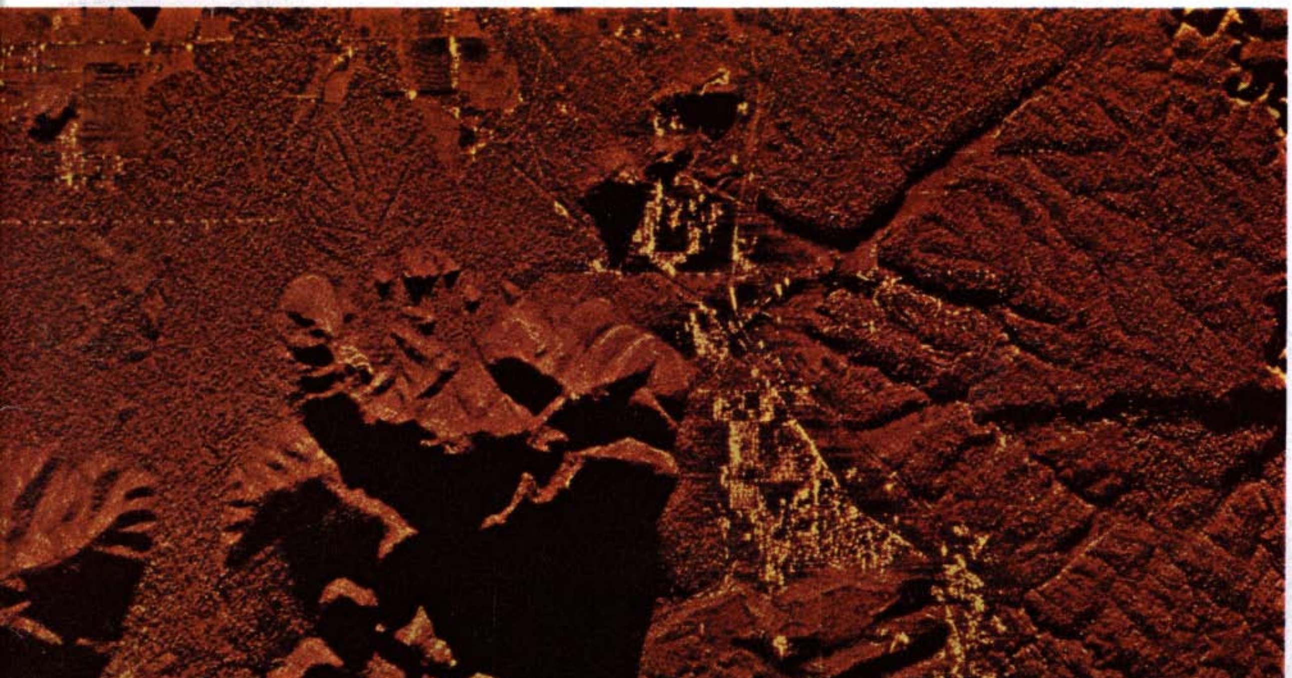
SCIENTIFIC AMERICAN



SIDE-LOOKING RADAR IMAGE

$1.50

October 1977

WHAT THE EXPERTS CALLED THE BEST LAST YEAR WASN'T GOOD ENOUGH FOR US.

"IT CANNOT BE FAULTED."
SA 9500-STEREO REVIEW
"AS NEAR TO PERFECT AS WE'VE ENCOUNTERED." TX 9500-POPULAR ELECTRONICS
"CERTAINLY ONE OF THE BEST . . . AT ANY PRICE." TX 9500-MODERN HI FI

SA9500II

TX9500II

Last year, the experts paid Pioneer's integrated amps and tuners some of the highest compliments ever.

But rather than sitting back and resting on our laurels, we set out to build even better amps and tuners. Amps and tuners that would not only surpass anything we'd ever built before, but anything anyone ever built before.

Our new TX 9500II tuner is a perfect example.

Its front end features three newly developed field effect transistors that work to let you pull in beautiful FM reception no matter how far you live from the transmitter.

It has separate bandwidths for weak and strong stations. Which means you can filter all the noise out of weak stations without worrying about overfiltering strong ones.

And where most tuners have multiplex circuits that accidentally cut out some of the frequencies that add depth and presence to the music, the multiplex circuit in the TX 9500II preserves them. So the music begins to sound as if it's coming live from your living room, instead of from some radio station miles away.

Naturally, after developing a tuner this advanced, we had no choice but to create an amplifier that could match it.

That amplifier is our new SA 9500II. A fully integrated amplifier that's capable of producing a continuous power output of 80 watts per channel RMS into eight ohms, from 20 to 20,000 hertz, with no more than 0.1% total harmonic distortion.

In other words, an amp that cannot only reproduce the entire audible frequency range of every conceivable instrument, but an amp that can reproduce *all* music with a total harmonic distortion level well under what the human ear can hear.

Obviously, both the SA 9500II and the TX 9500II are very sophisticated pieces of equipment. But all of the engineering skill that went into making them has gone into every tuner and amplifier in our new series II. No matter what the price, no matter what the specifications.

And that's something you don't have to be an expert to appreciate.

High Fidelity Components
PIONEER®
WE BRING IT BACK ALIVE.

Established 1845 SCIENTIFIC AMERICAN October 1977 Volume 237 Number 4

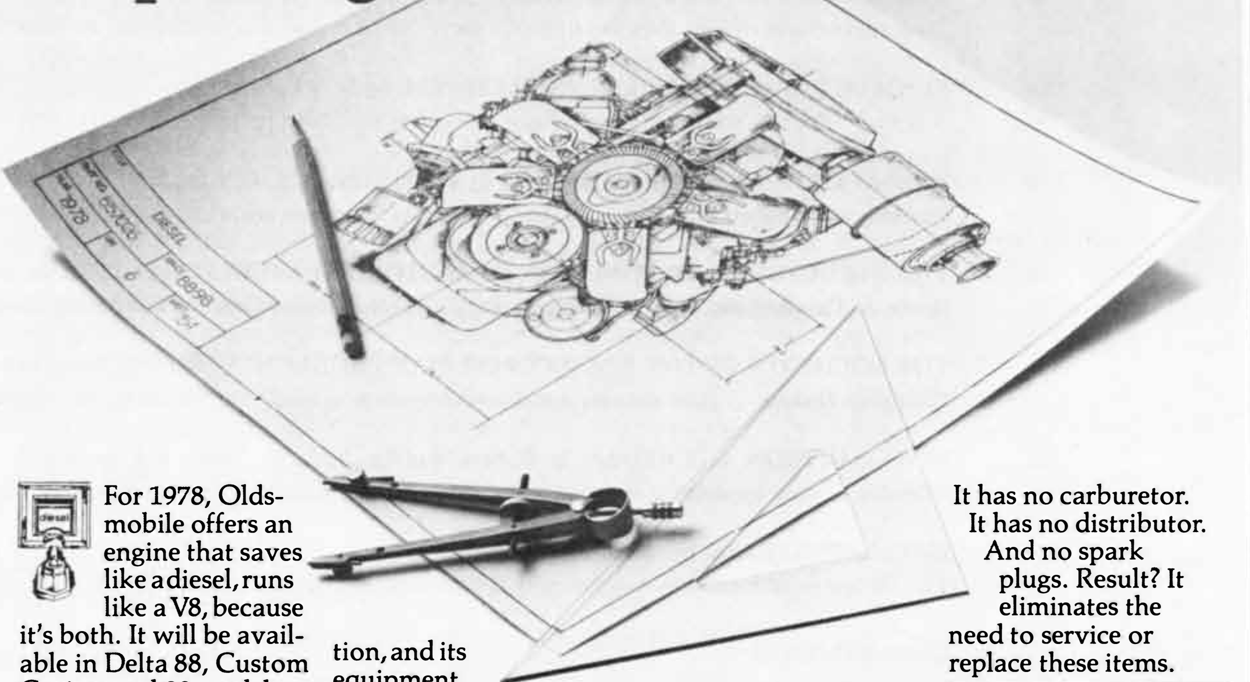
ARTICLES

DEPARTMENTS

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: $18 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $22 PER YEAR, ALL OTHER COUNTRIES.

ANNOUNCING A SIGNIFICANT ENGINEERING ADVANCE:

Oldsmobile introduces the world's first passenger cars with a diesel V8.



For 1978, Oldsmobile offers an engine that saves like a diesel, runs like a V8, because it's both. It will be available in Delta 88, Custom Cruiser and 98 models. And it combines the basic economy of a diesel with performance you might not expect from a diesel.

DIESEL ECONOMY

Diesel *means* fuel economy. That's what a diesel is all about:

1978 EPA MPG ESTIMATES WITH DIESEL V8 ENGINES

	Hwy.	City	Comb.
Delta 88	30	21	24
Ninety-Eight	30	21	24
Custom Cruiser	27	19	22

As you can see, these ratings are excellent, for full-size cars. EPA ratings are estimates; your mileage may vary depending on how and where you drive, your car's condition, and its equipment.

Something else to consider: You not only get *more miles per gallon* with a diesel, you also get *less cost per gallon*, based on traditionally lower prices for diesel fuel compared to unleaded gasoline.

V8 DESIGN

Diesels have been around since the 1890s. And response and acceleration have never been their strong suit. But you'll be pleasantly surprised by these new diesel-powered Oldsmobiles.

This new diesel V8 has one of the highest power-to-weight ratios among diesel cars on the road today.

It offers acceleration and a liveliness that the most powerful diesel foreign cars in America don't have. And it's quiet for a diesel, too.

NO SPARK PLUGS NO CARBURETOR

The new diesel V8 simply does not have many of the components that require tune-ups.

It has no carburetor. It has no distributor. And no spark plugs. Result? It eliminates the need to service or replace these items.

TEST DRIVE A DIESEL AT YOUR OLDS DEALER'S

Only your Olds dealer can offer you a selection of V8 power—diesel or gasoline—in a passenger car. No other dealer can make that statement.

In fact, we have made many strong statements about the new diesel V8. We invite you to put them to the test. Stop in at your nearest Oldsmobile dealer's and ask for the keys to an Olds 88, 98 or Custom Cruiser equipped with this remarkable diesel V8 engine.

In this day and age, it's certainly worth checking out.

Oldsmobile

Can we build one for you?

GM

THE COVER

The picture on the cover looks like an aerial photograph but is not. It is an image of the area of Flagstaff, Ariz., made by a side-looking airborne radar system. The antenna of such a radar system is mounted on the belly of the aircraft and points to the side. Microwave pulses emitted by the antenna strike the ground at a shallow angle; the reflections of the pulses are received by the antenna, displayed on the screen of a cathode-ray tube and recorded on photographic film, forming a radar hologram. The picture is then created by illuminating the hologram with monochromatic light and recording the reconstructed image on color film (see "Side-Looking Airborne Radar," by Homer Jensen, L. C. Graham, Leonard J. Porcello and Emmett N. Leith, page 84). The airplane making this image was 25 miles east of Flagstaff, flying on a course from left to right (north to south) at an altitude of 40,000 feet. The prominent elevation at left center in the picture is Elden Mountain. The cluster of yellow images extending from the top to the lower right center is the town of Flagstaff. The craterlike feature at bottom right is Dry Lake. Thin broken line of yellow images at the top left is a series of radar reflections from a high-voltage power line.

THE ILLUSTRATIONS

Cover photograph courtesy
of Goodyear Aerospace Corporation

Page	Source
29	*Sounds from Silence: Recent Discoveries in Ancient Near Eastern Music.* © 1976 Bīt Enki Publications
32	*A Natural History of Marine Mammals.* © 1976, Charles Scribner's Sons
35–41	Type Systems Inc.
43	© National Geographic Society–Palomar Observatory Sky Survey (*top*), Cerro Tololo Inter-American Observatory (*bottom*)
44–51	Lorelle M. Raboni
57	Jon Brenneis
58–70	Allen Beechel
78	Paul B. MacCready
85	Goodyear Aerospace Corporation and Aero Service Division, Western Geophysical Company of America
86–88	George V. Kelvin
89	Goodyear Aerospace Corporation and Aero Service Division, Western Geophysical Company of America

Page	Source
90–93	George V. Kelvin
94–95	Goodyear Aerospace Corporation and Aero Service Division, Western Geophysical Company of America (*top left*); George V. Kelvin (*top right*); Environmental Research Institute of Michigan (*bottom*)
102–103	Ben Rose
104–105	Ilil Arbel
109–121	Lorelle M. Raboni
123	Y. K. Rao, University of Washington
124	Tom Prentiss (*top*), Type Systems Inc. (*bottom*)
125–128	Robert Maddin, University of Pennsylvania
130	Type Systems Inc.
133–135	Ronald K. Siegel, University of California at Los Angeles
136–138	Alan D. Iselin
142	Michael Goodman
144	Michael Goodman (*top and center*), Ben Rose (*bottom*)
147–150	Michael Goodman

LETTERS

Sirs:

Bernard L. Cohen's article "The Disposal of Radioactive Wastes from Fission Reactors" [SCIENTIFIC AMERICAN, June] contains serious errors and potentially misleading statements. Considerations of space limit my discussion to a few flaws that seem important.

Cohen's comment (pages 23 and 24) regarding the present extensive use of plutonium as a fuel in light-water reactors is true, but for the unwary reader it is probably deceptive. The controversy concerning plutonium recycling has little to do with plutonium that sits securely within an operating reactor core. It concerns only the fate of the plutonium after its removal from the reactor.

The parenthetical remarks on the "linearity hypothesis" (page 24) assume that if that hypothesis is abandoned, it will be replaced by a threshold model or some other reduced-toxicity model. If hypotheses such as those of Karl Z. Morgan ("Suggested Reduction of Permissible Exposure to Plutonium and Other Transuranium Elements," *American Industrial Hygiene Association Journal,* pages 567–575; August, 1975) concerning alpha-active isotopes become generally accepted, however, then abandonment of the linear hypothesis could lead to an increase in nuclear-related health-hazard estimates, not "drastic reductions."

Cohen must substantiate his claim (page 28) that "ceramics or other more resistant materials" will be able to maintain their integrity and low leachability over long periods when they are subjected to thermal heating, intense irradiation, large overburden pressures and surrounding brine.

Cohen's discussion of ion-exchange processes that tend to retard radioactive-waste transport through aquifers (page 28) oversimplifies the matter. It has previously been criticized by F. von Hippel, D. G. Jacobs and J. E. Turner in *Physics Today* (pages 68–69, August, 1976, and pages 15 and 86, November, 1976).

The calculated probability of waste release (page 29) is based on an assumed "equal escape probability for atoms of waste and radium." This assumption is unwarranted, because one can devise scenarios of waste-repository failure that are not applicable to radium in the earth. For example, (1) the possibility exists that waste-generated thermal expansion or post-backfill mine subsidence might fracture overlying shale formations (depicted on page 25), resulting in unanticipated ground-water intrusion (see W. Hambleton, "The Unsolved Problem of Nuclear Wastes," *Technology Review,* pages 15–19; March–April, 1972), and (2) radium in the earth does not have one or more "sealed" mine shafts leading to it, and little work has been done on the feasibility of creating long-lived water-impervious seals. Moreover, the knowledge that New Mexico is currently arid is useful in terms of the relatively short half-lives of strontium 90 and cesium 137 but is not terribly germane to the half-lives of iodine 129 or actinides (more than 24,000 years). Failure to address such issues renders Cohen's probability calculation meaningless.

The author's belief (page 30) that nuclear fission cleanses the earth of radioactivity is incomplete. He neglects to mention that uranium deposits which were sitting safely (more or less) underground are dug up in order to fission them. In the process radium, radon and their daughter products are left exposed to the elements in tailing piles at ground level. On the basis of the linear hypothesis and an assumed global population of four billion one can calculate that each reactor-year of operation eventually gives rise to more than 300 latent cancer fatalities from this cause alone.

The comparison of chemical and radiological hazards (page 30) should note that chemical toxins can, if it is desired, be neutralized by appropriate chemical transformations. Radiological toxins remain hazardous regardless of their chemical form, provided one can postulate some plausible mechanism for "ad-

Scientific American, October, 1977; Vol. 237, No. 4. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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

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

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 35 cents.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name __________

New Address __________

Old Address __________

The Chicago Connection.

With it, the phone system of tomorrow is in Chicago today.

A while ago, we told you lightwave communications was just around the corner. Today, it's in the streets of Chicago.

For the first time, the human voice, business data and even video signals are being carried by lightwaves traveling over hair-thin glass fibers. Instead of electric current traveling over copper wire.

But without that little link you see on the opposite page, lightwave communications for such a wide range of services might still be an experiment in a lab. And without Western Electric technology, the link might still be a design on a blueprint.

The link is an outgrowth of an idea from the people at Bell Labs. While they were putting the major components of the lightwave system together, they had to find a way to splice the glass fibers and get the light across the splice.

A Simple Idea

The idea they came up with was deceptively simple. A coupling device made up of tiny grooved chips, smaller than the tip of your finger, that would guide the ends of the hair-thin fibers and butt them up in perfect alignment.

There was only one hitch. Making one chip was easy. But there was no machinery that could mass produce all the identical chips that would be needed for a lightwave system like the one in Chicago.

Making Ideas Reality

That's where Western Electric comes in. Turning ideas into technological innovations is nothing new at Western Electric.

Over the years, Western Electric has piled up an impressive list of innovations that have become manufacturing standards.

It was the first company in the world to manufacture the transistor.

It was the first to put the laser to work as a useful production tool.

And it is the company that went beyond conventional machining techniques to make the chips for Chicago's lightwave system.

Each chip is pure silicon crystal. Its internal structure (a criss-cross arrangement of intersecting planes) provides a built-in blueprint for regularly spaced grooves. And because the crystal's diagonal planes etch faster than its perpendicular planes, uniform grooves can be chemically cut into the chip.

By combining the science of chemistry and the art of lithography, Western Electric's Engineering Research Center developed a way to etch 12 ultra-precise, perfectly shaped, identical V-grooves on each chip. With each groove no wider than a hair and separated only by a hair's breadth from its neighbors.

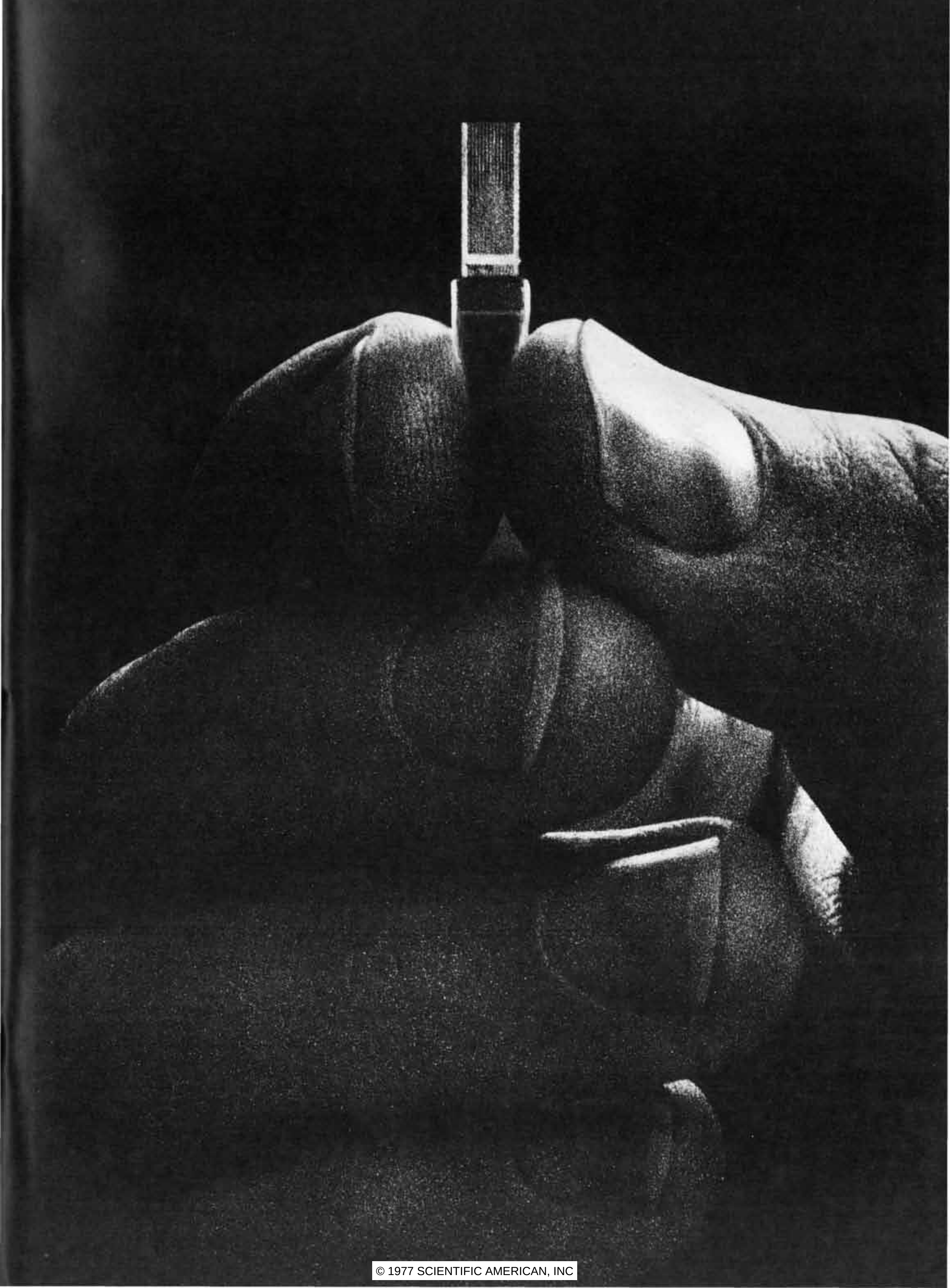
And, more importantly, they were able to reproduce these chips so that each one was a perfect double of the other.

Teamwork is the Key

The telecommunications revolution beginning in Chicago is another good example of how Western Electric and Bell Labs help put new technology into practical use for the Bell telephone companies, quickly and economically.

Their close relationship is an important reason why your telephone system is the most efficient and reliable communications system in the world. And it's a basic reason why innovations in technology are a common occurrence in the Bell System.

Western Electric



ministering" a dose. In addition, the absolute toxicity shown for nuclear waste after 100 years (10^7 lethal doses) is in contradiction with the information presented in the graphs on pages 28 and 29 (approximately 10^{10} lethal doses).

The discussion of the probability and nature of various future human activities (page 31) is entirely speculative. I might note, however, that "random exploratory drilling" may not be a helpful basis for prediction. Assuming some continuity to human nature, curiosity will long be with us. Thus the evidence of past drilling (which would be blatant if markers are left as warnings) might well increase the probability of future human activity at the same site....

Regarding the author's concluding statement, in the future no source of high-quality energy is likely to be "cheap."

Cohen also fails to assess the risks associated with predisposal transportation and handling operations.

Clearly we have radioactive waste, and we should get on the ball to put it someplace where it is unlikely to do us (or future generations) harm. My own preference would be to store high-level wastes or unreprocessed spent fuel in easily retrievable form in a man-made structure located deep within a salt mine or some other relatively stable geological formation, pending a more thorough evaluation of alternative disposal concepts than (barring some unforeseen conceptual breakthrough) can be obtained within the next couple of decades....

RICHARD SCLOVE

Department of Nuclear Engineering
Massachusetts Institute of Technology
Cambridge

Sirs:

My reply to the letter from Mr. Sclove will follow the order of his comments.

My article was based on a more detailed paper published in *Reviews of Modern Physics* (January, 1977), a reputable and thoroughly refereed journal (there were 14 referees on my paper). If the article contains "serious errors," Mr. Sclove should submit a paper to that journal about them.

I thought my one-sentence point about burning plutonium was interesting, and I do not see how it can be regarded as being deceptive. There are those who seem to believe that any use of plutonium, even its very existence, is sinful.

The position that the "linearity hypothesis" is much more likely to overestimate the harmful effects of small radiation doses than to underestimate them is endorsed by all official radiation-safety groups, including the National Academy of Sciences Committee on the Biological Effects of Ionizing Radiation (the BEIR Report), the U.S. National Commission on Radiation Protection and Measurements (NCRP Publication No. 43) and the United Nations Scientific Committee on the Effects of Atomic Radiation (*Ionizing Radiation: Levels and Effects*). The proposal that it may be otherwise, referred to by Mr. Sclove, was rejected by the International Commission on Radiation Protection in its decision not to lower permissible exposures. Under the circumstances I believe my one-sentence parenthetical remark was justified. In any case it did not form a part of my analysis.

My article made no statement about ceramics such as the one mentioned in Mr. Sclove's letter. My comment on ceramics is based on the work of G. McCarthy and Rustum Roy at Pennsylvania State University.

The criticism by von Hippel in *Physics Today* of my treatment of the effects of ion exchange on the transport of radioactive wastes was answered thoroughly in my reply to his letter (printed immediately following it). At the time von Hippel wrote that he was studying the problem further, and although he has written to me at least twice since then on other subjects, he has not mentioned that problem again. The Jacobs-Turner letter was essentially a more detailed treatment than mine rather than a criticism of mine, and there is little disagreement between us, as was explained in my reply printed with their letter.

The fracture of overlying shale by thermal or subsidence problems was studied by the Oak Ridge group and is not considered to be a serious problem. The shaft-sealing problem is being worked on and is not believed to be notably difficult. The shaft area is minuscule, it is quite distant from the buried waste in most designs and the salt is self-sealing. Moreover, there *are* holes in the ground (for example caves, mines and fissures) containing the radium to which the waste is compared. This is not to mention the comparison between the radium and the waste with respect to ground water, rivers, winds, animals, freeze-thaw cycles, vegetation working on surfaces and so on. My paper does *not* assume that the burial site is arid or will remain so. I do not agree that there is anything unfair in my comparing the escape probabilities of buried waste and the radium in the ground above it.

The discussion of tailings from uranium mining and milling is a rather separate issue that is discussed in other papers by me (for example *Bulletin of the Atomic Scientists*, February, 1976, page 61). That problem is being actively worked on and does not seem very difficult or costly to solve. Incidentally, the same calculation that Mr. Sclove cites as indicating 300 eventual cancer fatalities from that source yields a saving of 5,000 cancer fatalities from the burning up of the uranium.

The chemical toxicity of arsenic and barium is *not* essentially affected by ordinary changes in their chemical form. There was a mistake in the illustration on page 30 in that "(after 100 years)" should have read "(after 500 years)." The correct numbers appear in the text and in the illustrations on pages 28 and 29.

I do not understand or sympathize with Mr. Sclove's objection to my treatment of the drilling problem. Can he suggest a more appropriate treatment? Nothing frightful can result from drilling in the far distant future.

I see no reason why nuclear energy should not become cheaper in the future, and remain so for millenniums.

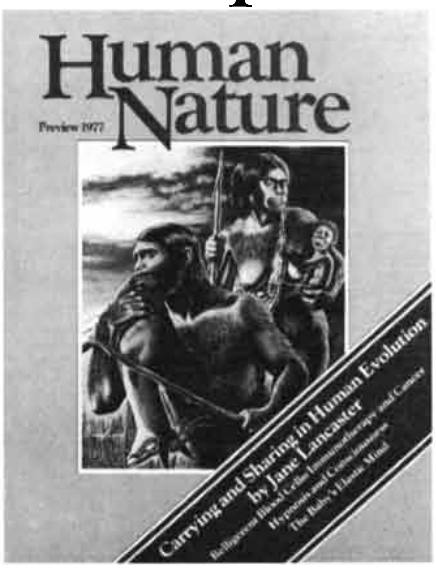
Dangers to the public from waste handling and transport have been studied in other papers and found to be of far less consequence than those discussed here. Radioactivity releases from transport accidents are estimated to cause an average of much less than .01 fatality for each year of all-nuclear power.

It is difficult to quarrel with Mr. Sclove's call for more research, and the research effort is being accelerated. It is also difficult for me to understand, however, the intensity of his concern for the wastes from nuclear power, which have never harmed anyone and which my paper shows will probably never cause as much as one eventual fatality per year, when we are killing at least 10,000 Americans each year with the air-polluting wastes from coal burning, our only viable alternative source of electric power. Moreover, I cannot understand his insistence that the most conservative assumptions be made and the most improbable events be considered in evaluating the dangers from nuclear wastes, when the opposite philosophy is followed in evaluating dangers from coal-burning wastes and all other risks in our society. There are papers in the scientific literature that estimate annual fatalities from coal-burning air pollution at many times 10,000. Those papers have never been refuted. Moreover, radiation effects are far better understood than the effects of air pollution are, and the radiation levels under consideration here are a minute fraction of the natural background of radiation, whereas in the case of air pollution the levels are many times the natural background. Is it too much to ask that there be some balance and perspective in the consideration of our waste problems?

BERNARD L. COHEN

Department of Physics and Astronomy
University of Pittsburgh
Pittsburgh, Pa.

There is new life on our planet



Announcing the birth of HUMAN NATURE...the first magazine to bring to life everything important about the humanity inside us and around us.

Human Nature is destined to be the most important new magazine of our time. It unboggles your mind about the most important invention on our planet: the human being.

Visually exciting and totally readable, Human Nature is the first magazine to take you on a journey—from genetic to global—through *all* the human sciences. You'll get possible answers and new looks at questions about ourselves: Why a child's early years may not be as important as we think • The meaning of money • How your immune system protects you against cancer • Teaching intuition • The physiology and culture of the kiss • Voodoo death • What hypnosis tells us about consciousness ...and more.

The effect is total—to make all aspects of life come together. So our human world becomes more meaningful, more understandable, more fascinating.

There has never been a general interest monthly like Human Nature. Because there has never been one publication to uncover the most important discoveries, ideas, and insights so essential to our quality of life. That's why Human Nature magazine is here. Finally. And the need has never been greater.

Now we can learn more about ourselves and our possibilities. And enjoy the entire experience in clear English and illuminating full color art and photography.

Human Nature is the first general interest magazine published by Harcourt Brace Jovanovich. Its renowned editorial staff is directed by the highly acclaimed author and psychologist Robert Ornstein. Together they collaborate with the most important authorities in all areas that touch human life.

Become a Charter Subscriber and be among the first to experience Human Nature. Each month you will lose yourself for a few hours. And find out about yourself for a lifetime.

Join us now and get Vol. I, No. 1 of Human Nature. As a Charter Subscriber, you save $6 off the newsstand price. We'll bill you just $15 for 12 monthly issues—and guarantee you a full refund at any time.

Human Nature **Charter Subscription Order** A7J041

Please enroll me as a Charter Subscriber. Send me the Premier Issue and bill me $15 for the year (12 monthly issues). I understand that I may cancel my subscription at any time and receive an immediate, unquestioned refund of my full subscription price.

Name ______________________

Address ______________________

City ____________ State ______ Zip ______

Send no money. Just fill in this coupon, clip it out and mail to:
Human Nature • P.O. Box 9100 • Greenwich, CT 06835

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

OCTOBER, 1927: "Clarence D. Chamberlin, the trans-Atlantic flyer, has added to his laurels by making the first ship-to-shore flight from the deck of a merchant vessel, the *Leviathan* of the United States Lines. Above the boat deck of the huge vessel, of nearly 60,000 tons register, a broad runway 114 feet long was built, extending from the starboard side of the first stack to the port side above the bridge. At the time of the test there was a wind of 15 miles per hour, and the *Leviathan* itself was steaming at about the same speed. Thus Chamberlin had ideal wind conditions. His Fokker biplane, equipped with a Wright Whirlwind engine, was lightly loaded, carrying only 900 pieces of mail in addition to the pilot and some two hours' supply of gasoline. The pilot only had to stick to the runway for a distance of 75 feet before his speed, together with the relative speed of the wind, sufficed to get him into the air. He covered the distance between a point at sea 80 miles east of the Ambrose Light and Curtiss Field at Garden City on Long Island in a little more than an hour."

"Cellulose, the principal constituent of wood fiber, may revolutionize photographic methods by its use in photographic films. A new process has just been developed by Philippe David, a collaborator of Alphonse Bertillon, the famous criminologist, by means of which cellulose takes the place of gelatin as a support for the sensitive silver salts. In the ordinary photographic plate or film the base of glass or celluloid is coated with a layer of gelatin in which are suspended the silver bromide particles. The gelatin layer is rather delicate, and great care must be taken with the films or plates before they are dry. Too much heat will melt the coating and spoil the picture. With the new films gelatin and its disadvantages are eliminated. Since the cellulose does not dissolve even in boiling water, the developing chemicals can be used at higher temperatures to speed up the process. The films can be developed in three to four minutes, fixed in two minutes and washed in 30 seconds, instead of the 15 to 30 minutes that the process now takes. It is anticipated that the new fiims and plates will prove especially valuable, both for still and motion pictures, in portraying news events."

"The engineers of the Westinghouse

THE AUTHORS

STEPHEN COLE, LEONARD C. RUBIN and JONATHAN R. COLE ("Peer Review and the Support of Science") are sociologists with a common interest in the sociology of science. Stephen Cole is professor of sociology at the State University of New York at Stony Brook. He was educated at Columbia University, where he received his Ph.D. in 1967. His recent research has focused on "how the various scientific fields differ in the way they are socially organized." Rubin is assistant professor at the New York University School of Social Work. He did his undergraduate work in industrial and labor relations at Cornell University and went on to obtain his Ph.D. in sociology at Stony Brook in 1975. His dissertation dealt with the acquisition of tenure in the fields of chemistry and sociology, a subject he plans to pursue. Jonathan Cole (Stephen's younger brother) is professor of sociology at Columbia and associate director of the university's Center for the Social Sciences. He was educated at Columbia, where he received his Ph.D. in 1969. His research has focused on the reward structure of science and the patterns of intellectual influence within the scientific community.

