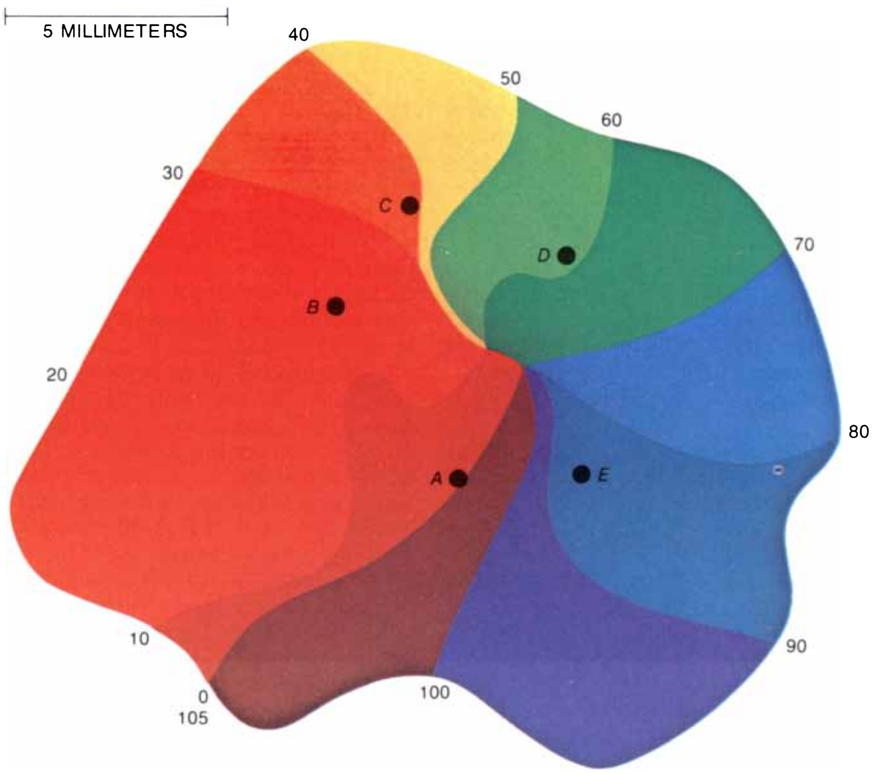
Before topology can be applied to heartbeat timing we must briefly consider the normal structure and function of the mammalian heart. The heart is an extraordinarily complex organ and a complete description of it is not possible here. The following much-simplified description concentrates on the properties of the heart that are pertinent to the topological argument. The heart consists of four chambers. The two atriums at the top of the heart serve as priming chambers. The larger ventricles below them are the main pumping chambers. Although in its normal operation the heart contracts more or less as a single mass, it can be noted when the action is observed closely that the contraction begins in the atriums and soon thereafter spreads to the ventricles.

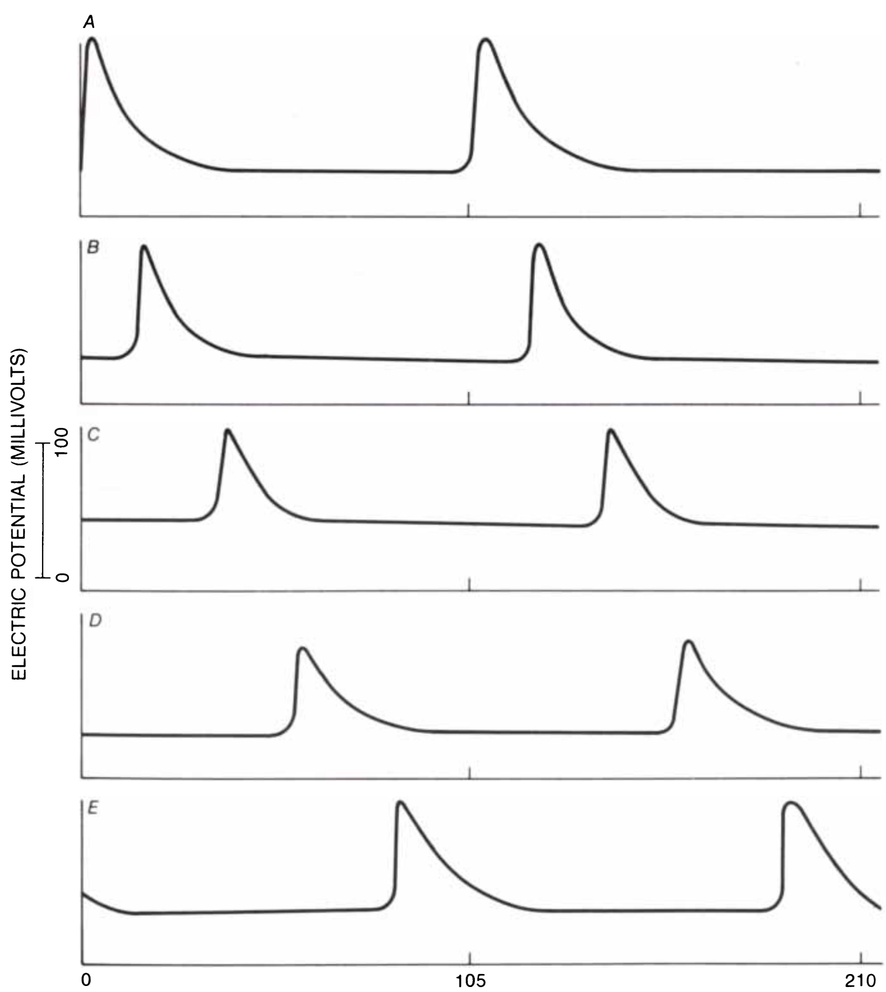
The coordination of the four chambers is maintained by the passage through the cardiac tissue of a wave of electrical impulses that travels quickly from cell to cell. The heart is composed mainly of several types of specialized muscle cells that have the form of long fibers. When a fiber is in the resting state, the electric potential inside the cell membrane is somewhat lower (more negative) than the one outside the membrane. The stimulus for each fiber to contract is a rapid increase in the potential inside the membrane, at the end of which the interior of the cell briefly becomes electrically positive with respect to the environment before the potential decreases again.

The changes in electric potential are mediated by the flow of several ions (atoms with a net electric charge) across the cell's outer membrane. The rapid increase in potential is called a depolarization, and it can propagate along the muscle fiber as an "action potential" much like the ones that convey information in the nervous system. After the passage of an action potential there is a refractory period of a few tenths of a second during which normal stimulation cannot elicit another action potential in the fiber.

The fibers in the heart are closely apposed to each other. In addition specialized organelles in the cell membrane of each fiber provide regions of low electrical resistance. These organelles make it possible for the triggering electrochemical impulse to pass from cell to cell. Therefore an impulse originating in one part of the heart can be transmitted quickly through the entire organ. Moreover, certain tissues in the heart have the capacity to spontaneously generate triggering impulses at regular intervals. The regions are called pacemakers and the one with the highest intrinsic rate imposes its rhythm on the entire heart.

The fastest spontaneous firings in the healthy heart come from a small mass of tissue near the top of the atriums known as the sinus node. The sinus node directly initiates the contraction of the atriums. Its impulse also propagates to the atrioventricular node, a second pacemaker lying between the right atrium and the right ventricle. From the atrioventricular node the impulse is relayed to the muscular wall of the ventricle by the Purkinje fibers. Purkinje fibers are adapted for very rapid conduction of electrical impulses. They extend throughout the ventricles in an arrangement that resembles the roots of a tree. When the pacemaking impulse passes from the Purkinje fibers into the ventricular muscle, a wave of contrac-





A ROTATING WAVE of electrical impulses can circulate indefinitely in the heart and could underlie fibrillation. In a fibrillating heart small regions of tissue contract rapidly and without coordination. If fibrillation persists for more than five minutes, the victim usually dies. The trigger for heart muscle cells to contract is an electrochemical impulse. Under certain conditions the impulse can travel in a circular path. The upper panel shows the path of a rotating wave induced experimentally in rabbit heart tissue by Maurits Allessie and his colleagues at the Center for Biomedicine at Maastricht in the Netherlands. Colored areas indicate the distance traveled by the wave front in 10 milliseconds; each complete rotation required 105 milliseconds. The traces in the lower panel show the electric potential at the five labeled points in the tissue. The periodic sharp increases in potential mark the passage of the rotating wave front.

tion spreads through the ventricles in milliseconds.

At rest the human heart contracts roughly once per second. The intrinsic beat in the sinus node that determines the resting rate can be accelerated or retarded by impulses carried by nerves coming from the brain, various ganglia and the internal organs. Such nerves infiltrate the entire heart, but their endings are particularly dense in the pacemaking system. For example, the vagus nerve can have a considerable effect on local rates of contraction and on the coordination of the entire heart.

The acceleration and retardation of the heart's contractions are usually smoothly synchronized. The mechanism can fail, however, if there is an infarction: an interruption of the flow of blood to a portion of the heart. Synchronization can also fail if the heart is exposed to unusually high levels of hormones or ions. Even without physical damage or chemical stress synchronization can fail if the heart is perturbed by substantial electrical stimuli, as in electrocution. The same result can be accomplished by a small stimulus if it is unfortunately timed. We shall see how such small stimuli can cause sudden cardiac death.

The term sudden cardiac death was coined in 1887 by John A. McWilliam of the University of Aberdeen, although the phenomenon had been known long before then. McWilliam noted that in some instances of sudden cardiac death the heart did not simply stop but entered into violent and disorganized activity. In the 19th century such violent activity was referred to as *delirium cordis,* or madness of the heart. The modern name, fibrillation, was introduced in the 1870's. According to Willis A. Tacker, Jr., and Leslie A. Geddes of Purdue University: "When a fibrillating heart is held in the hand, it feels like a wad of writhing worms. In many cases the rate of this random process is so rapid that the heart surface seems to shimmer. In other instances multiple waves of contraction and relaxation are clearly visible." If fibrillation persists for more than five minutes, death becomes almost certain. After a few gasping breaths respiration stops and the victim begins to turn blue because of a lack of oxygen in the blood. Convulsions and groaning sometimes accompany the stopping of respiration.

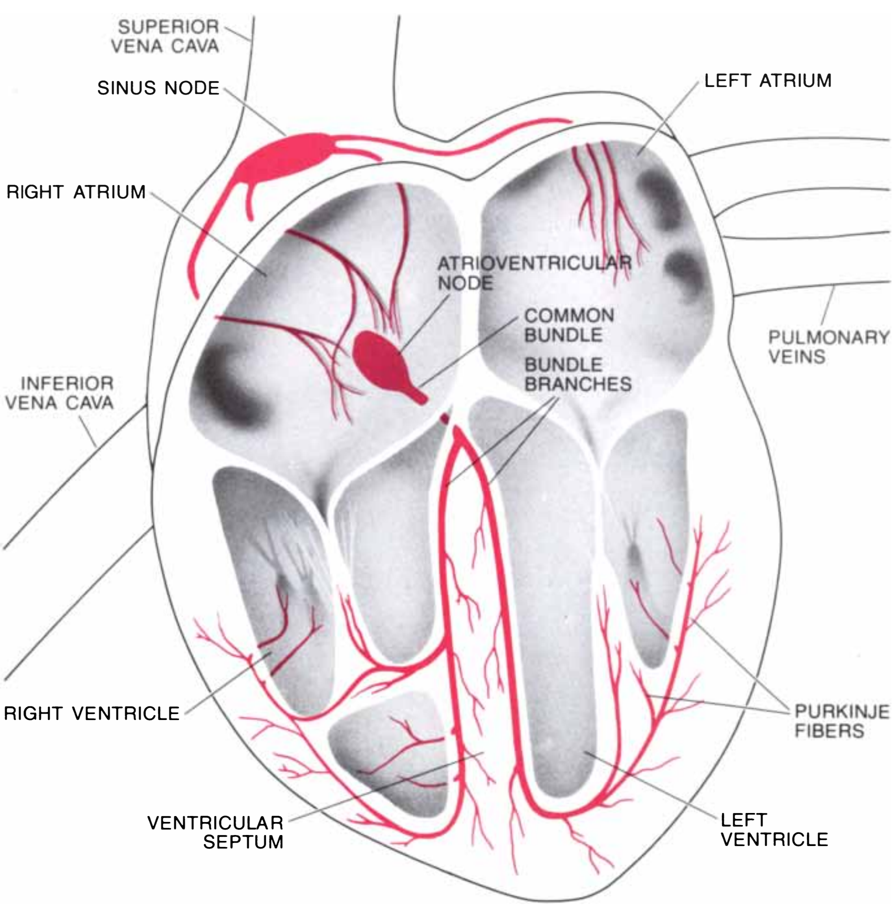
There is little agreement among physiologists on the mechanism of fibrillation. Studies of the membrane in single cells have been the source of much information about how electrical impulses are relayed in physiological systems. Such work has been particularly informative about the cells of the nervous system. Work with single cells, however, has been of only limited help in understanding fibrillation. The reason is that fibrillation is a disorder of the organization of the heartbeat rather than a malfunction of the individual fibers: it affects the timing of activity in the heart fibers with respect to one another. Indeed, it is probable that in the fibrillating heart each fiber is responding in the usual way, but the pattern of electrical impulses has been disturbed, resulting in a loss of synchronization.

What could cause the thousands of fibers in the human heart to suddenly lose the precise coordination they have maintained for a lifetime? Maurits Allessie of the Center for Biomedicine at Maastricht in the Netherlands, pointed out to me that in the paper Mines submitted to a medical journal just before his death he proposed that fibrillation could be caused by an electrical wave circulating in the heart muscle.

In his work Mines delivered electrical impulses to the heart, varying the timing of the stimulus to cover the interval between contractions. Most of the stimuli Mines applied had no lasting effect, and none initiated an unmistakable circulating wave in the heart tissue. He noted, however, that "under some conditions a stimulus of very brief duration may induce fibrillation . . . if properly timed. . . . The stimulus employed would never cause fibrillation unless it was properly timed." The precise moment between beats of the heart when a small stimulus can cause fibrillation was rediscovered many years after Mines's death; it is now known as the vulnerable phase.

My own interest in the vulnerable phase was aroused partly by my being a helpless witness to two sudden cardiac deaths. At the time I was working on the biological mechanisms of circadian rhythms, the intrinsic 24-hour cycles many organisms exhibit. Certain topological features of circadian timing systems had led me to predict that a single brief stimulus, if it was properly timed, could cause organisms that had been quite rhythmic to lose their regular rhythm. I had found such stimuli and demonstrated their effect, as other workers have done since.

It turns out that the topological properties of circadian clocks that make them susceptible to disruption by a sin-



HUMAN HEART is a four-chambered pump. The atriums are priming chambers. The ventricles are the main pumping chambers. The heart's contractions are synchronized by electrical impulses originating in the sinus node. The node, which has a regular intrinsic electrical discharge, serves as the pacemaker for the heart as a whole. Its impulse initiates the contraction in the atriums directly. The impulse is also relayed to the atrioventricular node, and from there it is conveyed to the muscular ventricles by specialized conducting fibers, among them the Purkinje fibers. Under ordinary conditions the heart contracts almost as a single mass.

gle stimulus are shared by physiological systems capable of regular electrical discharges, such as nerve cells, heart muscle fibers and even the heart as a whole. A topological property is a property of a geometric figure that persists when the figure is bent, twisted, stretched or otherwise deformed in a continuous manner. For example, an image seen through a distorting lens is topologically equivalent to the undistorted image.

Topology might be defined as the study of properties that remain unchanged in spite of quantitative changes (but not qualitative ones). Inferences based on topological features of a physical system can be quite effective in making predictions without reference to the precise quantitative aspects of the system. Topological reasoning was valuable in my work on circadian rhythms because little was known about their mechanism, and so quantitative models depended largely on guesswork. In contrast, so much quantitative information has been gathered on the coordination of the heartbeat that the assembled data can be overwhelming. By confining one's attention to certain topological properties of heartbeat timing it may be possible to cut through the mass of data and reach an understanding of how fibrillation begins.

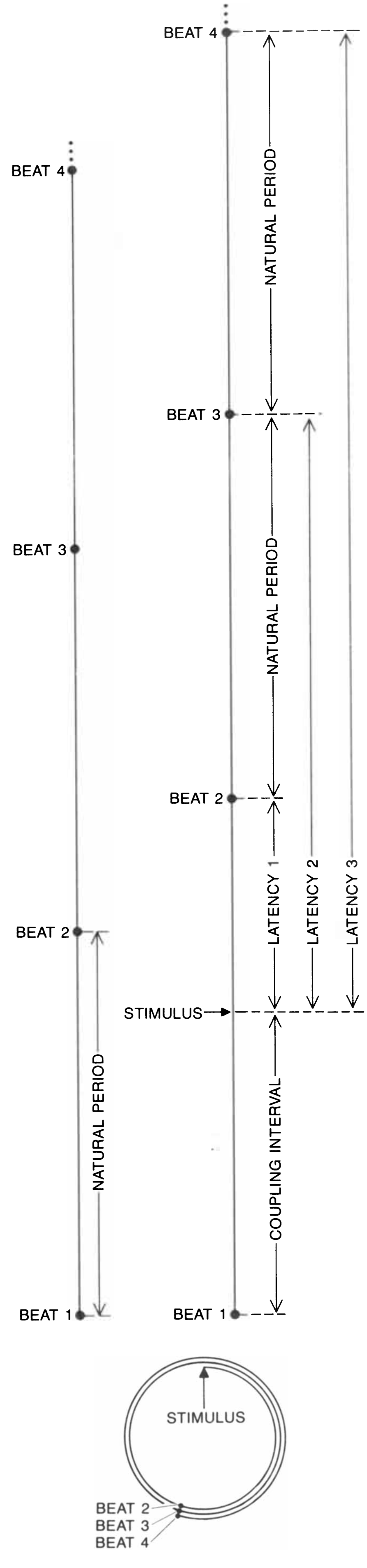
The topological qualities of the heartbeat that are of interest here were first suggested by observations made in about 1930 by J. C. Eccles of the University of Oxford. Like Mines, Eccles was examining the effect of the timing of an external stimulus on the subsequent contractions of the heart. Eccles, however, was concerned with the timing of the next few beats rather than with circulating waves or fibrillation. He stimulated the heart of the dog by activating the vagus nerve, which, as I noted above, is one of the regulators of the rate of contraction in the normal operation of the heart. Eccles observed that after a stimulus the heartbeat was perturbed but quickly returned to its usual rhythm. In many instances the heart recovered so rapidly that every complete cycle after the stimulus had the normal length; the effect of the stimulus was merely to offset the usual rhythm.