GEORGE W. CLARK ("X-ray Stars in Globular Clusters") is professor of physics at the Massachusetts Institute of Technology. He attended Harvard College and M.I.T., where he received his Ph.D. in physics in 1952 and has remained ever since. Clark's involvement with X-ray astronomy dates from 1959, when he worked as a consultant with Riccardo Giacconi at American Science and Engineering, Inc., in exploring the prospects for detecting celestial X-ray sources other than the sun. These studies were followed by rocket experiments, culminating in the discovery in 1962 of the first X-ray star, Sco X-1, by Giacconi and his associates. After working for several years on the origin of high-energy cosmic rays, Clark began his own research in X-ray astronomy in 1964, first utilizing balloon-borne detectors and later satellite-borne ones. At present he is in charge of the X-ray laboratory on the Third Small Astronomical Satellite (SAS-3), which is operated around the clock from a control center at M.I.T.

ROY F. SCHWITTERS ("Fundamental Particles with Charm") is associate professor of physics at Stanford University, working at the Stanford Linear Accelerator Center (SLAC). A native of Seattle, Wash., he attended the Massachusetts Institute of Technology, where he spent nine "very enjoyable" years pursuing undergraduate and graduate work in physics. After obtaining his Ph.D. in 1971, he moved to California to become a research associate at SLAC, where he helped design and build the magnetic detector for the SPEAR project. When not absorbed in the physics of electron-positron colliding beams, Schwitters is fond of mountain climbing, having spent his summers as an undergraduate working as a guide in Mount Rainier National Park.

HOMER JENSEN, L. C. GRAHAM, LEONARD J. PORCELLO and EMMETT N. LEITH ("Side-Looking Airborne Radar") share a long-standing interest in imaging radar systems. Jensen is vice-president in charge of radar surveys at Aero Service, a division of Litton Industries. He has a long history of involvement with airborne remote sensing, having helped conduct the first successful airborne-magnetometer experiments. Graham is manager of research and development engineering at the Arizona Division of the Goodyear Aerospace Corporation. He attended the University of Colorado at Boulder, where he received a bachelor's degree in engineering in 1950. He then joined the staff of Bell Laboratories and obtained a master's degree from the Stevens Institute of Technology; he moved to Goodyear in 1956. Porcello is manager of the Tucson, Ariz., division of Science Applications, Inc. He received his Ph.D. in electrical engineering from Michigan in 1963. From 1967 to 1972 he was director of the Radar and Electronics Laboratory there; he took up his present job in 1976. Leith is chief scientist of the Willow Run Laboratories at Michigan. He is best known for his work on the development of holography.

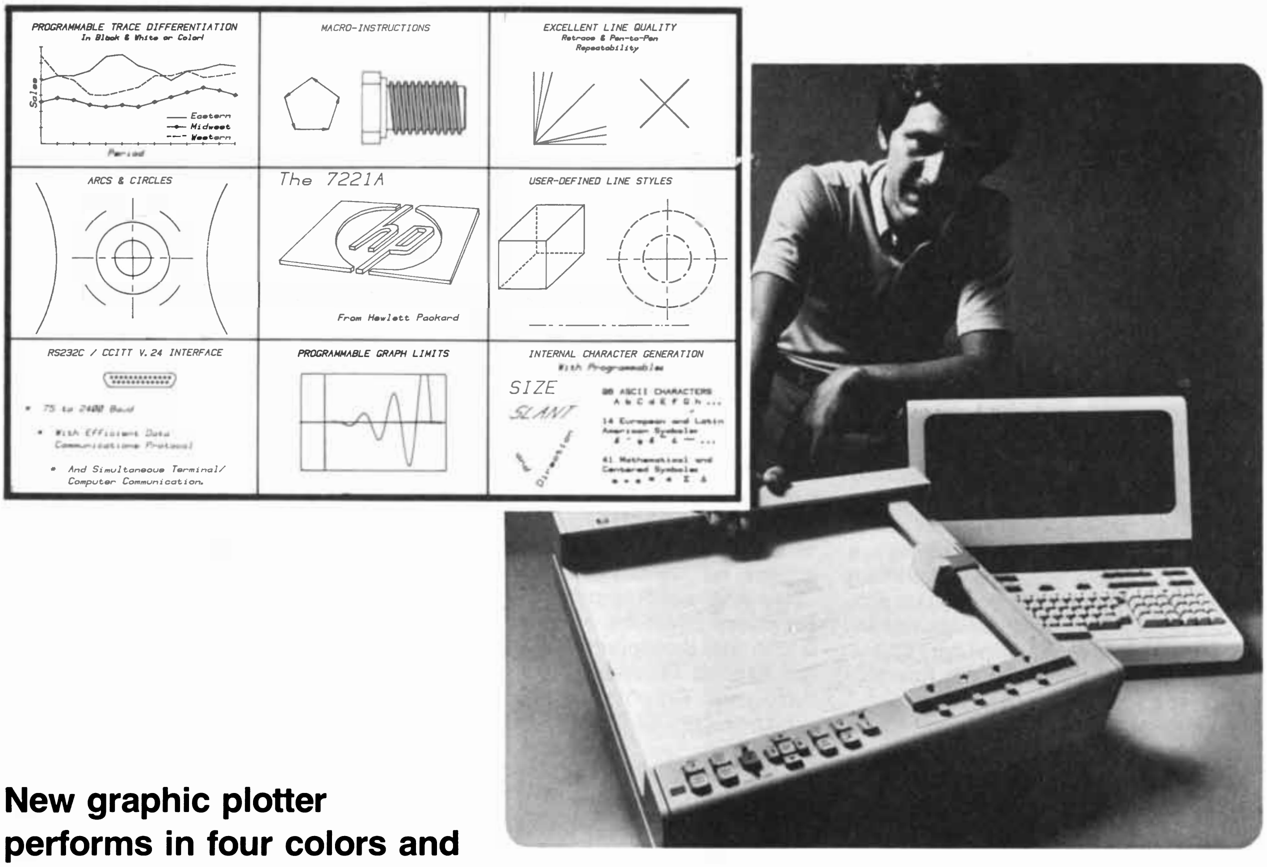
BRUCE A. CUNNINGHAM ("The Structure and Function of Histocompatibility Antigens") is associate professor of biochemistry at Rockefeller University in New York. He received his bachelor's degree in chemistry from the University of Dubuque and his Ph.D. in biochemistry from Yale University in 1966. He then served for two years as a postdoctoral fellow in Gerald M. Edelman's laboratory at Rockefeller, joining the faculty as an assistant professor in 1968. Cunningham's main interest has always been protein chemistry, which he has applied to determining the structure of antibody molecules and other molecules that are structurally, functionally or evolutionarily related to antibodies.

KENNETH APPEL and WOLFGANG HAKEN ("The Solution of the Four-Color-Map Problem") are professors of mathematics at the University of Illinois. Appel attended Queens College and the University of Michigan, where he obtained his Ph.D. in 1959. After two years on the technical staff of the Institute for Defense Analyses in Princeton, he joined the Illinois faculty in 1961. Although most of his work has been concerned with combinatorial problems in logic and group theory, he has made use of computers in solving combinatorial problems for more than 20 years. Haken, a native of Berlin, received his doctoral degree from the University of Kiel in 1953 and then went to work for the Siemens Corporation as a development engineer. In 1961 he published a solution of the "knot problem," and the following year he was offered a visiting professorship in mathematics at Illinois. From 1963 to 1965 he was a temporary member of the Institute for Advanced Study in Princeton, N.J. The article in this issue is based on one that will appear in a collection that is being prepared under the auspices of the Conference Board of Mathematical Sciences, and that will be published next year by Springer-Verlag.

ROBERT MADDIN, JAMES D. MUHLY and TAMARA S. WHEELER ("How the Iron Age Began") are an interdisciplinary group of archaeologists and metallurgists at the University of Pennsylvania. Maddin is professor of metallurgy and a staff member of the Applied Science Center for Archaeology. He received his bachelor's degree from Purdue University in 1942 and then entered the Air Force, where he served in Africa and the Far East. After the war he did graduate work in engineering at Yale University, receiving his doctorate in 1948. The following year he joined the faculty of Johns Hopkins University, and he moved to Pennsylvania in 1955. Muhly is associate professor of ancient history and the editor of *Expedition*, a magazine published by the University Museum. He was educated at the University of Minnesota and at Yale, where he received his Ph.D. in 1969. Wheeler is assistant professor of metallurgy and materials science. She attended Bryn Mawr College, where she obtained her Ph.D. in archaeology in 1973. After teaching a course on the city of Athens at Swarthmore College, she came to Pennsylvania in 1974.

RONALD K. SIEGEL ("Hallucinations") is associate research psychologist in the Department of Psychiatry and Biobehavioral Sciences and also adjunct associate professor of psychology at the University of California at Los Angeles. He did his undergraduate work at Brandeis University and received his Ph.D. from Dalhousie University in 1970. Before joining the UCLA faculty he spent two years as a postdoctoral fellow in pharmacology at the Albert Einstein College of Medicine in New York.

Hewlett-Packard advances in measurement and computation



New graphic plotter performs in four colors and cuts data communications costs dramatically.

The first thing you're likely to notice as the HP 7221 plotter zips into action is the very high quality and clarity of the plot it is producing. But what you don't see is perhaps even more impressive: the microprocessor control and built-in buffer memory by which the HP 7221 reduces time-share and data transmission costs.

Over the past few years, Hewlett-Packard designers have succeeded in introducing microprocessor control into a variety of instruments—a small investment that leads to big performance advances. With the HP 7221 remote terminal plotter, advances are obvious in two broad areas:

First, the new plotter operates with such speed and efficiency that the computer system need spend little of its precious time waiting while the plotter executes instructions. The reduction thus realized in on-line charges and transmission tolls often exceeds 50 percent. How does the HP 7221 do it?

- It receives computer instructions at any of eight speeds from 75 to 2400 baud, half or full duplex.
- It stores data and commands in a built-in buffer memory of 1156 characters, expandable to 3086—and plots the data off-line.
- It has 41 built-in instructions that simplify programming and increase its communications efficiency: e.g., for point digitizing and single-command arc and circle plotting.
- It allows the operator to define a sequence of instructions as a macroinstruction, to store as many as 64 macroinstructions at one time, and to recall any one of them at any time by a single command.
- It plots at any pen speed from 1 to 36 cm per second and typically at 3 characters per second.

Microprocessor control also makes possible a new level of plotting quality, and incorporates features that are especially useful in distinguishing the multiple traces of complex plots:

- programmable selection of four color pens.
- unlimited range of dash-line styles.
- internal character generation from six fonts with programmable control of size, slant, and direction.
- addressable pen moves as small as 0.025 mm for exceptional line quality and precision, even on mylar or acetate.
- programmable graph limits with error-free off-scale data handling.

At a base price of $4600*, the HP 7221 costs little more than remote terminal plotters that do not offer such signal advantages.

The first hand-held programmable calculator with a built-in printer and a memory that doesn't forget.

The new HP-19C combines HP's efficient logic system with advanced keystroke programming, a powerful range of functions, continuous memory that saves programs for reuse, and a thermal printer to record long calculations and to help you edit programs.

For its size—a convenient 165 by 88 by 40 mm—the HP-19C has a powerful array of capabilities:

- You can build and run programs up to 98 steps long, incorporating as many as four keystrokes per step (all multiple keystroke operations are fully merged into single program steps).
- Your programs can incorporate such advanced programming features as conditional branching, three levels of subroutines, and indirect and relative addressing. Editing keys let you correct or change your program with minimal effort.
- Preprogrammed functions include log, trig, mean and standard deviation, statistical summation, and rectangular/polar and angle conversions.
- Continuous memory saves programs and data even when you turn the calculator off—and so saves you the time and bother of reentering.
- The built-in thermal printer lets you see each calculation step clearly and conveniently, on paper tape, as it operates in one of its three modes:

MANUAL. Push the appropriate key and the printer will list the contents of program memory—or 30 storage registers, or 6 statistical registers, or the X register or stack registers. It will list any program in memory, for example, fully identifying it by step number, mnemonic code, and key code.

```
01 *LBL0  25 14 00
02    1         01
03 STO1      45 01
04    2         02
05 STO2      45 02
06    3         03
07 STO0      45 00
```

NORMAL. In this switch position, the printer will automatically produce a record of all keyed-in numbers and their functions.

```
18.00000000 ENT↑
68.00000000    ×
               LOG
               SIN
```

TRACE. The printer will also automatically record each step of an executing program by step number and result—as shown on the tape generated by the HP-19C in the photograph.

And, during manual calculations, it will automatically print out the keyed-in numbers, functions, and results as shown at left.

```
18.00000000 ENT↑
68.00000000    ×
1224.000000  ***
               LOG
3.087781418  ***
               SIN
0.053965869  ***
```

The HP-19C costs $345* including battery pack and recharger/AC adapter, two rolls of thermal paper, soft carrying case, comprehensive owners handbook/programming guide, and applications book. It's available at quality department stores, campus bookstores, or from HP. Call 800-948-4711 in the U.S.A. (in Nevada, call 323-2704 collect) for the name of the nearest retailer.

HEWLETT hp PACKARD

1505 Page Mill Road, Palo Alto, California 94304

For assistance call: Washington (301) 948-6370, Chicago (312) 255-9800, Atlanta (404) 955-1500, Los Angeles (213) 877-1282.

Mail to: Hewlett-Packard, 1505 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) HP-19C advanced printing programmable
( ) HP 7221 remote terminal plotter

Name ____________
Company ____________
Address ____________
City ____________ State ________ Zip ________

*Domestic U.S. prices only.

00749

MATHEMATICAL GAMES

On playing New Eleusis, the game that simulates the search for truth

by Martin Gardner

I shall always consider the best guesser the best Prophet.

Cicero, *De Divinatione*

Don't never prophesy—onless ye know.

James Russell Lowell, *The Biglow Papers*

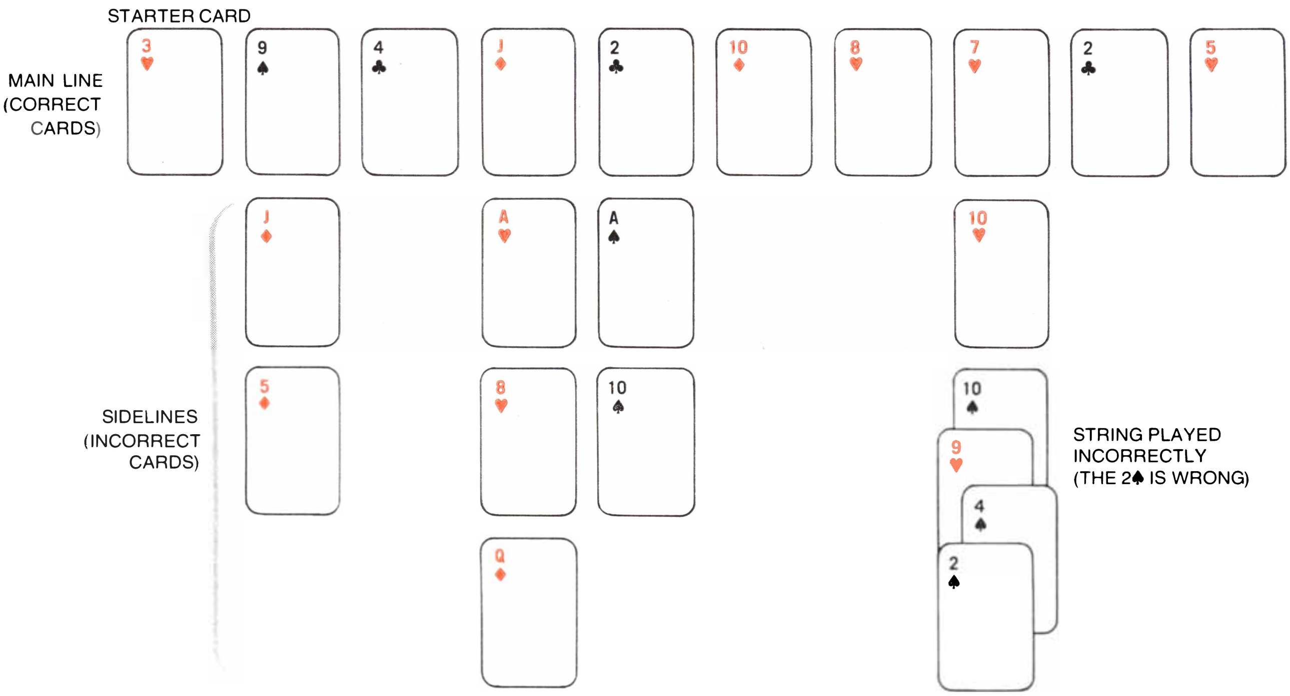
In June, 1959, I had the privilege of introducing in this department a remarkable simulation game called Eleusis. The game, which is played with an ordinary deck of cards, is named for the ancient Eleusinian mysteries, religious rites in which initiates learned a cult's secret rules. Hundreds of ingenious simulation games have been developed for modeling various aspects of life, but Eleusis is of special interest to mathematicians and scientists because it provides a model of induction, the process at the very heart of the scientific method. My first column on Eleusis was reprinted in *The 2nd Scientific American Book of Mathematical Puzzles & Diversions* (Simon and Schuster, 1961). Since then Eleusis has evolved into a game so much more exciting to play than the original version that I feel I owe it to readers to bring them up to date. I will begin, however, with some history.

Eleusis was invented in 1956 by Robert Abbott of New York, who at the time was an undergraduate at the University of Colorado. He had been studying that sudden insight into the solution of a problem that psychologists sometimes call the "Aha" reaction. Great turning points in science often hinge on these mysterious intuitive leaps. Eleusis turned out to be a fascinating simulation of this facet of science, even though Abbott did not invent it with this in mind. In 1963 Abbott's complete rules for the game appeared in his book, *Abbott's New Card Games* (hardcover, Stein & Day; paperback, Funk & Wagnalls).

Martin D. Kruskal, a distinguished mathematical physicist at Princeton University, became interested in the game and made several important improvements. In 1962 he published his rules in a monograph titled *Delphi: A Game of Inductive Reasoning*. Many college professors around the country used Eleusis to explain scientific method to students and to model the Aha process. Artificial intelligence scientists wrote computer programs for the game. At the System Development Corporation in Santa Monica research was done on Eleusis under the direction of J. Robert Newman. Litton Industries based a full-page advertisement on Eleusis. Descriptions of the game appeared in European books and periodicals. Abbott began receiving letters from all over the world with suggestions on how to make Eleusis a more playable game.

In 1973 Abbott discussed the game with John Jaworski, a young British mathematician who had been working on a computer version of Eleusis for teaching induction. Then Abbott embarked on a three-year program to reshape Eleusis, incorporating all the good suggestions he could. The new game is not only more exciting, its metaphorical level has been broadened as well. With the introduction of the roles of Prophet and False Prophet the game now simulates the search for any kind of truth. Here, then, based on a communication from Abbott, are the rules of New Eleusis as it is now played by aficionados.

At least four players are required. As many as eight can play but beyond that



A typical round of Eleusis at an early stage

In North Clarendon, Vermont, there are 450 people, and one Honda Dealer.

And because there's a Honda Dealer in North Clarendon, Vermont, there are a lot of Hondas in North Clarendon.

In fact, that's why you'll find a good many Hondas in towns like Ellicott City, Maryland with a population of 2,000. In McCook, Nebraska with a population of 8,300. Hurley, Wisconsin where 2,400 people live. There's a Honda Dealer in Dunbar, Pennsylvania with a population of 1,500.

The point is, there are Honda Dealers in small towns, big towns, medium-sized towns. All kinds of towns throughout the country.

And, as we continue to grow, we'll have even more dealers in even more towns.

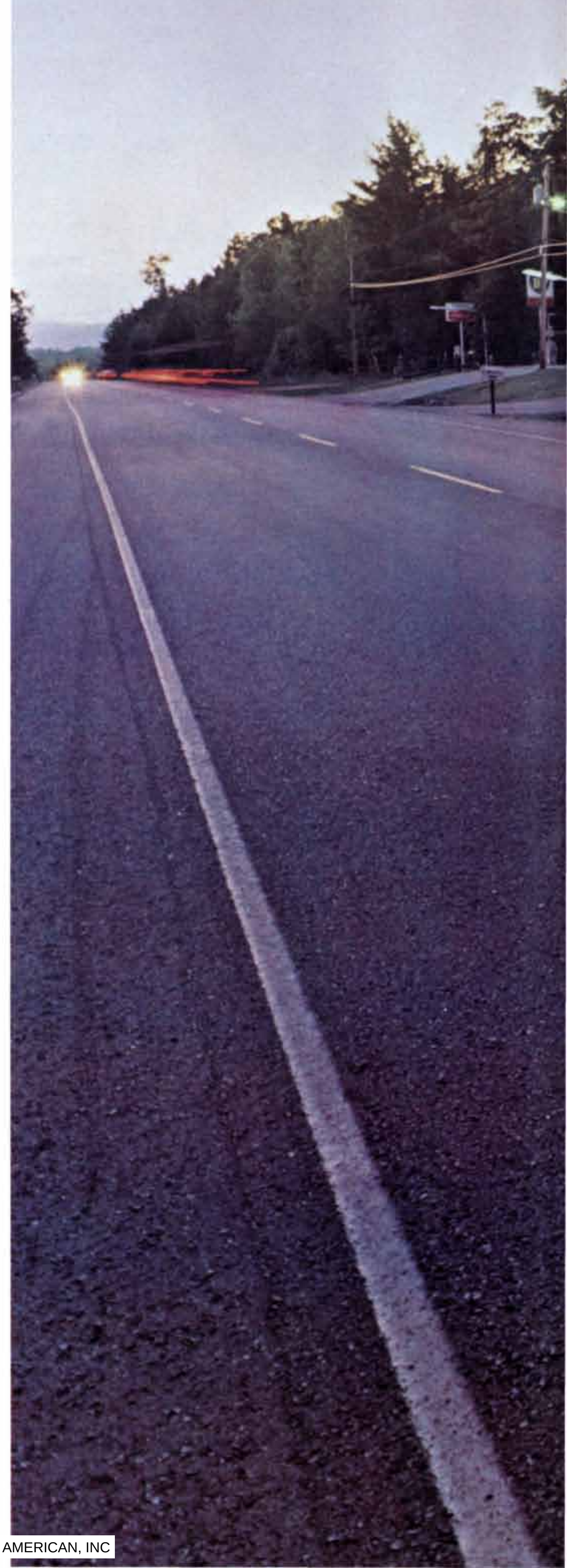
And all of our dealers are stocked with genuine Honda parts and accessories.

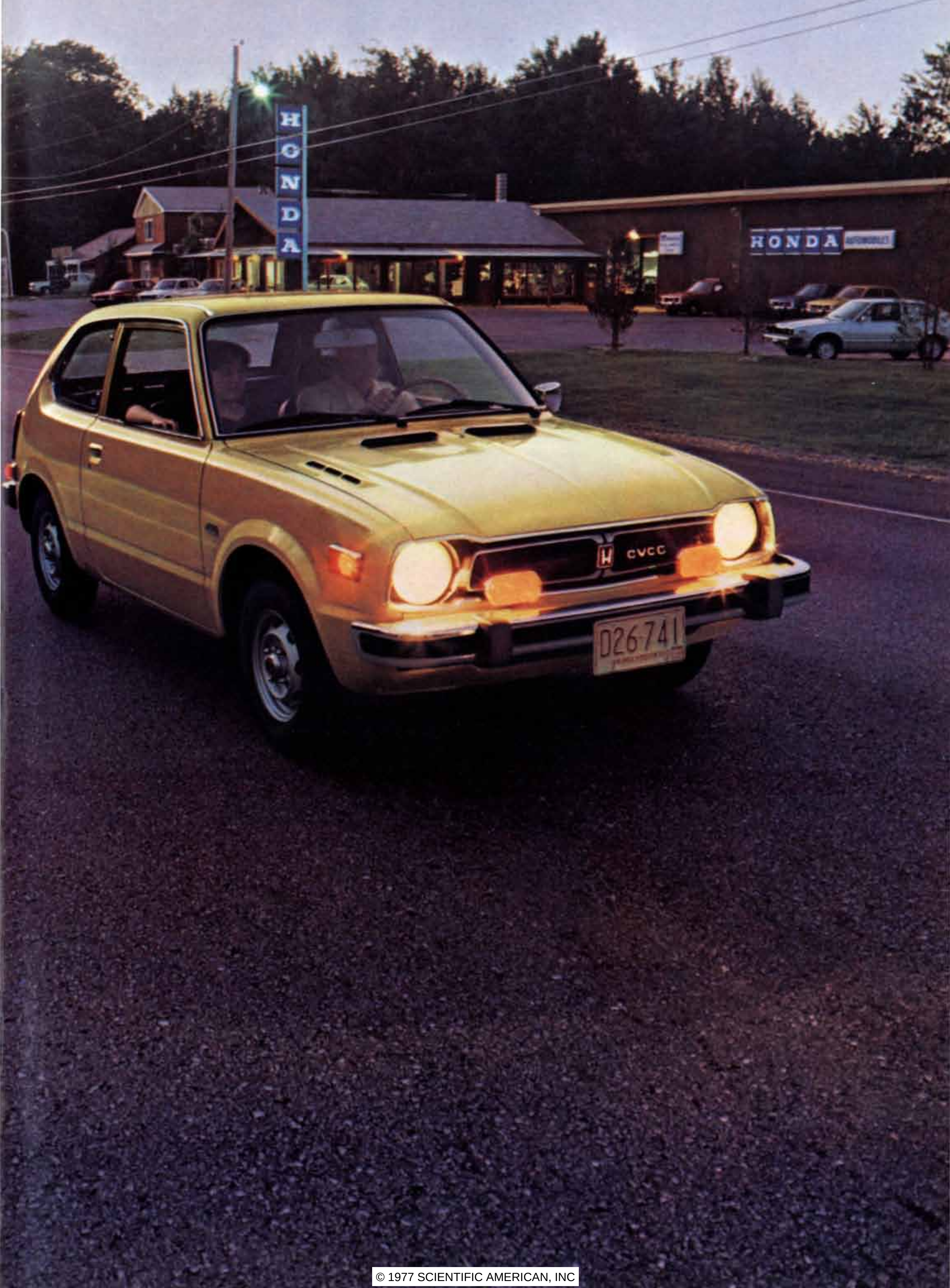
But even the best parts department can't always have every part you might want. So we have a computer that keeps track of every single Honda part that we have in the United States. And it can find a part with the speed of light.

So what it all boils down to is simply this: There are Honda Dealers all across the country. And they're there for one simple reason. To take care of your Honda.

That's why they're Honda Dealers.

HONDA
What the world is coming to.




HONDA
HONDA
CVCC
D26 741

the game becomes too long and chaotic.

Two standard decks, shuffled together, are used. (Occasionally a round will continue long enough to require a third deck.) A full game consists of one or more rounds (hands of play) with a different player dealing each round. The dealer may be called by such titles as God, Nature, Tao, Brahma, the Oracle (as in Delphi) or just Dealer.

The dealer's first task is to make up a "secret rule." This is simply a rule that defines what cards can be legally played during a player's turn. In order to do well players must figure out what the rule is. The faster a player discovers the rule, the higher his score will be.

One of the cleverest features of Eleusis is the scoring (described below) which makes it advantageous to the dealer to invent a rule that is neither too easy to guess nor too hard. Without this feature dealers would be tempted to formulate such complex rules that no one would guess them, and the game would become dull and frustrating.

An example of a rule that is too simple is: "Play a card of a color different from the color of the last card played." The alternation of colors would be immediately obvious. A better rule is: "Play so that primes and nonprimes alternate." For mathematicians, however, this might be too simple. For anyone else it might be too difficult. An example of a rule that is too complicated is: "Multiply the values of the last three cards played and divide by 4. If the remainder is 0, play a red card or a card with a value higher than 6. If the remainder is 1, play a black card or a picture card. If the remainder is 2, play an even card or a card with a value lower than 6. If the remainder is 3, play an odd card or a 10." No one will guess such a rule, and the dealer's score will be low.

Here are three examples of good rules for games with inexperienced players:

1. If the last legally played card was odd, play a black card. Otherwise play a red one.

2. If the last legally played card was black, play a card of equal or higher value. If the last card played was red, play a card of equal or lower value. (The values of the jack, queen, king and ace are respectively 11, 12, 13 and 1.)

3. The card played must be either of the same suit or the same value as the last card legally played.

The secret rules must deal only with the sequence of legally played cards. Of course, advanced players may use rules that refer to the entire pattern of legal and illegal cards on the table, but such rules are much harder to guess and are not allowed in standard play. Under no circumstances should the secret rule depend on circumstances external to the cards. Examples of such improper rules are those that depend on the sex of the last player, the time of day, whether God scratches his (or her) ear and so on.

The secret rule must be written down in unambiguous language, on a sheet of paper that is put aside for future confirmation. As Kruskal proposed, the dealer may give a truthful hint before the play begins. For example, he may say "Suits are irrelevant to the rule," or "The rule depends on the two previously played cards."

After the secret rule has been recorded, the dealer shuffles the double deck and deals 14 cards to each player and none to himself. He places a single card called the "starter" at the extreme left of the playing surface, as is indicated in the illustration on page 18. To determine who plays first the dealer counts clockwise around the circle of players, starting with the player on his left and excluding himself. He counts until he reaches the number on the starter card. The player indicated at that number begins the play that then continues clockwise around the circle.

A play consists of placing one or more cards on the table. To play a single card the player takes a card from his hand and shows it to everyone. If according to the rule the card is playable, the dealer says "Right." The card is then placed to the right of the starter card, on the "main line" of correctly played cards extending horizontally to the right.

If the card fails to meet the rule, the dealer says "Wrong." In this case the card is placed directly below the last card played. Vertical columns of incorrect cards are called "sidelines." (Kruskal introduced both the layout and the terminology of the main line and sidelines.) Thus consecutive incorrect plays extend the same sideline downward. If a player displays a wrong card, the dealer gives him two more cards as a penalty, thereby increasing his hand.

If a player thinks he has discovered the secret rule, he may play a "string" of 2, 3 or 4 cards at once. To play a string he overlaps the cards slightly to preserve their order and shows them to everyone. If all the cards in the string conform to the rule, the dealer says "Right." Then all the cards are placed on the main line with no overlapping, as if they were correctly played single cards.

If one or more cards in a string are wrong, the dealer declares the entire string wrong. He does not indicate which cards do not conform to the rule. The wrong cards are left overlapping to keep their identity as a string and the entire string, goes below the last card played. The player is then dealt twice as many cards as there are in the string.

The layout shown in the illustration on page 18 demonstrates all the rules of Eleusis mentioned so far. The dealer's secret rule for this layout is the first of the three given above.

Players improve their score by getting rid of as many cards as possible, and of course they can do this best if they guess the secret rule. At the start of a round there is little information to go on, and plays are necessarily random. As the round continues and more and more information is added to the layout the rule becomes steadily easier to guess.

It may happen that a player thinks he knows the secret rule but finds he has no card that can be legally played. He then has the option of declaring "No play." In this case he shows his hand to everyone. If the dealer declares him right and his hand contains four cards or less, the cards are returned to the deck and the round ends. If he is right and has five or more cards, then his cards are put back into the deck, and he is dealt a fresh hand with four fewer cards than he previously held.

If the player is wrong in declaring no play, the dealer takes one of his correct cards and puts it on the main line. The player keeps the rest of his hand and, as a penalty, is dealt five more cards. A player who thinks he has no correct play but has not figured out the secret rule should realize that the odds are against his using the no play option successfully. He would do better to play a card at random.

When a player thinks he knows the secret rule, he has the opportunity to prove it and increase his score. He does so by declaring himself a Prophet. The Prophet immediately takes over the dealer's duties, calling plays right or wrong and dealing penalty cards when the others play. He can declare himself a Prophet only if all the following conditions prevail:

1. He has just played (correctly or incorrectly), and the next player has not played.
2. There is not already a Prophet.
3. At least two other players beside himself and the dealer are still in the round.
4. He has not been a Prophet before in this round.

When a player declares himself a Prophet, he puts a marker on the last card he played. A chess king or queen may be used. The Prophet keeps his hand but plays no more cards unless he is overthrown. The play continues to pass clockwise around the players' circle, skipping the Prophet.

Each time a player plays a card or string the Prophet calls the play right or wrong. The dealer then either validates or invalidates the Prophet's statement by saying "Correct" or "Incorrect." If the Prophet is correct, the card or string is placed on the layout—on the main line if right, on a sideline if wrong—and the Prophet gives the player whatever penalty cards are required.

If the dealer says "Incorrect," the Prophet is instantly overthrown. He is declared a False Prophet. The dealer removes the False Prophet's marker and gives him five cards to add to his hand. He is not allowed to become a Prophet again during the same round, although any other player may do so. The religious symbolism is obvious, but as Abbott points out there is also an amusing analogy here with science: "The Prophet is the scientist who publishes. The False Prophet is the scientist who publishes too early." It is the fun of becoming a Prophet and of overthrowing a False Prophet that is the most exciting feature of New Eleusis.

After a Prophet's downfall the dealer takes over his former duties. He completes the play that overthrew the Prophet, placing the card or string in its proper place on the layout. If the play is wrong, however, no penalty cards are given. The purpose of this exemption is to encourage players to make unusual plays—even deliberately wrong ones—in the hope of overthrowing the Prophet. In Karl Popper's language, it encourages scientists to think of ways of "falsifying" a colleague's doubtful theory.

If there is a Prophet and a player believes he has no card to play, things get a bit complicated. This seldom happens, and so you can skip this part of the rules now and refer to it only when the need arises. There are four possibilities once the player declares no play:

1. Prophet says "Right," dealer says "Correct." The Prophet simply follows the procedure described earlier.
2. Prophet says "Right," dealer says "Incorrect." The Prophet is immediately overthrown. The dealer takes over and handles everything as usual, except that the player is not given any penalty cards.
3. Prophet says "Wrong," dealer says "Incorrect." In other words, the player is right. The Prophet is overthrown, and the dealer handles the play as usual.
4. Prophet says "Wrong," dealer says "Correct." In this case the Prophet now must pick one correct card from the player's hand and put it on the main line. If he does this correctly, he deals the player the five penalty cards and the game goes on. It is possible, however, for the Prophet to make a mistake at this point and pick an incorrect card. If that happens, the Prophet is overthrown. The wrong card goes back into the player's hand and the dealer takes over with the usual procedure, except that the player is not given penalty cards.

After 30 cards have been played, and there is no Prophet in the game, players are expelled from the round when they make a wrong play, that is, if they play a wrong card or make a wrong declaration of no play. An expelled player is given the usual penalty cards for his final play and then drops out of the round, retaining his hand for scoring.

If there is a Prophet, expulsions are delayed until at least 20 cards have been laid down after the Prophet's marker. Chess pawns are used as markers so that it is obvious when expulsion is possible. As long as there is no Prophet a white pawn goes on every 10th card placed on the layout. If there is a Prophet, a black pawn goes on every 10th card laid down after the Prophet's marker. When a Prophet is overthrown, the black pawns and the Prophet's marker are removed.

A round can therefore go in and out of the phase when expulsions are possible. For example, if there are 35 cards on the layout and no Prophet, Smith is expelled when he plays incorrectly. Next Jones plays correctly and declares herself a Prophet. If Brown then plays incorrectly, she is not expelled because 20 cards have not yet been laid down after the Prophet's marker.

A round can end in two ways: (1) when a player runs out of cards or (2) when all players (excluding a Prophet, if there is one) have been expelled.

The scoring in Eleusis is as follows:

1. The greatest number of cards held by anyone (including the Prophet) is called the "high count." Each player (including the Prophet) subtracts the number of cards in his hand from the high count. The difference is his score. If he has no cards, he gets a bonus of four points.

2. The Prophet, if there is one, also gets a bonus. It is the number of main-line cards that follow his marker plus twice the number of sideline cards that follow his marker, that is, a point for each correct card since he became a Prophet and two points for each wrong card.

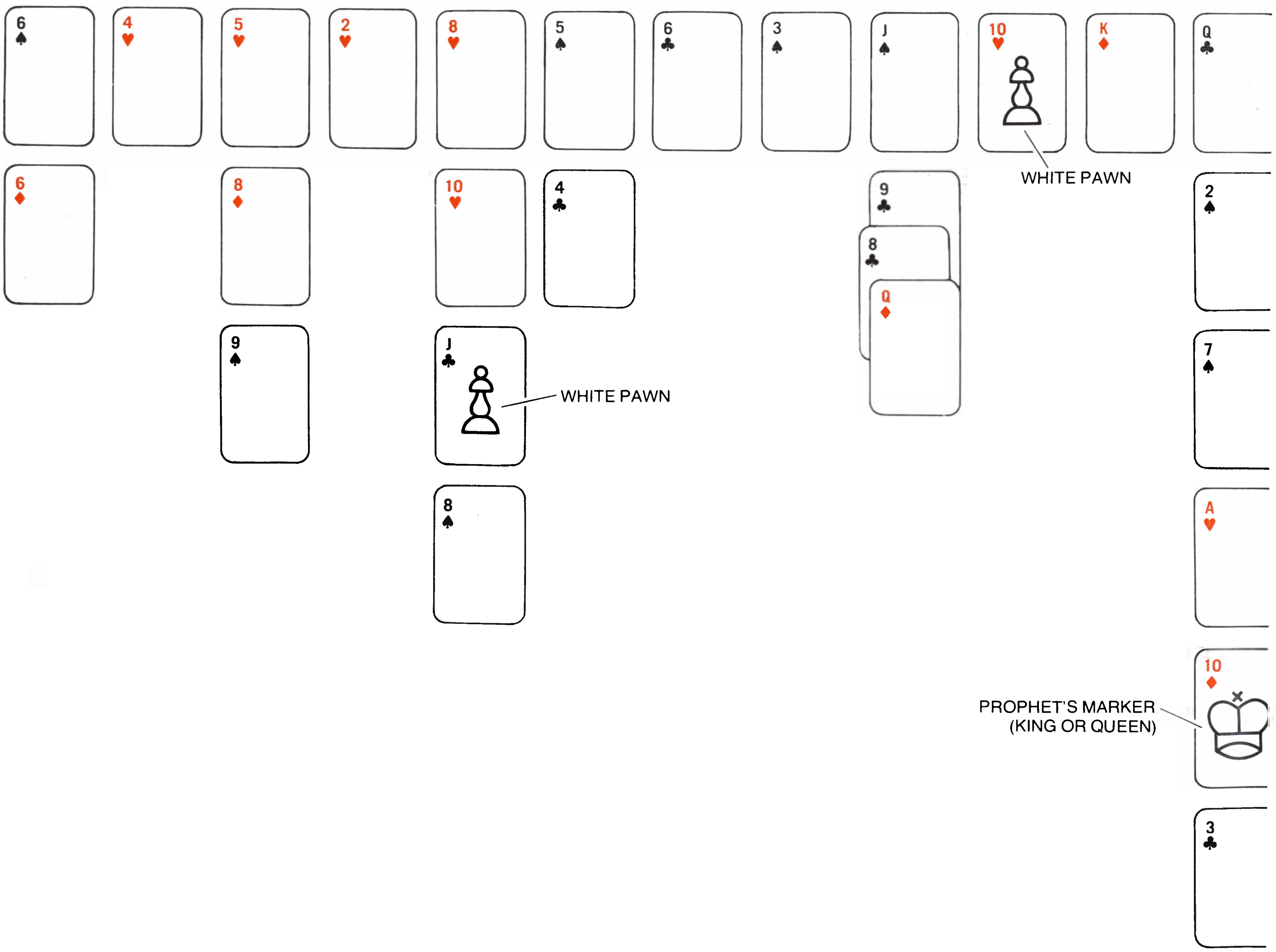
3. The dealer's score equals the highest score of any player. There is one exception: If there is a Prophet, count the number of cards (right and wrong) that precede the Prophet's marker and double this number; if the result is smaller than the highest score, the dealer's score is that smaller number.

If there is time for another round, a new dealer is chosen. In principle the game ends after every player has been dealer, but this could take most of a day. To end the game before everyone has dealt, each player adds his scores for all the rounds played plus 10 more points if he has not been a dealer. This compensates for the fact that dealers tend to have higher than average scores.

The layout in the illustration on these two pages shows the end of a round with five players. Smith was the dealer. The round ended when Jones got rid of her cards. Brown was the Prophet and ended with nine cards. Robinson was expelled when he incorrectly played the 10 of spades; he had 14 cards. Adams had 17 cards at the end of the game.

The high count is 17. Therefore Adams' score is 17 minus 17, or zero. Robinson's score is 17 minus 14, or three. Jones receives 17 minus zero, or 17, plus a four-point bonus for having no cards, so that her score is 21. Brown gets 17 minus nine, or eight, plus the Prophet's bonus of 34 (12 main-line and 11 sideline cards following his marker), making a total score of 42. This is the highest score for the round. Twice the number of cards preceding the Prophet's marker is 50. Smith, the dealer, receives 42 because it is the smaller of the two numbers 42 and 50.

Readers are invited to look over this layout and see if they can guess the secret rule. The play has been standard, so that the rule is confined strictly to the main-line sequence. I shall give the secret rule next month.



Layout at the end of a round of Eleusis includes a main line, several sidelines and various markers. White chess pawn

Some miscellaneous advice from Abbott should help inexperienced Eleusis players. Since layouts tend to be large, the best way to play the game is on the floor. Of course a large table can be used as well as miniature cards on a smaller table. If necessary, the main line can be broken on the right and continued below on the left.

Remember that in Eleusis the dealer maximizes his score by choosing a rule that is neither too easy nor too difficult. Naturally this depends both on how shrewdly the dealer estimates the ability of the players and how accurately he evaluates the complexity of his rule. Both estimates require considerable experience. Beginning players tend to underestimate the complexity of their rules.

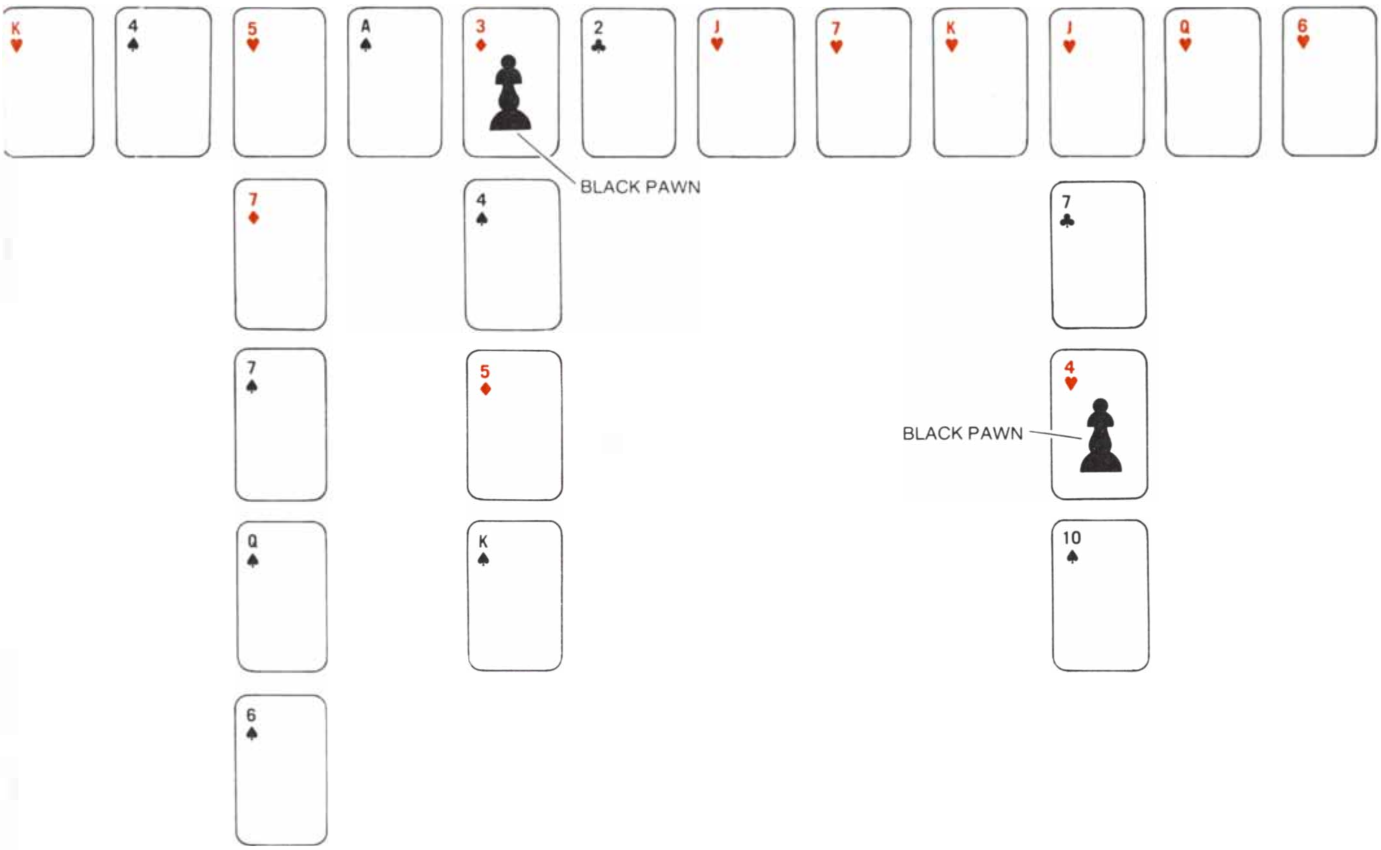
For example, the rule used in the first layout is simple. Compare it with: "Play a red card, then a black card, then an odd card, then an even card and repeat cyclically." This rule seems to be simpler, but in practice the shift from the red-black variable to the even-odd variable makes it difficult to discover. Abbott points out that in general restrictive rules that allow only about a fourth of the cards to be acceptable on any given play are easier to guess than less restrictive rules that allow half or more of the cards to be acceptable.

I shall not belabor the ways in which the game models a search for truth (scientific, mathematical or metaphysical) since I discussed them in my first column on the game. I shall add only the fantasy that God or Nature may be playing thousands, perhaps a countless number, of simultaneous Eleusis games with intelligences on planets in the universe, always maximizing his or her pleasure by a choice of rules that the lesser minds will find not too easy and not too hard to discover if given enough time. The supply of cards is infinite, and when a player is expelled, there are always others to take his place.

Prophets and False Prophets come and go, and who knows when one round will end and another begin? Searching for any kind of truth is an exhilarating game. It is worth remembering that there would be no game at all unless the rules were hidden.

Readers can send $1 to Robert Abbott and receive postpaid a more detailed set of instructions for playing Eleusis; they will also be placed on his mailing list and receive any future developments of the game. Abbott's address is Box 1175, General Post Office, New York, N.Y. 10001.

Last month I asked whether or not the straight-line shadow on the photograph of the McDonnell Planetarium in St. Louis corresponds to one of the straight lines that rule the surface of the structure's hyperboloid of revolution of one sheet. The answer is yes. At those times when the line containing the edge of the shadow intersects the sun, the edge of the shadow on the side of the planetarium coincides with a generating line of the surface. That line is a projection of a single point on the circumference of the circular top of the planetarium, not a projection of the entire circumference. When the sun is at any other elevation, the edge of the shadow will be a curved line and not a generating line of the surface.



are placed on every 10th card played in the round and black pawns on every 10th card played after a Prophet's marker

BOOKS

Computer crime, music of the ancient Near East, marine mammals and other matters

by Philip Morrison

CRIME BY COMPUTER, by Donn B. Parker. Charles Scribner's Sons ($10.95). COMPUTER SECURITY AND PROTECTION STRUCTURES, by Bruce J. Walker and Ian F. Blake. Dowden, Hutchinson & Ross, Inc. ($14.95). Every user of a computer teletype terminal expects to "sign on" as he begins his task. Often a mere tap of the carriage return key starts the process. The terminal prints the time and a stock greeting, and a few familiar requests for information, which you answer as usual with your coded credit identification and the password that allows access to a particular private file. Without much warning the system "crashes"; there is some defect, maybe a momentary power failure. You must sign on again, by a well-known procedure. It is a minor annoyance, not notable in any way. Sometimes, however, there is a sting! The first time you wrote out the password and identification data might have been a pseudo-sign-on; you merely entered those private matters into the waiting store of a clever programmer who had set a trap, without omitting the mock crash so that you would unsuspectingly sign on again, this time with the real computer for a working session. He has your passwords now; one day he may be able to exploit them. Away back in 1970 a couple of clever high school students accumulated a hundred passwords in just such an inconspicuous pitfall.

A world of personal histories, of financial assets and actual transactions, of corporate and private secrets, lies within reach of every terminal. There are 150,000 computers in service in this country; the number is tripling each five years. Parker, senior information-processing analyst at the Stanford Research Institute, has made himself since 1970 a full-time student of the problem of computer abuse. His book is a popular summary of what he has learned, the statistics of losses, the attitudes of the computing community, a study of the perpetrators as a class, and mostly a dozen cheerful case histories of remarkable criminal variety, nearly all in the U.S. Walker and Blake offer the other face of the coin. Theirs is a sober, compact monograph, full of jargon but not highly technical in argument, that analyzes and surveys the literature on threats and countermeasures for the programmer and manager; it will be attractive to both sides. These authors, both from the University of Waterloo in Ontario, include brief accounts of the security measures actually implemented within a dozen well-known individual computer systems.

At present a flood of back-up paper continues. Pulses sent by tape or by cable maintain the bank balances of the nation, but the checks and securities still flow as proof and for individual settlements. It looks as though negotiable paper will dwindle, and with it the payroll and bank-messenger heist; subtler crime will spread, crime that manipulates the invisible and the transient. Inside jobs will be the most frequent, as they are in the paper-and-ink world. The computers handle larger tasks; their attackers are more frequently teams in collusion. So far they have been mainly young amateurs, people knowing in the computer techniques, generally fortified by the rationalization of borrowing for a while, or of harming only some impersonal corporate entity.

Naturally we know only of the failed schemes. The estimate runs to a loss of $300 million per year, plausibly but not very firmly extrapolated from fewer than 300 cases. The incidence is taken as one theft per year per 2,000 computers in service and nearly $500,000 per case. But who can say? For years auditors of paper ledgers have known of the round-down fraud, a scheme to handle fractional pennies from many accounts, the totals showing no defects because everything balances. It is just that chance round-up and round-down are replaced by determinism. Many accounts that round down are honestly kept, but the others are preferentially accumulated to the account of the perpetrator. "How many programmers have retired to a life of leisure as their programs, long trusted and forgotten, continue to pump the pennies into their accounts at nearly the speed of light?" Avoiding indecent greed, this patient tapeworm of digital currency can remain all but invisible.

Whatever lies in the depths, the visible tip of the iceberg is crowded with witty rascals, whose activities owe less to counterfeit software or subversive hardware than to personal talent as confidence men. (A few women enter the cast too.) Take the youthful boss of Creative Systems Enterprises. His little company sold a few hundred thousand dollars worth of Western Electric's best gear. He obtained the supplies for resale free (all stamped "Released for Resale" in his own shop) by a long study of telephone-company procedures; he filed his orders by Touchtone in good company form and took delivery each morning in the company equipment yard in his company van, bought at company auction. It called for much thought but little computer expertise. After he had served his term, this wiseacre of course became a consultant on computer security. True to his style, he could "get into any time-sharing system in the country." But he did it for Parker's edification by conning his way in by means of a plausible line of persuasion addressed to the night operator at the computing service.

Here in word and picture is the Silicon Valley engineer who designed and wired a specialized computer into a money belt, able to help him count the cards during blackjack deals (1,000 words of read-only memory). The input switches were four snapping disks worked by his big toes; the display was tiny light-emitting diodes built into the inside of his eyeglass frame. These were too close to his eye to form any image, but the colors of the out-of-focus "balloons of light" were legible. Undercapitalized and unlucky, he lost his stake, underwent religious conversion and gave up gambling. He ended up reformed and a winner in print and on television.

The biggest fraud recounted was a wide conspiracy of high-living Los Angeles executives to generate false input. They took their insurance company out of the red by loading its disk files with 64,000 bogus policies. When their compliant auditors, a bit naive in a data-processing age, would ask for sample documentation, a special office went to work to create the paper for the next morning's inspection. Real money flowed in from stock sales. The big bust, tipped off by a fired employee, ended with 22 convictions and a rash of lawsuits, the stockholders and creditors holding a $2-billion bag.

The subtlest and most dangerous of computer abuses has not surfaced in illicit gainful use. Rather, it was a work of sheer bravado within a university time-sharing service. Tracking down a bug in central software, an alert programmer found by chance an odd sequence of instructions embedded in a 400-page program printout. The strange listing allowed "a terminal at one of the colleges... to be the ultimate and unquestioned ruler of the system." The logical virus was hidden within a program for editing tapes, offered generously for public use. Within half a year the editing

WHAT'S BETTER THAN SPEED READING?

SPEED LEARNING

(SPEED PLUS COMPREHENSION)

Speed Learning is replacing speed reading because it's easy to learn . . . lasts a lifetime . . . applies to everything you read . . . and is the only fully accredited course with the option of 3 college credits.

Do you have too much to read and too little time to read it? Do you mentally pronounce each word as you read? Do you frequently have to go back and re-read words or whole paragraphs you just finished reading? Do you have trouble concentrating? Do you quickly forget most of what you read?

If you answer "yes" to any of these questions — then here at last is the practical help you've been waiting for. Whether you read for business or pleasure, school or college, you will build exceptional skills from this major breakthrough in effective reading, created by Dr. Russell Stauffer at the University of Delaware.

Not just "speed reading" — but speed reading-thinking-understanding-remembering-and-learning

The new *Speed Learning Program* shows you step-by-proven-step how to increase your reading skill and speed, so you understand more, remember more and use more of everything you read. The typical remark made by the 75,000 slow readers who completed the *Speed Learning Program* was: "Why didn't someone teach me this a long time ago?" They were no longer held back by the lack of skills and poor reading habits. They could read almost as fast as they could think.

What makes Speed Learning so successful?

The new *Speed Learning Program* does not offer you a rehash of the usual eye-exercises, timing devices, costly gadgets you've probably heard about in connection with speed reading courses or even tried and found ineffective.

In just a few spare minutes a day of easy reading and exciting listening, you discover an entirely new way to read and think — a radical departure from anything you have ever seen or heard about. Research shows that reading is 95% *thinking* and only 5% eye movement. Yet most of today's speed reading programs spend their time teaching you rapid eye movement (5% of the problem) and ignore the most important part (95%) *thinking*. In brief, *Speed Learning* gives you what speed reading *can't*.

Imagine the new freedom you'll have when you learn how to dash through all types of reading material *at least* twice as fast as you do now, and with greater comprehension. Think of being able to get on top of the avalanche of newspapers, magazines and correspondence you have to read . . . finishing a stimulating book and retaining facts and details more clearly and with greater accuracy than ever before.

Listen-and-learn at your own pace

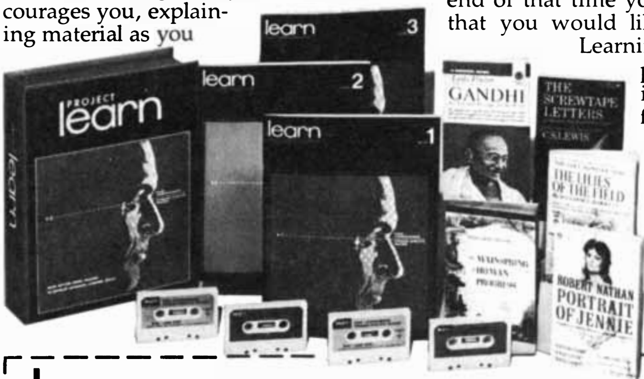
This is a practical, easy-to-learn program that will work for you — no matter how slow a reader you think you are now. The *Speed Learning Program* is scientifically planned to get you started quickly . . . to help you in spare minutes a day. It brings you a "teacher-on-cassettes" who guides you, instructs, encourages you, explaining material as you read. Interesting items taken from *Time* Magazine, *Business Week, Wall Street Journal, Family Circle, N.Y. Times* and many others, make the program stimulating, easy and fun . . . and so much more effective.

Executives, students, professional people, men and women in all walks of life from 15 to 70 have benefited from this program. Speed Learning is a fully accredited course . . . costing only 1/5 the price of less effective speed reading classroom courses. Now you can examine the same, easy, practical and proven methods at home . . . in spare time . . . without risking a penny.

Examine Speed Learning FREE for 10 days

You will be thrilled at how quickly this program will begin to develop new thinking and reading skills. After listening to just one cassette and reading the preface you will quickly see how you can achieve increases in both the speed at which you read and in the amount you understand and remember.

You must be delighted with what you see or you pay nothing. Examine this remarkable program for 10 days. If, at the end of that time you are not convinced that you would like to master Speed Learning, simply return the program and owe nothing. See the coupon for low price and convenient credit terms.



In most cases, the entire cost of your Speed Learning Program is Tax Deductible.

EARN PROFESSIONAL & COLLEGE CREDITS

Speed Learning is approved for credit by the following professional and educational institutions:

- AMERICAN MEDICAL ASSOCIATION
  15 Credit Hours – Category I
  Louisiana State University Medical Center
- FOUNDATION FOR ACCOUNTING EDUCATION
  20 Continuing Professional Education Credit Hours
- COLLEGE CREDIT
  3 credits from the National College of Business

Details and registration forms included with each program. Note: Many firms and government agencies have tuition assistance plans for employees providing full or partial payment for college credit programs.

ENDORSED BY
- American Management Associations
- Institute of Electrical & Electronics Engineers
- 1000s of executives, physicians and educators

SA-1

learn INCORPORATED 113 Gaither Drive, Mount Laurel, N.J. 08054

Please send me the Speed Learning Program at $79.95 plus $3.00 for handling and insured delivery.

Please check the method of payment below:

☐ Check or money order enclosed for items ordered.

☐ Please charge my credit card under their regular payment terms: ☐ Bank Americard ☐ Master Charge Interbank No. ______ ☐ American Express ☐ Diners Club

Card No. ______ Exp. Date ______

I understand that if after 10 days I am not delighted in every way, I may return the materials and obtain a full refund with no questions asked.

Name ______

Address ______

City ______ State ______ Zip ______

x Signature ______

If you don't already own a cassette player, you may order this Deluxe Cassette Recorder for only $49.95. (Includes handling and delivery.)

Check here to order ☐

New Jersey Residents add 5% sales tax

program was called up by a user at the highest level of privilege within the system. The insert, which had bided its time for that opportunity, then took control, placed itself within the system center, erased its tracks and returned control to the innocuous tape editor. The payoff would be complete only long afterward, at a preassigned moment of gleeful triumph, when the system would devote all its terminals to printing out the magnificent insert 100 times, then its graduate-student inventor's name, with "the worst failure in the system's history" to follow.

Perhaps the tale is folklore; not much evidence is given. The system designers nonetheless fear it. Elaborate architecture of certain new systems include a hierarchical ordering of capabilities. No user can employ a program at a higher control level than it is intended to reach, even if he is fully privileged. Only certain gates allow a change of level even when it is allowed. Passwords are double and frequently changed; each user is greeted with a message saying when and where he last signed on. Flexible and changing special routines—challenges and responses, for example—can be added. Military systems even sound an audible alarm at the console if an access attempt fails.

Wiretapping is of course a constant problem, and a very ingenious ciphering system, based on the work of Claude E. Shannon a generation ago, has been made the basis of a National Bureau of Standards design standard. It diffuses the message widely through a string of symbols. Nowadays it can be realized by a few chips—quite secretly, unless someone has surreptitiously exchanged the chips. Both of these books discuss physical damage to computers as well. Forget the traditional little magnets for tape sabotage; that is hard to manage. If you want to do in a computer, you cannot much improve on arson. More satisfyingly, you might emulate the Melbourne protesters who "shot a computer with a shotgun that did terrible things to it. It was a total loss." A tax-processing computer in Johannesburg was shot four times by a person firing through a window. Although it was dented, it gallantly continued to extract revenues for the state.

Somehow the system seems to be ahead, with the biggest battalions on its side. What remains is an aftertaste of Bertolt Brecht's bitter doubt: "Which is the biggest crime, to rob a bank or to found one?"

SOUNDS FROM SILENCE: RECENT DISCOVERIES IN ANCIENT NEAR EASTERN MUSIC, by Anne Draffkorn Kilmer, Richard L. Crocker and Robert R. Brown. Bīt Enki Publications, Post Office Box 9068, Berkeley, Calif. 94709 (paperbound book and one long-playing record, $16). ARCHAEOLOGY BY EXPERI-

As a dramatic introduction to membership in the
BOOK-OF-THE-MONTH CLUB®

The long-awaited new edition of the foremost reference work in its field

–completely revised and updated after a lapse of nearly a decade

Yours for only $12^50

[PUBLISHER'S LIST PRICE: $67 50]

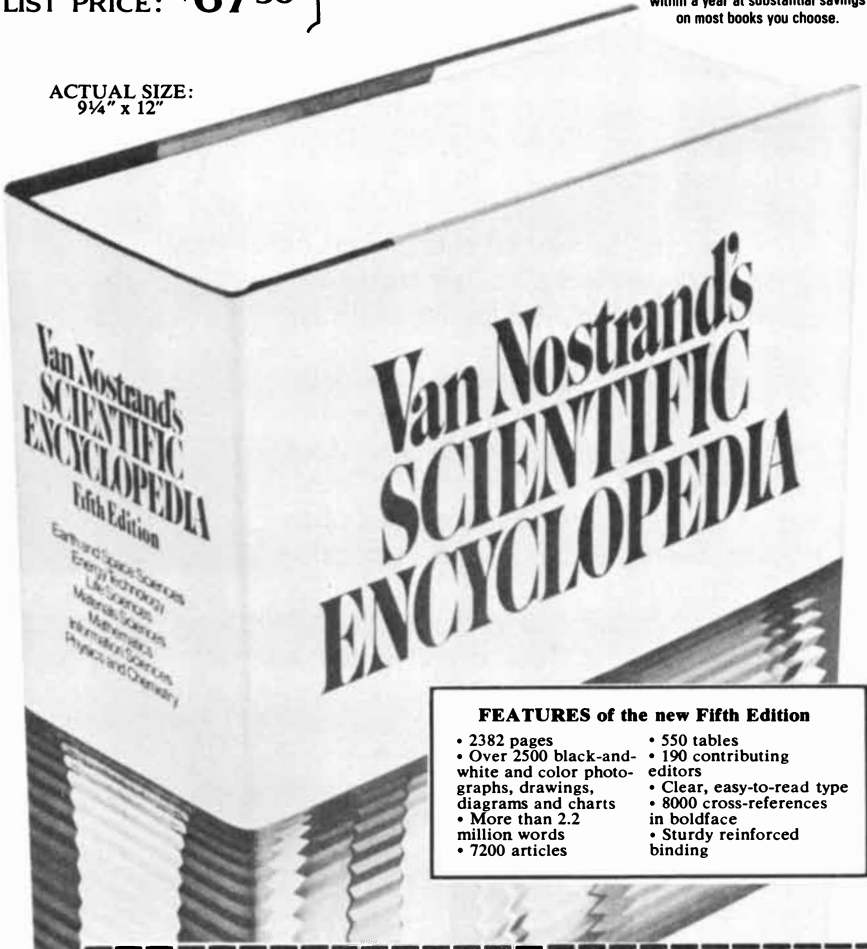
THE SUGGESTED TRIAL: You simply agree to buy four Club choices within a year at substantial savings on most books you choose.

FOR almost 40 years *Van Nostrand's Scientific Encyclopedia* has been acknowledged to be the single indispensable one-volume repository of 20th-century scientific, engineering and technological information. Written with extraordinary clarity, each successive edition of this monumental work has found a wider and wider audience of both professionals and non-professionals who rely on it to answer the myriad scientific questions that arise almost daily in reading, conversation and study.

The fifth and latest edition has just been published, and it is a worthy successor to its predecessors. Extensively revised (80% of the text and 90% of the illustrations are new), the new *Van Nostrand's Scientific Encyclopedia* incorporates all the important advances and discoveries of the past decade in every branch of science, including such diverse areas as computer technology, planetary exploration, systems analysis and oceanography. The section on Energy alone includes, among other entries: 26 pages on Coal; 13 on Natural Gas; 31 on Nuclear Reactors; 20 on Petroleum; 19 on Solar Energy.

The extraordinary opportunity to own this great reference work *at a saving of more than 80%* demonstrates the continuing benefits of membership in the Book-of-the-Month Club. If you remain a member after the trial, you'll keep up with all the important new books and enjoy immediate participation in the Club's unique Book-Dividend® plan—a sensible way to build a superb home library and save *at least 70%* of the publishers' list prices.

ACTUAL SIZE:
9¼" x 12"



FEATURES of the new Fifth Edition

- 2382 pages
- Over 2500 black-and-white and color photographs, drawings, diagrams and charts
- More than 2.2 million words
- 7200 articles
- 550 tables
- 190 contributing editors
- Clear, easy-to-read type
- 8000 cross-references in boldface
- Sturdy reinforced binding

AMONG THE HUNDREDS OF BROAD TOPICS COVERED

- Architecture
- Astronomy
- Biology and Biochemistry
- Chemical Engineering
- Civil Engineering
- Communications
- Computer Technology
- Environmental Science
- Geology and Hydrology
- Graphic Arts
- Guided Missiles
- Instrumentation
- Mathematics
- Mechanical Engineering
- Medicine
- Metallurgy
- Meteorology
- Navigation
- Nuclear Science and Engineering
- Oceanography
- Photography
- Physics
- Planetary Exploration
- Psychiatry
- Rocketry
- Statistics
- Transporation
- Zoology

Prices shown are publishers' U.S. prices. Outside the U.S., prices are generally somewhat higher.

FACTS ABOUT MEMBERSHIP

- You will receive the *Book-of-the-Month Club News,®* a literary magazine published by the Club fifteen times a year. The *News* describes the coming Selection and scores of Alternates, and will be sent to you approximately every three and a half weeks.
- If you wish to purchase the Selection, do nothing and it will be shipped to you automatically.
- If you do not want the Selection—or you would like one of the Alternates or no book at all—simply indicate your decision on the reply form always enclosed with the *News* and mail it so we receive it by the date specified.
- If, because of late mail delivery of the *News*, you should receive a Selection without having had at least ten days to decide whether you want it, that Selection may be returned at Club expense.
- All books distributed by the Book-of-the-Month Club are identical to the publishers' editions in content, format, size and quality.