Since 1930 a vocabulary has been developed to describe such changes in timing. The period from the initial beat to the application of the stimulus is called the coupling interval. The period from the stimulus to a subsequent beat is called the latency. Since there are many beats after the stimulus, there are many latencies: each successive latency is the period until a new beat. (It should be noted that the topological principles under discussion can be applied to several biological systems. In this context "beat" does not necessarily refer to the action of the heart. The beat is simply the event that marks the beginning and end of each cycle in the system under observation.)

Depending on the length of the coupling interval and the magnitude of the stimulus, the first beat after the stimulus can be either delayed or advanced. Both Mines and Eccles observed the effect of applying a stimulus with a constant magnitude at progressively later times in the cycle. When the coupling interval is varied in this way, the coupling interval and the latency can show either of two distinct relations. Which relation is observed depends on the magnitude of the stimulus.

One of the two patterns is called weak rescheduling. To understand weak rescheduling it is valuable to consider what happens when the stimulus is too small to affect the timing of the subsequent beat. Under these circumstances the sum of the coupling interval and the latency until the first beat after the stimulus must equal the natural period, or the interval between beats in the unperturbed system. Hence the coupling interval and the latency must vary inversely: as the stimulus is given progressively later in the cycle, the latency becomes progressively shorter. On a graph that relates latency to coupling interval all beats lie on a diagonal that slopes down to the right. As the coupling interval is increased smoothly from zero to the full natural period, the latency decreases smoothly over one full cycle. As a result every possible value of latency from zero to one full period appears once.

It is characteristic of weak rescheduling that every value of latency appears an odd number of times when the coupling interval is varied over its full range. The pattern of later beats result-



ELECTRICAL STIMULUS applied to the heart can advance or delay the next beat. In the unperturbed cycle (*vertical line at upper left*) the interval between beats is called the natural period. When an external stimulus is given to the heart (*vertical line at upper right*), the interval from the initial beat to the stimulus is called the coupling interval. The interval from the stimulus to a subsequent beat is called the latency. Since there are many beats after the stimulus, there are many latencies. Any complete natural periods in the latency can be removed by winding the line standing for the latency around a circle with a circumference equal to the natural period (*circle at bottom*). When this is done, the points standing for beats after the stimulus lie on top of one another. These later beats have all been advanced or delayed by the same interval and therefore successive beats are one natural period apart. The main effect of the stimulus is to shift the usual rhythm ahead or behind; the natural periodicity is maintained. The terminology employed to describe such shifts can be applied to any system that is capable of rhythmic firing; "beat" then simply means the event that marks the end of the cycle.

What if you chose as a technical

"At Union Carbide, an HP computer network automates production management and has paid for itself in less than two years."

Union Carbide's Linde Division in Tonawanda, N.Y., manufactures custom production equipment. Factory systems manager Doyce Coffman says, "When production demand fluctuates, it complicates our scheduling, product costing, and long range planning.

"To solve these problems we automated production and materials management with a network of HP 1000 and 3000 computers and 43 data capture terminals.

"Because these systems continuously monitor operations and respond in real time, we can identify and eliminate production bottlenecks on line. The payback? Improved efficiency factory-wide and a dollar savings that covered the cost of our entire HP network in less than two years."

What should you expect from a computer partner?

Quality—from researching customer needs to product development, manufacturing, marketing, after-sales service, and support!

Quality, by HP's definition, encompasses more than the product.

This high quality recently was recognized by the Union of Japanese Scientists & Engineers, who awarded a coveted Deming Prize to Yokogawa-Hewlett-Packard, HP's joint venture company in Japan.

YHP offers the entire spectrum of HP products and in 1981 accounted for Japanese sales of more than $200 million. Nearly 60 percent of the products YHP sells are imported from HP manufacturing operations

The Deming Prize: *for outstanding merit in the control of quality.*

Hewlett-Packard computer partner?

"At Likoff Cardiovascular Institute, an HP computer processes electrocardiograms in minutes instead of hours."

Likoff Cardiovascular Institute, in Philadelphia, is a leading center for the research and treatment of cardiovascular disease. Using an HP computerized ECG management system, the Institute processes 100,000 electrocardiograms a year, 70,000 from outside facilities.

Dr. Gary Anderson, heart station director, says, "By automating the ECG processing cycle, we reduced turnaround time per test from 2½ hours to just 4 minutes—and reduced patient costs by 10%.

"With the HP Management System, we can retrieve patient records instantly, including often critical comparison data in our evaluations. And because this system handles incoming ECGs on a priority basis, we're responding to emergencies much faster than before, with better interpreted ECGs at a lower cost."

in the US and abroad. Conversely, YHP manufactures and exports various products to HP customers throughout the world.

The total quality control approach for which YHP was awarded the Deming Prize is a point of pride throughout HP. Customers in every part of the world should expect *quality* as a fundamental element in a computer partnership with HP.

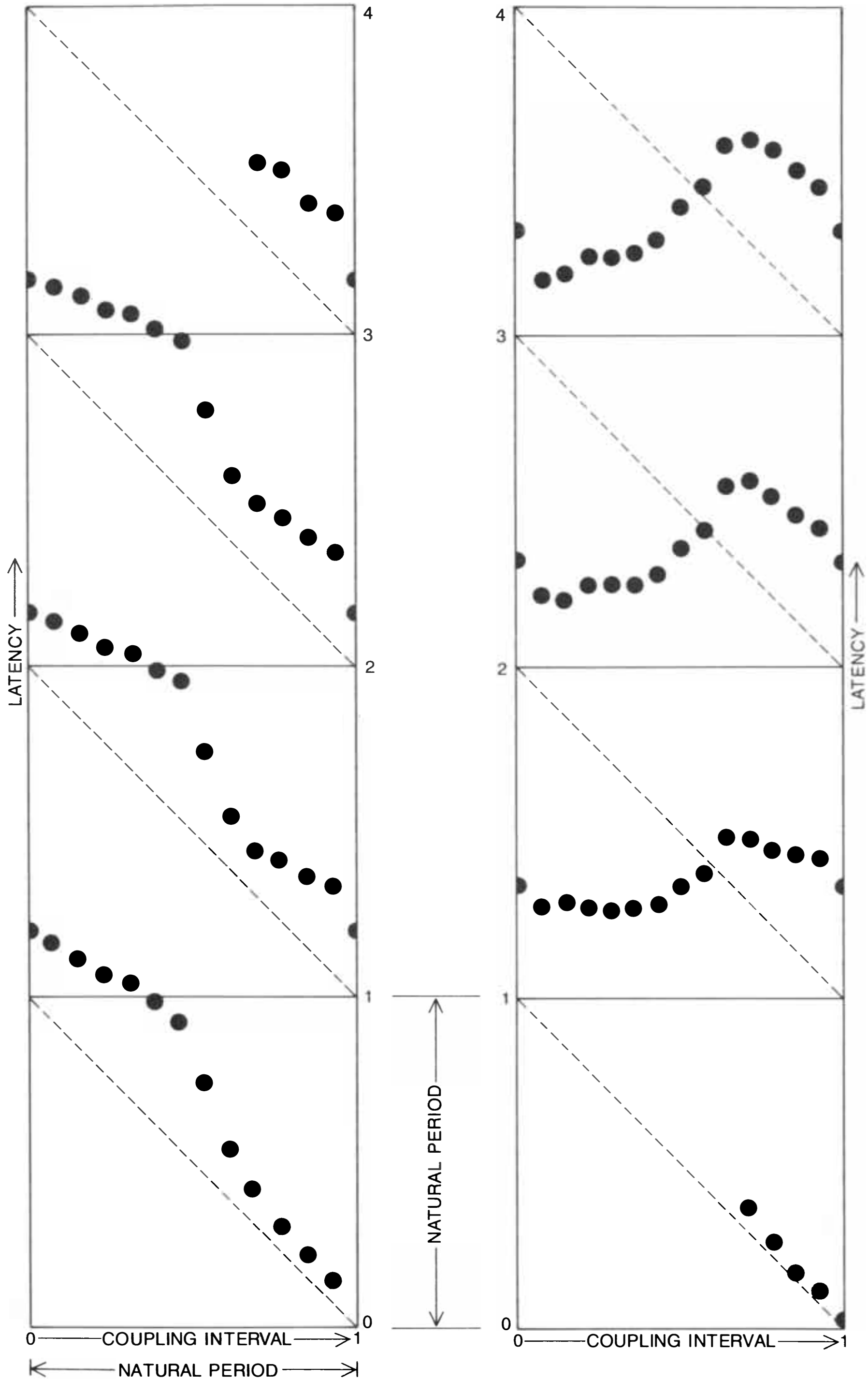
You should consider a working partnership with HP. Now.

For a free copy of our brochure explaining HP's practical, proven approach to meeting your long-term computer and information needs, write to A.P. Oliverio, Senior Vice President, Marketing, Dept. 207, Hewlett-Packard Company, P.O. Box 10301, Palo Alto, CA 94303.

When performance must be measured by results

HEWLETT PACKARD

0002341



WEAK AND STRONG RESCHEDULING are two topologically distinct patterns of beats following the application of an external stimulus. The data in the graphs come from work with rabbit hearts done by Jose Jalife and Joseph J. Salata at the Upstate Medical Center of the State University of New York at Syracuse. In each graph the dots stand for the beats that result when a stimulus with a particular strength is applied at a coupling interval varying from zero to one period. If the stimulus were too small to advance or delay the next beat, the coupling interval and the latency would vary inversely and the dots would fall on the diagonal in each square. If the stimulus is sufficiently strong, the next beat is advanced or delayed and the dots do not fall on the diagonal. If the strength of the stimulus is below a certain level, however, the latency still decreases smoothly over one natural period as the coupling interval is increased. This pattern is called weak rescheduling (*left*). If the stimulus is increased beyond a certain level, the latency undergoes no net increase or decrease as the coupling interval is varied through one period. The latency can vary by any amount as the coupling interval is increased, but it returns to the initial level when the coupling interval reaches one full period. This pattern is called strong rescheduling (*right*). In the example shown the latency decreases to a minimum, then increases to a maximum before returning to the initial value. That the two patterns are topologically distinct can be shown by joining the left and right edges of each graph to form a vertical tube. (Since the cycle repeats itself every period, the edges are equivalent.) The dots on the left graph then form a continuous helix and the dots on the right graph form three separate rings.

ing from a negligible stimulus is merely the limiting case of weak rescheduling. If the magnitude of the stimulus is increased, the next beat can be either advanced or delayed, depending on the timing of the stimulus. If the strength of the stimulus is not increased beyond a certain level, however, the latency still decreases smoothly and continuously as the coupling interval is increased through one full cycle. The two quantities are no longer inversely proportional; the plot of latency can curve above and below the diagonal as the coupling interval is increased. Nevertheless, each value of latency still appears an odd number of times.

If the strength of the stimulus is increased above a certain level, the second pattern, strong rescheduling, appears. In strong rescheduling the latency undergoes no net increase or decrease as the coupling interval is increased over one full period. The latency can increase or decrease by any amount as the coupling interval is varied, but it returns to the initial value when the coupling interval reaches one complete natural period. For example, when the coupling interval is zero (meaning that the stimulus coincides with a natural beat), the latency could be about one and a half natural periods. When the coupling interval is increased, so that the stimulus comes progressively later in the cycle, the latency decreases to a minimum, then increases to a maximum and ultimately returns to the initial value as the coupling interval approaches the natural period. As the coupling interval is varied through the complete cycle not all values of latency need appear. Furthermore, the values that do appear are seen twice: once as the latency is decreasing and once as it is increasing.

The curve described above is only one possible example of strong rescheduling. Many other curves are possible. What all such curves have in common is that each value of latency that appears does so an even number of times. Most values do not appear; the ones that do can appear any even number of times. The range of values of the latency that do not appear widens with increasing stimulus strength. In the limiting case of strong rescheduling only one value of the latency appears. The limiting case follows a stimulus so strong that it yields the same result regardless of when in the cycle it is applied; the next beat always comes, say, half a second after the stimulus.

In both weak and strong rescheduling the pacemaker resumes its normal periodicity after the advanced or delayed second beat. Weak rescheduling has long been known. Almost every rhythmic biological process shows weak rescheduling if the stimulus is sufficiently small. The topologically distinct strong

Over 300 manufacturers sell their microcomputers with Microsoft software.

There's a reason.

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

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

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

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard,™ products that materially increase the capabilities of Apple® and IBM PC computers. Plus, business and management software such as Multiplan,™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

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

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

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

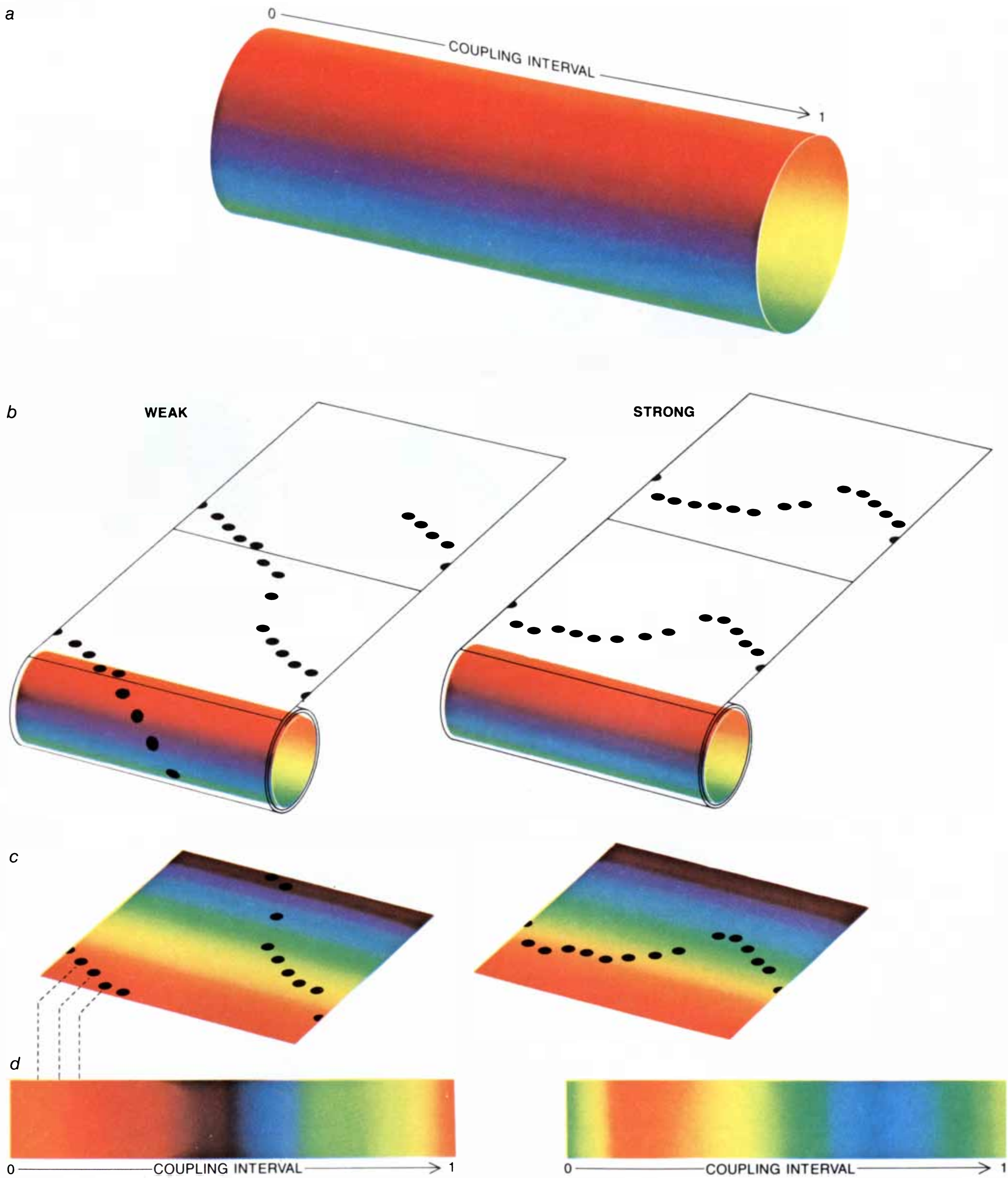
rescheduling, however, was recognized in neural pacemakers only five years ago. The data on strong rescheduling had been published some time before, but it took almost a decade for the distinct pattern to be recognized. The existence of strong rescheduling in neural pacemakers has now been confirmed by many workers, most recently by Jose Jalife and his student Joseph J. Salata of the Upstate Medical Center of the State University of New York at Syracuse. Jalife and Salata found both weak and strong rescheduling by stimulating the sinoatrial node of the rabbit and examining the subsequent impulses of the pacemaker.

As we shall see, the fact that the pacemaking apparatus of the heart is capable of strong rescheduling provides a crucial link in the argument leading from a single small stimulus to fibrillation. To understand how this could be so it is necessary to consider the latency in detail. As noted above, for each stimulus there are many latencies, corresponding to the many beats that follow the stimulus. For the purposes of topological analysis it is advantageous to subtract from each latency any complete natural periods included in it. There is a scheme for visualizing the result of this operation that leads directly to the topological argument.

Imagine a ring with a circumference equal to the natural period of the pacemaker. The ring is marked with a full cycle of colors arranged in the order of the spectrum that just fits around the circumference. Since the circumference of the ring is equal to the natural period, each color can be made to stand for a particular fraction of the period. For example, if red stands for a negligibly short latency, then yellow stands for a latency of about a third of the period, blue a latency two-thirds of the period and red again a latency one full natural period long.

Visualize the latency following a particular stimulus as being represented by a long, infinitesimally thin cord. One end of the cord corresponds to the time of application of the stimulus. A short distance away is the point corresponding to the first beat after the stimulus. The points marking the subsequent beats fall at intervals along the cord. The stimulus end of the cord is held against the red section of the colored ring and the rest of the cord is wrapped around the ring as many times as it will go.

When this procedure is carried out, the points representing beats after the stimulus lie on top of one another. Their position on the ring indicates the fundamental latency, that is, the time elapsed from the stimulus to each subsequent beat, minus any whole number of natural periods that can be subtracted. If the points happen to lie in the red part



LATENCIES CAN BE ASSIGNED COLORS by means of a cylinder with a full cycle of colors around its circumference; the colors are arranged in the order of the spectrum (*a*). The circumference of the cylinder equals one natural period. Therefore each hue can represent a fraction of the natural period. Red stands for a latency of zero or one full period. The graphs of weak and strong rescheduling shown in the illustration on page 148 are placed against the cylinder with the bottom edge of each graph (equivalent to a latency of zero) resting on the red band. The graphs are then wrapped around the cylinder; each square panel of the graph wraps exactly once around the circumference (*b*). In weak and strong rescheduling successive beats after the stimulus occur one period apart. Hence after the wrapping successive rows in each graph lie on top of one another. The procedure removes any complete natural periods from the latency. The cylinder with the graph wrapped around it is slit open along the edge of the red band and laid flat, yielding a row of dots on a square panel (*c*). The row of dots represents the latencies that result when a particular stimulus is applied with a coupling interval varying over the natural period. The colors on which the dots lie are arranged horizontally according to the coupling interval (*d*). The sequence at the left shows weak rescheduling. The sequence at the right shows strong rescheduling. If the stimulus were too small to affect the timing of the next beat, the wrapping would yield a horizontal sequence beginning with red and running smoothly through the full cycle of colors back to red.

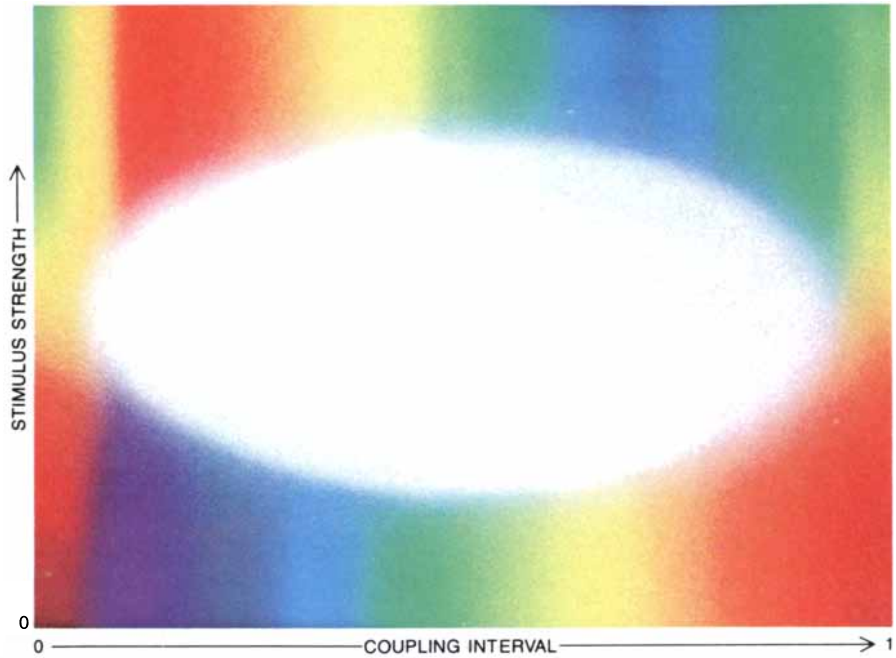
of the ring, the fundamental latency is equal to zero or the natural period of the system. If the points lie in the yellow part, the fundamental latency is about a third of the natural period. Because the colors of the ring correspond to fractions of the natural period, the latency has been assigned a color that stands for its length.

Any stimulus that yields weak or strong rescheduling merely offsets the natural cycle of the pacemaker. Any such stimulus, depending on its timing, can therefore be associated with a fundamental latency and hence with a color on the ring of latencies. Now consider what happens when the colors that stand for latencies are put in the appropriate places on a rectangular graph with coordinates representing the coupling interval and the strength of the stimulus.

The coupling interval is shown along the horizontal axis at the bottom of the rectangle, with zero at the left and one full period at the right. The strength of the stimulus is shown along the vertical axis at the left edge of the rectangle, with zero at the bottom. The points along the lower edge of the rectangle represent the result of applying a negligible stimulus at a coupling interval ranging from zero to one full period. As I have indicated, such a stimulus yields latencies that decrease smoothly through one natural period. Therefore when each point along the lower edge of the rectangle is assigned a color according to its fundamental latency, a full cycle of colors in sequence appears.

Along the upper edge are the points that stand for the result of applying a very strong stimulus at a coupling interval that varies over the full cycle; these are the conditions leading to strong rescheduling. As we have seen, strong rescheduling can result in a latency that begins at a value of about one and a half natural periods, decreases to a minimum and then increases to a maximum before returning to the initial level. Thus the upper edge of the rectangular graph could be colored in a sequence that includes green at the left, orange, green and blue in the middle and green again at the right.

The colors along the left and right edges of the rectangle represent the result of applying a progressively stronger stimulus at a coupling interval of zero (at the left) and of one full period (at the right). There could be many different sequences of colors along the vertical edges, but as we shall see the exact arrangement is inconsequential from a topological point of view. It should be noted, however, that the pattern must be the same on both edges. The reason is that a coupling interval of zero is equivalent to a coupling interval of one full period, and so stimuli given at those intervals must generate the same pattern of latencies.



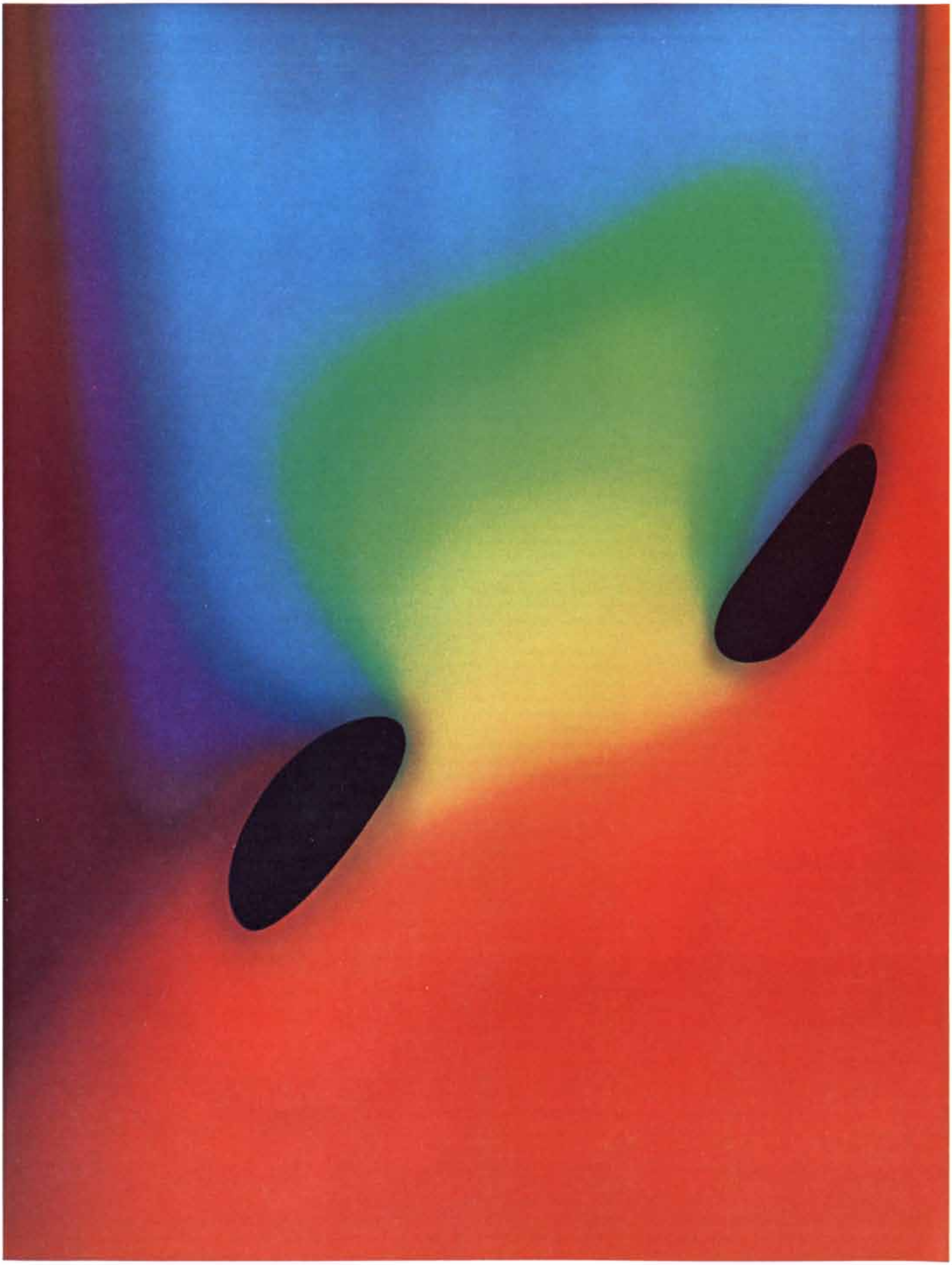
RECTANGULAR GRAPH can be partially filled in with colors that stand for latencies. Each point on the graph stands for a specific combination of coupling interval and stimulus strength. The color at the point designates the resulting latency. The lower edge of the rectangle includes the full cycle of colors that stands for the effect of applying a negligible stimulus at a coupling interval varying from zero to one period. The left and right edges show the effect of applying a progressively stronger stimulus at a coupling interval of 0 (at the left) or 1 (at the right). The upper edge represents strong rescheduling. Moving around the perimeter of the rectangle, one cycle of colors from red back to red is traversed. Beginning at the lower left corner and moving right, the full cycle of colors is encountered along the bottom of the graph. Up the right edge is a smooth sequence from red to green. Across the top is a sequence from green through blue, green, orange and back to green. Down the left edge is a smooth sequence from green to red: the same sequence as on the right edge but in the reverse order. Thus there is one full cycle of colors on the perimeter of the rectangle. When the perimeter has such a cycle, a circle can be inscribed somewhere in the rectangle on which the colors of the cycle appear once each and in sequence. A topological theorem called the nonretraction theorem shows that it is impossible to associate every point inside the circle with a point on the circle itself and also preserve the original continuity of the points within the circle. Therefore the theorem implies there is at least one point in the interior of the graph that cannot be associated with a color. In other words, a particular combination of stimulus strength and coupling interval does not yield a consistent latency. Such a stimulus represents a mathematical singularity. Its effect on the regular rhythm cannot be predicted, but it will not result in a mere offset of the pacemaker's cycle.

The edges of the rectangular graph have now been colored. To continue the visual exercise imagine a wheel much like the ring of latencies, with a full cycle of colors applied to its rim. The wheel is placed at the lower left corner of the graph and then moved all the way around the perimeter of the rectangle, beginning by moving to the right along the lower edge. As the wheel is moved it is turned so that the color on the wheel matches the color appearing underneath it on the graph.

What net advance will the wheel have made by the time the circuit is complete? Along the lower edge of the graph the wheel makes one complete forward revolution as it traverses the latencies corresponding to a negligible stimulus. Going up the right edge, the wheel turns from red to green. Along the upper edge the wheel turns from green through blue, green and orange before returning to green; hence the net forward rotation is zero. Going down the left edge of the rectangle, the pattern of colors on the right edge is encountered again but in reverse order. Therefore, whatever the effect of going up the right edge was, it is reversed by going down the left edge.

It is apparent that the upper edge of the rectangle and the vertical edges make no ultimate contribution to the turning of the wheel: their turns are self-canceling. When the wheel returns to the red area at the lower left corner of the rectangle, the only contribution to its turning that has not been reversed is the contribution of the lower edge of the graph. Thus the wheel makes one complete forward revolution as it traverses the perimeter of the graph of latencies.

What about the points in the interior of the rectangle? One might expect to find that every possible combination of coupling interval and stimulus strength is associated with some definite fundamental latency. It would then follow



that every point in the rectangle could be assigned a color. Furthermore, one would expect the latency to change continuously with any small change in coupling interval or stimulus strength, so that the colors in the interior of the rectangle would flow smoothly into one another without discontinuities. It is the surprising prediction of topology that such a smooth coloring of every point is not possible.

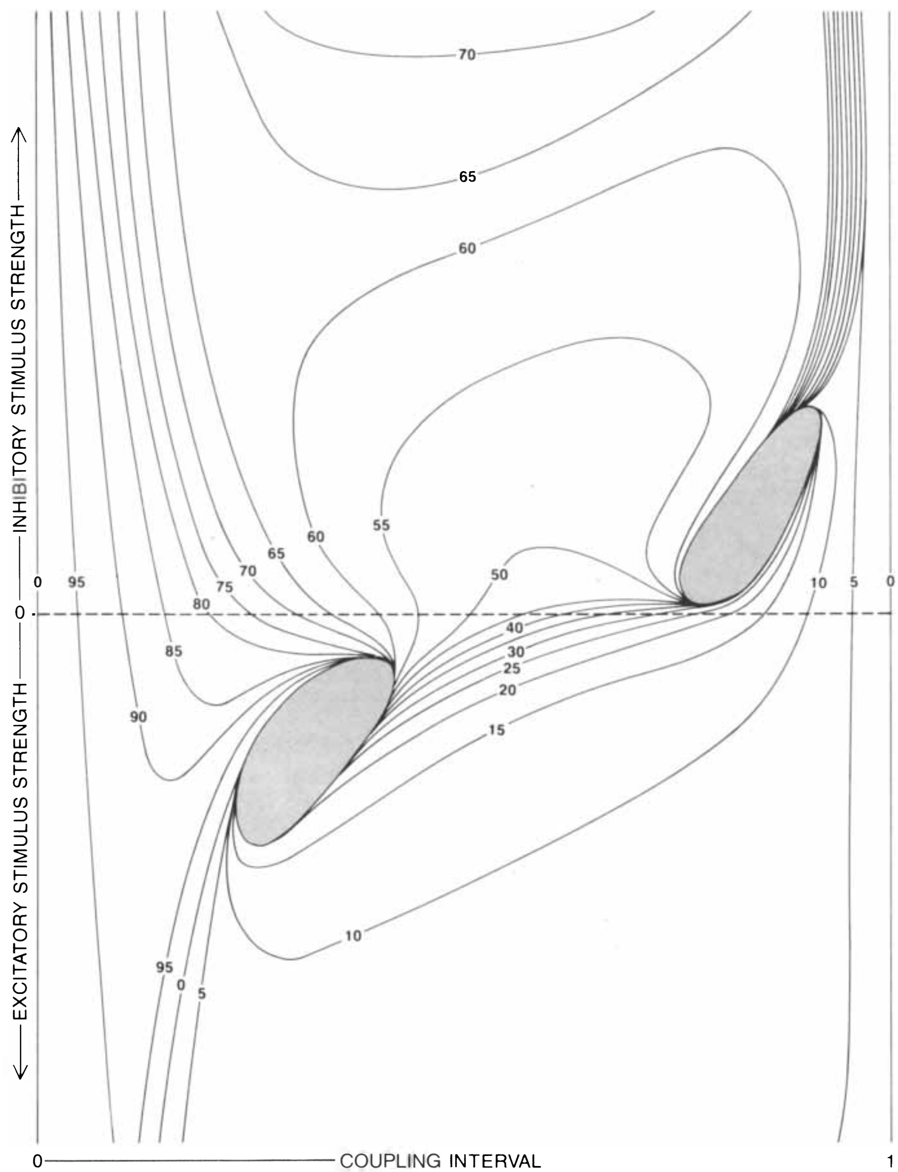
The topological theorem called the nonretraction theorem shows that it is impossible to completely retract the points on the surface of a disk onto the boundary of the disk and at the same time ensure that if two points are adjacent on the surface, they will also be adjacent when they are retracted onto the boundary. Only if at least one point is not retracted can the rest be rectracted while maintaining their original continuity. A physical example of the implications of the nonretraction theorem is the fact that a thin film of soap in a circular holder cannot shrink onto the holder unless the film is pricked. The topological significance of the pricking is that it removes one point from the disk, allowing all the other points to find appropriate places on the boundary and maintain their continuity. When the singular point is removed, the rest of the film can be continuously retracted onto the holder.

As we saw in the exercise with the wheel, the perimeter of the rectangular graph includes a complete cycle of colors. Some colors, however, appear more than once on the perimeter. Green, for example, appears at least four times: once on the lower edge and three times along the upper edge. It can be shown that if there is a full forward cycle on the perimeter of a rectangle, it is possible to inscribe in the interior of the rectangle a circle on which the colors of the cycle appear once each and in order. The nonretraction theorem implies that all the points inside the circle cannot be retracted onto the circle and also maintain their original continuity. It is possible to imagine the points inside the circle flowing smoothly toward the circle itself, perhaps in a pattern where each point moves toward the nearest point on the circle. No matter what the pattern of flow, however, there must be some point whose neighbors diverge: that point cannot itself find a place on the circle.

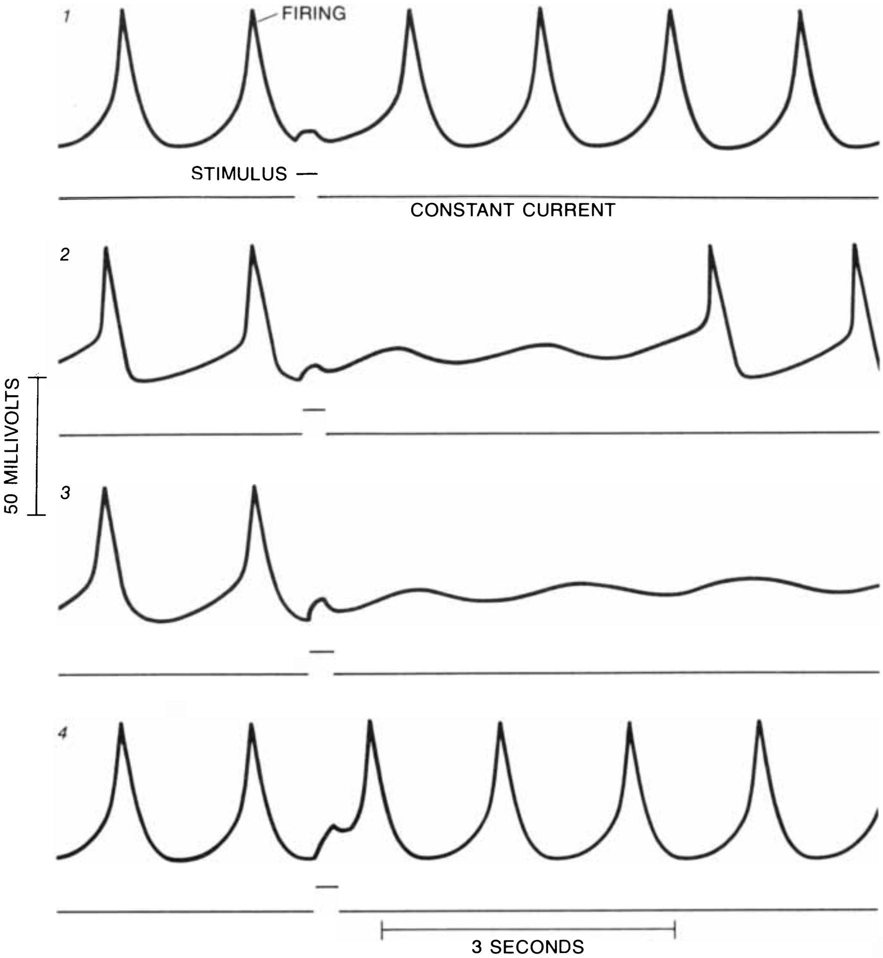
If some point inside the circle cannot be associated with a point on the circle, it follows that there is some point inside the circle (and hence inside the rectangle) that cannot be given a color. Since each point in the rectangular graph stands for a particular combination of stimulus strength and coupling interval and each color stands for a fundamental latency, the nonretraction theorem implies there is some stimulus that, if it is given at a particular time in the cycle, does not result in a fundamental latency.