BOOK-OF-THE-MONTH CLUB, INC.
Camp Hill, Pennsylvania 17012 **7-A170-10**

Please enroll me as a member of the Book-of-the-Month Club and send me *Van Nostrand's Scientific Encyclopedia* billing me $12.50. I agree to buy at least four Selections or Alternates during the first year I am a member, paying in most cases special members' prices. My membership is cancelable any time after I buy these four books. A shipping charge is added to all shipments.

12

MR.
MRS. } ....
MISS
[Please print plainly]

Address .... *Apt.* ....

City ....

State .... *Zip* ....

"They just don't like they used to."

We build them better.

Take safety. Underneath the hood you'll find a safety feature you wouldn't dream of finding in most cars:

Holes.

They're specially designed using a computer, so the front end will help absorb energy in the event of a collision.

We also developed such things as an "independent stabilizer rear axle," which increases the stability of the car on rough roads.

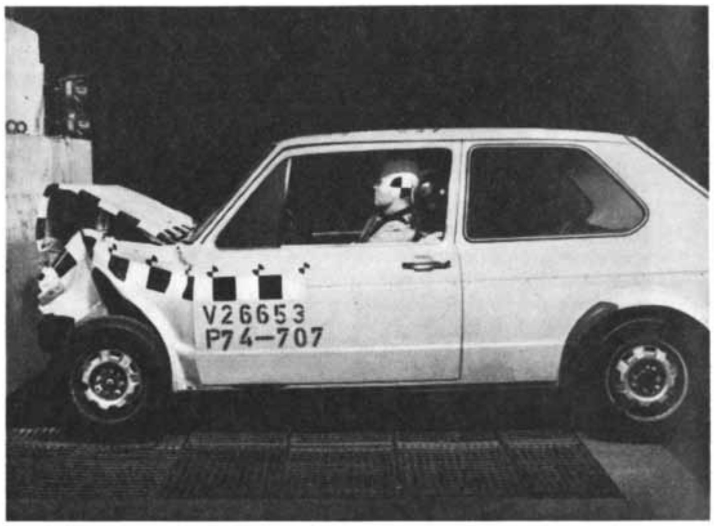
And for our deluxe Rabbit, we developed special seat belts that actually put themselves on when you close the door.

Hardly a detail escaped our attention. Even the ignition key is padded for safety.

What about room? It's probably not the first thing you think about when you say "VW."

Yet today, Volkswagens actually have more combined interior room and trunk space than most other cars in their class.

Rabbits are so roomy they're being used as taxicabs in Lexington, Kentucky.



In fact, Rabbits have more interior room than 25 other cars you could buy. And more trunk space than a Cadillac Seville.

They're also surprisingly quick, with acceleration from 0 to 50 mph faster than a Triumph Spitfire.

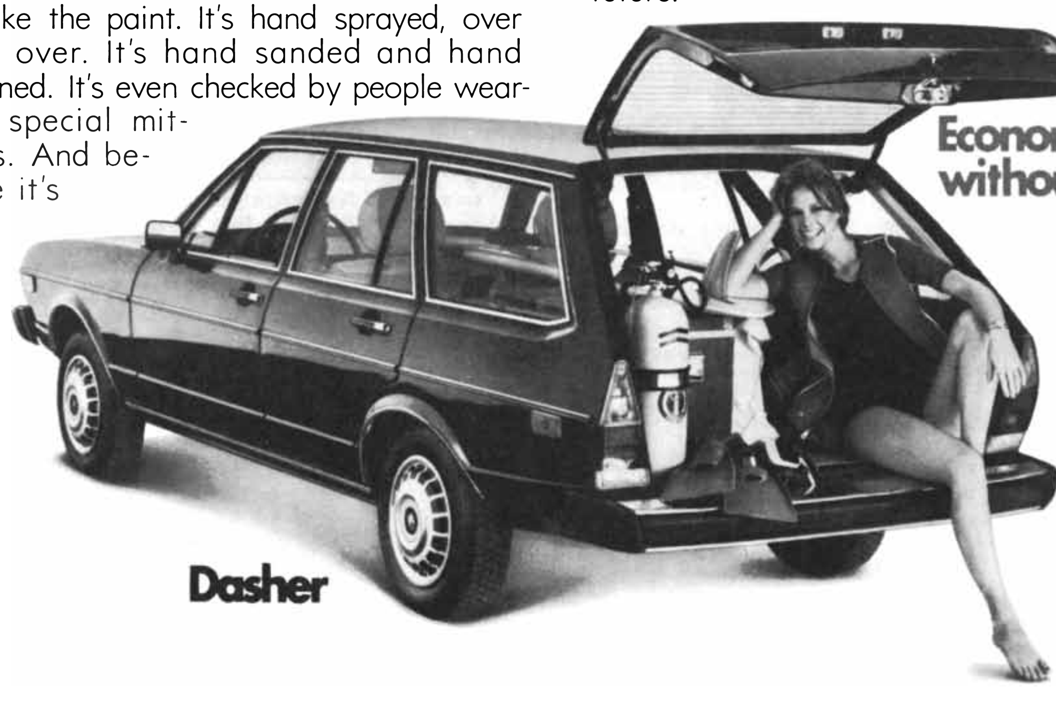
Which brings us to another subject: performance.

Our Scirocco can take you from 0 to 50 in an amazing 7.5 seconds. And last year it was the Trans Am Champion in the under-2-liter class.

What's more, our newest Volkswagens are among the few cars in the world that

Rabbit


43
BOSCH
SCIROCCO
GOODYEAR
Scirocco


Dasher

Nimrud musicians of the eighth century B.C., *from* **Sounds from Silence**

MENT, by John Coles. Charles Scribner's Sons ($3.50). When 50 years back Sir Leonard Woolley was digging out the "Great Death Pit" in the Royal Cemetery of Ur of the Chaldees, he noticed one day only a couple of small holes in the ground near where the pick-man was working. "Something unusual about their shape" spoke to him—his intuition for where to work carefully was famous—and he stopped the digging. They poured plaster of Paris through one hole into the void below it. The form that was revealed was the plaster cast of a royal harp, its wood long since perished, so that only a copper-and-shell ornament remained sticking to the plaster. Nearby there was a lyre, called the Silver Lyre because it had once been ornamented with the metal. We have the precise form of those instruments; the lyre is shown here delicately reconstructed out of the ornamenting fragments as it stands in the British Museum.

Robert R. Brown, a physicist at Berkeley, has made a careful replica of the lyre, after the prototype buried some 4,500 years ago. He used birch and spruce, and he strung the instrument with gut, fastened over a bridge to the sound box. At the top the strings are wrapped around a crossbar with oak pegs twisted in like little belaying pins. These allow controlled variation of the string tension, so that the lyre can be tuned rather sensitively. There is plenty of pictorial evidence on ivory, stone, clay and gold to support the method of tuning and of playing this ritual instrument, both hands plucking the strings. There, however, spade and trowel stop. No dig can recover the vanished melody, long since lost in the noise of time.

Yet the melody is not past plausible reconstruction. Documents of clay remain that suggest what music was like 1,000 years before Pythagoras himself. The physicist-lyremaker has been joined by two Berkeley faculty colleagues, assyriologist Kilmer and musicologist Crocker, to bring back those sounds from silence. The three have made a well-illustrated brochure and a careful long-playing record to allow us to understand and to share their work. The record includes a slow-paced demonstration of the tuning of the lyre and its scale (which turns out to be a very normal-sounding diatonic *C* major) and a livelier account of the proposed music and lyric of an ancient song.

The song is presented in two versions, one for the Ur lyre and a tenor voice and the other for a smaller and somewhat later lyre from Megiddo, in a register suited to a soprano voice. The voice and its string accompaniment sound comfortable and not at all outlandish; the music recalls a quiet fireside hymn. The words are in the non-Semitic Hurrian tongue; we cannot translate them well, but they are a cult hymn concerning the goddess Nikkal, wife of the moon god. One phrase is construed "Thou lovest them in [thy] heart," and the closing phrase is "something close to 'Born of thee'."

The text and musical notation for the song come from a clay tablet pictured on the cover of the brochure (and also drawn by Professor Kilmer to bring out the cuneiform notations). That tablet was among hundreds excavated in ancient Ugarit on the Syrian coast; it was published in 1968. It bears four long lines of the song's words written in Hurrian, and beneath them six lines (and a descriptive colophon) in Akkadian, a dead Semitic language much used in the ancient cities of Ur, Assur and Nippur. In 1970 it was recognized that the obscure language and numbers of the Akkadian text were in fact a musical notation, and there followed several attempts to decipher the score. Kilmer's is the latest, and it is described in this work in a semipopular form.

Four other tablets supply the basis for the inferences of these real tune-sleuths. The first known is a catalogue of hymn titles in Middle Assyrian. Its text was well known, but only in the present study were the technical musical terms in it recognized. A late text from Ur itself supplied a list of strings, relating names to positions. Then in a "mathematical text" a cyclical list of strings and intervals appears. A dozen scholars of music and cuneiform prepared the ground. The 19 preserved lines of the "tuning text" repeat clearly sequential directions for the tuning of a lyre. It is pretty plain that the octave is spanned in seven intervals, whole and half steps as our own.

The only consistent reading allows recognition of the various modes and identification of the mode in which the song itself is written. The actual pitch is of course uncertain, but a match to the lyres and voice is plausible. The key turns out to be without sharps or flats: say *C* major. The crucial step is taken by the count of text syllables and the musical notation; that notation lists with the name of each tone-pair a small number. The burden of the work is that the matter is neatly resolved by the startling conclusion that the tablet records an upper melodic line for the voice and a lower line for the string accompaniment. The root of harmony taps quite deep into the past.

Of course the story is not pat. The texts span almost 1,500 years. The deep lyre from Ur sounds rather unlike the higher strings of Megiddo; the hymn of the Hurrian cult was not in fact known in Ur. Allowing such a long span, it is the oldest music we now know. This is a rare excursion; it is not suited for background music, but for anyone who will attentively read and listen, who can bring to the work the bare elements of theory—what is meant by a minor third or a diminished fifth—it sings a tune full of wonder. Perhaps old Sir Thomas Browne was right when he said: "What song the Sirens sang, or what name Achilles assumed when he hid himself among women, though puzzling questions, are not beyond all conjecture."

The resuscitation of vanished music is an example of experimental archaeology up to date, to be sure with the indispensable use of the word writ plain on clay. Dr. Coles's rich assembly of work in a similar spirit is not new; it appeared four or five years ago, but it is too germane to be overlooked. Archaeology today rightly enjoys high public favor, but it suffers as well a spate of meretricious works, by authors who make a pretty story—and a pretty penny—by some outlandish invocation of mysteries to explain away the brilliant heritage of past cultures. This volume is a powerful antidote to the von Däniken toxins, even though once in a while the entirely unexpected will be found under the spade.

The theme that unites the 200 researches cited in this excellent broad survey over all periods and continents is the idea that one can learn how they did it then by trying the scheme out now. Could the Mayans build a masonry city, the Britons raise Stonehenge, the Easter Islanders erect their top-hatted images? One needs only plan a bit, recruit a few energetic modern Mayans or Wiltshire youths or sturdy islanders and try out a sample task of hauling, quarrying or high rigging. Of course there are pitfalls. It is dangerous to extrapolate an obelisk and an entire pillared temple along the Nile from two hours of pounding with a stone on some granite ledge in the quarry. You need a larger sample than that. Sometimes the pace of work is hard to gauge; the Indians who were set to work near Uxmal in 1965 were promised piecework rates, "but the speed of their work so exceeded my expectations that I had to halt the experiment after thirty minutes." Great Uxmal was not built in a day, we can reckon, but—summing all the types of work involved, the quarrying, transporting and building—after seven and a half million man-days. The labor was "readily within the capabilities" of the estimated local population, whether within a decade in a burst of effort for the biggest single monument or extended to the entire complex within the 250 years of activity at the site.

Smaller tasks too can be studied. The text distinguishes three classes: food production, heavy industry and light industry. There is tree clearance, plowing, the storage and consumption of food, architecture and boat building, and work "at the bench" in stone, wood, bone, metal, hide, clay and even music. Naturally one uses old tools and plausible methods and materials. Ingenuity and skill must be gained, and probably a single shot at a task is a premature basis for strong conclusions. Still, if we can do it, so might they have done it, as the inference must run.

Brave experimenters have eaten a cave mineral, mirabilite; it was salty and laxative, the basis for an advertising claim, it may be, but not the establishment of ancient needs. The rock paints of the African bushman? The mediums tried were varied; wax resin will not make fine lines. Hyrax droppings are no good. Plant juices seem to hold too little pigment. Bile is impermanent. Honey had many problems; it is attacked by ants and by rain. The most successful of all mediums (and still the standard one) was egg-yolk tempera; it yielded long, thin lines with any pigment tried. Of all the local supplies the ostrich egg seems the real basis for those fine paintings, "as near to the truth as prehistoric investigations are likely to get."

The Danish pioneer prehistorian Christian J. Thomsen, founder of the system of Stone, Bronze and Iron ages, was one of the first to try experimental musical archaeology. He tested six bronze *lurer* in the 1830's. What a sound they make, these great twisted horns in pairs! Like the trumpet of Tutankhamun blown in 1939 in radio broadcasts by bandsman James Tappern of the 11th Hussars to sound a trumpet voluntary (but using a modern mouthpiece conceded to be essential to the performance), the sounds out of the past serve both as real evidence and as valuable, if not always authentic, rallying points for scholars and the public alike. More experiments, entire series of them, lasting centers for systematic trials and public demonstrations, growth in incisiveness and scale—these are the recommendations of this University of Cambridge archaeologist, and the fascinated reader gives easy assent.

TO BE AN INVALID: THE ILLNESS OF CHARLES DARWIN, by Ralph Colp, Jr., M.D. The University of Chicago Press ($17.50). "I would sooner be the wretched contemptible invalid, which I am, than live the life of an idle squire," Darwin wrote. No idle squire, he worked in his own quarters on his private income from the time he returned from around the world in the *Beagle* at 27 until the year he died, from anginal attacks and heart failure, at 73. His work is no invalid's meager harvest but makes an inspired solid, sustained and seminal catalogue. And his personal life, by now mythic in the delight and the richness with which our century has recounted it, is not any ill, unhappy story. We know it from the days of a bad student but a good naturalist, to the *Beagle* adventure, to the big comfortable house at Down and the long domestic evenings of song, piano and backgammon. "You all know well your Mother.... She has been my greatest blessing. [She] never failed in the kindest sympathy towards me. [She has been] wise adviser and cheerful comforter.... From your earliest days to now that you are grown up, you have all, sons and daughters, ever been most pleasant, sympathetic, and affectionate." He ended famous, well-to-do, happy among the seven of his 10 children who survived childhood, a little coy in his pleasure that "his name ought to last a few years" and that he "should have influenced... the beliefs of scientific men on some important points."

Still, he was chronically ill, with half a dozen long trips for Victorian hydropathy (baths, douches, sitz baths and the like), a series of well-known and expensive physicians (some with views that induce astonishment in the modern reader) and a number of year-long periods in life during which he was hardly able to eat, sleep or work well. It was not all loss; he almost never spent time on the demanding ceremonies of London middle-class society. A perceptive observer wrote of Darwin at 60: "Upstairs Mrs. Darwin, Miss D. and Mr. Charles Darwin himself—yellow, sickly, very quiet. He has his meals at his own times, sees people or not as he chooses, has invalid's privileges in full, a great help to a studious man." The nature of his illness comes clear in this physician's meticulous study as never before. Blurred by the reticence of various editors and biographers, it is here made plain in doctorly style, the sources as far as possible manuscript letters and notebooks themselves.

Darwin writes to Joseph Hooker: "I should suppose few human beings had vomited so often during the last five months.... I seldom throw up food, only acid and morbid secretion.... During more than a month I vomited after every meal and several times most nights." Palpitation of the heart, eczema of the hands and perhaps lips were frequent in addition to stomach disorder. The puzzle is of course diagnosis. What was Darwin's illness? His physicians did not know, nor did he, in spite of much self-study, even to a daily journal he carefully kept for more than five years, his "diary of health." In Dr. Colp's book Darwin's illness is followed in detail over his lifetime, displaying a wealth of Victorian medical practice and social custom, a most readable addition to all we now have of this man's life.

The last part becomes a 40-page medical analysis. There is by now a very full set of discordant theories. Consider only "suppressed gout," popular among his own doctors, whatever it could mean today; the result of his dreadful seasickness aboard the *Beagle;* porphyria, a congenital metabolic disease, hypothesized from the symptoms recorded in his grandmother; chronic arsenic poisoning, from self-medication or the pharmacological enthusiasms of a couple of his physicians ("Fowler's solution"); mere hypochondria, half-conscious play-acting for his tender wife; Chagas' disease, a trypanosomal infestation, assumed to be contracted in Chile from the repeated bites Darwin describes he received there from the horrid blood-sucking beetle now known to be its common vector; brucellosis; even allergy to pigeons!

It will be not much of a surprise to hear that the author, a practicing Columbia psychiatrist, opts for the much-favored recent diagnosis of psychosomatic illness. Dr. Colp is a most reasonable man and a realist. He recognizes that without electrocardiograms and tests for protozoa, liver function, uric acid, porphyrin levels, even an autopsy, we cannot really exclude an organic cause. He nonetheless shows that no present diagnosis fits everything we know, and he comes down for psychological stress, mainly brought on by Darwin's growing recognition of the struggle he would perforce enter into once his evolutionary views became known. Colp avoids excesses of minute interpretation, such as Darwin's variable use of capital letters on the word "father" in a few places, which others have tried to make significant. Rather, he depends on study of Darwin's own complaints and on the time sequence of the bouts of illness.

This does not all work so well either. Maybe there was more than one illness over the years. We have yet to learn enough about the mind-body interaction to diagnose even patients who are still alive. The entire story is sympathetic, and we take from it something at depth about those times and indeed our own. The Latinity of psychiatrists is left open to doubt; there is evidence in the footnotes that they know German better! That is a mere cavil about one odd remark; this is altogether a first-rate small authoritative volume for the many who enjoy any facet of the Darwin epic or merely the strong scent of Victoria's reign.

A NATURAL HISTORY OF MARINE MAMMALS, by Victor B. Scheffer. Illustrations by Peter Parnall. Charles Scribner's Sons ($7.95). Nearly 50 delicate stippled drawings by a well-known illustrator embellish and document this deft little text by a biologist-writer. Scheffer has written a couple of much denser books in which he took the reader through an imagined year in the life of a typical individual animal. Now he turns his pen to a synoptic style; he seeks to introduce the reader easily to the diversity of living species of marine mammals. His devices are workable and skillfully utilized. He builds an evolutionary framework, chapter by chapter telling the adaptations of these cousins of ours, their feeding, senses, intelligence measures, reproduction and growth, diseases and other hazards (mainly us). He ends with an account of where to see the animals ranging free and a technical listing of the species.

The freshest material here is an account of the six groups of marine mammals. They are all, of course, representatives of footed mammals who long ago gave up hoof and claw for web foot, flipper or fluke. Should you doubt it, consider the whales and porpoises occasionally born (perhaps once in 100,000 births) with useless, partly formed vestigial legs. Baleen whales *in utero* have tiny teeth, doomed never to harden and break through the gum. It is easy to order the groups by the shapes of their feet, correlated with the recency with which they set to sea. Last to enter were the sea otters, which have normal front paws; there is only one species of them, and they came from crab-eating land otters only five million years ago. The walking or eared seals, which can still use their hind flippers in support, are next; they derive from a bearlike carnivore of some 30 million years ago. The crawling seals, which drag their hind flippers, arose from an otterlike ancestry about at the same time. Then come the sirenlike manatees and dugongs, whose hind legs are entirely gone after 55 million years, and the quite legless cetaceans, toothed or baleen-jawed, 60 million years down the evolutionary stream from some ancient hoofed forebear. Are those present powerful "dabblers in the sea," the polar bears, getting ready to leave the land?

Brain weight is tabulated and discussed, along with neocortical structures and orientation toward play. "But perhaps, after all, the question of intelligence . . . should not be raised. Instead of wondering whether porpoises can reason, or worry, or plan ahead, . . . perhaps we should simply admire their complex brains—those mysterious organs which allow them to cope with a dim, cold, watery world in which we ourselves could not long survive." We may have something to learn from fellow creatures that lack the ability "to drive harpoons through living flesh."

HAILSTORMS AND HAILSTONE GROWTH, by Narayan R. Gokhale. State University of New York Press ($30). The intricacy of the snowflake faces the physicist with a challenge; to this day it is not at all clear how those symmetrical crystals grow. But the hailstone looks like a frozen raindrop, and the theorist might regard his work as done. This volume by an expert is about the only book to collect what we know and do not know about hailstorms and what they drop on the fields below. The theorist is not found wanting; he can pretty well explain the hailstone. It is by no means a single frozen droplet. Here one is in a splendid color shot: a thin section through a hailstone illuminated by white light through crossed polaroids displays a patchwork of hues as complex as an abstract painting.

The key to the structure is the nature of hailstone growth. Take a cold embryo hailstone a few millimeters across, in fact a frozen drop. It grows by accreting cloud droplets as it falls through an updraft. There are two main kinds of growth. One is dry growth, the droplets individually freezing on the cold stone as they appear. This tends to make bubbly opaque ice—small crystals in random orientation. Wet growth occurs at higher temperatures, when the latent heat of fusion is enough to melt the outermost layers. The ice is formed in a wet skin of liquid water, and tends to freeze clear in larger well-oriented grains. Since these processes depend on temperature and hailstone radius, it turns out that a layered grainy structure like the one usually seen (first described by that same Allessandro Volta we know from the voltaic pile) can develop even under smooth variation of updraft conditions. More complex layering, even stones with icicle-like protrusions, are observed; the real turbulent cloud is no simple environment. This work reflects much experience with artificial growth of hail in a hail tunnel, first built by Swiss researchers and now found in several laboratories worldwide.

Hail is damaging to aircraft and particularly to crops. Much effort is spent on the detection and measurement of hail by radar; these issues are taken up in detail. Still more effort is spent on trying to forecast and to suppress hail. The forecasting is getting better; severe hail is the companion of fierce thunderstorm updraft conditions. Suppression by exploding rockets has long been claimed; nowadays Russian workers use accurate artillery and rockets to inject silver iodide dust into the hail-forming volume. Commercial hail suppression by ground-based silver iodide generators is the popular American counterpart. The profession at large has still to be convinced of the success of either approach.

Severe turbulence and cold, wet clouds go together under continental conditions. There is not much hail in the polar regions; the thermal energy is too small for big storms. Nor is there much in the Tropics; melting below the clouds makes rain. Prime hailstone weather occurs in continental mid-latitudes, late of a summery afternoon. The worst places are perhaps the Great Plains of the U.S., the Po Valley of Italy, with its tender crops of grape and fruit, and one tropical exception: a high-altitude region of inland Kenya with possibly "the highest incidence of hail in the world." The commonest severe storms seem to grow stones one to two centimeters in diameter in all these places. Such a stone falls with a terminal speed (calculated from various drag measurements in wind tunnels and from scaled-up model hailstones falling in the open air) of 20 to 30 meters per second, a mile a minute and more. One experimenter timed oblong and knobby objects made of polyethylene as they fell down a mine shaft with a vertical drop of 1.5 kilometers! In 1970 a whopper of a hailstone was picked up in Coffeyville, Kan., prime hailstone country, that weighed in at just under a pound and three-quarters. The cost of hailstorm damage in the U.S. runs to about $300 million a year; at the caliber of the Coffeyville stone matters could be a lot worse.

This is a remarkably complete and expert reference, but it is high-priced for a book set by typewriter and not justified on the right-hand margin. We welcome the one fine color plate.

The extinct marine mammal **Desmostylus,** ***from*** **A Natural History of Marine Mammals**

THE SEAGRAM'S GIN AND JUICE SECRET?

Seagram's Extra Dry. The Perfect Martini Gin. Perfect all ways.

Seagram Distillers Co., N.Y.C. 80 Proof. Distilled Dry Gin. Distilled from Grain.

Established 1845 SCIENTIFIC AMERICAN October 1977 Volume 237 Number 4

Peer Review and the Support of Science

A statistical analysis of the evaluative procedures on which the National Science Foundation bases its funding decisions provides no evidence to substantiate recent public criticisms

by Stephen Cole, Leonard Rubin and Jonathan R. Cole

For more than 25 years the National Science Foundation has played a major role in the expenditure of public money for the support of science in the U.S. Currently the NSF accounts for about 20 percent of the funds distributed by the Federal Government for basic scientific research and more than 30 percent of the Federal funds allocated for such research at universities. The NSF awards its grants on the basis of a decision-making process commonly known as peer review. The term is derived from the fact that the Government officials responsible for deciding which investigators receive grants rely on the evaluations of other investigators in the same discipline.

In recent years the peer-review system has been attacked for a variety of reasons by certain members of both the scientific community and the Congress. Hearings on the alleged inequities of the peer-review system were held two years ago by a subcommittee of the House Committee on Science and Technology.

In an effort to assess the validity of the public criticisms of the peer-review system raised in the Congressional hearings and elsewhere we have been engaged for more than a year in a sociological study of the operation of the peer-review system at the NSF. This study, which is being conducted for the National Academy of Sciences, is supported by grants from the NSF; we have nonetheless had complete autonomy from the NSF in conducting our research. Our results to date have yielded little evidence in support of the main criticisms that have been made of the peer-review system. On the contrary, we have tentatively concluded that the NSF peer-review system is in general an equitable arrangement that distributes the limited funds available for basic research primarily on the basis of the perceived quality of the applicant's proposal. In particular, we find that the NSF does not discriminate systematically against noneminent scientists in the ways that some critics have charged. This is not to say, of course, that there are not errors in individual cases.

How does the NSF peer-review system work? To begin with, a scientist who wants to obtain NSF funds prepares a written proposal describing his past research, his qualifications and the new research he intends to do if he receives funds from the NSF. This proposal is usually submitted to the NSF through the scientist's institution, in most cases a university.

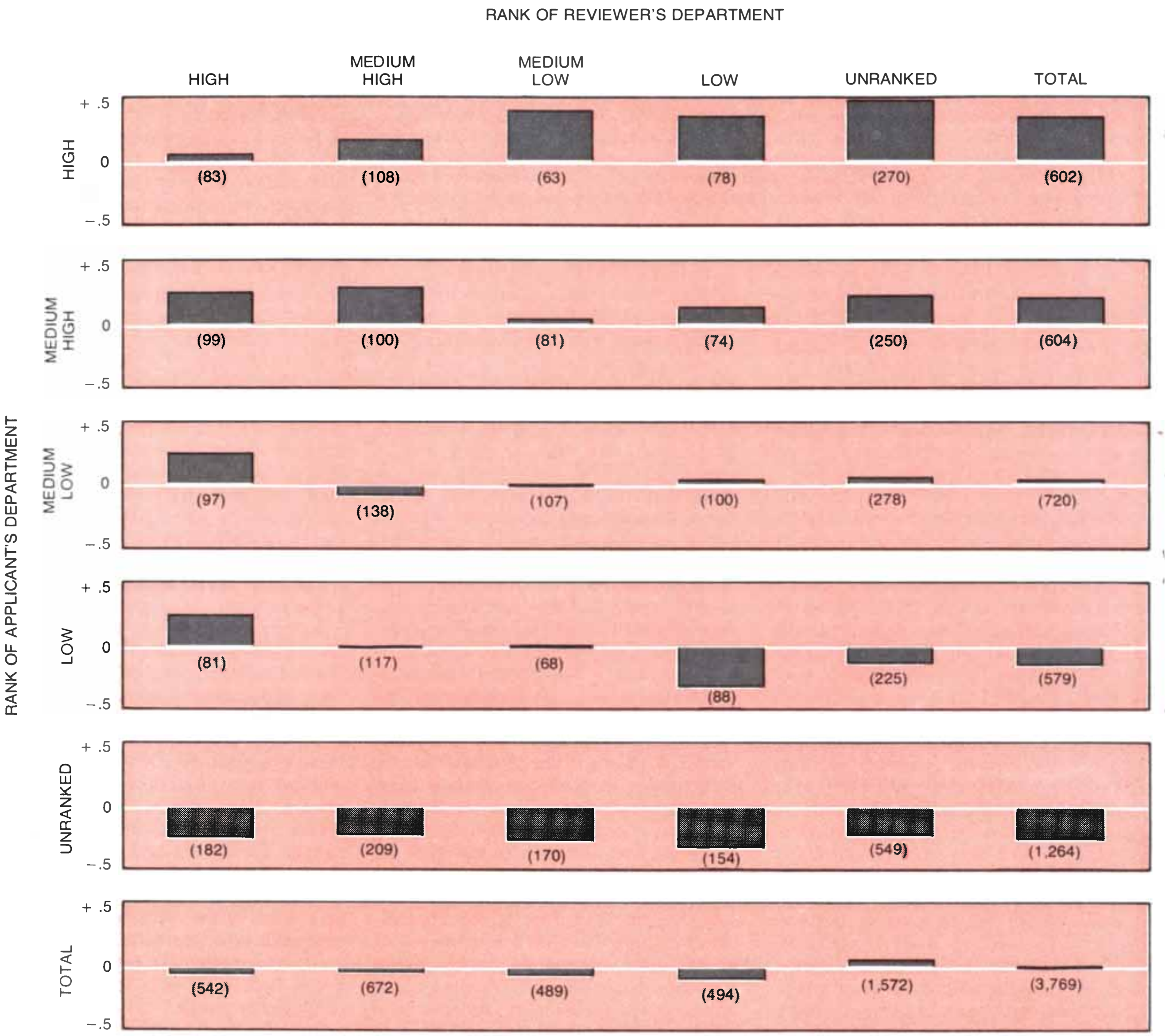
The staff of the NSF is divided into approximately 80 program areas corresponding to the various scientific disciplines and subdisciplines. (The chemistry section, for example, is divided into eight different programs.) When a research proposal comes to the NSF, it is assigned to the appropriate program and is thereafter handled by an employee of the NSF called the program director. On receiving a proposal the program director generally looks it over to determine its specific subject area. He then selects a number of reviewers who are sent the proposal by mail. The reviewers are asked to rate the proposal as being excellent, very good, good, fair or poor and in support of their rating to present written comments evaluating the proposal. In some programs an independent evaluation of the proposal is also made by a panel of scientists who meet with the program director three times a year in Washington.

The NSF explicitly states to its reviewers the criteria that should be applied in evaluating the proposals. The main criteria are (1) the significance of the scientific investigation described in the proposal, (2) the ability of the applicant to carry out the proposed research and (3) the capacity of the applicant's institution to support the type of research in question. Where all these factors are roughly equal, another set of criteria, including the geographic location of the applicant's institution, may be considered. Heavy emphasis is placed on the quality of the work described in the proposal and on the past research performance of the applicant.

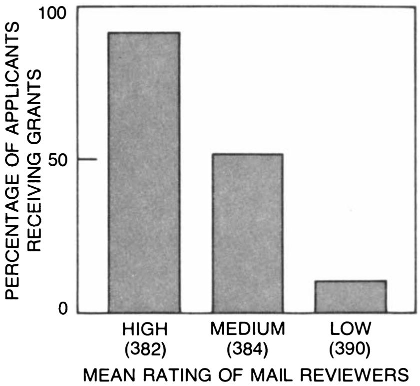
The most fundamental criticism made of the NSF peer-review system is that it leads to inequitable decisions. Critics charge that scientists who are most capable of advancing science are sometimes denied grants and that scientists who are doing less significant work are given grants. Former Representative John B. Conlan of Arizona, for example, asserted at the Congressional hearings that peer review is essentially an elitist system run primarily for the benefit of a clique of eminent "old boys." He said: "I know from studying material provided to me by the NSF that this is an 'old boy's system' where program managers rely on trusted friends in the academic community to review their proposals. These friends recommend their friends as reviewers.... It is an incestuous 'buddy system' that frequently stifles new ideas and scientific breakthroughs, while carving up the multimillion-dollar Federal research-and-education pie in a monopoly game of grantsmanship."

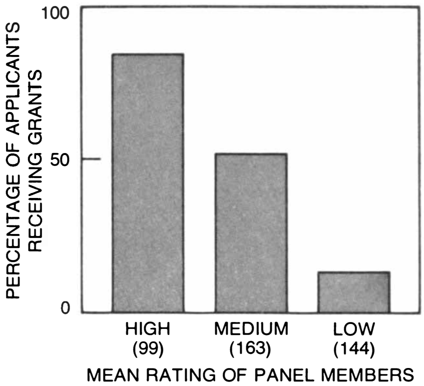
Critics in and out of Congress maintain that the main organizational condition that gives rise to this unfair distribution of support is the extraordinary power in the hands of the program directors to decide who should get funds. The program director is alleged to be at the center of the old-boy network in which reviewers favorably evaluate the proposals of their friends, eminent scientists favorably review the proposals of other eminent scientists and funds are denied to scientists who are not part of the exclusive old-boy system.