The point in the rectangle that cannot be colored is called a singular point, and the corresponding combination of stimulus strength and coupling interval is called a singular stimulus. The singular stimulus will not be particularly large, since it falls somewhere within the rectangle. The upper edge of the rectangle stands for stimuli that yield strong rescheduling and the lower edge for stimuli that yield weak rescheduling. Therefore the singular stimulus must have a strength between the strengths of the stimuli that yield the two kinds of rescheduling.

As noted above, the fact that a stimu-



"BLACK HOLES" corresponding to singular stimuli (the ones that do not yield consistent latencies) were found by the author's student Eric Best in a computer simulation of the rhythmic firing of the squid giant axon. A mathematical model was employed to simulate the response of the axon to various stimuli. The stimulus strength and the coupling interval were varied, and the resulting latencies were assigned colors. The squid axon is sensitive to both excitatory and inhibitory stimuli. Therefore the simulation yielded two rectangles like the one in the illustration on page 153. The two rectangles can be joined along the line that represents a stimulus of zero strength. The result is the colored graph at the left. The black area in each rectangle represents stimuli after which the axon does not resume rhythmic firing. Surrounding each black area is a full cycle of colors in sequence according to the duration of the latency. The colors stream in an orderly way from one black area to the other. When points on the graph that have the same latency are connected, the result is a set of contour lines like the ones on a topographical map. The contours from the squid-axon simulation are shown above. When such contours represent latencies, they are called isochrons. The number on the isochron indicates a percentage of the natural period. The topology of the colored graph would be unchanged if the graph were laid out on a region of the heart. If this were done, the isochrons would represent the latencies of the cells in the region of tissue, that is, the interval after which each cell would fire.



SINGULAR STIMULI have been shown to exist in heart pacemaker tissue by Jalife with Charles Antzelevitch of the Masonic Medical Research Laboratory. They worked with Purkinje fibers from the heart of the dog. In each panel the upper trace represents the electric potential of the fiber; the peaks indicate firing. The lower trace represents the small constant current applied to the fibers to make them fire regularly. The raised segment of the lower trace indicates the application of a small additional stimulus lasting for 200 milliseconds. If the stimulus is applied soon after the beat, the only effect is to delay the next beat slightly (*1*). If the stimulus is given a little later, the fiber stops firing but resumes its rhythm (*2*). If the stimulus is given still later, the rhythmic firing stops and is not resumed (*3*). The time when a stimulus can have such an effect is called the vulnerable phase. Stimuli with the right strength given in the vulnerable phase correspond to the black regions in the illustration on the preceding two pages. If the stimulus is given after the vulnerable phase, the next beat is advanced slightly (*4*).

lus has a consistent fundamental latency means it yields an offsetting of the usual rhythm of the pacemaker without eliminating the system's natural periodicity. The nonretraction theorem implies that there is at least one stimulus that does not yield such an offsetting. Whatever the result of applying such a stimulus is, it does not preserve the pacemaker's normal rhythm. The nonretraction theorem does not indicate what the actual result of applying such a stimulus to a physiological system might be. The subsequent beats could be unpredictable or nonexistent. The particular combination of coupling interval and stimulus strength that yields such a result represents a mathematical singularity: a "hole" in the pattern of timing.

It should be noted that the finding of the singularity depends on the existence of strong rescheduling. If the pacemaker showed only weak rescheduling, the colors on the upper edge of the rectangle would vary through one full cycle, as those on the lower edge do. In the exercise with the wheel the movement along the top would cancel that along the bottom: the wheel would make no net forward motion. As a result the nonretraction theorem could not be applied.

The topological patterns of weak and strong rescheduling apply to simple systems as well as to complex ones such as the human heart. To determine whether the predicted singularity could actually exist in a physiological system my student Eric Best employed a computer to simulate the action of a very simple system capable of periodic "beating": a single nerve cell. The simulation was based on a well-known set of differential equations formulated by Alan L. Hodgkin of the University of Cambridge and Andrew F. Huxley of University College London to describe the propagation of the action potential along the giant axon (an exceptionally large nerve fiber) of the squid. Best used the Hodgkin-Huxley equations to construct a model of rhythmic firing in the squid axon. He then simulated the delivery of a range of stimuli while varying the coupling interval. The resulting latency was recorded.

The giant axon of the squid is sensitive to both excitatory and inhibitory stimuli. Therefore the simulation yielded two rectangles of the kind described above, one rectangle for each type of stimulus. The two rectangles can be joined along the horizontal axis corresponding to a stimulus of negligible intensity. The measured latencies were marked in the two rectangles and contour lines were drawn connecting points with the same latency, much as contour lines on a topographical map connect points with the same altitude. Contour lines that represent latencies are called isochrons. When the isochrons had been drawn in the rectangles, the contours were filled in with colors indicating fractions of the natural period.

The results of the plotting and coloring were quite exciting. In both rectangles the isochrons converge on a "black hole": a singularity where the latency is not defined. A full cycle of isochrons radiates from the perimeter of each black hole, a fact that is of considerable significance in connection with fibrillation. When the two rectangles are joined along the horizontal axis, the isochrons stream in an orderly way from one singularity to the other. Although the nonretraction theorem guarantees only the existence of a singular point, the black holes in the simulation were substantial in area. The reason is that the conditions of the simulation were adjusted to overcome the difficulty of finding a singular point. As we shall see, the conditions in the heart also lead to the expansion of the singularity into a substantial region.

Best's findings have since been confirmed in actual biological systems by other workers. John M. Rinzel of the National Institute of Arthritis, Metabolism, and Digestive Diseases and Rita Guttman of the Woods Hole Marine Biological Laboratory found both excitatory and inhibitory singularities in the squid giant axon. Jalife and Charles Antzelevitch of the Masonic Medical Research Laboratory found excitatory and inhibitory singularities that terminate the spontaneous firing of the Purkinje fibers of the dog and the sinus node of the cat.

It appears that the periodicity of any pacemaker capable of both weak and strong rescheduling can be eliminated by a brief stimulus with the right magnitude and timing. The contraction of the human heart, however, is spatially distributed, and its arrhythmias, or departures from the usual rhythm, often have the form of rapidly circulating waves rather than unpredictable timing or failure to fire. In particular, fibrillation is a spatial disorganization of the normal contraction. Can the discovery of topologically singular points, fascinating as it is, throw any light on the cause of fibrillation?

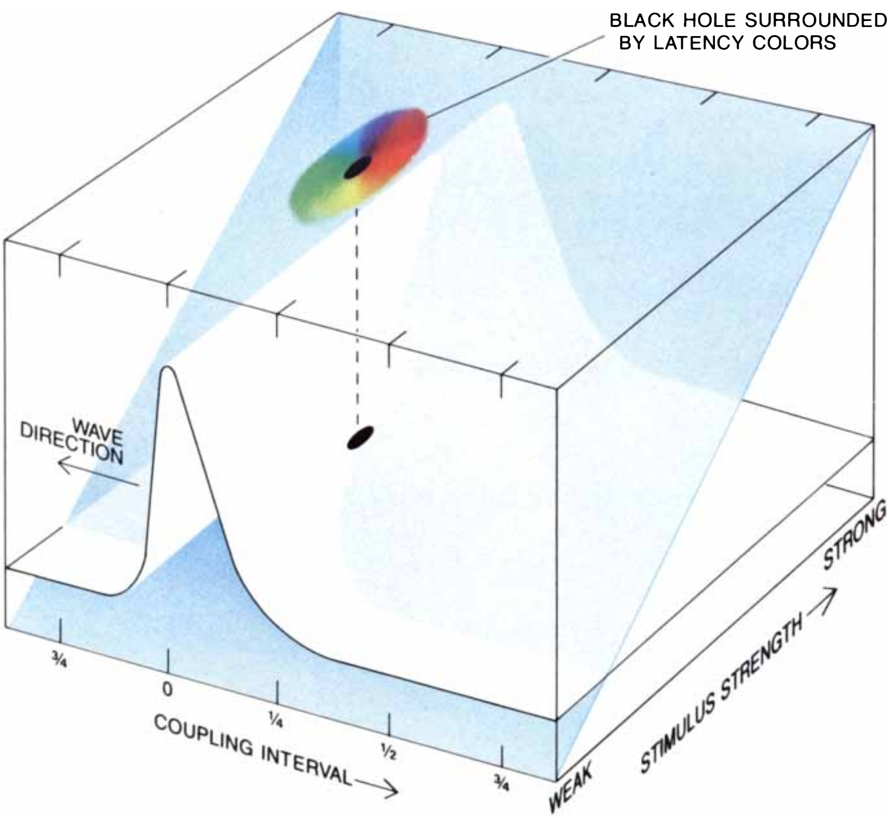
There is a connection between the singularities and circulating waves, but in order to understand the connection it is necessary to consider how waves of electrical impulses propagate in the heart. In heart muscle the electrochemical impulses that trigger a contraction propagate without attenuation. Each fiber restores the passing signal to its full strength before the signal is transmitted to a neighboring cell. Therefore if the signal could be deflected into a circular path, it could continue propagating around the heart indefinitely. The phenomenon is referred to in modern clinical terminology as reentry.

The continuous circulation of an electrical wave in the heart can become established only if the circular path of the wave is too long to be traversed within the refractory period that follows each action potential. If the period of the circulating wave were shorter than the refractory period, the wave would disappear after a single circuit because it would then encounter fibers still unable to fire. It is now known that waves can circulate on an appropriate path in the heart. Indeed, waves circulating around an obstruction such as the opening of a blood vessel or a patch of dead fibers have been shown to underlie several forms of arrhythmia.

It was long thought such waves could not circulate in unperforated tissue because they would quickly cancel themselves by propagating across the tissue in the center of the circular path. Recent work in which computers have been utilized to simulate the action of heart tissue has shown, however, that a wave can circulate even in continuous tissue such as healthy ventricular muscle. Gordon K. Moe of the Masonic Medical Research Laboratory, J. A. Abildskov of the University of Utah Medical Center at Salt Lake City and Werner C. Rheinboldt of the University of Maryland at College Park were the first to simulate an excitable medium like heart muscle and show that circulating, or rotating, waves can exist. V. I. Krinskii of the Institute of Biophysics of the U.S.S.R. Academy of Sciences and his colleagues have found that a single rotating wave, or rotor, can multiply, fragment and spread if the tissue has inhomogeneous areas such as those that frequently result from obstructions of the flow of blood to the cardiac muscle. Such rotors can circulate faster and in a much smaller space than had been thought possible. This work yields the best theoretical model of fibrillation.

Allessie and his colleagues Felix I. M. Bonke and F. J. G. Schopman were the first to observe rotors in actual heart tissue. The Dutch workers were able to induce a wave circulating about 10 times per second through an unperforated piece of rabbit tissue. They employed an array of moving electrodes to record the impulses in the tissue. The position of the wave front was plotted as a sequence of isochronal contours. As the wave rotates, the inner endpoint of each isochron describes an irregular loop less than a centimeter across. Similar isochronal maps of ventricular fibrillation have been obtained by M. J. Janse and F. J. van Capelle of the Hospital of the University of Amsterdam.

As I have mentioned, the nonretraction theorem does not predict what kind of arrhythmia might be caused by a stimulus in the singular region. The theorem predicts only that such a stimulus exists. The topological argument is therefore quite limited in predictive power. On the other hand, the topological principles have a more general interpretation than I have given them so far. The principles of pacemaker timing given above could apply to a spatially distributed system as well as to a single contracting mass. In a spatially extended system the isochronal contours in the rectangular graphs could be mapped onto a region of tissue. They would represent the electrical activity distributed over the tissue at a given moment rather than the accumulated results of stimu-



AUTHOR'S HYPOTHESIS of how fibrillation could begin includes two gradients: one of stimulus strength and one of timing. The transverse overlapping of the gradients results in a mathematical singularity in the heart. The illustration shows a small rectangular area of cardiac tissue in schematic form. The wave passing through the tissue from right to left is an impulse from the heart's pacemaker. As the wave passes, the fibers fire in turn. The fibers to the left of the peak fired almost a full period ago and will soon fire again. Those at the peak are just firing. The fibers immediately to the right of the peak have just fired. Thus all values of the coupling interval from zero to one period are laid out along the front of the rectangle. One of the nerves that modulate the action of the heart has endings that are densely concentrated near the rear of the rectangle. The endings become sparser with the distance toward the front of the rectangle. As a result a stimulus from the nerve will be strongest at the rear of the rectangle; its strength will decrease with the distance toward the front of the rectangle (*tilted plane*). Thus the latency diagram, with coordinates of stimulus strength and coupling interval, is laid out on the rectangle of tissue. It follows that somewhere in the rectangle there is a combination of stimulus strength and coupling interval that does not yield a consistent latency. Radiating from the singularity is the complete sequence of isochrons. When a stimulus from the nerve strikes the rectangle, the fibers around the singularity begin to fire in the order of their latencies, which corresponds to the sequence of isochrons. This is an ideal setup for starting a rotating wave.


IBM
IBM

The IBM Personal Computer XT. More power to the person.

Plenty of muscle. That's what the new IBM Personal Computer XT means to a person with heavyweight data to manage. Because one of the XT's many strong points is a 10-million-character fixed disk drive that helps give you the power to pump more productivity into your business.

What's so special about a fixed disk? Exactly that. It's *already* fixed inside the system, with the capacity to store the facts, figures, names and numbers you need to work with. (Rather than go from diskette to diskette, store up to 5,000 pages of text or up to 100,000 names and addresses in one place.)

Yet there's more built into the XT than its fixed disk. Reliability and quality are built in as well. Plus more than 30 years of IBM experience. A new level of price/performance. And a remarkable compatibility of both software *and* hardware with the original IBM Personal Computer. So, with the introduction of XT comes a special tool designed to help you be more productive in high-volume applications.

WHAT'S THE DIFFERENCE?

IBM Personal Computer	IBM Personal Computer XT
BASE SYSTEM*	**BASE SYSTEM***
User Memory 64KB (expandable to 640KB)	**User Memory** 128KB (expandable to 640KB)
Auxiliary Memory Up to two 5¼" 160KB/180KB or 320KB/360KB diskette drives optional	**Auxiliary Memory** One 10 megabyte fixed disk drive and one 5¼" 360KB diskette drive standard

*An expansion unit can also be added to both 16-bit (8088) systems for 6 more system expansion slots. Added to the IBM Personal Computer, it can house two 10 megabyte fixed disk drives. Added to the IBM Personal Computer XT, it can house one additional fixed disk drive for a total of 20 megabytes.

Another tool for modern times to keep you going strong.

To find out where you can see the IBM Personal Computers, call 800-447-4700. In Alaska or Hawaii, 800-447-0890. IBM®

lating a single fiber or a single pacemaker in many different trials over a period of time. The logic of the topological argument would be unchanged by such a transformation. Each point in the graph would still correspond to a combination of stimulus strength and coupling interval, but the point would also represent a position in the system. The color of the point would give the latency at that position. In addition suppose the spatially distributed rescheduling action of the human heart turns out to resemble the rescheduling of the squid axon as described in the Hodgkin-Huxley equations and simulated by Best. The crucial part of the resemblance would be the existence around each singularity of the full cycle of isochrons: the "rainbow" sequence of colors that surrounds the black hole.

How could such a rainbow be formed on the surface of the human heart and what would its consequences be? Consider a small rectangular piece of the muscular wall of the ventricle. Successive electrical waves from the pacemaker move continually across the tissue. For the sake of visualization it is convenient to imagine that the wave fronts are passing from right to left through the rectangle. Now suppose the length of the rectangle is roughly equal to the wavelength of the impulse. There will be one peak of the wave in the rectangle. The fibers at the peak are just firing. The fibers slightly to the left of the peak fired almost a full cycle ago and are about to fire again. The fibers to the right of the peak fired a short time ago as the peak passed; the fibers farther to the right fired longer ago. Thus all possible coupling intervals from zero to one natural period are laid out along the front of the rectangle.

Spread throughout the tissue are endings from several major nerves. The endings, however, are not evenly distributed. The strength of a nerve impulse is proportional to the density of the endings. Suppose the region at the back of the rectangle has a higher density of endings than the region at the front. An impulse from one of the nerves will therefore decrease in strength from the back to the front of the rectangle.

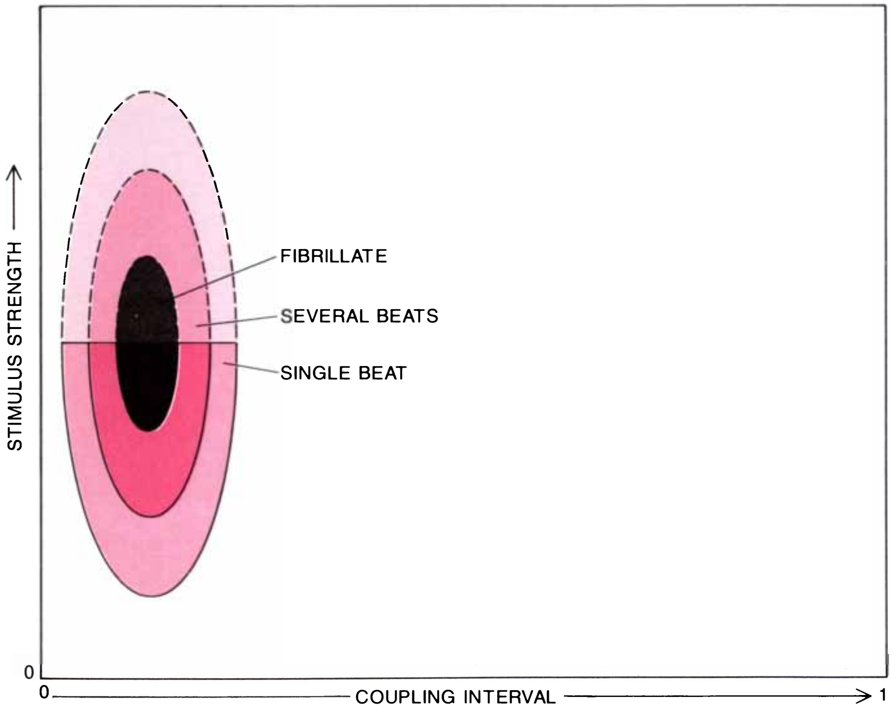
Now, when a nerve impulse enters the rectangle, the gradient of its strength will be transverse to the gradient of coupling intervals along the path of the pacemaking wave. The result of such an arrangement is that the rectangular graph of coupling interval and stimulus strength is laid out in an orderly way on the rectangle of heart muscle. If these conditions are met, then somewhere in the block there must be a black hole, a point where the coupling interval and the impinging stimulus do not give rise to any well-defined latency.

If, as I have hypothesized, the rescheduling action of the heart resembles in its essentials that of the squid axon, then arrayed around the black hole there is a rainbow of isochron colors standing for a complete sequence of latencies. Each isochron runs through cardiac fibers that fire after the same delay following the stimulus from the nerve. Thus the stimulus could yield a rotating sequence of firings around the point that represents the singularity.

As visualized here, the rainbow of isochron colors would be confined to about a square centimeter, which is roughly the area occupied by the rotors observed in heart tissue. It is difficult to imagine a better setup for starting a rotating wave than such a black hole surrounded by the rainbow of latencies. Furthermore, if the cardiac tissue includes many inhomogeneities, the rotor would be shattered into many small waves. By this process the rotor generated by a rescheduling singularity could ultimately be transformed into the catastrophic arrhythmia of fibrillation.

The gradients of timing and stimulus strength that exist in profusion in the human heart could make it easier for fibrillation to start. To understand how this could be so, consider the following argument. The topological argument guarantees only the existence of one singular point: a unique combination of coupling interval and stimulus strength. If the human heart were made up of an ideal tissue in which all impulses were transmitted instantaneously and all fibers contracted simultaneously, then it would be necessary to duplicate that combination exactly to cause fibrillation. The black hole would be an infinitesimally small point.

The heart of a large mammal such as man, however, is very different from such an idealized organ. As noted above, the timing of firings and the strength of stimuli vary considerably even across short distances. As a result many combinations of stimulus strength and timing can give rise to a rotor. For example, if a particular stimulus is too strong to generate a rotor at the point where it is applied, some distance away from that point it may have the right strength. Similarly, if the timing of the stimulus is not correct at the point of application, it may well be right a short distance away.



TARGET shows the stimuli that can cause fibrillation in intact hearts. The upper part of the graph is based on qualitative clinical data; the lower part is based both on such data and on more precise information from work with animals. In the outer ring are stimuli that cause an extra beat when they are applied to the heart. In the middle ring are stimuli that cause several extra beats. In the inner area are the stimuli that cause fibrillation. The topological argument presented here guarantees the existence of only one combination of stimulus strength and coupling interval that could start a rotating wave at a specific point on the heart. In the actual heart, however, at any one moment there are gradients of timing and stimulus strength. Such gradients increase the range of stimuli that can give rise to a rotating wave somewhere on the heart. If the applied stimulus is too large to give rise to a rotating wave, it could have the right strength a short distance away. If its timing is not right at the point of application, it could be right a short distance away. The many combinations of stimulus strength and timing that could give rise to a rotating wave in an intact human heart are shown in the bull's-eye of the target.

Imagine again a graph with the coupling interval on the horizontal axis and the stimulus strength on the vertical axis. On the graph are plotted the stimuli that can cause fibrillation. For the heart composed of an ideally uniform tissue the target created in this way is a single point. For the actual heart the target is enlarged considerably and is therefore much easier to hit. Work with animals shows there is indeed a substantial region of the graph where fibrillation can result; stimuli outside the target region do not cause fibrillation. The principle is exemplified in cases of accidental death by electrocution. It has also been employed to stop the beating of the heart during delicate open-heart operations.

In his final paper George Mines made three generalizations about fibrillation that have since been borne out by much clinical experience. First, fibrillation can be triggered by an untimely electrical impulse, that is, one coming in the vulnerable phase. Second, fibrillation frequently involves the reentry of circulating impulses. Third, fibrillation is favored by almost any departure from spatial uniformity in the target tissue or in the stimuli from the nerves that infiltrate the tissue. The topological argument given here elucidates the connections among these three empirical generalizations.

For the moment this is as far as either clinical observation or topology can take us. If the topological mechanism proposed here is correct, its clinical implications could be somewhat bleak. The heart is continually bombarded by electrical impulses from many sources. The possibility of a singularity resulting from such impulses cannot be eliminated without eradicating the strong rescheduling property of the pacemaker's membranes. If this is not possible or desirable, then the challenge to preventive medicine is to make the black holes as small and as inaccessible as possible.

Some of the unanswered questions suggested by the topological analysis may well lead to the most interesting clinical results. It appears that all human hearts are subject to bombardment by electrical impulses. Only some hearts, however, develop fibrillation. Under what conditions does a rescheduling singularity turn into a rotor? When is a rotor transformed into many scattered smaller waves? When does it simply die out into harmless synchronized activity? Can the conditions necessary to transform the singularity into fibrillation be identified and possibly controlled by drugs that alter the electrical properties of the cardiac tissue? These are questions that have long concerned cardiologists. By calling attention to them in a mathematical context topology may ultimately make a significant contribution to overcoming sudden cardiac death.

THE AMATEUR SCIENTIST

What causes the "tears" that form on the inside of a glass of wine?

by Jearl Walker

When wine is left standing in a glass for a few minutes, a curious array of drops is likely to appear on the inside of the glass above the surface of the liquid, particularly if the wine has a fairly high alcoholic content. The phenomenon is known as "tears of strong wine." It was examined as early as 1855 by James Thomson, a British engineer and physicist, who concluded that it did not arise from the condensation of water on a cool surface, since the drops appeared even when the glass was at room temperature. He thought the drops formed as a result of variations in the surface tension of the wine as alcohol evaporated from a thin film of wine on the wall of the glass.

The formation of the tears has since figured in the general study of circulation patterns in liquids varying in surface tension. Although such motions have been studied for more than a century, the physical events are still not understood in detail. The motions are interesting to a physicist because they serve as a tool in the study of surface tension and the stability of circulation patterns in liquids. They are also important because they can be found in many industrial and biological processes.

Although Thomson's explanation for the tears of strong wine was basically correct, his work was apparently ignored until 1869, when the entire subject of motion resulting from surface tension was reviewed by Gustav L. van der Mensbrugghe in a Belgian journal. The priority of the research was challenged two years later in Italy by Carlo Marangoni. Thomson's early work was virtually forgotten, and Marangoni's name is now attached to liquid motions driven by surface tension.

Thomson's explanation was a simple one based on the observation that when alcohol is added to water, the surface tension of the water decreases. Hence wine has less surface tension than pure water. If the wine is exposed to air, the alcohol in it continuously evaporates, creating areas of higher tension on the surface. Such an area pulls on the adjoining liquid and starts it in motion.

The more visible formation of tears on the wall of the wineglass is due to the same kind of variation in surface tension. The surface of the liquid is curved at the wall because the surface tension at the interface of the wine and the glass pulls the wine up the wall a little way.

Alcohol evaporates from this film, increasing the surface tension of the liquid there and causing more wine to be pulled up. Thomson noticed that the additional liquid tends to form a thick ring at the top of the film. As the alcohol evaporates and the surface tension in the ring increases, the liquid begins to contract into small drops. More liquid is pulled up. More alcohol evaporates. The drops get larger. Finally a drop gets heavy enough to slide back down into the wine. Soon another drop forms in its place because evaporation continues in the film and fresh liquid is constantly pulled up. The process continues until so much alcohol has been removed that the variation in surface tension gets too small to keep the cycle going.

Thomson tested the role of evaporation by corking a partly filled vial of wine. After shaking the mixture he examined the film on the interior wall. No tears of wine formed and no motion was apparent in the film except normal drainage. Then he removed the cork, so that fresh air entered the vial. Under these conditions, he reported, "a liquid film is instantly to be seen creeping up the interior of the vial with thick or viscid-looking pendent streams descending from it like a fringe from a curtain." When the cork was in place, shaking so saturated the air in the bottle with alcohol that no more evaporation was possible. Fresh air restored both evaporation and the cycle whereby liquid was driven into motion by the variation in surface tension.

I demonstrated the formation of tears with several kinds of alcoholic beverage. First I poured a small amount of 94.6-proof gin into a watch glass. (The concentration of alcohol by volume is half the proof number.) Four tears formed almost immediately near the rim of the glass. The region between a tear and the rest of the gin was visibly wet and occasionally showed motion. I doubt that I saw gin flowing up the sides of the watch glass, since gin is transparent; it is more likely that I was seeing dust motes caught up in the flow. To improve the visibility I dusted the surface of the gin with lycopodium powder. The microscopic spores of the powder rest on top of the gin instead of sinking. Talc and other household powders that are not readily wetted by water and alcohol can serve in place of lycopodium.

With the powder in place as a tracer I could follow much of the flow of gin up the side of the glass. In general the motion was irregular, but below a spot where a tear formed the flow was mainly upward. (The irregular flow elsewhere results because the film is evaporating and thus varying in surface tension from place to place.)

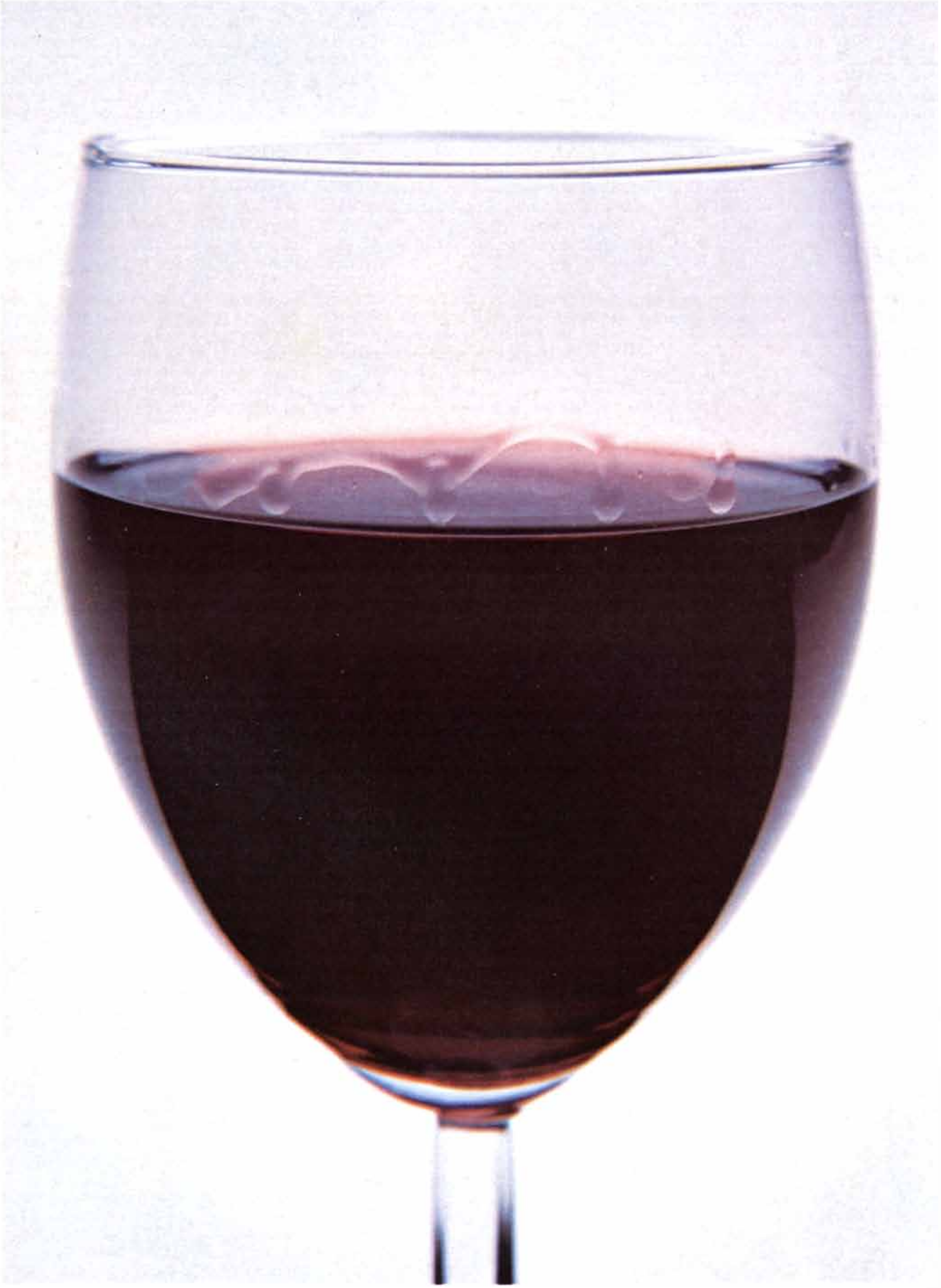
A tear would eventually slide abruptly down into the gin or would descend gradually until it touched it. In the latter case the drop jerked a little way back up the wall after touching the gin. The reason is mainly that the drop loses liquid when it touches the surface. It is then lighter and more responsive to the surface tension pulling upward than to its weight pulling downward. The jerk is also promoted by the sudden difference in surface tension at the interface of the drop and the gin.

A drop touching the gin injected a small jet of liquid into the gin. The jet is visible because it has an index of refraction different from that of the rest of the gin and therefore bends the light rays crossing through it. The difference in refractive index probably results from two factors: the jet has both more alcohol and greater density than the gin around it. The jet is denser because the liquid has been cooled by the evaporation that causes tears to form.

I next poured about a cup of 80-proof rum into a porcelain bowl and added a small amount of chocolate flavoring at one spot near the edge of the liquid. Beautiful purple patterns raced through the rum. The colored lines were quite lively, apparently driven by strong variations in the surface tension at the top of the mixture and by circulation systems within it.

I put the bowl in a warm oven (its temperature somewhat below 93 degrees Celsius, or 200 degrees Fahrenheit). Tears soon formed on the inside surface of the bowl. With a flashlight and a small magnifying glass I examined the edge of the rum. Small, undissolved particles left from the flavoring served to trace the liquid motion. The particles darted toward and away from the edge, revealing a vigorous and complex circulation of liquid there.

Food coloring turned out to be a bet-



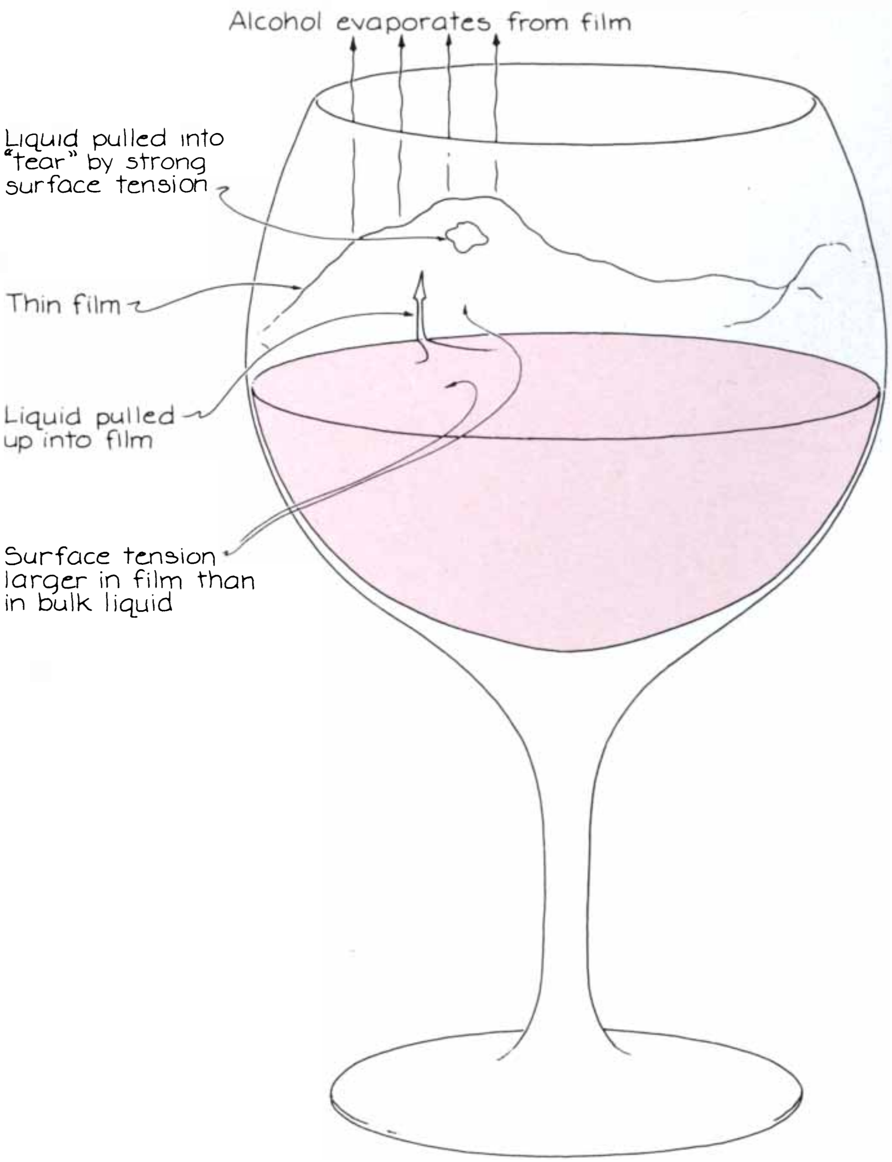
Movements caused in vodka by surface tension, as traced by red wine ter tracer. A small drop added to isopropyl alcohol in a porcelain cup colored one section of the perimeter. I watched as a small colored section of the alcohol reached the perimeter of the cup and began climbing through the invisibly thin film of alcohol and water lining the wall, finally entering a tear that had already formed on the wall.