Further abuse is said to be possible because the reviews received by the program director are only advisory, leaving him free to ignore them, and because the program director can predetermine the outcome by selecting a biased group of reviewers. The critics argue that knowledgeable program directors deliberately select reviewers who will be either hard or easy on a particular proposal. Even if



STATISTICAL ANALYSIS of 3,769 peer-review ratings given by various mail reviewers to 1,200 applicants for basic-research grants from the National Science Foundation in the fiscal year 1975 was aimed at testing the "old boy" hypothesis, which holds that the proposals of eminent scientists are apt to be rated more favorably by eminent reviewers than by other reviewers. The ratings in the 10 different program areas studied were first converted into standard scores in the following manner: Within each field the mean rating was set at zero, and the rating received by an applicant was then expressed in terms of the corresponding number of standard deviations above or below the mean rating. A high number means a comparatively favorable rating, and vice versa. Both the applicants and the reviewers were separately classified according to the prestige of their current academic department, as determined in an independent survey. Thus the entry in the upper left-hand corner of the table signifies that there were 83 reviews by reviewers in high-ranked departments of proposals submitted by applicants from high-ranked departments; on the average these reviews yielded ratings that were .05 of a standard deviation above the mean. Since it appears that proposals from applicants in high-ranked departments are actually rated lower by reviewers from high-ranked departments than by reviewers from lower-ranked departments, in this sample at least the data offer no support for the old-boy hypothesis. The analysis does show that applicants from high-ranked departments are slightly more likely to receive favorable ratings than are those from unranked departments, but there is no evidence that this outcome is the result of inequitable treatment.





NSF PROGRAM DIRECTORS appear to rely heavily on the evaluations of the peer reviewers in deciding whether or not a research proposal is to be funded. As the chart at left shows, among the 382 applicants who received comparatively high ratings from the mail reveiwers 92 percent were awarded grants, whereas among the 390 receiving low mean ratings only 10 percent received grants. Similarly, as the chart at right shows, among those proposals that received comparatively high ratings from an independent panel of peer reviewers 84 percent were funded, whereas among those that received low panel ratings only 12 percent were funded. Evidently peer-review ratings are the most important determinant of the program director's decision.

the program director feels compelled by the reviews to support a proposal he dislikes, he can effectively stifle the research by reducing the size of the budget. The program director can supposedly do so because there are no effective checks on his power either inside or outside the NSF. In short, there is no appeals system to challenge the decisions made by the program director.

Critics assert further that the NSF cloaks its activities in secrecy in order to protect the old-boy system, refusing to allow Congressmen or others to see verbatim reviews or to learn the names of the reviewers of particular proposals. This protective shield of confidentiality enables the old-boy system to function unchecked and prevents effective oversight of the NSF by Congress. The ultimate consequence is that the peer-review system actually stifles innovative research, since the eminent scientists who serve as reviewers are likely to reject ideas that differ from their own.

In our study of the peer-review system we decided to limit ourselves at first to an examination of how peer review works in just those NSF programs responsible for the funding of basic research. We have not studied peer review in the NSF's applied-research programs or in its educational programs. Furthermore, we chose a sample of only 10 basic-research programs for detailed study: algebra, anthropology, biochemistry, chemical dynamics, ecology, economics, fluid dynamics, geophysics, meteorology and solid-state physics. Because our intensive analysis included only about an eighth of the NSF's basic-research programs our results may not be generalizable for the entire organization. We are currently conducting follow-up studies of other programs.

Our investigation has combined both qualitative and quantitative sociological techniques. We began by conducting 70 in-depth interviews with scientists involved at all levels of the peer-review system, including program directors, former program directors, mail reviewers, review-panel members and supervisory-level NSF officials. We also scrutinized more than 250 specific research proposals, read all of the peer-review comments on those proposals and examined all of the correspondence between the applicant and the program director. In some cases in which our analysis of the applications raised specific questions about how the peer-review system worked in that particular situation we went back and reinterviewed program directors with the files in hand.

In addition, we conducted a quantitative analysis of 1,200 applicants to the NSF in the fiscal year 1975. (Roughly half of the applicants were ultimately awarded grants.) The purpose of the quantitative study was to identify those characteristics that were correlated with the receipt of a grant from the NSF. Were Representative Conlan and the other critics of peer review correct in their assertion that eminent scientists have a great advantage in the competition for funds and that less eminent scientists, particularly younger ones, are at a serious disadvantage? We shall try to answer this question by summarizing below some of the results obtained so far in our study.

One of the main charges of the critics is that the NSF program director can predetermine the outcome of the peer-review process by sending a proposal to scientists who he knows in advance are biased either in favor of the proposal or against it. We shall call this view the old-boy hypothesis. Presumably the proposals of eminent scientists who are members of the old-boy network are sent to other eminent scientists who give their eminent colleague a favorable evaluation. In return, of course, the reviewers expect reciprocity when their proposals are sent to other members of the old-boy club. Equally important, the proposals of less eminent scientists, who are not part of the network, are sent to scientists who will give them lower evaluations than they deserve. Although we have no direct evidence that the program directors either do or do not select reviewers with a certain outcome in mind, we can see if the outcomes are consistent with the old-boy hypothesis. Are the proposals of eminent scientists actually rated more favorably by eminent reviewers than by other reviewers?

To test this hypothesis we classified both the applicants and the reviewers according to the prestige of their current academic department, as determined by a survey conducted in 1969 by the American Council on Education. The ratings given to the applicants by the reviewers in the 10 programs we studied were standardized separately before being combined into one large table [*see illustration on preceding page*]. For example, there were a total of 83 cases in which an applicant from a high-ranked department had his proposal reviewed by someone who was also from a high-ranked department. The number associated with this particular applicant-reviewer pair (+.05) indicates the average rating (in standardized units) given by high-ranked reviewers to proposals from high-ranked applicants. The higher the number, the higher the rating.

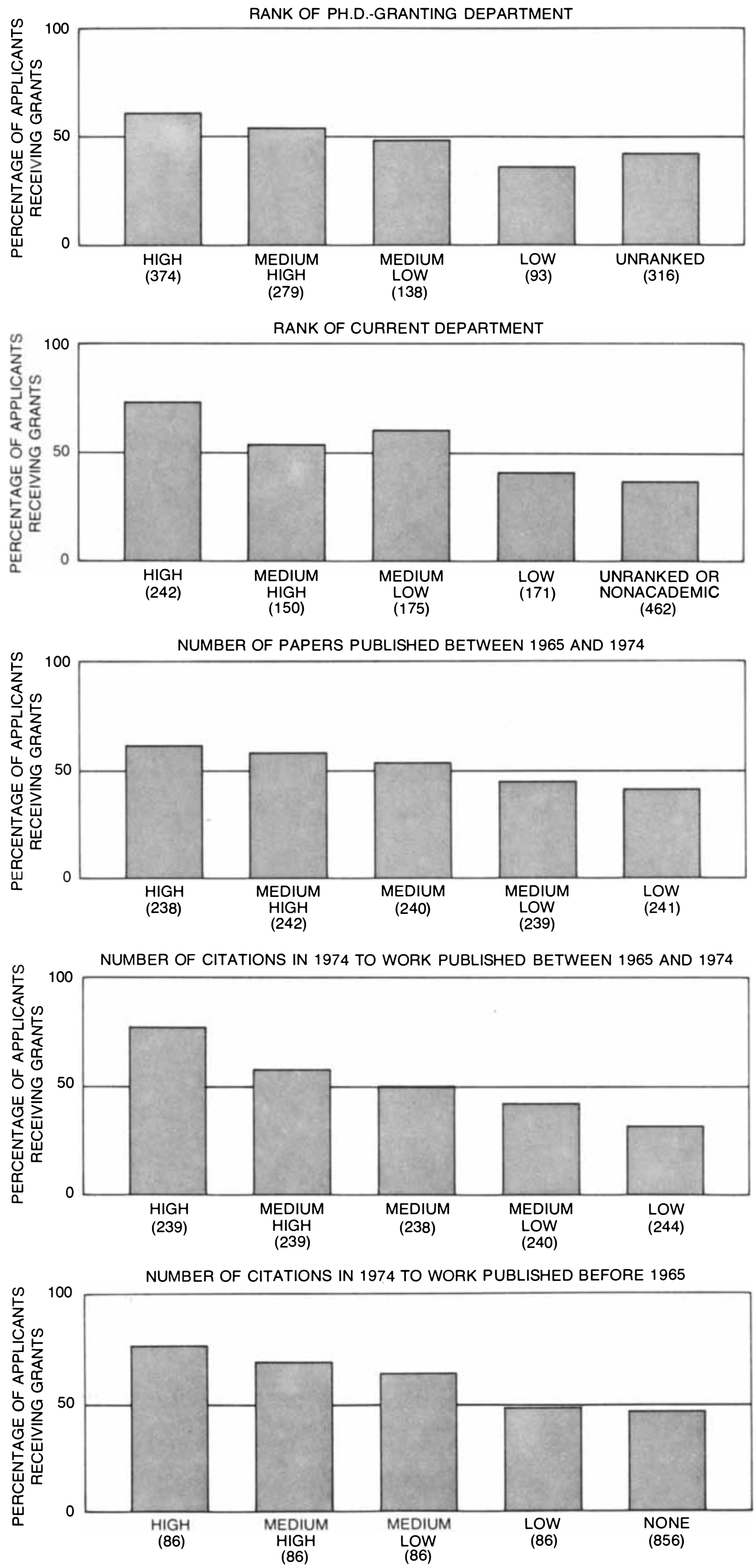
In general we found that applicants from high-ranked departments received slightly better reviews of their proposals than applicants from medium-ranked and low-ranked departments. Furthermore, it appeared that high-ranked reviewers tend to be slightly more lenient with proposals than low-ranked reviewers are. These results, in and of themselves, cannot be interpreted as offering support for the old-boy hypothesis. For example, the fact that eminent scientists tend to get higher ratings could simply be a result of the higher quality of their proposals or of the belief on the part of the reviewers that the eminent scientists are in fact better able to carry out the proposed research.

In order to explore the matter more deeply we next conducted a statistical analysis of variance that compared the observed mean rating for each applicant-reviewer pair with the expected mean rating, assuming no bias. The results of this analysis indicated that in general reviewers from high-ranked departments were not disproportionately favoring proposals from applicants in similarly high-ranked departments. We conducted this analysis separately for each of the 10 programs. In only one program were reviewers at high-ranked departments detectably more lenient toward the proposals of their colleagues at similarly high-ranked departments.

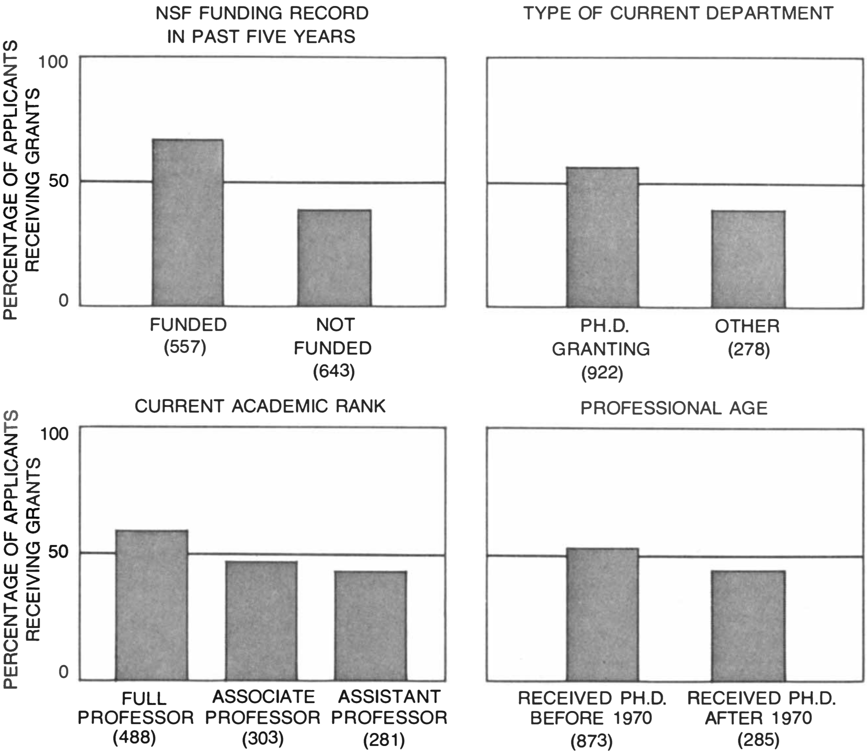
Another statistical analysis of variance tested the reviewers' bias in terms of geographic location and of the relative eminence of the reviewer and the applicant. It showed no significant tendency for scientists in one part of the country to favor proposals from colleagues in their own region or for eminent scientists to favor the proposals of eminent scientists over the proposals of less eminent scientists. Thus even if it were true that the program directors at the NSF were attempting to manipulate the outcome of the peer-review process by their selection of reviewers (and our qualitative findings indicate that it is unlikely), the quantitative data suggest that they have not been successful.

One reason it is difficult to test the validity of the old-boy hypothesis is the absence of conceptual clarity in the charge. What is referred to by the old-boy label? There are at least three possibilities. The term could refer to investigators with a common view of their field who will only appraise favorably work that is done by people with similar views. It could refer to networks of friendships: scientists who know one another, who "grew up" together or attended the same schools and who tend to fraternize and also to favor one another's proposals. It could refer to social position: scientists at a given level of eminence might tend to favor the proposals of others who are similarly situated in the hierarchy of science, even if they have no personal contact with them. Critics of the peer-review system never specify clearly which form of old-boyism is undermining the peer-review system. The data reported here allow us to examine the assertion that persons of similar rank, similar intellectual background and similar repute favor one another's proposals, but we do not have in hand data for examining forms of old-boyism that may be connected with friendship patterns.

How do the characteristics of the applicants affect the peer-review ratings they receive? Critics of the peer-review system say that regardless of the quality of proposals eminent scientists enjoy an advantage over those who are



CHARACTERISTICS of successful applicants for NSF grants in 1975 are summarized in these bar charts. Among the characteristics represented here are rank of Ph.D.-granting department (*top*), rank of current department (*second from top*), number of scientific papers published between 1965 and 1974 (*middle*), number of citations to work published between 1965 and 1974 (*second from bottom*) and to work published before 1965 (*bottom*).



OTHER CHARACTERISTICS of successful applicants for grants in 1975 are represented in these charts. Characteristics include past five years' funding record (*top left*), type of academic department (*top right*), academic rank (*bottom left*) and professional age (*bottom right*).

not eminent. In the final analysis, these critics contend, the peer-review system results primarily in eminent scientists at high-ranked departments having an unfair advantage in grant approval over less eminent scientists at lower-ranked departments. To test this "rich get richer" hypothesis we combined the applicants from all 10 programs into one large standardized sample. The 1,200 applicants in the sample were characterized by nine variables that established their status in the social system of science. Each of these characteristics was then tested separately to see if it provided evidence in support of the rich-get-richer hypothesis.

For example, we characterized the applicants according to the graduate departments from which they obtained their doctoral degree to see if scientists that come from prestigious Ph.D.-granting departments tend to get higher ratings than those who come from less prestigious departments. The applicants were also classified according to their current academic departments in order to test the assertion that applicants in high-ranked departments have an undeserved advantage over applicants in low-ranked departments. We classified the applicants according to their current academic rank in order to see if assistant professors are any less likely to receive grants than associate professors or full professors. In addition we classified all the applicants according to their professional age, their published scientific works, the number of citations of their published works and whether or not they had received NSF funds in the past.

The rich-get-richer hypothesis would suggest the existence of strong correlations between all of these variables and the ratings the applicants received on their proposals. There are, indeed, reasons other than old-boyism for this expectation. For one thing scientists who in the past had done research that other scientists had valued highly could reasonably be expected to write proposals that would be more likely to be rated highly. Moreover, since the NSF explicitly instructs reviewers to regard past performance as one of the major criteria in determining a rating, reviewers could be expected to give higher ratings to scientists with a superior "track record."

The data, however, provide little support for the rich-get-richer hypothesis. Our results show only weak or moderate correlations between each of the nine "social stratification" variables and the ratings received on proposals. The most highly correlated variable was the number of citations in the 1975 *Science Citation Index* of work published between 1965 and 1974. Even this rough measure of the significance of recently published work is not correlated very strongly with the ratings, explaining only 6 percent of the variance in the ratings. The correlations between the other variables and the ratings are all surprisingly low, explaining only an additional 5 percent of the variance in the ratings. In the end 89 percent of the observed variance in the ratings is left unexplained by the nine variables.

These results ran so counter to our expectations that at first we suspected they might have been caused by some methodological error. A thorough review of our correlation and regression procedures, however, left the results intact. In fact, the validity of our findings has been corroborated by a recent study conducted by members of the NSF's own chemistry section. Their independent analysis yielded results that were virtually identical with our own. It is difficult to avoid the conclusion that there is no substantial correlation between peer-review ratings received by applicants for NSF grants and statistical indicators of their professional status or past scientific performance. Scientists whose published work is frequently cited were only slightly more likely to receive favorable ratings than scientists with only a few citations or none.

It still appeared possible, however, that the weak correlations we observed could have resulted from a lack of agreement among the reviewers. For example, if an applicant with a large number of citations of his work received very favorable ratings from some reviewers and very unfavorable ones from others, that could account for a weak or nonexistent correlation between citations and ratings. How much agreement was there among the various reviewers of a given research proposal?

To answer this question we first determined the mean standard deviation of the reviewers' ratings, a quantity that can be taken as an approximation of the degree of agreement in a given field. This number varied from a low level of .31 in algebra to a high level of .69 in ecology and meteorology. (A low mean standard deviation corresponds to a high degree of consensus, and vice versa.) This approach could itself be flawed, however, if one were to fail to take into account the mean rating of the reviewers in each field. Clearly if there is a general tendency in a field to restrict the range of evaluations to either high or low scores, there would be less chance for variations in the ratings. We therefore relied on a statistic called the coefficient of variation, which is simply the mean peer-review rating divided by the mean standard deviation. In general we found that there was a good deal of agreement among the mail reviewers in all 10 fields and little systematic variation among the fields. The coefficient of variation ranged from a low of .13 in economics to a high of .30 in ecology.

To test further the notion that the weak correlations we observed resulted from a lack of agreement among the reviewers, we examined the correlations between the mean rating received by a proposal and several characteristics of the applicant. If the weak correlations had resulted from a lack of agreement among the reviewers, the associations between mean ratings and individual characteristics would be substantially higher, since mean scores are almost invariably more strongly correlated with any given variable than are individual scores. When the mean rating was used as the dependent variable in a statistical regression analysis, we obtained results similar to those obtained in our original analysis. The highest correlation was found between citations of recent work and the mean rating, followed by the correlation between past funding history and the mean rating. Although this method of analysis had the effect of increasing the amount of variance explained by the characteristics of the applicants from 11 percent to 16 percent, the great bulk of the observed variance in the ratings remained unexplained. The new analysis supported the conclusion that the weak correlations observed were not a result of a lack of agreement among reviewers.

In short, these data suggest that the mail reviewers are not strongly influenced by the professional status of an applicant in evaluating a proposal. On the contrary, they appear to be much more likely to be influenced by their perception of the quality of the research proposed. One crucial question remained: How is the program director's funding decision related to the reviewers' ratings on the one hand and to the characteristics of the applicants on the other?

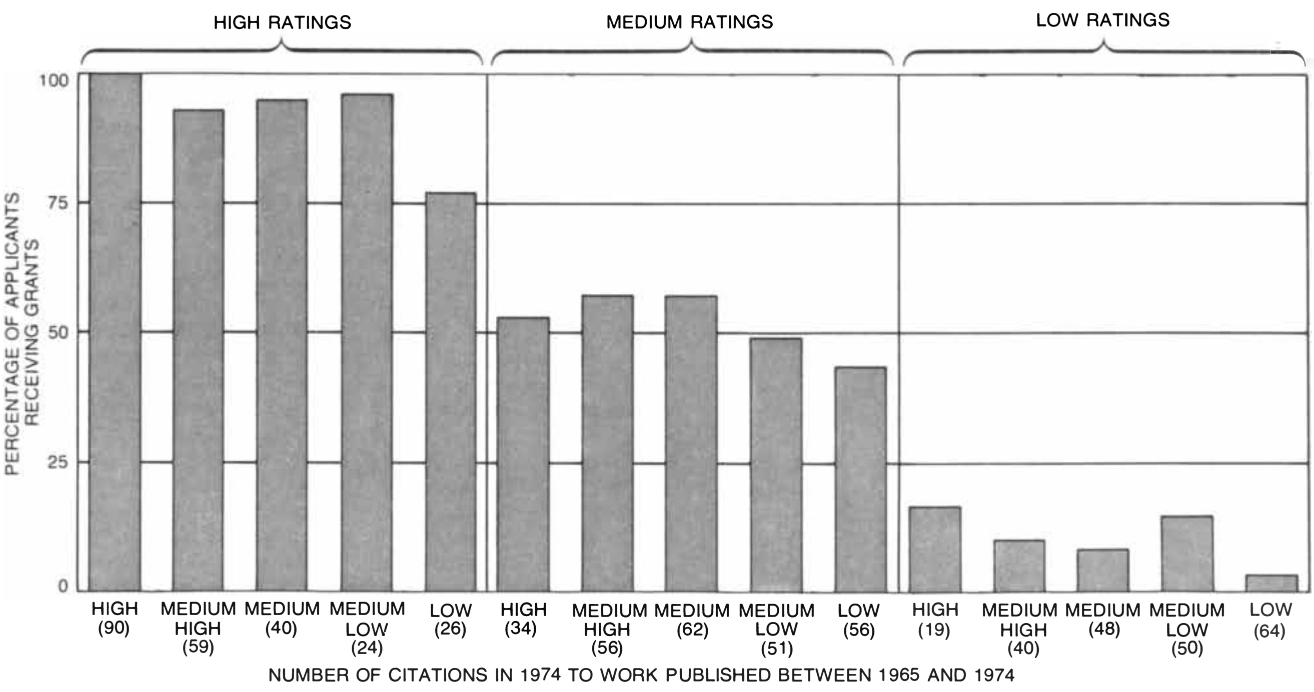
Critics of the peer-review system contend in effect that the decisions of the NSF program directors depend more on who you are than on what you propose to do. So far our data have tended to refute this version of the old-boy hypothesis. Before this refutation can be established conclusively, however, we must establish that the peer-review ratings are the single most important determinant of the program director's funding decision and that the characteristics of the applicants have little independent effect on the outcome.

The NSF states clearly that the reviews by either the mail reviewers or the panel members are advisory and the program director has the final responsibility for deciding whether or not a proposal is to be funded. Our data show that the program directors in fact rely very heavily on the evaluations of the peer reviewers. For example, among those applicants who received comparatively high mean ratings from the mail reviewers 92 percent were awarded grants, whereas among those receiving low mean ratings only 10 percent got grants. Among the group who received mean ratings in the middle ranges about half were awarded grants. Similarly, among those applicants who received comparatively high panel ratings 84 percent were funded, and among those who received low panel ratings only 12 percent were funded [*see illustration on page 36*].

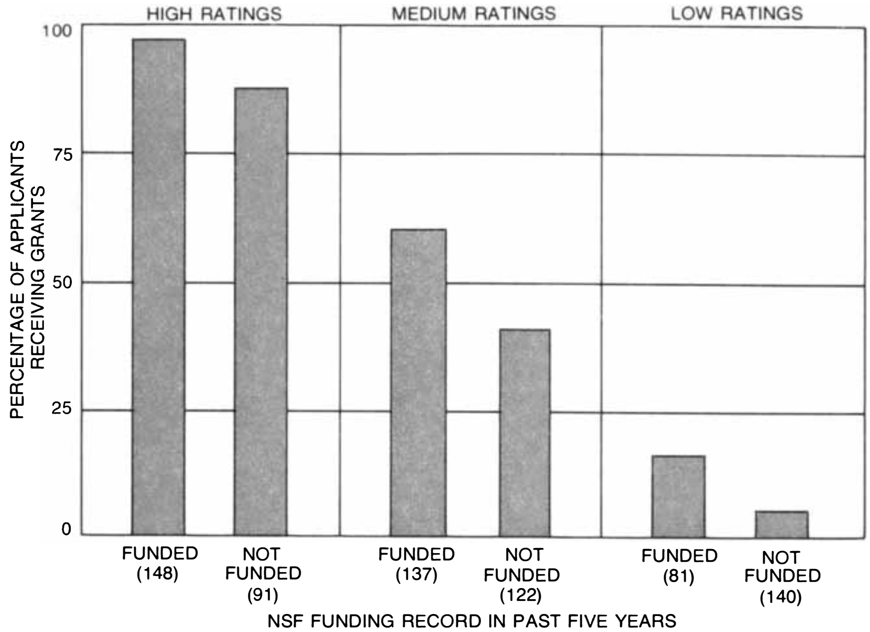
What types of scientists were successful in receiving grants from the NSF in 1975? Of those applicants who obtained their degrees from the highest-ranked graduate departments 62 percent were awarded grants, compared to 38 percent of those who were graduated from the lowest-ranked departments. Similarly, 74 percent of the applicants currently employed in the highest-ranked departments were funded, compared with 38 percent currently in either unranked departments or nonacademic institutions.

Recent NSF funding history and frequency of citations of recent work both had a moderate influence on the probability of receiving a grant. Among applicants receiving the most citations to recently published work roughly three-quarters were awarded NSF grants; among those receiving the least citations of recent work less than a third received grants. The number of papers published and the number of citations of work published before 1965 were less strongly associated with the receipt of a grant. Other attributes of the applicants, such as their professional age or their academic rank, had a minor effect on the probability of receiving a grant.

The effect of professional age on the probability of receiving an NSF grant is



INDEPENDENT EFFECTS of a scientist's past achievements on the probability of receiving an NSF grant are represented in this chart and the one on the next page. The applicants were divided into three groups: those who received comparatively high mean ratings from mail reviewers, those who received medium mean ratings and those who received low mean ratings. Within each category the probability that particular scientists—in this case those with different numbers of citations of their recent work—would receive grants was then calculated. The results show that scientists whose work is frequently cited have a slight competitive advantage in the competition for funds.



SIMILAR ACCUMULATIVE ADVANTAGE is indicated, among those scientists whose proposals received medium or low peer-review ratings, for applicants who had been funded by the NSF in the past five years. Again, a good record appears to produce a slight advantage.

particularly noteworthy. When we began our study many scientists indicated that they believed it was more difficult for younger scientists to obtain NSF funds. Our interviews with program directors, on the other hand, revealed that they perceived just the opposite. Because there is a commitment on the part of the NSF to help young, talented scientists get started, several program directors said that in the case of roughly equal peer reviews they would prefer to fund younger applicants. As it happens, the perceptions of both the applicants and the program directors are mistaken. The data we have gathered indicate that professional age has almost no effect on either the peer-review ratings or the final funding decision.

The overall pattern of our data suggests that scientists with an established track record, many scientific publications, a high frequency of citations, a record of having received grants from the NSF and ties to prestigious academic departments have a higher probability of receiving NSF grants than other applicants do. Nevertheless, the granting process is actually quite open, and there is nothing approximating a scientific caste system. Even among the most frequently cited scientists who apply for support an appreciable number do not receive grants, and among the group with the fewest citations to their work a significant number do receive grants. There is no evidence that scientists who have received grants in the past are guaranteed continued support, or that those without a past funding record have no chance of obtaining current NSF funding. Indeed, given the heavy emphasis the NSF places on past performance as one of the two most important criteria in evaluating research proposals, it is somewhat surprising that measures of past scientific performance do not show a stronger influence on the probability of receiving a grant.

It should incidentally be noted that the data presented here allow us to answer two distinct questions. The first is: How well do the social characteristics of scientists and their previous record predict peer-review ratings and the probability of funding in general, that is, when we examine the entire sample of applicants? The second is: Are there substantially different probabilities of receiving high ratings or a favorable decision for the most eminent applicants compared with the least eminent applicants, that is, when we compare relatively small subsets of the sample? The answers can be different depending on which of these two questions we ask.

For the sample as a whole status differences are not good predictors of ratings. Consider a concrete example of what we mean by focusing again on the relation between the rank of an applicant's current department and the final funding decision. First recall that 55 percent of all 1,200 applicants received NSF grants; if one had to predict whether an individual applicant had received a grant, to predict in every case that he had received one would make one right on 55 percent of the applicants and wrong on 45 percent. The question is: How does knowledge of the rank of an applicant's department increase the ability to predict whether he received a grant? To estimate this we examine each of the five classifications of departmental rank. In the two bottom categories, where a majority did not obtain support, we would guess that all applicants did not receive grants; in the other three categories, where a majority received support, we would do better to guess that all received support. That would result in correct predictions in 63 percent of the cases. When we subtract from this total the proportion (55 percent) that we would have guessed correctly without any information about the individual's departmental affiliation, we get an estimate of the increase in predictability that results from knowledge of rank of department: in this case an increase of 8 percent, which is not an extraordinary increase in predictability.

Suppose, on the other hand, we want to know whether scientists in the highest-ranked departments have a better chance of receiving NSF support than those in unranked departments or in a nonacademic setting. If we compare the percentage difference between these extreme subgroups, we find a substantial 36-point difference. In other words, some percentage differences do appear large in the extremes, but that does not mean the characteristic is a good predictor of a decision for the entire sample. Of the variance that can be accounted for in funding decisions, the peer-review rating is by far the best predictor.

The well-documented social process referred to by sociologists of science as "accumulative advantage" would lead one to expect that eminent scientists have a better-than-average chance in the competition for NSF funds. Accumulative advantage in this context means that a scientist who has been rewarded at one stage in his career has an enhanced probability of being rewarded at a later stage, regardless of the quality of his scientific work in the interim. The concept explains in part the increasing inequality in rewards that is observed as an age cohort of scientists moves through time.

According to the concept of accumulative advantage, the initial social status of a scientist influences the probability of his obtaining a variety of forms of recognition, including the esteem of his colleagues, an association with centers of excellence in the academic world and the resources and facilities necessary for productive scientific work. For example, young scientists who are trained in the best university science departments, and particularly those who have been apprenticed to leading scientists, have a better chance than less well-placed students of equal ability to secure first jobs at prestigious institutions. Once established in these positions they have a better chance than their peers to obtain support for their research. With greater support they have an enhanced opportunity for making significant scientific discoveries and publishing the results. And once the results are published they have still greater chances for future success. To the extent that this process works to the advantage of scientists who are initially well placed in the social system of science it also works to the disadvantage of their peers who are not so fortunate.

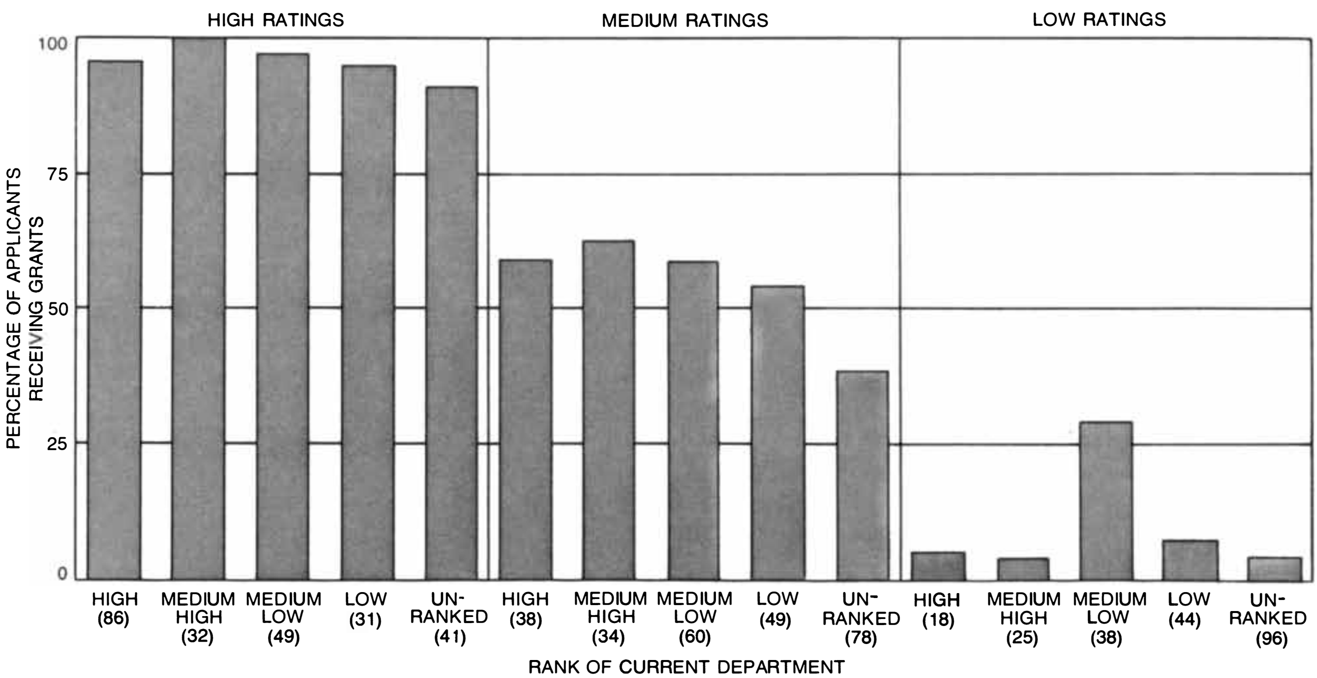
By taking the mean peer-review rating received by an NSF research proposal as a rough measure of the quality of the proposal we attempted to determine the independent effect of a scientist's past achievements on his receiving a grant. We first divided the applicants into three groups: those who received comparatively high mean ratings, those who received medium mean ratings and those who received low mean ratings. Within each category we calculated the probability that scientists who had had different numbers of citations of their recent work would receive grants. We then considered only the group of proposals that received the highest peer-review ratings. Of this group 100 percent of the quintile with the highest number of citations were awarded NSF grants. In the lowest quintile 77 percent received grants. This finding leads to two conclusions: (1) the mean peer-review rating is a far more important determinant of whether a scientist receives a grant than is the number of citations of his recent work, and (2) within each category of mean ratings the number of citations of recent work has only a slight influence on the probability of approval.