I next made a simple but crucial test of Thomson's hypothesis that the drops appear only because of a difference in the surface tension in the film on the wall. Suppose the alcohol contains no water. Then the surface tension would hardly vary as the alcohol evaporated on or near the wall. Indeed, it would not vary at all except for the changes caused by the fact that the liquid cools as it evaporates.

Working with a pair of identical cups, I poured a quantity of 70 percent isopropyl alcohol into one cup. In the other I poured half as much alcohol and then added enough water to bring the surface level up to that of the first cup. I gave both cups a brief swirl to wet the sides. Many tears formed quickly above the diluted sample. Only a few small tears appeared above the concentrated sample. Thomson was right: water is necessary for a variation in surface tension.

Any factor that promotes evaporation also aids the formation of tears. A warm environment, direct sunlight and a wide, shallow container help. I made a large-scale demonstration of tear formation employing a serving platter with sides that curved upward. A tall glass with a small amount of liquid works poorly because the evaporation is slow, unless you first wet the entire inside wall. Then when the alcohol is poured into the glass, the liquid film rapidly climbs the wall, rising about as high as the remaining wetness. The appearance of tears several centimeters above the alcohol mixture seems almost magical. A watcher might say the tears are actually drops of condensation. To prove their origin you can color the liquid at the bottom of the glass. The tears will be colored too.

You might like to investigate the formation of tears with various alcoholic drinks. Beers made in the U.S. do not work, at least in my experiments, apparently because they contain too little alcohol. Can you find liquids other than alcohol and water that form tears?



The forces that create "tears of strong wine"

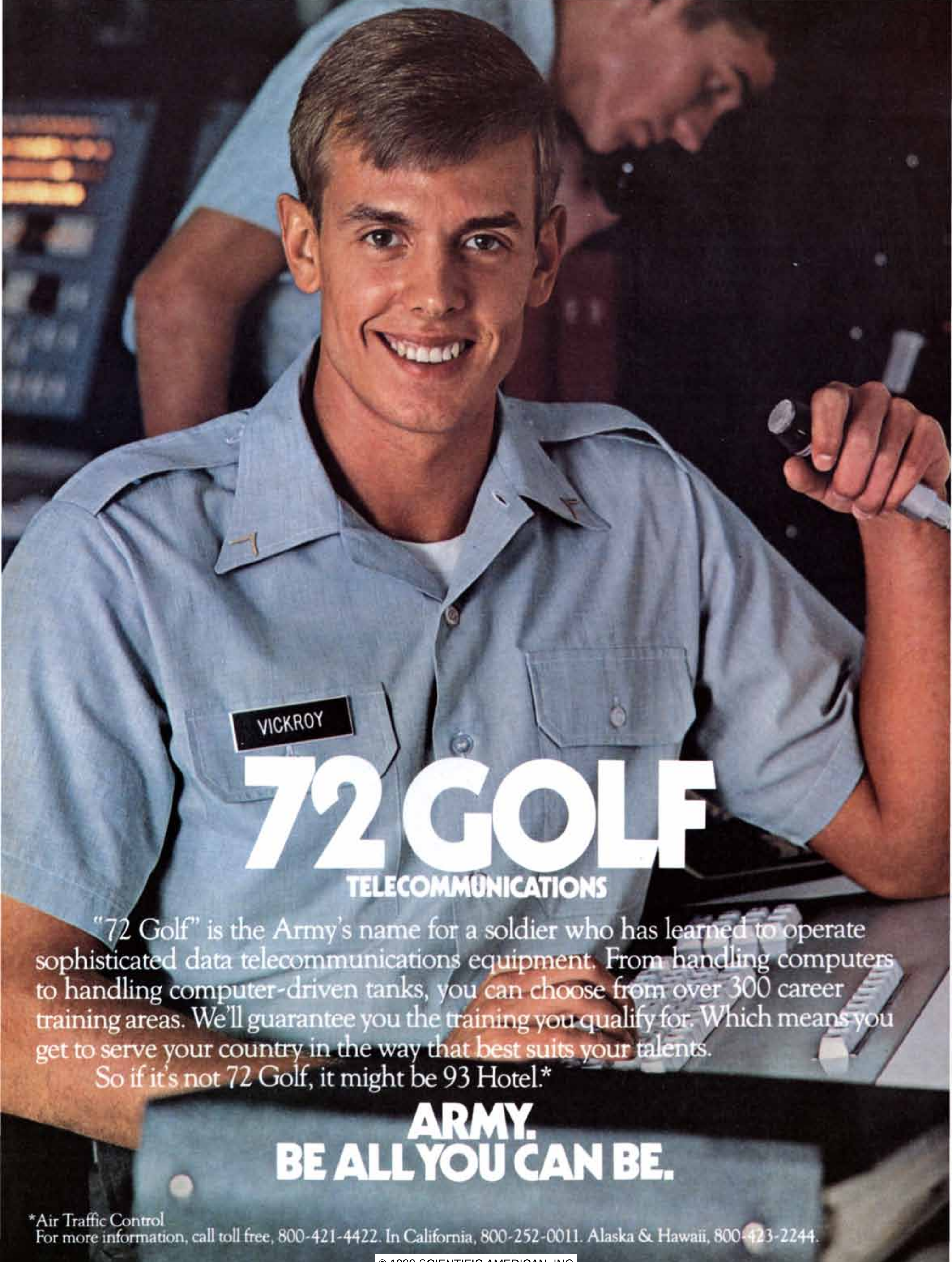
As I was investigating the tears of strong wine I was writing the article about Middle Eastern coffee that appeared in this column last month. The drink consists of water, sugar and finely ground coffee grains, all brewed together and served in a cup along with the grains. I prepared so much of the stuff that I began to pour it into bowls instead of drinking it. When I was cleaning up one morning, I discovered a pattern in a bowl of coffee left from the previous night. Along the edge of the liquid, just below the surface, lay a neat array of dark lanes of coffee silt. They and the intervening clear lanes were each a few millimeters wide.

I was stumped by the pattern. Not one of the hundreds of unfinished cups of coffee I had left lying around had ever shown it. Is Middle Eastern coffee somehow special? I prepared another batch and left it in the same kind of porcelain bowl. I saw nothing interesting in the period of rapid evaporation from the hot surface, but in rechecking over the next several hours I began to notice the same pattern developing. Although I sat patiently with a flashlight and a magnifying glass, I could not actually see the lanes growing. Since I had tears of strong wine on my mind, I wondered if the patterns could be related. No, the liquid showed little movement up the side of the bowl.

Perhaps this pattern has been reported before, but I can find no mention of it. A silt such as the residue in Middle Eastern coffee is required, as is sugar. I prepared two bowls of the coffee, one bowl with sugar and one without. The next morning the one with sugar had a fine pattern and the other did not.

Evaporation is also needed. I put a bowl of the coffee in a warm oven and quickly got the pattern. Erasing it with a gentle swirl, I covered the bowl with plastic food wrap. After a while I shone a flashlight through the plastic. The pattern was there. I gave the bowl another swirl and waited again. The pattern would not re-form.

Apparently the pattern developed the first time because the water was able to evaporate into the air above the coffee. Thereafter the saturation of the air with water vapor prevented further evaporation. The high concentration of water vapor trapped on the inside of the plastic wrap was demonstrated when I opened the oven door for a look; the cool air from the kitchen caused rapid


VICKROY
72 GOLF
TELECOMMUNICATIONS
"72 Golf" is the Army's name for a soldier who has learned to operate sophisticated data telecommunications equipment. From handling computers to handling computer-driven tanks, you can choose from over 300 career training areas. We'll guarantee you the training you qualify for. Which means you get to serve your country in the way that best suits your talents.
So if it's not 72 Golf, it might be 93 Hotel.*
ARMY.
BE ALL YOU CAN BE.
*Air Traffic Control
For more information, call toll free, 800-421-4422. In California, 800-252-0011. Alaska & Hawaii, 800-423-2244.

condensation on the inside of the plastic.

I then set about investigating the formation of the pattern by preparing many samples of Middle Eastern coffee and pouring them into a wide beaker and an assortment of watch glasses. The vertical wall of the beaker enabled me to examine the sedimentation of the grains in the coffee with the aid of a strong flashlight. I saw nothing interesting in the beaker and expected nothing but the usual pattern in the watch glasses. Surprisingly, every batch in the watch glasses failed to form a pattern. I substituted plastic lids shaped like the watch glasses but still could not produce the patterns. I was stymied.

Back with the porcelain bowls, I began to search for circulation patterns near the edge of the liquid by adding some kind of tracer. In one trial I injected a small amount of food coloring by inserting the tip of a hypodermic syringe just below the coffee level. In another trial I dusted the top surface with lycopodium powder. I could not discern any circulation, but when I left the bowl at room temperature overnight, the pattern appeared.

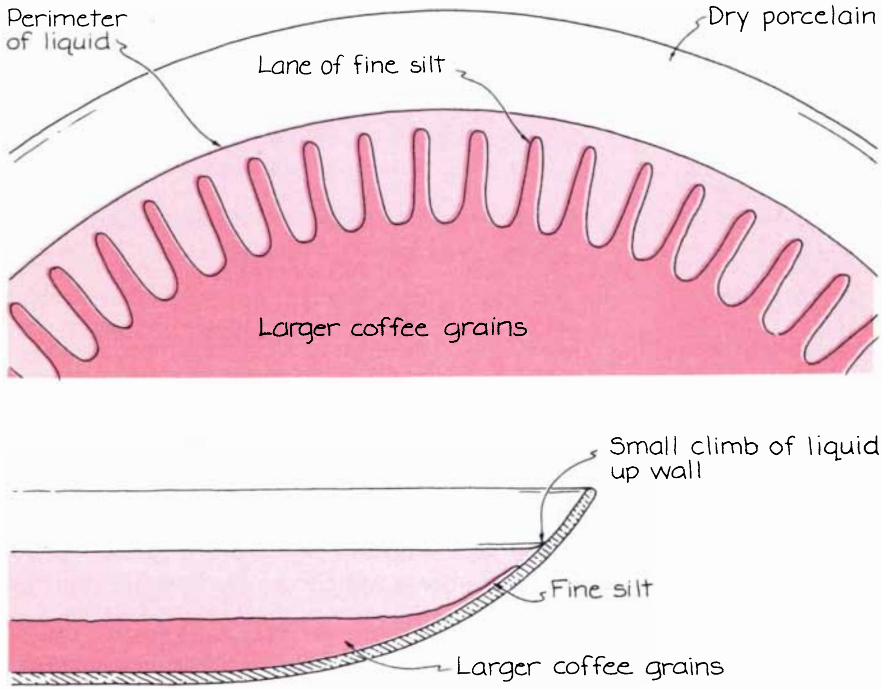
I tried for similar patterns with other mixtures. Neither tea leaves nor grains of sand produced a pattern in sugar water. Apparently the pattern requires fine silt lying on the bottom; tea leaves and sand are too big.

With a normal mixture of Middle Eastern coffee in a bowl I scraped the silt at one side of the bowl and then waited for the pattern to form. It did form but not in the area I had cleared. This result gave me a useful clue to the nature of the pattern. Apparently the silt had to be near the edge of the liquid. Whatever causes the pattern does not appreciably transport the silt up the wall of the bowl and toward the liquid's edge.

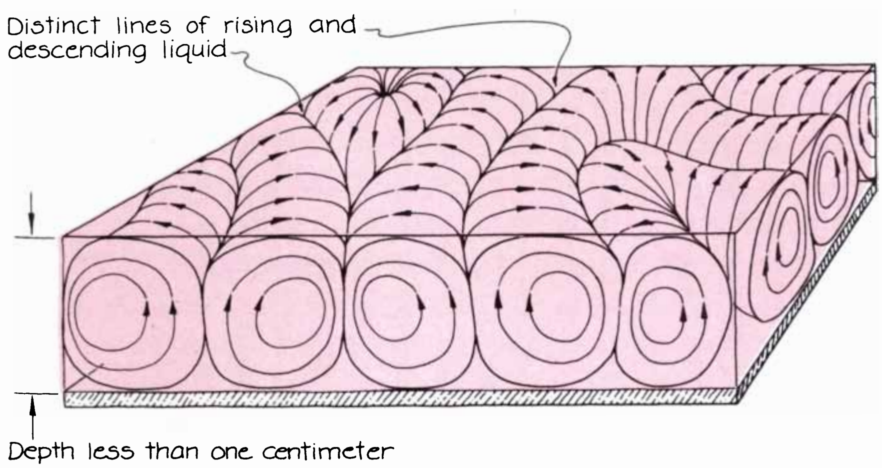
I made a mixture of Middle Eastern coffee and gin with a small amount of red food coloring as a tracer. The alcohol was poured in slowly after the coffee had cooled for a while. Soon tears formed on the upper reaches of the wall but the coffee pattern failed to appear.

From this test, however, I received my next clue to the cause of the pattern. Whenever a tear slid down the wall and into the drink, it cleared a lane through the sediment just below the spot where it entered the liquid. In this way the drops soon produced a pattern resembling the patterns I had seen only with the coffee. The spacing in the pattern was about 10 times too large, and the pattern varied a lot because the points where the tears entered the liquid shifted.

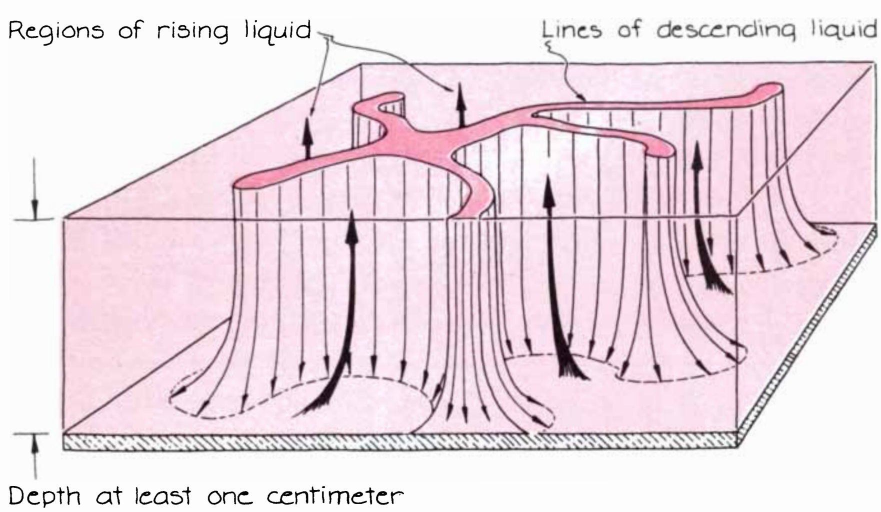
Perhaps a circulation system at the shallow edge of normal Middle Eastern coffee generates the pattern in the silt. Two mechanisms can drive such a system in a naturally evaporating liquid. (Natural evaporation takes place without an additional source of heat, such as



A serrated pattern on the edge of silt from Middle Eastern coffee



Circulation in a shallow liquid



Circulation in an evaporating liquid

a stove, to supply the energy for vaporization.) The liquid might move because of the variation in surface tension over its surface. It might also move because the liquid in an upper level becomes denser than the liquid under it. A circulation system is created when the conditions altering surface tension and density recur as fresh liquid is brought in.

For example, circulation cells can be seen in a cup of hot coffee. Hot liquid from the bottom rises to the top, where it cools by evaporation as it flows a short distance over the surface. Then it sinks in narrow, crooked lines to the bottom. The cycle continues until the coffee gets too cool to sustain it. The cells in hot coffee are randomly shaped and constantly changing. In some circumstances and with certain liquids they have ordered shapes and can be quite beautiful.

Natural evaporation in liquids drives circulation cells because the evaporative cooling increases the density and surface tension of the liquid. Either the increased surface tension pulls in fresh liquid or the denser liquid sinks and is replaced by fresh liquid. It is usually difficult to determine which of these mechanisms is primarily responsible for a circulation system.

According to research published a number of years ago by J. C. Berg and Michel Boudart of the University of California at Berkeley and Andreas Acrivos of Stanford University, neither mechanism operates unless the liquid is deep enough. For water the required depth is one centimeter. Shallower water displays no circulation systems; deeper water has the large circulation cells that can be seen in hot coffee.

Several other liquids investigated by Berg, Boudart and Acrivos had circulation systems even when they were quite shallow. The investigators monitored each liquid as the depth was increased. In all liquids except water circulation systems first appear when the depth of the liquid is about two millimeters. The systems are "two-dimensional, wormlike roll cells." As the depth is increased to one centimeter the cells increase in width and the lines of descending liquid become more distinct. At a depth of one centimeter the circulation cells resemble those in hot coffee. The regions where the liquid rises to the surface are then indistinct.

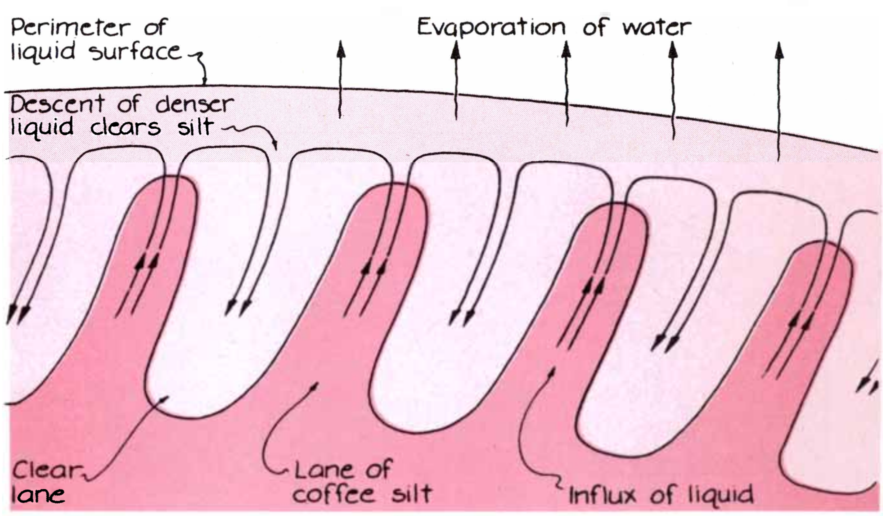
Why is water so different? Trace contaminants from the air, the container or even the experimenter spread a monolayer (a layer one molecule thick) over the surface, preventing any circulation generated by variations in surface tension. Then only circulation generated by variations in density is possible. Apparently a monolayer of contamination can stabilize water shallower than one centimeter.

This finding disappointed me. It implied there should be no circulation system near the shallow edge of evaporating Middle Eastern coffee. Yet the coffee silt revealed a pattern regular enough to have been made by hand. The coffee was certainly not pure water. It contained a lot of sugar and oil, and it was exposed to the air for hours. The top surface was undoubtedly coated with contamination at least one molecule thick. In some places I could even see tiny pools of oil.

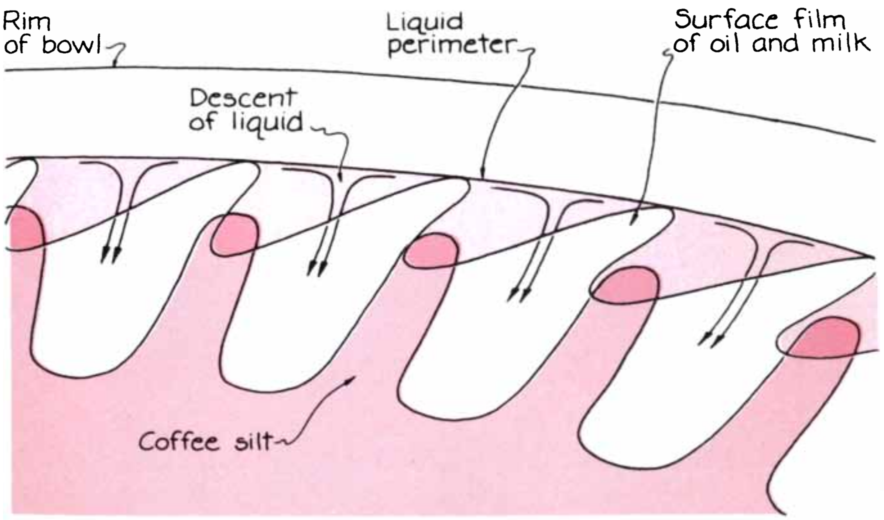
Ultimately I succeeded in putting all the clues together. The pattern requires evaporation, but because of contamination it cannot arise solely from a variation in surface tension. The pattern appears in the fine silt just below the edge of the liquid. If the silt is too far below the edge, no pattern develops. It forms only on a surface of moderate slope, not on the vertical wall of a beaker and not on the shallow slope of a watch glass.

Contamination over the surface of the coffee surely interferes with the natural evaporation of the water from the bowl but may not reduce the evaporation at the edge as much. When some of the water evaporates from an area of liquid at the edge, the remaining liquid is made denser because the concentration of sugar and oil is then higher. This section sinks. As it descends it sweeps out a lane in the silt. Other liquid flows in to replace it. Since the circulation depends on evaporation, it is normally quite slow. Part of the flow can be toward the edge along a lane of silt. If the surface is not too well stabilized by contamination, part of the flow can be along the surface.

As evaporation continues more of the clear lanes are swept clean of silt. Some of the intervening lanes of silt may be extended toward the edge by the gentle circulation system, but the diffuse flow does not carry the silt far. The pattern is enhanced by the sugar, which makes the section of liquid left by the evaporation of water near the edge denser. The pattern appears only on moderate slopes. If the slope is too steep, the circulation has no chance to sweep out lanes of silt. If



How the circulation in coffee silt causes serrations



The formation of a pattern in a surface film the slope is too gentle, the liquid layer near the edge is too shallow for circulation. I suppose the layer of contamination is then too close to the layer of silt to allow any stabilized flow. The descent of the dense liquid from the edge may also be too gradual for the formation of a pattern.

I did one final experiment, mixing a teaspoon of milk into my usual Middle Eastern coffee. The oil in the drink and in the milk made a film over the surface. As a pattern developed in the silt along the edge, this film formed into a similar pattern. Above the silt lanes was the film, extending to the liquid edge. Above the clear lanes the surface had no film. This arrangement seems to fit exactly the circulation system I had visualized at the edge of the evaporating coffee.

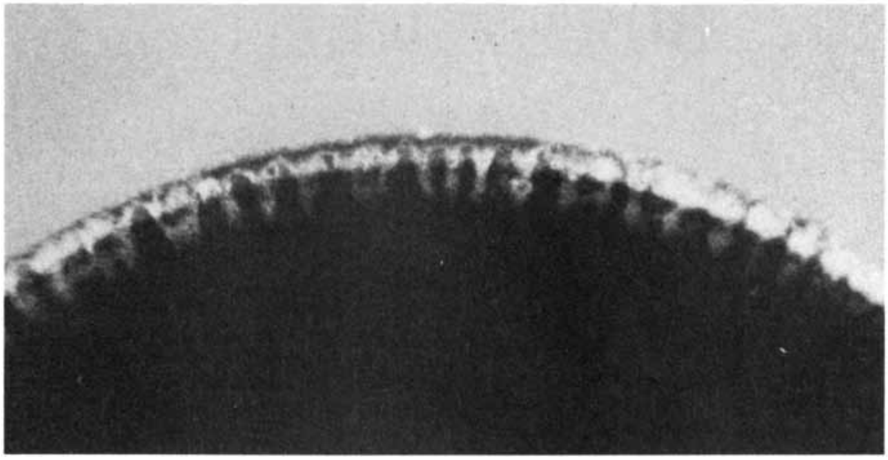
Another observation may support this analysis. I had made instant coffee with milk (the drink is often called white coffee) and had not finished it. The next morning a pattern of white lines streaked the liquid surface. Another pattern of distinct radial lines appeared in the liquid left in the spoon with which I had stirred the coffee. The depth of the liquid did not seem to matter. It was about a centimeter in the cup but only a few millimeters in the spoon.

To investigate these patterns I made another preparation of white coffee and poured the mixture into several watch glasses and one metal spoon. I dusted one watch glass with lycopodium powder. Off and on for about 10 hours I checked the samples, particularly near the edge, where I employed a strong flashlight and a magnifying glass. Large patterns of white lines appeared after several hours of evaporation, but I could detect no consistent motion of the lycopodium powder. In one watch glass the pattern resembled the veins in a leaf. In another it consisted of parallel lines across the entire surface. One pattern had what looked like an upswelling because the white lines extended outward from a single point. If I gently disturbed one of these patterns with the tip of a spoon, it usually reestablished itself within an hour or so.

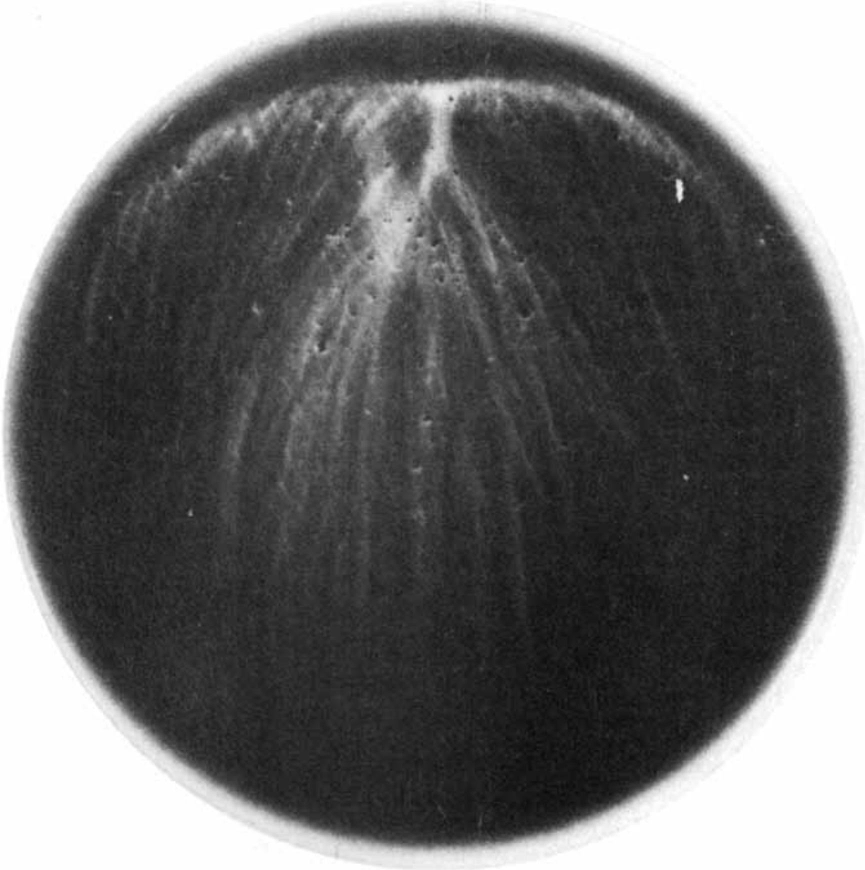
Near the edge of the liquid in a watch glass fine-scale patterns appeared. The hair-thin lines of dried milk were roughly perpendicular to the edge. In the spoon the shallow coffee developed the same kind of fine-scale patterns along the edge of the liquid until evaporation had reduced the depth to less than a millimeter. Sometimes the radial lines in the spoon became pronounced and more widely spaced. The coffee itself was apparently not necessary, but it did facilitate my observations. Thin layers of milk evaporating from a watch glass displayed the same kind of fine-scale patterns along the liquid edge.

I believe the two types of patterns reveal the circulation systems in the evaporating liquid. The larger patterns in coffee cups resemble the wormlike cells. As the liquid flows across the surface and then descends, milk particles collect over the lines of descent. Hardening there, they mark the paths of descent of the coffee. In addition they may help to stabilize the location of the circulation cells, otherwise the cells would be no more stable than those in hot coffee. The cells also appear to be coupled to the gentle circulation of air over the liquid surface. When a bowl of coffee is exposed to the cool air moving downward from a cold window, the same milk patterns develop.

The fine-scale patterns along the edge resemble the patterns in coffee silt and may be due to the same circulation system. Once these milk lines are established, they might stabilize the circulation along the edge. As the water continues to evaporate from the liquid, the white lines harden on the container. When white coffee has fully evaporated from a watch glass, I can hold the glass in sunlight and see a beautiful radial pattern of lines.



A camera's view of a serrated pattern in coffee silt



How a pattern in white coffee looks to the camera

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

THE PRISONER'S DILEMMA. Anatol Rapoport and Albert M. Chammah. University of Michigan Press, 1965.

TWO-PERSON GAME THEORY. Anatol Rapoport. University of Michigan Press, 1966.

THE EVOLUTION OF COOPERATION. Robert Axelrod and William D. Hamilton in *Science*, Vol. 212, No. 4489, pages 1390–1396; March 27, 1981.

SMART WEAPONS IN NAVAL WARFARE

INTRODUCTION TO SHIPBOARD WEAPONS. Carl D. Corse, Jr. Naval Institute Press, 1975.

SHAPING THE GENERAL-PURPOSE NAVY OF THE EIGHTIES: ISSUES FOR FISCAL YEARS 1981–1985. Congressional Budget Office. U.S. Government Printing Office, 1980.

THE FALKLANDS CAMPAIGN: THE LESSONS. Secretary of State for Defence. Her Majesty's Stationery Office, 1982.

VIBRATIONS OF THE ATOMIC NUCLEUS

NUCLEAR SPECTROSCOPY. Aage Bohr and Ben R. Mottelson in *Nuclear Structure*, W. A. Benjamin, Inc., 1975.

NUCLEAR VIBRATIONS. G. F. Bertsch in *Nature*, Vol. 280, No. 5724, pages 639–644; August 23, 1979.

GIANT RESONANCES IN NUCLEI. J. Speth and A. van der Woude in *Reports on Progress in Physics*, Vol. 44, Part 3, pages 719–786; 1981.

DAMPING OF NUCLEAR EXCITATIONS. G. F. Bertsch, P. F. Bortingnon and R. A. Broglia in *Reviews of Modern Physics*, Vol. 55, No. 1, pages 287–314; January, 1983.

MICROBODIES IN THE LIVING CELL

PEROXISOMES (MICROBODIES AND RELATED PARTICLES). Christian de Duve and Pierre Baudhuin in *Physiological Reviews*, Vol. 46, No. 2, pages 323–357; April, 1966.

LOCALIZATION OF NINE GLYCOLYTIC ENZYMES IN A MICROBODY-LIKE ORGANELLE IN *TRYPANOSOMA BRUCEI*: THE GLYCOSOME. Fred R. Opperdoes and Piet Borst in *FEBS Letters*, Vol. 80, No. 2, pages 360–364; August, 1977.

THE HYDROGENOSOME. Miklós Müller in *Symposia of the Society for General Microbiology*, Vol. 30, pages 127–142; 1980.

METABOLIC PATHWAYS IN PEROXISOMES AND GLYOXYSOMES. N. Edward Tolbert in *Annual Review of Biochemistry*, Vol. 50, pages 133–157; 1981.

PEROXISOMES AND GLYOXYSOMES. Edited by Helmut Kindl and Paul B. Lazarow in *Annals of the New York Academy of Sciences*, Vol. 386, pages 1–550; 1982.

MODERN PORK PRODUCTION

SWINE SCIENCE. M. E. Ensminger. Interstate Printers and Publishers, Danville, Ill., 1970.

SWINE PRODUCTION IN TEMPERATE AND TROPICAL ENVIRONMENTS. Wilson G. Pond and Jerome H. Maner. W. H. Freeman and Co., 1974.

SWINE PRODUCTION. Jake L. Krider and William E. Carroll. McGraw-Hill Book Company, 1975.

SILENT PARTNERS: FARMERS, THEIR ANIMALS AND CONSUMERS. W. G. Pond, Lacuna Books, Interlaken, N.Y., 1978.

NEW INORGANIC MATERIALS

STRONG SOLIDS. Anthony Kelly. Oxford University Press, 1973.

INORGANIC POLYMERS. N. H. Ray. Academic Press, 1978.

MAKING THE MOST OF MATERIALS: MATERIALS AND ENERGY CONSERVATION IN THE MATERIALS PROCESSING INDUSTRIES. Science Research Council, 1979.

FLEXURAL STRENGTH AND POROSITY OF CEMENTS. J. D. Birchall, A. J. Howard and K. Kendall in *Nature*, Vol. 289, No. 5792, pages 388–389; January 29, 1981.

FLEXIBLE INORGANIC FILMS AND COATINGS. D. G. H. Ballard and G. R. Rideal in *Journal of Materials Science*, Vol. 18, No. 2, pages 545–561; February, 1983.

COMPUTER-INTENSIVE METHODS IN STATISTICS

MULTIVARIATE ANALYSIS. K. V. Mardia, J. T. Kent and J. M. Bibby. Academic Press, 1979.

COMPUTERS AND THE THEORY OF STATISTICS: THINKING THE UNTHINKABLE. Bradley Efron in *SIAM Review*, Vol. 21, No. 4, pages 460–480; October, 1979.

BOOTSTRAPPING A REGRESSION EQUATION: SOME EMPIRICAL RESULTS. D. Freedman and S. Peters. Technical Report No. 10, Department of Statistics, University of California, Berkeley, 1982.

THE JACKKNIFE, THE BOOTSTRAP AND OTHER RESAMPLING PLANS. Bradley Efron. SIAM Monograph No. 38, Society for Industrial and Applied Mathematics, 1982.

THE VARIABILITY OF ACID RAINFALL. B. Eynon and P. Switzer. Technical Report No. 58, Department of Statistics, Stanford University, 1982.

THE SOCIAL INFLUENCE OF THE MOTTE-AND-BAILEY CASTLE

MANUEL D'ARCHÉOLOGIE MÉDIEVALE. M. de Bouard. Sedes, 1975.

ENGLISH CASTLES. R. Allen Brown. Batsford, Ltd., 1976.

LE CHÂTEAU DANS LA FRANCE MÉDIÉVALE: ESSAI DE SOCIOLOGIE MONUMENTALE. Gabriel Fournier. Aubier Montaigne, 1978.

SUDDEN CARDIAC DEATH: A PROBLEM IN TOPOLOGY

SUDDEN CARDIAC DEATH: THE MAJOR CHALLENGE CONFRONTING CONTEMPORARY CARDIOLOGY. Bernard Lown in *American Journal of Cardiology*, Vol. 43, No. 2, pages 313–328; February, 1979.

PHASE RESETTING AND ANNIHILATION OF PACEMAKER ACTIVITY IN CARDIAC TISSUE. Jose Jalife and Charles Antzelevitch in *Science*, Vol. 206, No. 4419, pages 695–697; November 9, 1979.

SUDDEN CARDIAC DEATH—1978. Bernard Lown in *Circulation*, Vol. 60, No. 7, pages 1593–1599; December, 1979.

THE GEOMETRY OF BIOLOGICAL TIME. Arthur T. Winfree. Springer-Verlag, 1980.

COMPUTER SIMULATION OF ARRHYTHMIAS IN A NETWORK OF COUPLED EXCITABLE ELEMENTS. Frans J. L. van Capelle and Dirk Durrer in *Circulation Research*, Vol. 47, No. 3, pages 454–466; September, 1980.

FIBRILLATION AS A CONSEQUENCE OF PACEMAKER PHASE-RESETTING. A. T. Winfree in *Cardiac Rate and Rhythm*, edited by L. N. Bouman and H. J. Jongsma. Martinus Nijhoff, 1982.

THE AMATEUR SCIENTIST

THE MARANGONI EFFECTS. L. E. Scriven and C. V. Sternling in *Nature*, Vol. 187, No. 4733, pages 186–188; July 16, 1960.

INTERFACIAL PHENOMENA. J. T. Davies and E. K. Rideal. Academic Press, 1963.

NATURAL CONVECTION IN POOLS OF EVAPORATING LIQUIDS. J. C. Berg, M. Boudart and Andreas Acrivos in *Journal of Fluid Mechanics*, Vol. 14, pages 721–735; 1966.

GERMAN INNOVATION

A GLOBAL PERSPECTIVE

A Business Profile by Andrew Hargrave

Henkel: Listening to the Customer.

The Henkel Group is West Germany's biggest privately-owned company. It has a considerable lead in detergent sales in Germany over such international giants as Procter and Gamble and Unilever and is the runner-up in Europe where it holds the top place in the sale of adhesives. Henkel is also the world's largest consumer and processor of natural oils and fats as base materials for chemicals; it is, with a turnover of DM 9.1 billion last year, Germany's fourth largest chemical concern, after the "Big Three", Bayer, Hoechst and BASF.

The parent company has its headquarters at Düsseldorf-Holthausen where it employs 13,036 people and houses the central research and development unit. With its 100 or so subsidiaries, Henkel manufactures over 8,000 products in seven main product groups (see next page) in 44 countries and employs 33,583 people world-wide. Annual investment totals about DM 400 million.

Although Henkel is perhaps best known in Germany for Persil, that remarkable detergent which last year celebrated its 75th anniversary, chemicals are now responsible for more than half the group's turnover. It is also worth noting that in a world-wide recession with its severe impact on the chemical industry, including Henkel, the group still managed to record an increase in sales of 3.5 per cent last year.