We next considered the cases of those scientists whose proposals received low ratings. A substantial majority of all the proposals in this category were declined, but the number of citations made little difference. Within the group of proposals that received low ratings 16 percent of the scientists with the most citations received grants, compared with 3 percent of those who received the fewest citations.

The foregoing data offer some limited support for the concept of accumulative advantage. Scientists whose recent work has been frequently cited have a measurable advantage in the competition for current funds; this advantage is, however, very slight. The process of accumulative advantage is somewhat more evident among those scientists whose research proposals received medium peer-review ratings but who had been funded frequently by the NSF in the past five years. Among scientists whose proposals received medium ratings, for example, 61 percent of those who had been funded within the past five years were awarded a current grant, whereas only 41 percent of those who had not received funds from the NSF in the past five years were awarded a current grant. Clearly a good funding record gives rise to a slight competitive advantage.

We also examined the independent effect of an applicant's current academic department on the probability of his being awarded an NSF grant. Here the story was somewhat different. The rank of a scientist's current department apparently has almost no effect on the probability of his receiving a grant independent of the peer-review ratings received by the applicant's proposal. Of the scientists in the highest-ranked departments whose proposals received comparatively low ratings 6 percent were awarded grants, a figure no different from that found among applicants in lower-ranked departments. In the competition for current funds, therefore, a scientist's past performance as measured by citations of his work and his recent NSF funding record does lead to a very slight accumulative advantage, but his academic affiliation does not appear to give him any advantage.

The results of our study of the operation of the peer-review system in the basic-research programs of the NSF are consistent thus far with other recent findings in the sociology of science, which suggest that the scientific enterprise is an exceedingly equitable, although highly stratified, social institution in which the individuals who produce the work that is most favorably evaluated by their colleagues receive the lion's share of the rewards. Further study of the equity of research-fund distribution will address two basic problems not yet considered. In the first phase of our study we relied on the peer-review ratings elicited by the NSF program directors as an indicator of quality and found those ratings were strongly related to the actual funding decision; now we are submitting proposals to independent review panels in order to obtain independent appraisals of their quality. Finally, having learned that peer-review ratings are strong predictors of funding decisions, we are interested in whether or not they also are good predictors of future scientific performance, and so we are studying how the ratings and recent research performance compare as predictors of future research performance.



NO INDEPENDENT EFFECT was detectable in this similar statistical analysis, which measured the influence of an applicant's current academic department on the probability of being awarded an NSF basic-research grant. Apparently current academic affiliation does not give an applicant any competitive advantage independent of the peer-review ratings that were received by his research proposal.

X-Ray Stars in Globular Clusters

The dense central cores of some such clusters favor the formation of X-ray-emitting double stars in which a neutron star or black hole accretes matter from a star still consuming its nuclear fuel

by George W. Clark

Over the past 15 years X-ray telescopes carried aloft by balloons, by rockets and most recently by satellite observatories have located some 100 highly luminous and variable X-ray stars among the 100 billion stars that form the disk of our galaxy. A typical X-ray star radiates thousands of times as much power in the X-ray region of the spectrum as the sun radiates at all wavelengths. With a frequency of occurrence of roughly one per billion stars, such objects are among the rarest and most fascinating of all celestial phenomena. As a result of the relatively precise observations provided by X-ray telescopes aboard satellites it has now been established that at least six variable but nonetheless persistent X-ray stars are located within globular star clusters in the galaxy.

Globular clusters are self-gravitating systems consisting on the average of about 100,000 stars. Some 130 of them are known at present and perhaps 70 more may be hidden from us by intervening dust clouds. They populate the galactic "halo": the spherical region that is bisected by the disk of the galaxy. All told the globular clusters consist of some 10 million stars. Thus the presence in them of six X-ray stars represents a frequency of occurrence of roughly one in two million, a ratio 500 times larger than that of the X-ray stars in the disk of the galaxy. The higher frequency must indicate either a much higher rate of formation of X-ray stars in globular clusters or a greater longevity of such objects, or both.

More recently the X-ray telescopes aboard satellites have discovered more than 30 X-ray "bursters": a special kind of X-ray star that emits brief but brilliant bursts of X rays lasting up to several seconds. In some cases the bursts repeat at fairly regular intervals measured in hours or days. One unique rapid burster emits its bursts in rapid-fire sequence at rates as high as several thousand per day. In a typical 10-second burst the X-ray flux can exceed a week of the sun's total energy flux at all wavelengths. Persistent but highly variable X-ray emissions have been detected from several of the slow bursters, thus making it quite evident that the bursts represent an extreme form of variation characteristic of a particular kind of X-ray star.

Most of the bursters are located within the galactic disk and are concentrated toward the galactic center. Three slow bursters have been found, however, to lie in globular clusters that were previously known to contain persistent X-ray stars. The single rapid burster lies in the direction of a previously unknown globular cluster that is nearly hidden by interstellar dust. Hence the rapid burster brings to seven the number of X-ray stars in globular clusters. In the case of at least one of the slow bursters located within globular clusters the bursts are observed only when the more persistent X-ray emission is fairly weak, indicating that the burster and the persistent source are one and the same object.

The observations made so far of the X-ray stars in globular clusters are not yet extensive enough to establish whether or not all of them on occasion emit bursts. On the other hand, several of the X-ray sources in the galactic disk have been observed for many weeks of accumulated time by one or more of the satellite observatories without evidence of bursts. Thus it appears certain that some, and perhaps most, of the persistent X-ray stars not located within clusters are also not bursters. It is therefore evident that the special circumstances that give rise to X-ray stars in globular clusters also favor the production of the kind that produce bursts.

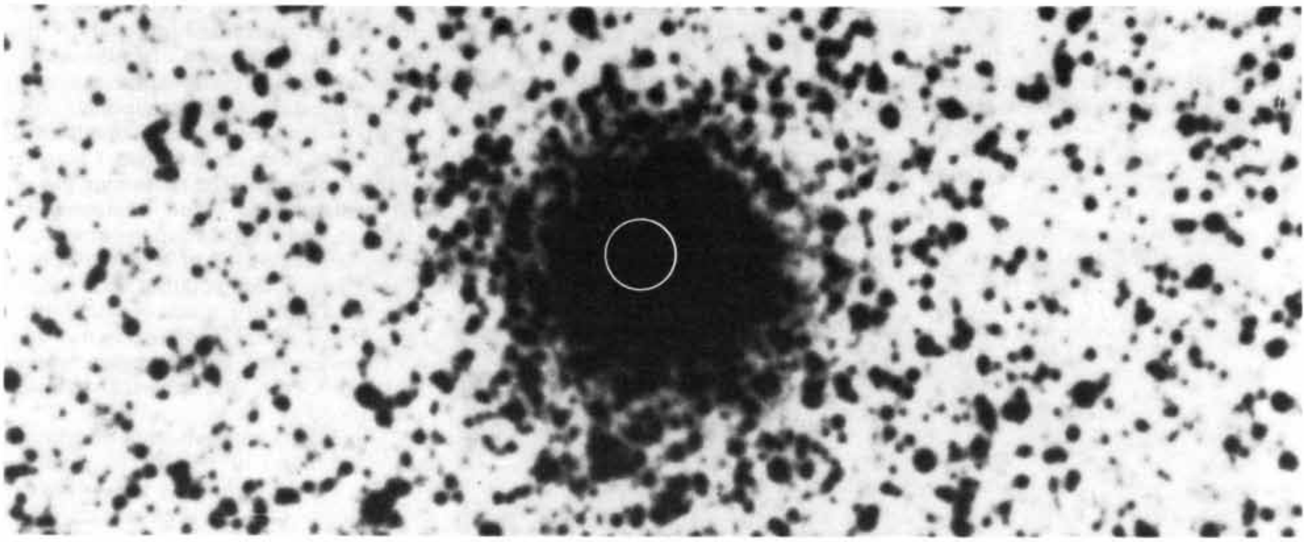
What kind of object can generate X rays at the prodigious power levels of X-ray stars? One can say with considerable certainty that X-ray stars are dense remnants of stars that have exhausted their supply of nuclear energy and have collapsed under the attractive force of their own gravity. Their X-ray emission arises from a process of accretion in which the gravitational potential energy of matter falling toward the collapsed star is converted into heat at temperatures of many millions of degrees and radiated as X-ray photons, which have energies lying mostly between 1,000 and 50,000 elecltron volts (between one and 50 KeV). Hence in order to make an X-ray star nature must combine a dense stellar remnant and a supply of matter that feeds the accretion process at an appropriate rate.

Both constituents are provided by a binary system consisting of a star still burning its nuclear fuel in close proximity to a burnt-out collapsed star. Under certain conditions matter will flow from the nuclear-fuel-burning star over to the compact star to sustain the accretion process. X-ray bursts may be caused by periodic interruptions in the flow of matter, which could pile up and then suddenly be dumped onto the surface of the collapsed partner. The dense members of these partnerships are either neutron stars, objects about as massive as the sun but only about 10 kilometers in radius, or black holes, objects that are entirely contained within a critical radius at which the gravitational field is so intense that not even light can escape from it. The critical radius for a body of one solar mass is about three kilometers, and it is proportionately larger for more massive objects.

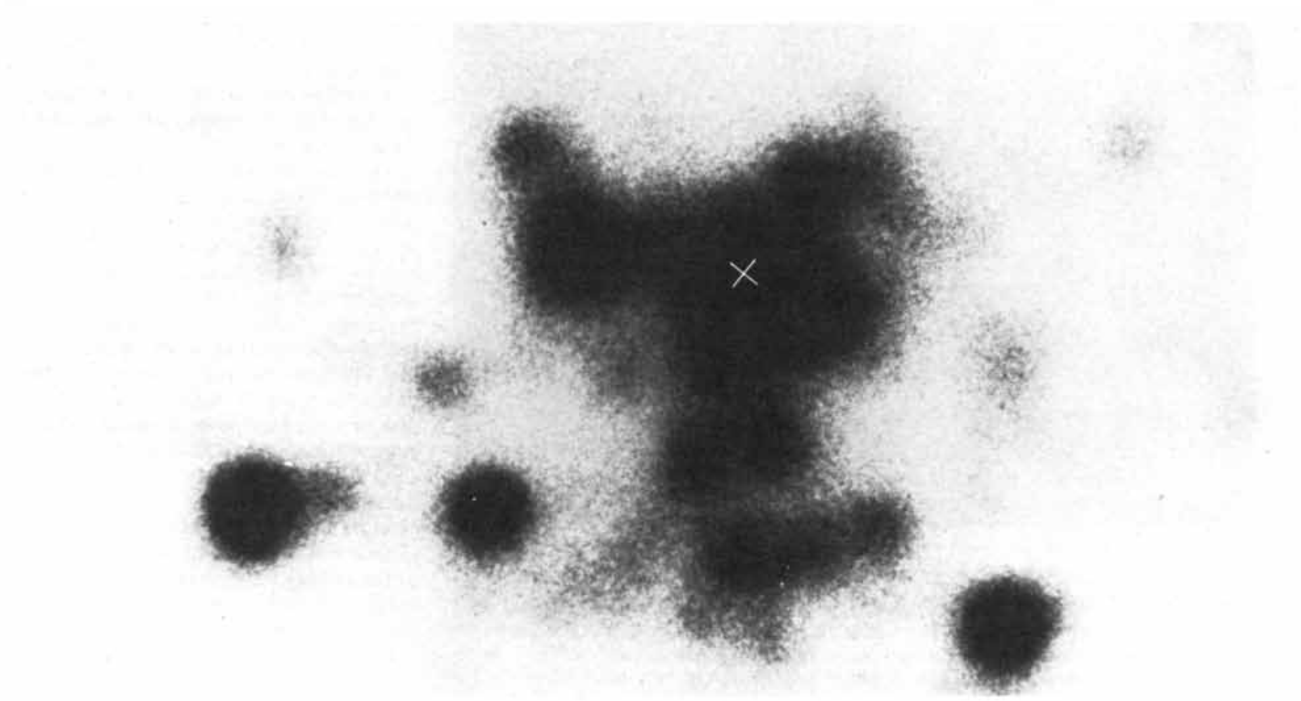
Although the existence of black holes remains a matter of controversy, there seems to be no other way for a body of sufficient mass and sufficiently low angular momentum to escape an ultimate collapse into a black hole. As we shall see, the evidence is clear that some X-ray stars are rotating neutron stars with powerful magnetic fields that control the flow of accreting material, thereby causing extremely regular and revealing pulsations in their X-ray emission. Some X-ray stars show no regular pulsations whatever and may be neutron stars with weak or negligible magnetic fields or may be black holes. Such is the case with the X-ray stars in globular clusters. Thus the study of such objects yields information both about the nature of nonpulsating X-ray sources and about the evolution of the clusters where they are most abundantly created.

Globular star clusters are the oldest and among the most beautiful objects in the galaxy. They have also played a crucial role in the effort to define the size and shape of the galaxy as a whole. They were first recognized to be gravitationally bound systems of stars by William Herschel in the 1780's. A century later Solon I. Bailey discovered many stars in the outer regions of several nearby globular clusters that vary periodically in brightness. Variables with regular periods on the order of a day constitute the well-known class called RR Lyrae variables. Since they all have nearly the same average luminosity they are useful as standard candles for distance measurements. Using the RR Lyrae variables in globular clusters as an index of distance, Harlow Shapley concluded in 1917 that the clusters are distributed more or less spherically around a point in the Milky Way lying in the direction of the constellation Sagittarius. Shapley proposed that the center of the halo of globular clusters is also the center of the entire galactic system.

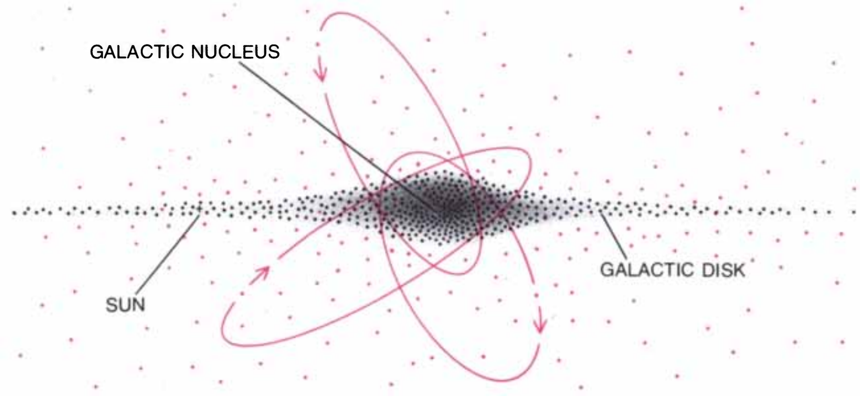
Shapley's work laid the foundation for the present view that the globular clusters were formed 10 to 13 billion years ago during the gravitational collapse of the protogalaxy, a vast cloud of



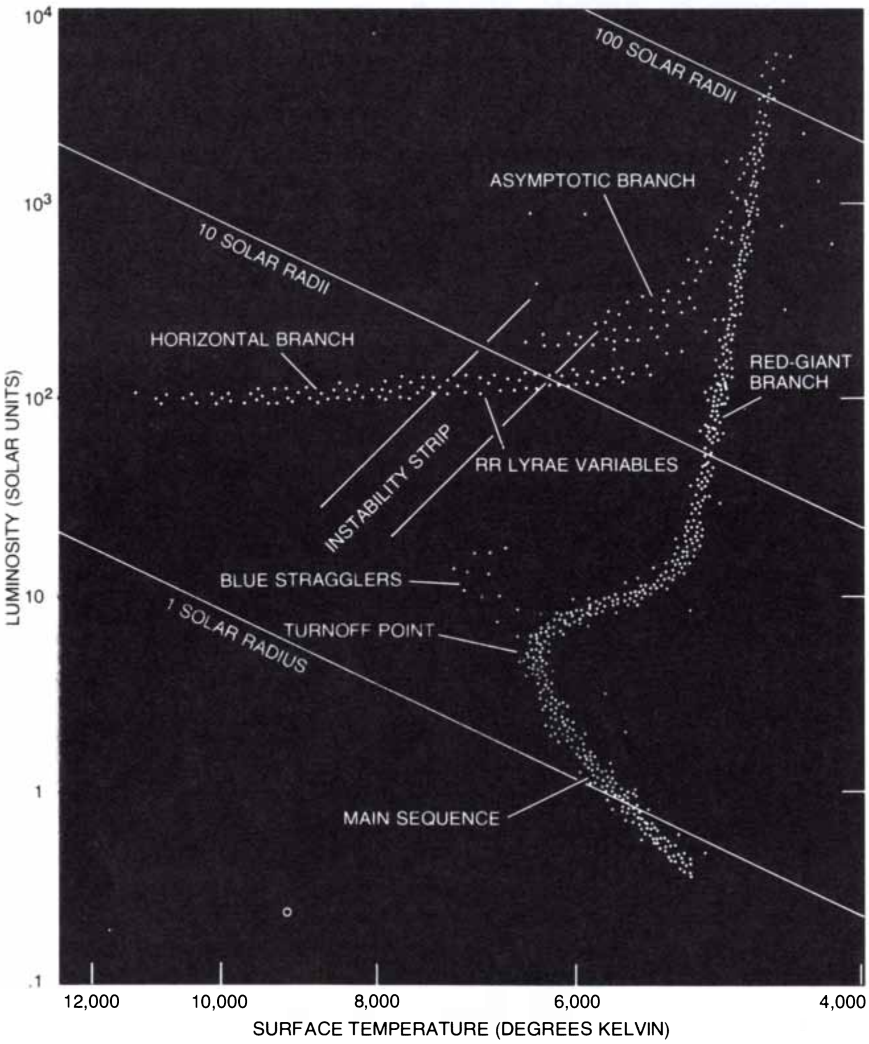
BRIGHTEST OF SEVEN X-RAY STARS discovered in seven different globular clusters lies within the error circle, 20 arc seconds in radius, inscribed on this negative print of a photograph made with the 48-inch Schmidt telescope on Palomar Mountain. The cluster is NGC 6624; the X-ray star within it is 3U 1820–30. Such clusters each consist of some 100,000 stars. The position of 3U 1820–30 was determined by the Third Small Astronomy Satellite (SAS-3). Center of the cluster is packed with stars, as is shown in the illustration below.



CENTRAL CORE of the globular cluster NGC 6624 appears in this high-resolution short-exposure photograph that was made by William Liller of the Harvard College Observatory with the four-meter reflector of the Cerro Tololo Inter-American Observatory in Chile. A dozen red-supergiant stars are crowded into a region that is only 10 arc seconds in diameter. At the cluster's estimated distance of 20,000 light-years that angular diameter corresponds to one light-year. For every red supergiant there are probably some 200 stars that are too faint to register in the photograph. Thus the estimated density of the stars at the core of the cluster is such that their mass per cubic light-year would be 3,000 times the mass of the sun. Such a density of stars would be a million times the density of stars in the vicinity of the sun.



SCHEMATIC VIEW OF OUR GALAXY shows how globular star clusters (*color*) populate a spherical "halo" around the central disk of the galaxy (*seen here from the edge*). A typical globular cluster completes one orbit around the center of the galaxy in about 300 million years. On each passage through the disk the cluster is swept free of accumulated gas and dust. The diameter of the disk is some 100,000 light-years. The sun is 27,000 light-years from center of disk.



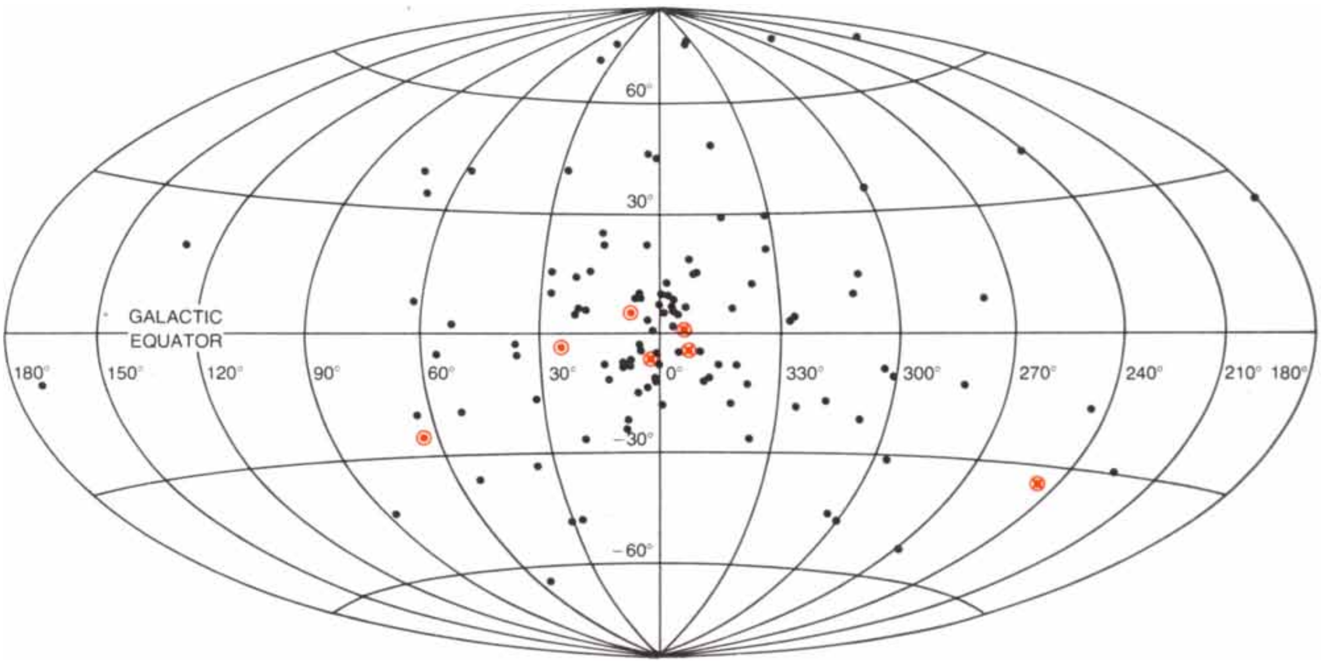
TEMPERATURE-LUMINOSITY DISTRIBUTION of the stars in a typical globular cluster is plotted on a Hertzsprung-Russell (H-R) diagram. A star begins its life on the "main sequence" of the diagram, "burning" by thermonuclear reactions the hydrogen in its core. It remains on the main sequence as long as the supply of hydrogen in its core lasts. When the hydrogen is gone, having been converted by the thermonuclear reactions into helium, the star moves off the main sequence and travels rapidly up the red-giant branch of the diagram, burning the hydrogen that remains in its outer shell. At the tip of the red-giant branch the helium in the core of the star becomes hot enough to burn, generating a violent "helium flash," after which the star shows up on the horizontal branch. Later it will enter the asymptotic branch, burning hydrogen and helium in separate shells. Many stars blow off outer layers and end up as white dwarfs or neutron stars (*off chart at lower left*). A few may become black holes.

gas consisting of hydrogen and helium. As the protogalactic cloud contracted local regions of higher density became self-gravitating and condensed within a relatively short time into globular star clusters, each consisting of millions of stars. Once the clusters were formed they moved through the protogalactic cloud as coherent bodies in orbits determined by the general gravitational field of the galaxy and with orbital periods of several hundred million years. As they picked up speed in their orbital motion they were stripped of all matter that had not already condensed into stars. Thus no new stars appeared after the initial burst of star formation; the range of ages of the stars in a given globular cluster may be no more than a few tens of millions of years, or less than 1 percent of their total age. The remaining gas of the protogalactic cloud continued to contract, and within a time comparable to the orbital period of a cluster it collapsed into a thin rotating disk within which formed the stars of the Milky Way. The clusters continued to move in large orbits that fill the region where they were born, the spherical halo of the galaxy.

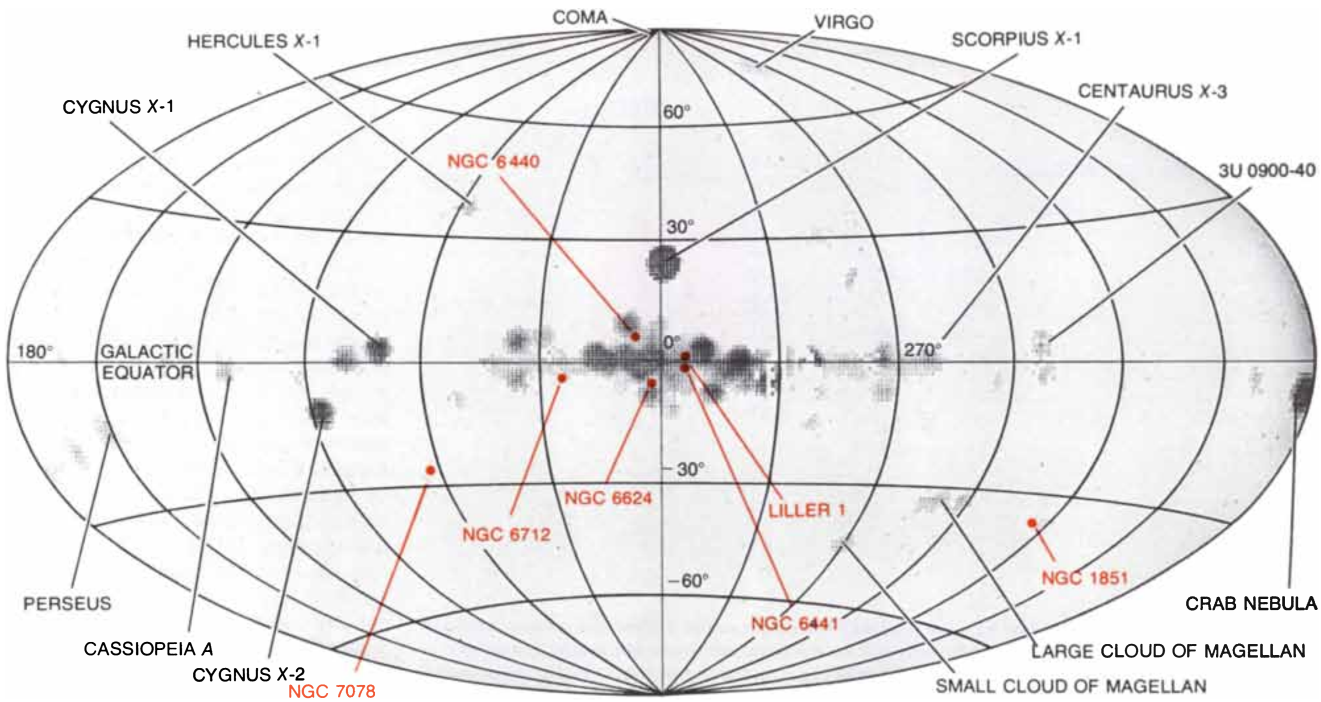
Much of the initial mass of the globular clusters has been lost. Even within the first galactic orbital period every star with a mass of more than several solar masses completed its life cycle and exploded as a supernova that spewed out the heavy elements that had been synthesized by thermonuclear reactions in its core. This material contaminated the primordial hydrogen and helium of the contracting protogalactic cloud with an increasing concentration of the heavier elements, with the result that the clusters formed later and now lying closer to the center of the galaxy are enriched in those elements.

The less massive stars that initially populated the original clusters, stars with a mass down to about the mass of the sun, completed their life cycle without exploding, ultimately evolving into white-dwarf stars. The stars of very low mass and bodies the size of the planet Jupiter and smaller were swept out of the clusters by tidal forces as they passed repeatedly through the galactic disk or were ejected by a process akin to evaporation in which stars acquire escape velocity through random gravitational interactions with other stars in the cluster. There is reason to believe that a substantial fraction of the entire mass of the galaxy is in the form of such low-mass stars, either faint or nonluminous, that were once in globular clusters. What is left in a typical cluster is a total mass equal to the mass of about 100,000 suns, mostly in the form of stars of .4 to .8 solar mass that are still burning their nuclear fuel. The rest of the cluster consists of faint or optically nonluminous remnants of the more massive stars that consumed their nuclear fuel and collapsed. X-ray observations have now begun to provide direct information about some of these dense and optically dark members of globular clusters.

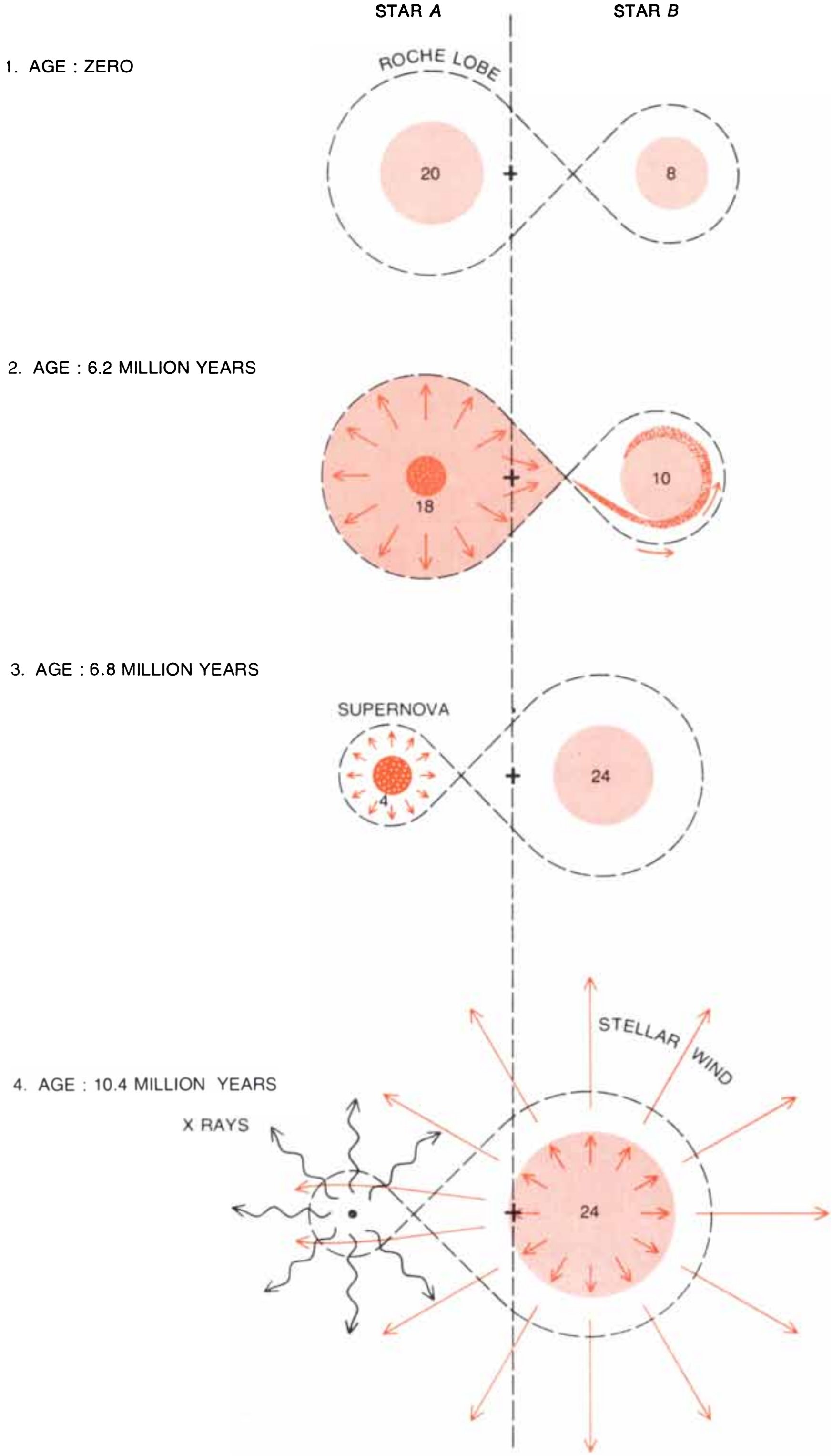
Let us now examine in more detail the processes of star formation and evolution in order to see how X-ray stars can arise both within the galactic disk and within globular clusters. Star formation begins with the contraction of a protostellar cloud of gas and dust. In the galactic disk about half of the stars are born single. The rest are born as double or multiple systems. In most cases the members of such systems are far enough apart for them to evolve as single stars, unaffected by their companions. Some binary stars are so close together, however, that the evolution of both is drastically altered by the transfer of matter,



DISTRIBUTION OF GLOBULAR CLUSTERS is shown in an all-sky map plotted in galactic coordinates. Some 130 globular clusters are known. Presumably another 70 or so are hidden from view by the dust in the galactic disk. The seven clusters containing X-ray stars are circled in color. The four clusters from which bursts of X rays have now been detected are marked additionally with a colored cross.



X-RAY MAP OF SKY was made by computer with data obtained with the Massachusetts Institute of Technology's X-ray detector aboard the Seventh Solar Orbiting Observatory (OSO-7). The detector has a three-degree field of view, so that the images of pointlike sources appear as blurred circles with three-degree radii. Many of the X-ray sources are binary systems in the galaxy proper, in which one member is either a neutron star or possibly a black hole, for example Hercules X-1, Scorpius X-1, Cygnus X-1, Cygnus X-2, Centaurus X-3 and 3U 0900–40. The Crab Nebula and Cassiopeia A are nebulous supernova remnants that emit X rays; they are also within our galaxy. X-ray sources have also been detected in two neighboring galaxies: the Small and Large Clouds of Magellan. In the constellations Virgo, Perseus and Coma are clusters of galaxies 30 million to 200 million light-years away that contain detectable X-ray-emitting galaxies and intergalactic matter. The six sources that are labeled with NGC numbers are in globular clusters in our galaxy, as is Liller 1.



EVOLUTION OF A CLOSE BINARY SYSTEM is believed to follow the scheme depicted here. At the outset star *A* and star *B*, both still on the main sequence, are massive members of a typical binary system. Because star *A* is much more massive (20 solar masses) than star *B* (eight solar masses) it rapidly consumes the hydrogen in its core, starts burning helium, turns off the main sequence and expands into a red giant. After 6.2 million years mass from star *A* begins to overflow its Roche lobe (the limit at which its matter is gravitationally bound) and is accreted by star *B*. After 6.8 million years star *A* explodes as a supernova, leaving behind a neutron star. After 10.4 million years star *B* has become a supergiant and begins emitting mass in the form of a strong stellar "wind," part of which is accreted by the neutron star. The infalling mass gives up energy in the form of X rays. This scheme is based on computations that were made by Camiel de Loore and Jean-Pierre de Grève of the University of Brussels.

and under certain conditions such a binary system can become a powerful X-ray source. First, however, let us trace the evolution of single stars.

A protostellar cloud contracts at a rate that is regulated initially by the transformation of gravitational energy into heat. Eventually, however, the temperature and pressure at the core of the contracting cloud become high enough to initiate the fusion reactions that convert hydrogen nuclei into helium nuclei. The resulting release of nuclear energy effectively halts the contraction of the star and causes its temperature and density to settle into a nearly steady state characterized by a particular combination of surface temperature and total luminosity determined primarily by the total mass of the star.

A chart plotting surface temperature against luminosity for many stars of different masses that are burning hydrogen in their core shows them clustered along the line known as the main sequence of the Hertzsprung-Russell, or H-R, diagram. The more massive a star, the faster it evolves and the greater are its luminosity and surface temperature.

During the core-burning phase the position of a star in the H-R diagram moves only a small distance upward along the main sequence, so that the correspondence between initial mass and position along the main sequence depends only slightly on age. When all the hydrogen in the core of the star is consumed, the star enters a period of complex adjustments in its temperature and density and begins to burn the hydrogen in a shell around the core. During this phase helium continues to accumulate in the core, the rate of energy production increases, the outer envelope expands, the surface temperature decreases and the star becomes a red giant. When such a star is plotted on the H-R diagram, it turns off the main sequence and moves rapidly up and to the right along the red-giant branch.

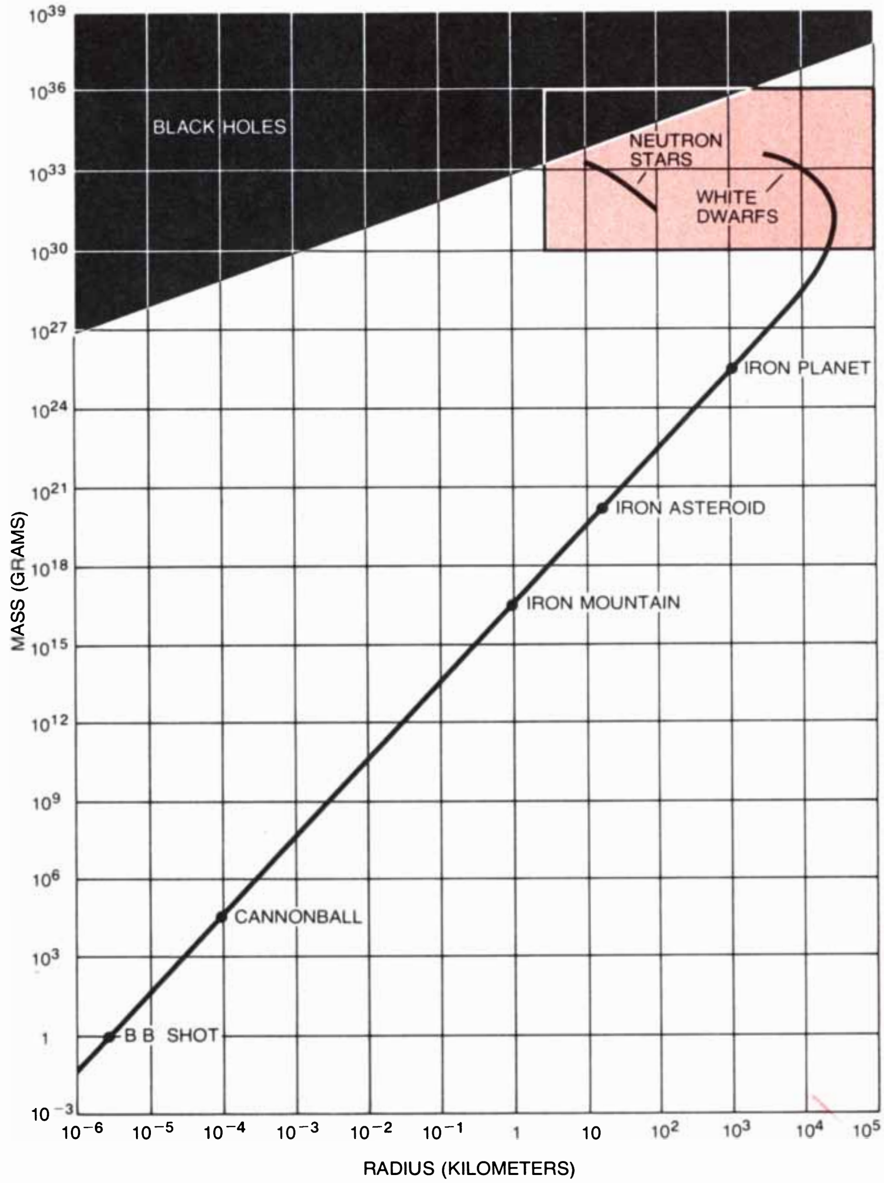
Having spent perhaps 90 percent of its total nuclear-fuel-burning life sitting quietly on the main sequence, the star now undergoes drastic changes in the processes by which it generates nuclear energy, climaxed by violent disruptions of its structure. Finally all that is left is an expanding nebula of ejected gas and a dense remnant of fusion products that has exhausted its available nuclear energy. Without an internal source of energy the remnant loses the pressure of thermal agitation and radiation and shrinks to a state of extremely high density.

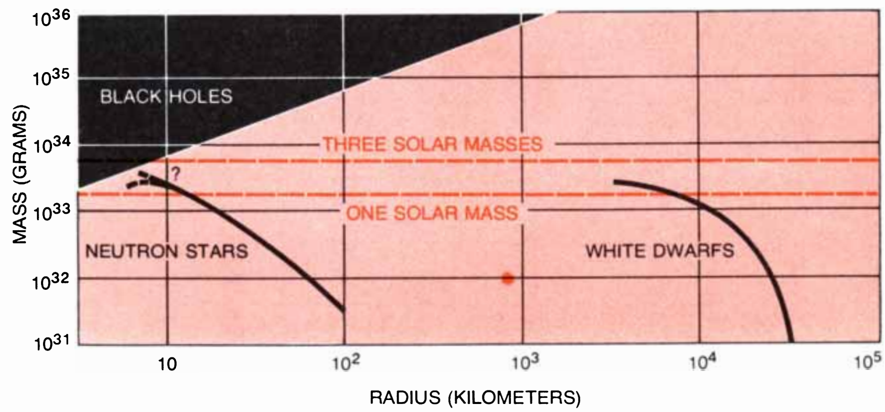
The first example of a dense remnant star was discovered in 1862 by Alvan Clark, the leading telescope maker of his time. During a factory test of a new 18½-inch lens, the largest yet made, he caught a glimpse of a faint star lying close to Sirius, the brightest star in the sky, while Sirius itself was hidden by the edge of a distant building. The faint object, only a ten-thousandth as bright as Sirius, turned out to be a binary companion of Sirius whose existence had been inferred by Friedrich Bessel in 1844 from measurements of perturbations in the bright star's motion. Analysis of the orbits showed that the companion, Sirius *B*, is almost as massive as the sun, even though it is only .2 percent as bright. When Walter S. Adams showed in 1914 that the spectrum of Sirius *B* is that of a white-hot star, it became clear that the star is faint because it is small, roughly the size of the earth. Thus the density of Sirius *B* is more than 100 kilograms per cubic centimeter.

R. H. Fowler explained the structure of Sirius *B* and other white dwarfs in 1926 on the basis of the newly invented quantum mechanics. His theory invoked the Pauli exclusion principle, which limits the density at which electrons with a momentum less than some specified value can be packed into a given volume of space. At this limiting density the electrons are said to be in a state of degeneracy. Fowler suggested that white dwarfs are compressed by gravity to the limiting density for electrons and that such stars are in effect supported against further contraction by the pressure of their degenerate electrons. The detailed theory, published by Subrahmanyan Chandrasekhar in 1931, showed that this source of support is stable up to a critical mass of 1.4 solar masses for a degenerate dwarf composed of helium or carbon. Because of its small surface area such a star can continue to shine as a faint but visible white dwarf for billions of years on the strength of its stored heat.

If a degenerate core should exceed the Chandrasekhar limit, it will collapse. What happens next was suggested in 1933 by Walter Baade and Fritz Zwicky, who were searching for an explanation of supernovas, which for a brief period shine as brightly as entire galaxies. They proposed that such an explosion is caused by the collapse of an overmassive stellar core in the last stage of its evolution and that the collapse causes the electrons and hydrogen nuclei (protons) to combine into neutrons, thus creating a neutron star. Since neutrons also obey the Pauli principle, they too can assume a degenerate state, and hence they can prevent the star from collapsing further. A neutron star of one solar mass becomes stable at a radius of about 10 kilometers and a density of some 10^{14} grams per cubic centimeter.

The gravitational potential energy released by the contraction of matter during the formation of a neutron star is converted partly into rotational kinetic energy as the collapsing star spins faster and faster to conserve its angular mo-





MASS-RADIUS RELATION for a ball of cold iron (*upper diagram*) follows a straight line up to a mass of about 10^{30} grams. The region between 10^{31} and 3×10^{33} grams, shown enlarged in the lower diagram, is occupied by white-dwarf stars, which are supported against collapse primarily by the pressure of electrons in the "degenerate" state, a quantum-mechanical concept that explains why electrons densely packed together with a certain limited amount of energy resist further compression. The addition of mass beyond 3×10^{33} grams (1.4 solar masses) causes the electrons to combine with protons to form neutrons, giving rise to a neutron star. It too is supported against collapse by degeneracy of its constituent particles. Theory indicates that no stable state exists for objects beyond three solar masses and that addition of mass beyond a critical value will cause a neutron star to collapse into a black hole. Lower diagram is based on a discussion of dense states of matter by Steven Weinberg of Harvard University.

mentum. A comparable amount of energy escapes in a blast of neutrinos and in the catastrophic expansion of the star's outer envelope, which forms the typical expanding cloud of a supernova remnant. The energy stored in the rotation of the neutron star escapes gradually in streams of high-energy particles.

In some way not yet understood the streaming particles generate beams of electromagnetic radiation, which as the neutron star rotates are observed as the characteristic periodic signal of a pulsar. Some 150 pulsars have been discovered in the past 10 years. The fastest one rotates 30 times per second and is the stellar remnant of the supernova that gave rise to the Crab Nebula. Any object that is held together exclusively by its own gravity will fly apart at such a rotation rate unless its density exceeds 10^{12} grams per cubic centimeter. (Terrestrial objects such as tops and fly wheels that rotate more rapidly are held together by much stronger electromagnetic forces.) Since 10^{12} grams per cubic centimeter is more than a million times the density of a typical white dwarf, there is no doubt that the Crab pulsar is a neutron star. Considering the rate at which such events occur in our galaxy one must conclude that several hundreds of millions of neutron stars have formed and are now drifting about in the galactic disk, invisible except for the few that are close enough and young enough to be detected as pulsars.

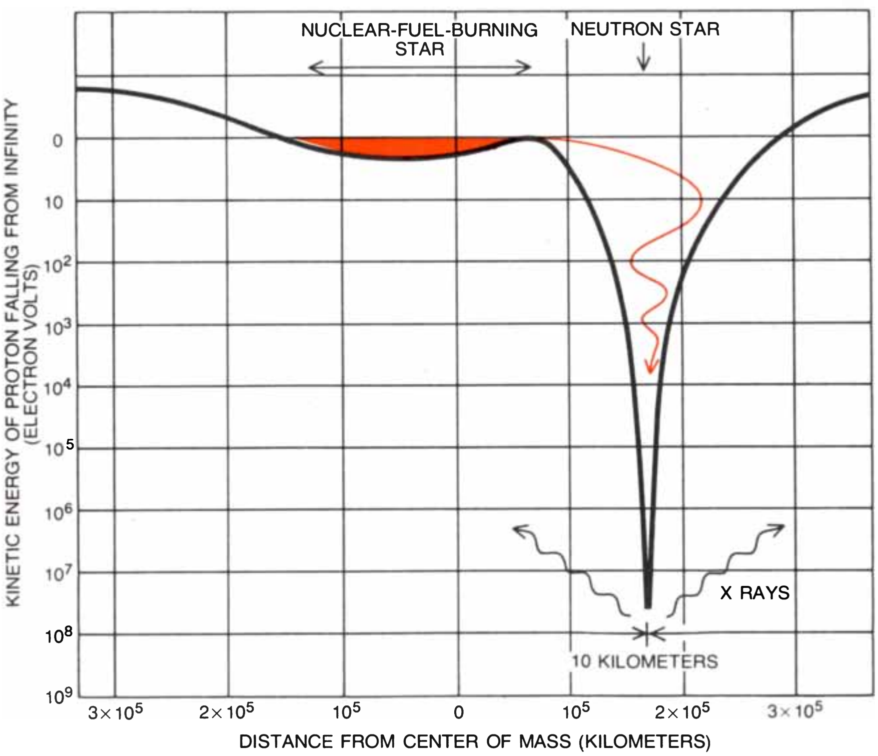
The behavior of matter in aggregations from which all available nuclear energy has been extracted can be summarized by plotting the radius of a spherical body against its mass [*see illustration on preceding page*]. For the sake of simplicity one can consider aggregations of iron, which lies at the bottom of the curve of nuclear binding energy, indicating that no more nuclear energy can be extracted at any temperature or pressure. From iron "BB" pellets to iron planets there is little deviation from the curve for constant density. When the mass exceeds .01 solar mass, however, there is substantial compression, and the radius actually decreases from about 20,000 kilometers to about 10,000. This is the mass range of the white dwarfs. At the Chandrasekhar limit, which for iron is at 1.2 solar masses, the curve is flat and there is no stable radius in its vicinity above or to the left. A degenerate stellar core that grows beyond this mass must collapse to the regime of neutron stars described by a curve ranging downward from a radius of a few hundred kilometers to one of less than 10 kilometers, corresponding to a mass ranging upward from .1 solar mass to a few solar masses.

Above two solar masses there are only unconfirmed extensions of theory. Calculations by J. Robert Oppenheimer and G. M. Volkoff, based on the general theory of relativity, placed an upper limit of .7 solar mass on the mass of a degenerate star composed of noninteracting neutrons. More recent calculations that take into account the interactions of neutrons raise the upper limit to about three solar masses. The most persuasive theory predicts that a degenerate star with a mass in excess of this limit that is entirely depleted in nuclear energy will collapse to a black hole with a radius in kilometers that is approximately equal to three times the ratio of its mass to one solar mass.

This, then, is the general scheme of stellar evolution that can lead to the formation of isolated white dwarfs, neutron stars and possibly black holes. Before considering how such objects are implicated in the formation of X-ray stars, we must consider two important questions. How old are the stars in clusters? What is the range of masses for main-sequence stars that evolve into neutron stars?

Answers to the first question began to emerge in the 1950's, when advances in electronic computers made feasible the numerical modeling of stellar structure. Since one cannot directly observe the interior of a star (except perhaps the sun, whose internal processes are being studied through attempts to measure the flux of solar neutrinos), one must extract from such a model a prediction of external appearances that can be checked against the observable properties of stars: mass, luminosity, surface temperature, surface composition and age. The ideal places to check such predictions are in star clusters, both the ancient globular clusters found primarily in the galactic halo and the younger, smaller and rather amorphous clusters found in the regions of active star formation in the galactic disk.

The important property of a cluster for this purpose is that its stars were all formed at nearly the same time from material of nearly the same composition. Thus the differences between the observed external properties of the stars in the cluster are due largely to differences in their initial masses. Since the rate of evolution increases with mass, the H-R diagram of a cluster is in effect a stop-action display of the evolutionary states of many stars of the same age and composition but different initial masses. The test of a model is whether the H-R diagram of stars computed with various initial masses can be fitted to the H-R diagrams of real clusters by adjusting the age and the initial composition. Calculations carried out in the 1960's by Icko Iben, Jr., and Robert T. Rood showed that the H-R diagram of model stars satisfactorily fits the H-R diagram of stars in typical globular clusters if it is assumed that the stars have an age of 13



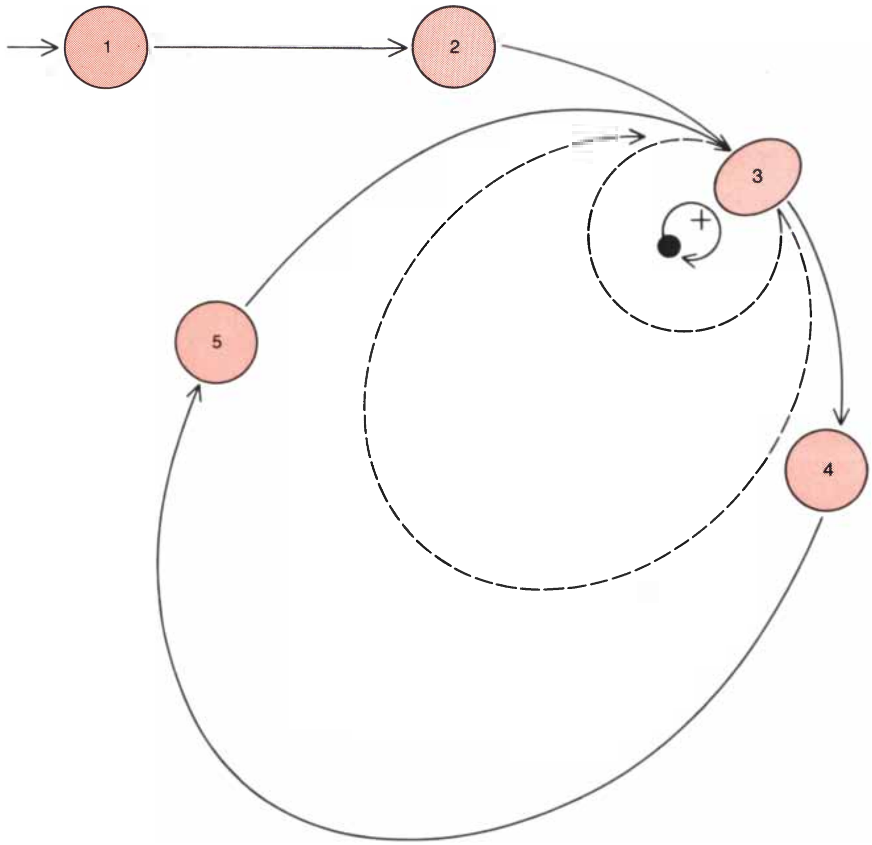
POTENTIAL-ENERGY PLOT OF X-RAY BINARY depicts two potential wells: a broad, shallow one containing the mass of the star still in the nuclear-fuel-burning stage, and a narrow, deep one containing the mass of the companion neutron star. Any matter that falls from the surface of the large star into the potential well of the neutron star acquires 100 million electron volts of kinetic energy per proton, which can be released as heat and radiated as X-ray photons with energies of 1,000 to 100,000 electron volts. The protons spiral inward as they fall.

billion years and an initial composition of hydrogen mixed with 29 percent helium and .1 percent heavier elements. According to Iben and Rood's model, globular cluster stars of .8 solar mass reach the turnoff point on the main sequence at an age of just under 10 billion years.

Theoretical and computational problems still leave a large uncertainty in the range of the initial masses of stars that are the progenitors of neutron stars. A lower limit is obtained from the observation that there are white dwarfs in the Hyades, a nearby galactic cluster whose brightest members form the *V* in the constellation Taurus. In this cluster one finds main-sequence stars up to four solar masses, which implies that white dwarfs are descended from stars with initial masses up to some value greater than four solar masses. Model computations show that the nuclear-fuel-burning life of a star of four solar masses is less than 100 million years.

Let us turn now to massive "close" binary systems, where the transfer of matter influences the evolution of the two companions. Early in their life the companions evolve as isolated stars with minor perturbations, if any. Each star is contained within its own well of gravitational potential. There is a radical change, however, when star *A*, the more massive member of the close binary pair, ends the stage in which it is burning the hydrogen in its core and begins burning the hydrogen in its shell. As we have seen, the star then turns off the main sequence and expands as it ascends the red-giant branch of the H-R diagram. At some point the outer envelope of star *A* overflows the potential barrier that separates it from its partner, star *B*, and mass is transferred to star *B* on a grand scale. The result is that star *A* is stripped of its outer envelope, leaving only its highly evolved core to continue to the later stages of evolution and supernova explosion.

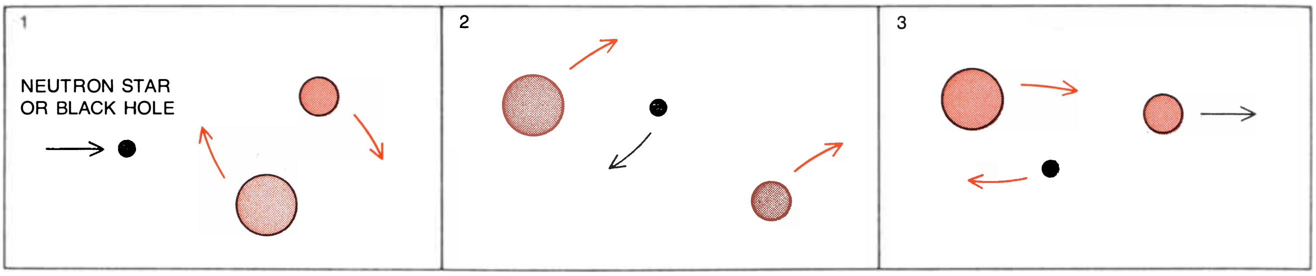
Meanwhile star *B*, having inherited the envelope of star *A*, now has sufficient mass to hold the system together through the eventual supernova explosion. Thereafter the system is a binary consisting of star *B* and the remnant of star *A*, which is now a neutron star in the company of its newly massive companion. Star *B* continues to evolve and soon begins to feed matter back onto the neutron star by emitting high-speed particles (the stellar "wind") or by expanding until its envelope overflows its gravity barrier. On this reverse trip matter is drawn into the gravitational potential well of star *A*, which is now so deep that a freely falling proton would arrive at the surface of the neutron star with a kinetic energy of more than 100 million electron volts (100 MeV). This energy is 14 times the energy released per proton when hydrogen and helium fuse. Indeed, the process is so efficient that an accretion rate of only one solar mass per billion years is sufficient to provide all the power radiated by a typical high-luminosity X-ray star. Moreover, at this low rate of accretion the density of the inflowing material is small enough to allow X rays emitted near the neutron star to escape from the system with little loss by absorption on their way out. Scenarios such as this one have been intensively studied by Edward P. J. van den Heuvel and his co-workers at the University of Amsterdam and by several other groups. Their work provides plausible explanations for a wide variety of X-ray stars that are found outside globular clusters.



ONE MODE OF BINARY-STAR FORMATION in the dense cores of globular clusters visualizes the capture of a typical nuclear-fuel-burning star (*color*) by a neutron star or a black hole. During a close approach of the two stars the tides raised dissipate enough of the mechanical energy of the pair to leave the two objects bound in orbit. Tidal losses at subsequent approaches cause original eccentric orbit to become more and more circular. This kind of capture is proposed by Andrew Fabian, James Pringle and Martin Rees of University of Cambridge.

Actually the flow of accreting matter toward a neutron star is retarded by shock waves, magnetic fields, centrifugal forces and radiation pressure. Instead of falling radially inward the matter forms a rotating "accretion disk" of hot plasma: a gas of charged particles. The material spirals inward through the accretion disk and is finally guided to the surface at the magnetic poles of the neutron star, with each proton losing its energy not to a single 100-MeV gamma ray but ultimately to many X rays with energies mostly in the range below 20 KeV. The magnetic poles are usually not on the axis of rotation, so that the regions of intense X-ray emission rotate off-axis. As a result the X-ray emission seems to pulse periodically when it is observed from a distance. The detailed study of these pulsations has left no doubt that the X-ray sources are rotating magnetized neutron stars.

The pressure of radiation flowing outward from an accreting neutron star controls the rate of accretion. The force of gravity on a uniformly inflowing plasma is balanced by radiation pressure if a star of one solar mass is 50,000 times more luminous than the sun. At a higher luminosity such a star would blow off its outermost shell. The limitation that this effect places on the luminosity of massive stars was pointed out by A. S. Eddington in the 1920's. The relevance of the Eddington limit to the accretion rate of X-ray stars is complicated by the probable channeling of accretion flows and by directional variations in the emission of X rays. Nevertheless, the most luminous X-ray stars tend to crowd the Eddington limit for stars of one solar mass, indicating that they are accreting matter at close to the maximum allowable rate. If the accretion rate exceeds the limit at which the generated radiation can escape or the excess



ALTERNATE CAPTURE MECHANISM proposed by Jack Hills of the University of Cambridge resembles the exchange collisions observed in nuclear reactions. Incoming neutron star or black hole displaces low-mass star in a binary system. Displaced star flies off.

matter can be blown away, the matter must pile up around the neutron star and smother its X-ray emission.

The notion that X-ray stars are neutron stars accreting matter in binary systems was first published by I. S. Shklovskii in 1967 in an attempt to explain the phenomena associated with Sco X-l, an X-ray star in the constellation Scorpius discovered five years earlier in a rocket observation made by Riccardo Giacconi and his associates at American Science and Engineering, Inc. In 1965 the American Science and Engineering group, working in collaboration with the X-ray-astronomy group at the Massachusetts Institute of Technology and optical astronomers at the Hale Observatories and in Tokyo, had identified the optical counterpart of Sco X-l as a faint variable blue star. No clear indication of a binary system, however, was perceived in either the X-ray or optical behavior of Sco X-l until quite recently. Meanwhile the First Small Astronomy Satellite (*Uhuru*), launched in 1970 and carrying an X-ray observatory developed under the direction of Giacconi, discovered four binary X-ray sources. They were unmistakable because one companion regularly eclipsed the other as the two traveled around each other. Moreover, two of the binaries pulsed with periods that showed Doppler variations due to binary orbital motion. More than a dozen pulsating X-ray binaries are now known, with a remarkable variety of pulsation rates, orbital periods, eclipse characteristics and variations in luminosity. This variety appears to be explainable in terms of the available ranges in the initial characteristics of the binary systems: the masses of the two companions and their orbital elements.

One characteristic all the eclipsing X-ray binaries share is youth. None can be much older than the sum of the main-sequence ages of two stars of which one has an initial mass large enough (greater than four solar masses) to give rise to a neutron star and the other an effective initial mass equal to what the secondary star has attained at the end of the first episode of mass transfer. In general this implies an age of less than a few hundred million years, so that such systems could not have survived the 13 billion years that have passed since stars began to form in the globular clusters. Therefore the sequence of events in binary evolution that provides such a plausible explanation for the variety and frequency of X-ray stars in the galactic disk does not hold for the X-ray stars in the globular clusters, except perhaps for an extremely narrow and improbable range of starting conditions.

I became aware of this dilemma in 1975, when my colleague and then graduate student, Thomas H. Markert, discovered that the X-ray sky survey map he was preparing from satellite data showed a previously unreported X-ray source at a position far from the galactic plane and coincident with the globular cluster NGC 1851. The data had been gathered with an X-ray detector that the M.I.T. X-ray-astronomy group had developed for the Seventh Orbiting Solar Observatory (OSO-7), which was launched in September, 1971.

One of our goals was to prepare a comprehensive map of the sky at X-ray wavelengths. A similar survey had already been carried out with *Uhuru* and had been published in 1974 in the form of an extensive list of X-ray sources: the "3U" catalogue. Among the entries were several variable sources identified with globular clusters. Our survey with OSO-7 confirmed three of them and discovered two more.

These five objects clearly constituted a class of X-ray stars that did not fit the evolutionary scheme that explained the other X-ray binaries. Jonathan Katz of the Institute for Advanced Study had previously pointed out that the ratio of X-ray luminosity to mass of all globular clusters taken together was at least two orders of magnitude greater than the same ratio for the galaxy as a whole. Now we could count five X-ray sources among the estimated 10^7 stars in globular clusters compared to roughly 100 X-ray sources among the total of 10^{11} stars in the entire galaxy. Thus the preferential occurrence of X-ray stars in the globular-cluster star population, which I noted at the beginning of this article, became even more impressive. The problem was therefore to find not only some new explanation for the origin of X-ray stars in globular clusters but also one that would account for their much greater relative frequency of occurrence.

The concept of an X-ray star as an ultradense stellar remnant powered by accretion is so compelling that it was reasonable to assume that the same basic mechanism was involved in the globular-cluster X-ray sources. Many neutron stars and possibly black holes must have been left by the more massive and shorter-lived stars that were born when the clusters first formed. Some of the neutron stars may have been hurled out of their cluster by the violence of the supernova that created them, but many undoubtedly remained gravitationally bound, although perhaps initially in orbits of large amplitude.

In general the mechanical energy of the stars in a globular cluster tends to be distributed equally among the stars, as it does among molecules of a gas in a closed container. Moreover, the larger the mass of a star of given energy, the smaller is the mean distance between the star and the center of the cluster. Hence the remnant neutron stars and black holes, being more massive than the average star, must have congregated near the center of the cluster as their energies reached equilibrium. There they have remained for billions of years, dark objects waiting to be turned on as X-ray stars by acquiring a source of material for accretion.

The most likely hypothesis concerning the source of accretion material is in a sense the most obvious one: the material must be supplied by binary companion stars that are still in the nuclear-fuel-burning stage. Since, as we have seen, such companions cannot have survived from the time the stars in the cluster were born, they must have been acquired by capture, It turns out that six of the seven clusters that contain X-ray stars or bursters, or both, are among the 20 clusters with the highest central density, which would appear to favor capture processes capable of creating binaries in their core.

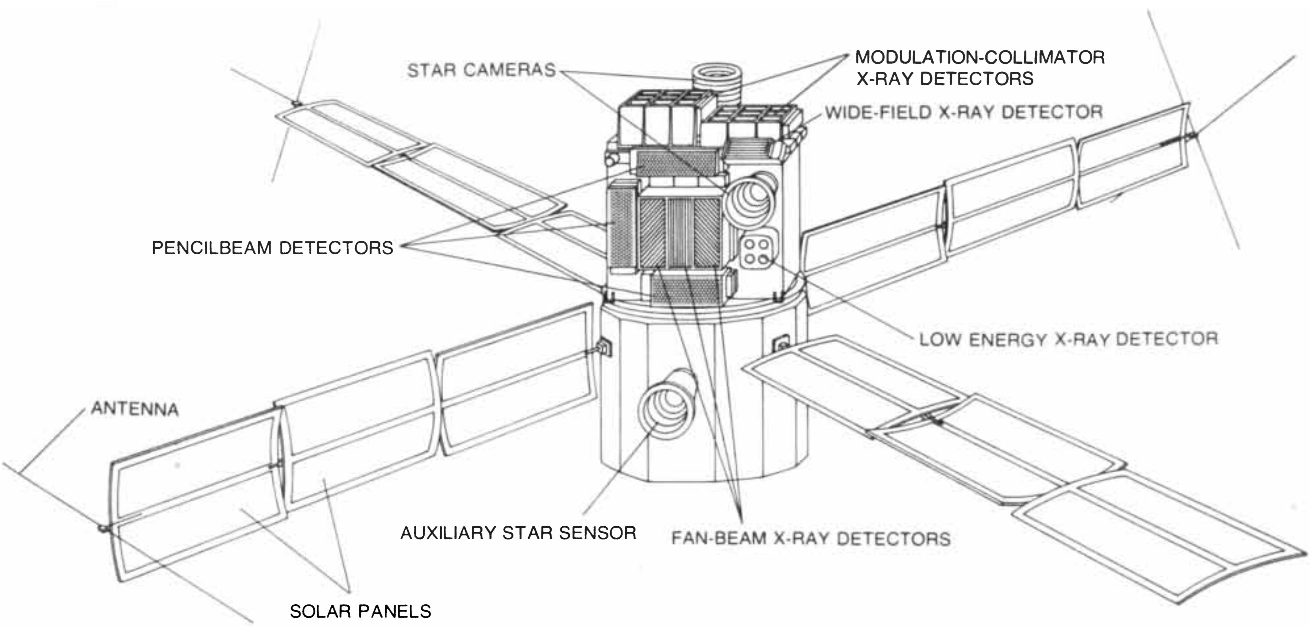
In principle two stars can capture each other through a close encounter among three stars in which one carries away so much kinetic energy that the remaining two are left closely bound. Even in the densest globular-cluster core, however, the frequency of close encounters between two stars is small, and the frequency of encounters among three stars is so small that none could be expected in the 13-billion-year life of a cluster.