As the latest annual report comments: "Our company serves different customers in many countries; we are therefore less affected by fluctuation in economic activity... We pursue a balanced expansion of domestic and foreign business on the one hand, and of consumer products and chemical specialties on the other..."

Since its establishment in 1876, Henkel has been a family-owned company. Dr. Konrad Henkel, grandson of the founder Fritz Henkel (1848–1930) and Chief Executive Officer from 1961 to 1980, is today Chairman of the Supervisory Board and of the Shareholders' Committee. The close connection between the management and the businesslike-thinking family and the fulfillment of the corporate policy principles have both contributed to the continuously successful development of the company.

A broad international spread has no doubt contributed to the competitive strength of this 107 year-old company. Foreign subsidiaries and associates in the U.S., Japan, Brazil, France, Italy, Great Britain, Spain, Mexico, Indonesia, the Philippines, Malaysia and elsewhere have more than doubled their sales in the past ten years and increased their share in the group's overall sales (excluding exports) from under 40 per cent to nearly 60 per cent. More than 10 per cent of total world-wide sales originate in the U.S.

Strategy, present and future

Henkel's business philosophy is summed up thus by Dr. Helmut Sihler, the Chief Executive Officer: "The thrust of our innovative efforts comes from the needs of our customers." Then again: "We are specialists in applied chemistry." This twin-track philosophy is confirmed by recent examples. Some years ago, the German Federal Ministry of the Interior decreed that, for ecological reasons, the phosphate content of detergents must be reduced by 25 per cent by October 1981 and halved by January 1984. As a result of research amounting to millions of DM over ten years and carried out, in part jointly by Henkel and independent institutes contracted by the company, the phosphate substitute SASIL (**S**odium **A**luminum **Sil**icate) has been developed which, the company says, is "toxicologically and ecologically safe" as well as efficient. SASIL has by now, well ahead of the Ministry's deadline, replaced 50 per cent of the phosphate content in all Henkel's detergents and is being used on licence in many other brands the world over.

As a major user of natural oils and fats, Henkel is constantly on the lookout all over the world for new sources of this raw material for the production of fatty acids, fatty alcohols as well as a wide variety of the company's fat based products.

Researchers have found that the plant **cuphea** normally growing wild in tropical places such as Mexico or Brazil also grows in temperate climates and may therefore be suitable for crops yielding fatty acids of coconut fat type in the northern hemisphere, including Europe and the U.S. Thus, this important raw material could be eventually obtained virtually on the doorstep of Henkel's main production centers.

At the Minneapolis, U.S., research unit Henkel is experimenting with naturally based polymers as contained in **guar beans** as a source of chemicals for a more efficient recovery of crude oil. Guar beans are grown in Texas, near the oil fields, but especially in India and Pakistan. "Known chemicals not so far used for this purpose" are being tested as base materials for paints and lacquers for metals used in the automotive industry – yet another example of Henkel's twin-track philosophy of customer service combined with chemical specialist expertise.

The Persil story and others

Although more than half of Henkel's turnover is in chemicals, Persil remains a "flag-ship" of the company. The brand name itself stems from the ingredients **Per**borate and **Sil**icate.

When it first hit the market in 1907, Persil was a break-through. At a time when the weekly washing was a back-breaking, hand-mauling, exhausting chore, a soap powder requiring no washboard, no scrubbing, no bleaching soda was a revelation and a godsend to the long-suffering housewife.

Among others, it also laid the foundation for laundry automation. Two world wars temporarily interrupted the triumphant march of Persil, but each time it was relaunched with fresh improvements, with new and further improved versions coming out periodically ever since.

Diversification has added a host of new products to the Henkel range. They include Germany's first nationally marketed and completely phosphate-free heavy-duty liquid detergent "Liz" and the highly concentrated rinse-softener "Fasa".

Some diversifications which realize today high shares in Henkel's sales have developed as a consequence of traditional Henkel activities. So it was only natural, for example, to offer consumer goods like detergents, cleaning agents and disinfectants with which Henkel had shown its competence in cleanliness and hygiene, also to institutional users just like commercial laundries, food service establishments and hospitals. This volume user business of Henkel is also characterized to a high degree by innovations. Henkel's knowhow in applied technology, its cleaning agents and unparalleled disinfectants combined with especially developed machines and metering equipment offer complete systems for the solution of specific consumer problems. That applies also to Henkel's business with inorganic chemical products and especially with P3-products for industrial surface treatment application.

Production and distribution of adhesives were two of the first diversifications of the company anyway. Built up besides the detergent business in the twentieth, Henkel's adhesives production has developed into one of the most important ones in the world. Today the company is successful in the do-it-yourself sector – e.g. with the glue stick "Pritt" which is meanwhile marketed in 60 countries all over the world – and in the field of industrial adhesives as well. Products like polyurethane adhesives and specialties for film laminating, sandwich construction and rubber-to-metal compounds again show the competence of Henkel as a "specialist in applied chemistry."

Oil from the sun

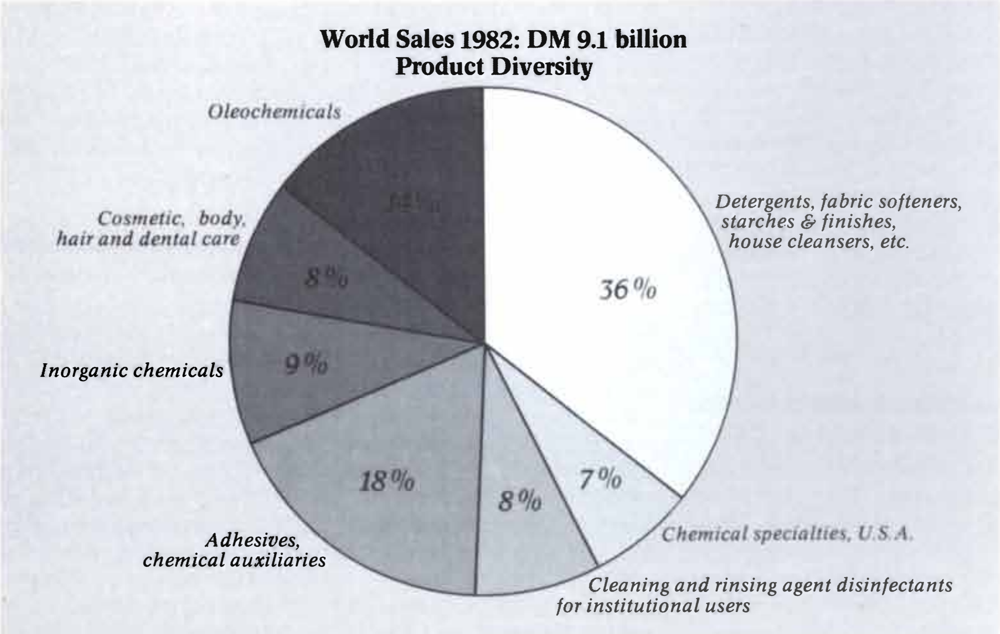
Henkel's factory complex at Düsseldorf-Holthausen is the largest production site of the group where, above all, detergents and cleaning agents, cosmetics, household care products, inorganic products, adhesives, and organic products are manufactured. The Holthausen complex also houses Henkel's oleochemicals plant, probably the world's largest integrated plant based on natural oils and fats. From these raw materials Henkel produces fatty acids, glycerin and derivatives which are used as basic materials for the production of adhesives, leather, textile and food additives, cosmetics, pharmaceuticals, detergents, mining chemicals, etc. both in the Henkel Group and in other industries.

Henkel had already been using natural fats and oils, notably in Persil; but it was not until 1930 that it became interested in a firm which first developed fatty alcohols on a commercial basis from whale oil and animal fats. A couple of years later Henkel decided to take over the firm today called Deutsche Hydrierwerke (Dehydag), an operation which has since assumed major significance, particularly as a result of the oil shocks of the past decade. Natural oils and fats have not only become competitive on price but also because of low energy use which is only about one-sixth of that consumed by the petrochemical process. Since then the use of natural crops – soyabeans, coconut, sunflowers, rapeseed, palm oil as well as of mutton, beef, etc. fats – has risen sharply and, as indicated above, the search is on for fresh crops in temperate climates as sources of fatty acids.

All such crops are renewed constantly in the natural cycle of seasons by the sun's radiation – hence "oil from the sun."

Henkel, as already indicated, is the world's largest user of natural oils and fats, consuming about 350,000 tons a year and converting them into some 600,000 tons of derivatives. They are then processed into basic materials used both in the Henkel Group and in

Dr. Helmut Sihler, President and Chief Executive Officer of Henkel KGaA, the parent company of the Henkel Group (right). Dr. Stefan Schulz-Dornburg, Executive Vice President of Henkel KGaA and President of Henkel of America



many other industries, altogether worth around DM 1.5 billion a year. During the current decade Henkel intends to invest DM 500 million in fat chemistry alone, including new applications and the search for suitable crops. These oil-from-the-sun base materials are now processed in a dozen of its production units throughout the world, including Europe, the U.S., Central and South America and Asia.

In the sundrenched regions of South-east Asia, crops yielding natural oils are already enjoying substantial expansion. For instance in the Philippines (one of the important supplier countries for Henkel), coconut oil yield will be doubled in the 90s. Palm oil, another important raw material for the company, experiences a similar development. In Malaysia, for example, the arable area for palm oil sown has been expanded and the yield per hectare has been enormously increased.

Despite the temporary fall in crude oil prices due to glut, consumers of natural oils and fats, including Henkel, are confident that the policy of relying on natural oils and fats for raw materials rather than on hydrocarbons will continue to pay dividends, particularly in the long run.

The U.S. connection

The structure and character of Henkel's American business is somewhat different from that in Europe and some other parts of the world. (It has not, for one thing, entered the detergent market.) There is, on the other hand, an extensive interface between technology developed in Germany and that originating in or acquired through takeovers in the United States.

Indeed, more recent acquisitions have substantially expanded the range, scale and turnover of Henkel's U.S. group – the Henkel Corporation, based in Minneapolis. Overall U.S. sales, says Dr. Stefan Schulz-Dornburg, member of the management board, have increased from $ 40 million in 1975 to around $ 400 million at present, more than ten per cent of Henkel's world-wide turnover.

The first major step to create new dimensions in Henkel's U.S. operations was the acquisition of the chemical division of General Mills in 1977. In 1980 Henkel added Amchem Products, Inc., a leading manufacturer of chemicals for metal treatment, and in 1981 Bonewitz Chemical Services, Inc., the U.S. producer of Henkel P3-industrial cleaners, to its portfolio.

The Henkel business philosophy being geared to the needs of existing and potential customers of course applies to its U.S. operations as well. Activities were originally limited largely to products used in the textile and leather industries, specialty chemicals, glues and industrial adhesives. Acquisitions have enabled the group to enter vast new markets, including the automotive, metal, printing and oil recovery and fracturing industries.

For example, Dr. Schulz-Dornburg points out the General Mills knowhow of food-additives, including raw materials such as gluten – a protein ingredient of wheat – is now allied to the interest for raw materials in other plants, such as the guar bean grown in Texas and Asia. Guar based polymers have proved to be a very efficient and cost-effective chemical agent for crude oil recoveries and mineral extraction.

Amchem's expertise in metal surface treatment, combined with Henkel's own in paints and lacquers, has led to the testing of new processes with a known chemical not hitherto used for the same purpose. Dr. Schulz-Dornburg believes these experiments promise cheaper, safer and more effective ways of treating metal surfaces, particularly in the automotive industry.

The knowhow exchange within the world-wide Henkel Group extends beyond the actual product range: as, for instance, the recent meeting of experts from Henkel's inorganic chemicals division, from Amchem and Bonewitz and from Henkel's two engineering subsidiaries to discuss the development of new equipment, instruments and other tools of the trade.

Henkel Group: Basic Facts

	1972	1981	1982
World Sales (DM billion)	3.9	8.8	9.1
Share of foreign sales (%)	43	62	64
Share of foreign manufacture (%)	37	54	56
No. of employees (at year's end)	33,145	34,082	33,583*

** of the 33,583 employees, 16,244 (or 48.5 per cent of total) are employed outside the Federal Republic. Of the remainder, employed in Germany, about 13,086 are located at the Düsseldorf-Holthausen headquarters and factory complex.*

Over 2,000 Henkel employees are engaged in research and development which absorbs around DM 200 million a year.

Staff development

As a specialist in chemical applications, the Henkel Group devotes a good deal of its resources to research and development – but also to the general development of employees. In 1981, the company's 610 apprentices in Düsseldorf were undergoing vocational training programmes in 18 separate areas. Most of the 218 young people who completed such programmes in 1981 became permanent employees.

Of the adult employees in Germany, 2,200 attended seminars while a further 3,000 employees participated in what is termed "personnel development programmes," with technical training as well as language included in the curriculum.

In other words, fully one-third of the group's total work force in Germany benefit from training of one sort or another at anyone time.

Future outlook

There is no doubt that companies such as Henkel, heavily engaged as they are in the consumer market, depend greatly on fluctuations of demand as much as on changing consumer habits and fashions. Henkel's confidence in the future is, however, based not only on an eventual recovery in that demand and a close watch on habits and customs but on the closely-knit internal structure: the constant "two-way traffic" between marketing and research, resulting in products that the customers want and buy.

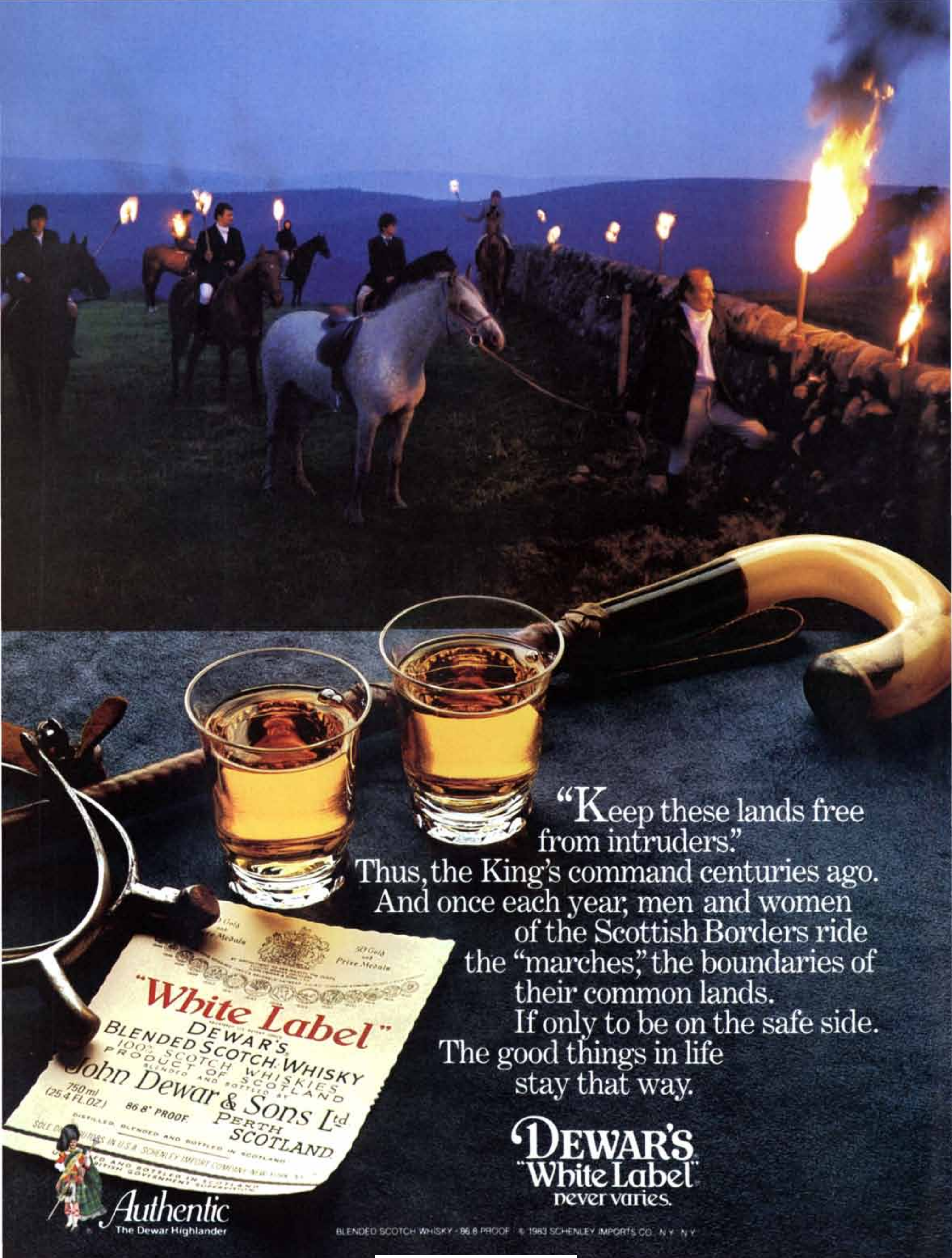
The sound financing of the Henkel Group is essentially contributing to its future competitiveness. The considerable investments are financed from the company itself, the equity ratio is more than 40 per cent of the balance sheet total. A very high proportion for German companies.

The strategy of the logically regional and product diversification has contributed to the stability of the Henkel Group already in the past. Dr. Helmut Sihler: "We shall advance in those fields where our strong points lie and where we know all about markets, users and their problems or where we can quickly gain this knowledge. This is our lead in competition as a 'specialist in applied chemistry'."

Henkel

Henkel KGaA
Henkelstrasse 67
D-4000 Düsseldorf-Holthausen
Telephone: (211) 797-1
Telex: 8 5 817-0

Henkel Corporation
Braemar Office Park
7900 West 78 Street
Minneapolis, Minnesota 55435
Telephone: 830-7831
Telex: 290 472


"Keep these lands free from intruders."
Thus, the King's command centuries ago.
And once each year, men and women of the Scottish Borders ride the "marches," the boundaries of their common lands.
If only to be on the safe side.
The good things in life stay that way.
DEWAR'S
"White Label"
never varies.
"White Label"
DEWAR'S
BLENDED SCOTCH WHISKY
John Dewar & Sons Ltd
PERTH
SCOTLAND
750 ml
(25.4 FL.OZ.)
86.8° PROOF
Authentic
The Dewar Highlander
BLENDED SCOTCH WHISKY · 86.8 PROOF · © 1983 SCHENLEY IMPORTS CO., N.Y., N.Y.