Two mechanisms requiring only encounters between two stars were soon proposed for the creation of capture binaries. One, put forward by Andrew Fabian, James Pringle and Martin Rees of the University of Cambridge, rests on the fact that the stars in a globular cluster, considered as particles in a gravitational trap, are relatively "cool" in their kinetic motions. Thus the dissipation of even a small fraction of the total kinetic energy of two stars that experience a close encounter with each other can result in their mutual capture. If a star burning nuclear fuel has a close encounter with a neutron star or a black hole, a tide will be raised on the nuclear-fuel-burning star that dissipates mechanical energy. In a certain range of circumstances enough energy will be dissipated for the two objects to become gravitationally bound to each other, initially in a highly elliptical orbit with a small distance of closest approach. Tidal losses of mechanical energy at succeeding close approaches will gradually grind the orbit down to a nearly perfect circle with a radius small enough to favor the efficient transfer of mass at a later time. The stage is now set for a long wait until the nuclear-fuel-burning star evolves off the main sequence, expands and begins to pour matter into the gravitational well of its compact and formerly dark companion.

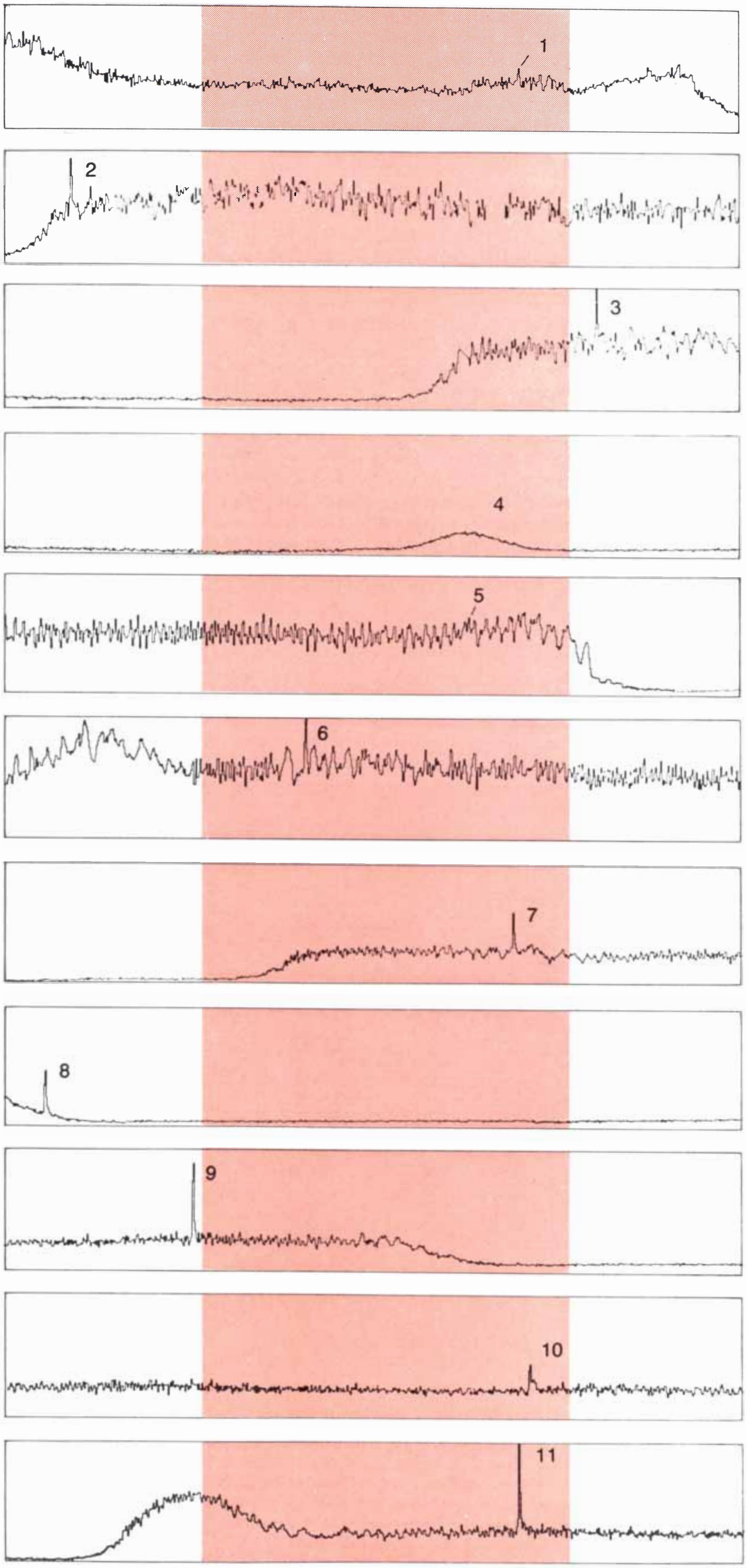
The other capture mechanism, worked out by Jack Hills at the University of Cambridge, is analogous to the exchange collisions known in atomic physics. Hills analyzed by computer simulation many close encounters between, on the one hand, a primordial close binary system consisting of two low-mass stars typical of the main-sequence population in globular clusters and, on the other, a heavy stellar remnant such as a neutron star or a black hole. He showed that the result will frequently be an exchange collision in which the heavy remnant replaces one of the low-mass stars, hurling the replaced star out of the system and leaving the neutron star or black hole in orbit with a supply of matter for accretion.

With plausible assumptions about the number of neutron stars both the tidal-dissipation and exchange-collision hypotheses predict the formation of several close capture binaries in the most condensed globular clusters, where the central stellar density exceeds 1,000 stars per cubic light-year. When one of these binaries is turned on, it could emit X rays much longer than the binaries of the galactic disk because its supply of accretion material is a slowly evolving low-mass star whose expansion phase after leaving the main sequence lasts for hundreds of millions of years. The slow expansion allows matter to be transferred at the modest rate required to power the X-ray emission without smothering it. On this basis the observed long-term variability of globular-cluster X-ray stars is explained as the result of a variability in the accretion flow due to residual orbital eccentricity or to nonuniform expansion of the kind that occurs late in the life of red giants. The apparent absence of pulsations that might be expected to result from the rotation of neutron stars may be explained by the decay of magnetic fields in very old neutron stars. In fact, one may speculate that the occurrence of bursts may be related to the absence of strong magnetic fields in such ancient stars.

An entirely different hypothesis was put forward by Jeremiah P. Ostriker and John N. Bahcall of Princeton University and by Joseph Silk and Jonathan Arons of the University of California at Berkeley. They proposed that the X-ray sources are massive black holes, presumably formed by the coalescence of stars in the dense core of a cluster that has experienced a catastrophic central collapse. It is a remarkable fact that the mathematical and computational difficulties of solving such many-body problems are so formidable that even today the question of whether such a catastrophic coalescence can occur at all has not been definitely answered. Some



THIRD SMALL ASTRONOMY SATELLITE (SAS-3) carries a package of X-ray detectors and star cameras designed and built by the X-ray-astronomy group at the Massachusetts Institute of Technology. Images of star fields provided by the cameras make it possible to determine the orientation of the observatory with an accuracy of plus or minus 30 arc seconds. Through computer analysis of the X rays recorded by the modulation-collimator detectors as the craft rotates around the *Z*-axis X-ray sources can be located with an accuracy of plus or minus 20 arc seconds. The satellite was launched in May, 1975, from a platform on the Equator off the coast of Kenya. Operated by the Aerospace Research Center of Italy, the platform was used in 1970 to launch the pioneer X-ray-astronomy satellite *Uhuru*. SAS-3 is operated by controllers at Goddard Space Flight Center of National Aeronautics and Space Administration under the hour-by-hour direction of the M.I.T. X-ray-astronomy group. Tape-recorded data are beamed from the satellite to the earth and commands sent from the earth to the satellite every 104 minutes when SAS-3 passes over a tracking station operated by NASA in Ecuador.



workers believe that the first moderately massive close binary that forms in the core of a cluster will act like an egg beater and whip up the surrounding stars, halting any coalescence before it can progress. Indeed, the members of a cluster are so lightly bound that a decrease in the orbital separation of one close binary could supply enough mechanical energy to whip up an entire core.

The proponents of the massive-black-hole hypothesis nonetheless postulate the existence of such objects at the center of highly condensed clusters. They show that the gravitational well of such a cluster is deep enough to hold the gas streaming out in stellar winds and the matter being lost in sporadic episodes by normal nuclear-fuel-burning stars. The ejected material falls to the center of the cluster and feeds the black hole through the accretion process, thereby generating X rays. Since the infalling material is in the form of a tenuous interstellar gas, the black hole must be quite massive (in the range of 100 to 1,000 solar masses) to draw it in at the rate required to generate the observed X-ray power.

There matters stood as observers searched for evidence to support one or the other of the two hypotheses: the capture-binary model and the massive-black-hole "vacuum sweeper" model. Periodic pulsations or eclipses in a globular-cluster X-ray source would clinch the model of a capture binary as the X-ray source. None has been observed. A sharp peak in the overall brightness of the cluster near the center would favor the black-hole hypothesis; so would evidence of luminous gas near the center. The data, which are difficult to acquire, are not yet decisive. Consistent with both hypotheses, however, is the evidence that the cluster X-ray sources are found preferentially in clusters with a high central density. The existence of a high central density was clearly shown for one globular cluster, NGC 6624, in

SERIES OF X-RAY BURSTS was recorded by SAS-3 in May, 1975, from 3U 1820–30, the X-ray star in globular cluster NGC 6624. Ten bursts were recorded with a mean interval of 4.4 hours. Presumably there were 11 bursts in all; the fourth in the series apparently occurred while the earth was blocking the satellite's view. Each of these 11 traces contains a 40-minute portion of the X-ray record from 11 orbits of SAS-3 centered on the "expected" recurrence interval of 4.4 hours. (The expected fourth burst is of course missing.) The portion of the record tinted in color extends $10\frac{1}{3}$ minutes to the sides of the mean interval and represents one standard deviation (the "jitter") in the mean frequency of the 10 recorded bursts. Note that burst No. 8 was caught just as source was disappearing from view behind the earth. The orientation of the satellite was changed after burst No. 6 to reduce interference from other X-ray sources.

short-exposure plates made with the one-meter Wise telescope in Israel by Neta A. Bahcall in 1975. At the core of the cluster there is an impressive blob of unresolved star images about four arc seconds in diameter.

A most exciting and unexpected development then occurred. Late in 1975 Jonathan E. Grindlay and Herbert Gursky of the Center for Astrophysics of the Smithsonian Astrophysical Observatory and the Harvard College Observatory, scanning data from the American X-ray experiment on the Astronomical Netherlands Satellite (ANS), discovered evidence of a brief burst of X rays from NGC 6624. This globular cluster contains the variable X-ray star 3U 1820–30, which had been found earlier in the *Uhuru* survey and is one of the brightest variable sources near the center of the galactic disk. The position defined by *Uhuru* gave assurance, however, that 3U 1820–30 is inside the globular cluster NGC 6624 and is not a disk object. Claude R. Canizares of the M.I.T. X-ray-astronomy group, working with OSO-7 data in 1974, had found a remarkable 20-minute outburst in the intensity of 3U 1820–30. With interest now focused on X-ray stars in globular clusters, it was natural that 3U 1820–30 should be a prime target when new satellites capable of being pointed at specified objects for long periods became available. The star was put on the observing schedule of ANS for March and September of 1975, and it was one of the first targets for a refined position measurement by the M.I.T. X-ray observatory on the Third Small Astronomy Satellite (SAS-3), using special techniques developed under the direction of Herbert W. Schnopper and Hale V. Bradt. SAS-3 was launched in May, 1975.

The ANS observations in March of 1975 found 3U 1820–30 in a state of high average luminosity without spectacular variability. The SAS-3 observations in May quickly yielded a refined measurement of the position that narrowed the uncertainty to an error circle with a radius of 40 arc seconds; the circle included the center of the cluster, where we expected the X-ray star to lie. During the May SAS-3 observation the luminosity of 3U 1820–30 was low compared to the peak value found by *Uhuru*.

At that time Garrett Jernigan, then a graduate student who was responsible for developing a major part of the analytical techniques in the SAS-3 project, was scanning the "quick look" data, which are received at M.I.T. and displayed in computer-drawn charts within a few hours of being recorded by the detectors in orbit. He noticed two sudden increases in the X-ray counting rates, each lasting about 10 seconds. The particular instrument being used in the observation was a modulation-collimator detector mounted on top of the satellite with its axis pointing in the direction of the satellite's rotation. The detector is sensitive to X rays with energies from two to 11 KeV. The field of view defined by the collimator consists of alternate bands of high and low sensitivity over a rectangular area roughly 24 degrees on a side.

As the satellite rotates the transmission bands of the collimator pass over the sources in the broad field of view, modulating their counting rates in patterns that can be reduced by computer analysis to correlation maps that show each source as a central peak with concentric rings around it. With such brief bursts, however, the position of the source could not be determined. They could have come from anywhere in the 24-degree-by-24-degree field of view or even from an instrumental "glitch" in the newly launched satellite, in whose flawless functioning we were only beginning to believe. In any case the slides Jernigan prepared in order to display the data indicating the two sudden increases languished while we dealt with a number of more easily recognized discoveries that were coming from the SAS-3 observations.

The ANS observation was repeated in September, and on inspecting the data a few months later Grindlay and Gursky made their surprising discovery of the X-ray burst in NGC 6624. Following a telephone call from Grindlay, J. Heise of the Space Research Laboratory of the Astronomical Institute in Utrecht soon found a second burst. Both came from a narrow sliver of sky that contained 3U 1820–30 and the globular cluster. The detectors were sensitive to X rays from one to 30 KeV. Assuming that the X rays were emitted in all directions by a source in the globular cluster, the total energy in each burst was more than 10^{39} ergs. Noting that the bursts rose to their peak intensity in less than a second and then decayed gradually over the next 10 seconds, Grindlay and Gursky suggested that the X-ray photons of a burst are released in a two-second "spike" and then reverberate by being scattered in a surrounding cloud of hot plasma. The gradual decay was interpreted as the result of the reverberation.

Grindlay and Gursky's model also provided an explanation for the fact that the spectrum of the bursts became "harder," that is, richer in X-ray photons of higher energy, as the bursts decayed. It postulated that the surrounding cloud was so hot that on the average the scattered X-ray photons gained energy. According to Grindlay and Gursky's estimate, the temperature of the cloud had to be so high that it could be held together only by a central body of 1,000 solar masses, which would have to be a massive black hole. Working independently along similar lines Canizares showed that a much cooler cloud could produce both the reverberation and the observed spectral hardening and would require a central body of only one solar mass to hold it together. In any case the discovery of this peculiar and totally unexpected phenomenon in a globular cluster greatly intensified the enthusiasm of many for the massive-black-hole concept.

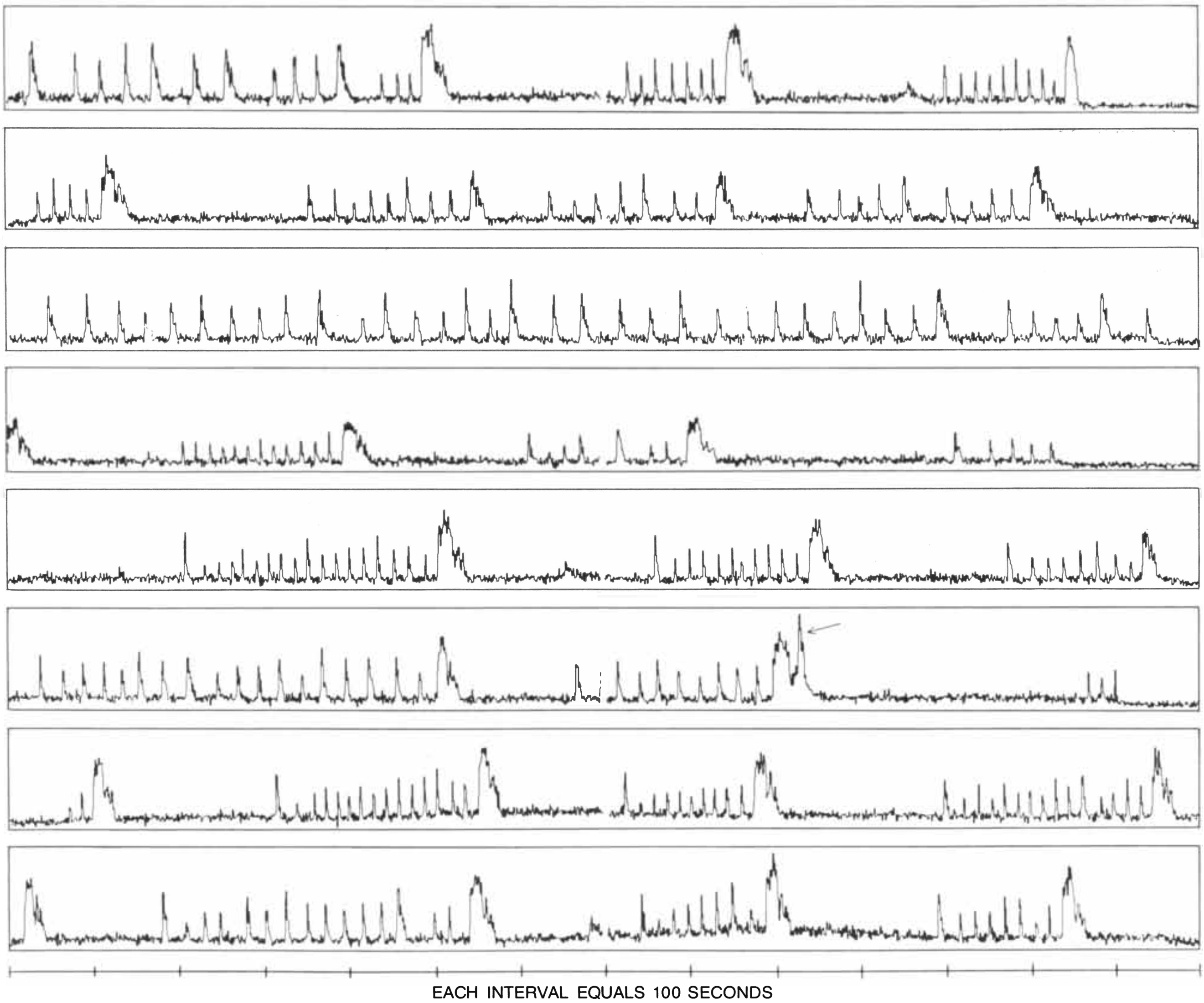
The announcement of the burst discovery caused us to reexamine our SAS-3 data. We soon discovered that the two sudden increases noted in the quick-look sample of data by Jernigan in May were actually two in a series of 10 bursts recorded in the complete "production" data that had been received from the National Aeronautics and Space Administration several months after the observation. We discovered to our astonishment that the bursts recurred almost periodically at a rate of one every 4.4 hours, with one burst in the series apparently missing when NGC 6624 was eclipsed by the earth as the satellite moved in its orbit. Particularly provocative and mystifying was the fact that the intervals between successive bursts were not exactly equal but had a spread of about 5 percent. The burst phenomenon has four remarkably different characteristic times: a rise time of about a second, a decay time of about 10 seconds, a recovery time of roughly 10^4 seconds and a "jitter," or variation, in the recovery time of plus or minus 600 seconds.

The jitter appeared to be of special significance. We wondered whether it is caused by the equivalent of a loose screw in an otherwise accurate clock mechanism or is the result of variability in the trigger level or power supply of a relaxation oscillator. A mechanism of the first kind was immediately proposed by John Bahcall and Ostriker, who attributed the bursts to the effects of a neutron star in orbit around a massive black hole with an accretion disk. They suggested that the neutron star periodically crashes through the accretion disk, generating pulses with a highly stable average period but with a jitter caused by a random flapping of the disk.

A mechanism of the second kind was also immediately forthcoming from Fred Lamb of the University of Illinois, who suggested that plasma flowing toward the surface of a magnetized neutron star could be trapped in a magnetospheric reservoir, where it would pile up until a certain critical density is exceeded. Then the reservoir would spring a leak, and a blob of plasma would suddenly fall to the surface of the star and release its gravitational energy as an X-ray burst. The leak would then "heal," and the reservoir would fill up again. On this model the recurrence rate is governed by a variety of factors such as the rate at which the reservoir refills.

All this was happening during a meeting of the High Energy Astrophysics Division of the American Astronomical Society at M.I.T. in January, 1976. Grindlay and Jernigan each reported their recent results on the bursts from NGC 6624. Richard Belian, Doyle Evans and Jerry Conner of the Los Alamos Scientific Laboratory described their observations of X-ray bursts from a source in the constellation Norma by the Vela satellites, with which mystifying bursts of somewhat higher energy had been discovered some years earlier [see "Cosmic Gamma-Ray Bursts," by Ian B. Strong and Ray W. Klebesadel; SCIENTIFIC AMERICAN, October, 1976]. Unknown to any of the American workers at the time was the fact that high-energy X-ray detectors on the Russian Kosmos 428 satellite had recorded two brief increases in counting rate during an observation that had lasted for a few hours in 1971. In a paper published in the January, 1975, issue of *Soviet Astronomy Letters* the increases were attributed to "hard X-ray bursts" from two different cosmic sources. A translation of the paper appeared in our reading room in March, 1976. A detailed analysis of the properties of these bursts and of the circumstances of their observation indicates, however, that they originated in the earth's magnetosphere.

During the American Astronomical Society meetings the SAS-3 group was conducting another scheduled observation of the X-ray source 3U 1820–30 in NGC 6624 from the operations room in a nearby building, this time using a different set of detectors with pencil-beam and fan-beam collimators that define the field of view. Keeping the center of these fields of view close to the target by means of the satellite's gyroscope control system, which we adjusted according to our analysis of the star-camera data, we recorded a high X-ray intensity and a few bursts. The bursts were puzzling because they did not register in all the detectors, as one would expect if they had all come from a single source in the direction (NGC 6624) toward which the pencil-beam field of view was pointed. Instead they appeared in one detector with a fan-beam collimator whose narrow field of view happened to pass over the center of the galaxy. In the next scheduled observation Walter H. G. Lewin and Jeffrey Hoffman of our group led a team of SAS-3 investigators in a study of a nearby pulsating X-ray star, GX1 + 4. They too observed bursts in a single fan-beam detector from at



RAPID-FIRE SEQUENCE OF BURSTS was recorded from MXB 1730-335, an X-ray source located in a previously unknown globular cluster lying close to the center of the galactic disk. The interval between the bursts is sometimes as short as 10 seconds. The recovery time after each burst is roughly proportional to the size of the burst. The peak intensity, however, is nearly the same from one burst to the next. This behavior suggests that the source resembles a relaxation oscillator with a governor, probably represented by radiation pressure, that limits the amplitude. The arrow in sixth trace indicates a burst from a nearby slow burster that is not located in a globular cluster.

least two sources, of which one could be near the galactic center and the other could not. Soon afterward the SAS-3 detectors were pointed directly at the galactic center, and Lewin and his colleagues discovered three new bursters within .2 degree of the center.

Several weeks later, on the track of the fourth new burster, Lewin pointed the detectors at a position about five degrees from the galactic center, where his analysis of earlier single-detector burst events indicated the presence of another burster. Having by now got used to slow bursters with repetition intervals ranging from hours to days, we were all flabbergasted to see a machine-gun sequence of bursts with intervals as short as 10 seconds. It was soon evident that the more energetic bursts were followed by longer intervals, and Saul A. Rappaport of our group soon demonstrated that the correlation is a roughly linear one. Over the next two months Lewin and his colleagues observed the behavior of this unique object, MXB 1730–335, and David Hearn measured its position with five-arc-minute accuracy. Soon after the intense work on the rapid burster began the other burster, which had originally been sought close by, was located .7 degree away.

In spite of the remarkable variability of its recurrence patterns, the rapid burster quickly demonstrated regularities revealing much about the nature of its mechanism. The approximately linear relation between the bursters and the following interval is like that expected of a relaxation oscillator such as a neon-lamp flasher, which has a constant trigger level but a variable depth of discharge. After each flash the oscillator must recover for a length of time proportional to the depth of discharge before it can flash again. Evidently the rapid burster has an energy reservoir that is filled at a more or less steady rate with the gravitational potential energy of infalling matter. At some critical value of accumulated mass the energy reservoir springs a leak and a burst of X-ray energy is generated. Another clue to the mechanism of the rapid burster is that all of its bursts, regardless of their total energy flux, rise to nearly the same peak intensity. Only the duration of the bursts changes by a large factor. This indicates that the energy-conversion process is self-limiting, perhaps as a result of the effect of radiation pressure and heating on the motion of the plasma falling through the leak in the reservoir.

After the position of the rapid burster was announced William Liller of the Harvard College Observatory made a long-exposure red-sensitive plate of the star field with the new and powerful four-meter telescope recently placed in operation at the Cerro Tololo Inter-American Observatory in Chile. The plate revealed a faint nebulous object within the error circle of the rapid burster. Liller concluded that the object is probably a previously unknown and highly obscured globular cluster.

Subsequent infrared observations made by Douglas E. Kleinmann of the Harvard College Observatory in collaboration with Susan G. Kleinmann and Edward Wright of M.I.T. confirmed that the object is indeed a globular cluster with a heavy concentration of stars at its center. The plate made by Liller showed no trace of a globular cluster at the position of the nearby slow burster. Nevertheless, there were now two globular clusters known to contain X-ray burst sources. A bandwagon began to roll for the notion that the sources of all the bursters were massive black holes in the core of globular star clusters. Incidentally, the rapid burster turned off in April, 1976, and did not turn on again until 13 months later.

We now undertook another observation of NGC 6624 with SAS-3. This time we found 3U 1820–30 to be emitting bursts but from a state of low brightness. Over a period of four days we observed 22 bursts with intervals between bursts that began near 3.4 hours and gradually decreased to 2.2 hours. (As before, we missed some bursts when the source was eclipsed by the earth.) Over the four days the persistent intensity of 3U 1820–30 gradually increased until it had reached about a third of its normal bright value. At that point the bursts stopped, apparently because of some change in conditions related to the increase in the persistent intensity.

The new observation demonstrated several key facts. First, the correlation between increasing brightness and the cessation of bursts removed any remaining doubt that 3U 1820–30, lying at or near the center of the globular cluster NGC 6624, is the source of the bursts. Second, the change in the intervals between bursts demonstrated that the burst mechanism is not a good clock, such as one controlled by orbital or rotational motion, but is a relaxation oscillator with a variable power supply. As for the nature of the mysterious blob in the plate of NGC 6624 made by Neta Bahcall, an extraordinarily clear short-exposure photograph of the central region, made by Liller with the four-meter Cerro Tololo telescope, shows that it is composed of red-supergiant stars crowded together in a region only a light-year in diameter [*see bottom illustration on page 43*]. Allowing for the faint stars (perhaps 200 per red supergiant), one arrives at an estimate for the central density that is about 3,000 solar masses per cubic light-year.

By June, 1977, 29 repetitive-burst sources had been located, 17 of them by SAS-3, nine by the experiment of the NASA Goddard Space Flight Center on OSO-8 and three by ANS, *Uhuru* and the Vela satellites. Of these only four seem to be located in globular clusters. Several are definitely not in globular clusters. Thus the conclusion became clear: bursters, like X-ray stars in general, occur preferentially but not exclusively in globular clusters.

Recent results from the British X-ray satellite *Ariel 5* have thrown a monkey wrench into the hypothesis that only centrally condensed globular clusters contain X-ray stars. Kenneth A. Pounds and his associates of the University of Leicester discovered an X-ray star apparently associated with the relatively open globular cluster NGC 6712. SAS-3 measurements have confirmed the association with a position determination that places the source within one arc minute of the cluster center. Thus our current concepts of the nature of cluster X-ray stars, which assume they arise in dense cluster cores, may be in some difficulty. On the other hand, we may be witnessing in NGC 6712 the egg-beater effect in the process of dispersing the cluster that experienced its fatal central collapse some hundreds of millions of years ago. There is no observational evidence, however, that such a collapse gives rise to a massive black hole. Recent work by Stuart Shapiro of Cornell University and Alan P. Lightman of Harvard University indicates that the collapse would be halted and reversed after the formation of a small black hole (10 to 100 solar masses) or perhaps even a few close binaries.

The prospects are good for rapid advances in this area of high-energy astronomy. On August 12 of this year the first NASA High Energy Astronomy Observatory, HEAO-A, was launched. Provided with instruments substantially more sensitive than those on earlier satellites, it will quickly extend the survey of X-ray emission from globular clusters to sources that are more than 10 times fainter than the present observational limits of *Uhuru* and SAS-3.

Next year HEAO-B will be launched, carrying the first image-forming X-ray telescope to be put aboard a satellite. The instrument will have an angular resolution of two arc seconds and will achieve a thousandfold increase in sensitivity. One of the prime objectives of the mission will be systematic measurements of the positions of the X-ray sources within globular clusters. The deviation of the positions from the center of the clusters will provide a statistical measure of the average mass of the X-ray stars and should settle the question of whether or not they are massive black holes. Moreover, the sensitivity of HEAO-B will even allow a survey to be made of the X-ray emission from globular clusters in other galaxies. The 100 clusters of the Andromeda galaxy will be easy targets, and the 10,000 globular clusters of the more distant giant elliptical galaxy M87 are within reach.

Fundamental Particles with Charm

The search for particles with this quantum-mechanical property has been a preoccupation of high-energy physics. A few such particles have now been seen in the debris of electron-positron annihilations

by Roy F. Schwitters

At the level of elementary particles the properties of matter are remarkably few. A particle can have mass, or energy, and it can have momentum, including the intrinsic angular momentum called spin. It can have electric charge. There are more arcane properties, such as the one called strangeness, but not very many. In most cases a list of half a dozen attributes completely describes a particle. Nothing more can be said about it.

Because there are so few fundamental properties of matter the discovery of a new one is a major event in physics. Such a new property has recently been found: it has the whimsical name charm. The atoms of ordinary matter have no charm; the property can be observed only in the debris of high-energy collisions between particles.

The first hint that charm exists came in 1974, when particles were discovered that have charm in a hidden or latent form. More recently particles with overt charm have been detected. These new particles are unquestionably among the most important discoveries of high-energy physics in the past decade. What is more, in unraveling the story of charm physicists have learned much about the structure of ordinary matter.

Most of the particles observed in nature fall into one of two classes, the leptons and the hadrons. The leptons include just four known particles: the electron, the muon and two kinds of neutrino. The electron and the muon both have an electric charge of -1, and they are also essentially alike in all other properties with the exception of mass. (The muon is about 200 times heavier than the electron.) The neutrinos have no electric charge and seem to have no mass. There are also four antileptons, which are identical to the corresponding leptons in certain respects, such as mass, but have other properties that are exactly opposite those of the leptons. For example, the antielectron, or positron, has an electric charge of $+1$.

Leptons are considered elementary because they cannot be broken down into smaller entities. They have no measurable size and give no hint of any internal structure. Hadrons, on the other hand, are complex objects, and there is strong evidence they do have an internal structure. More than 100 kinds of hadron have been identified, the most familiar being the proton and the neutron, the constituents of atomic nuclei. Both the multiplicity of the hadrons and their properties as individual particles set them apart from the leptons.

A theory that is now widely accepted seems to explain these differences. It does so by stating that hadrons are not elementary particles at all but composite entities consisting of a few simpler constituents called quarks. In many of their properties the quarks are thought to be quite similar to leptons; for example, they ought to be simple and pointlike particles. There is no question, however, that the quarks are in a class apart from the leptons. Interactions between quarks are dominated by a force that does not affect leptons at all.

Physicists recognize four basic forces in nature; in order of increasing strength they are the gravitational, the weak, the electromagnetic and the strong forces. Gravitation influences all particles and its range is unlimited, but its effects on subatomic particles are negligible. The weak force also affects all kinds of matter. Although the weak force is many orders of magnitude stronger than gravitation, it is still feeble enough for it to be observable only when stronger interactions are inhibited.

The electromagnetic force acts exclusively on particles that have an electric charge; among those particles are the electron and the muon and all the quarks. Electromagnetic forces bind atoms together and are responsible for almost all the gross properties of matter, including chemical properties.

It is the strong force that distinguishes between leptons and hadrons or, according to theory, between leptons and quarks. None of the leptons is sensitive to the strong force in any way; only quarks and hadrons (which are assumed to be made up of quarks) feel its influence. Quarks can interact with leptons through the weak and electromagnetic forces, but with each other they interact almost exclusively through the strong force. That force is more than 100 times stronger than the electromagnetic force, and at energies studied today it is about 10^{10} times stronger than the weak force.

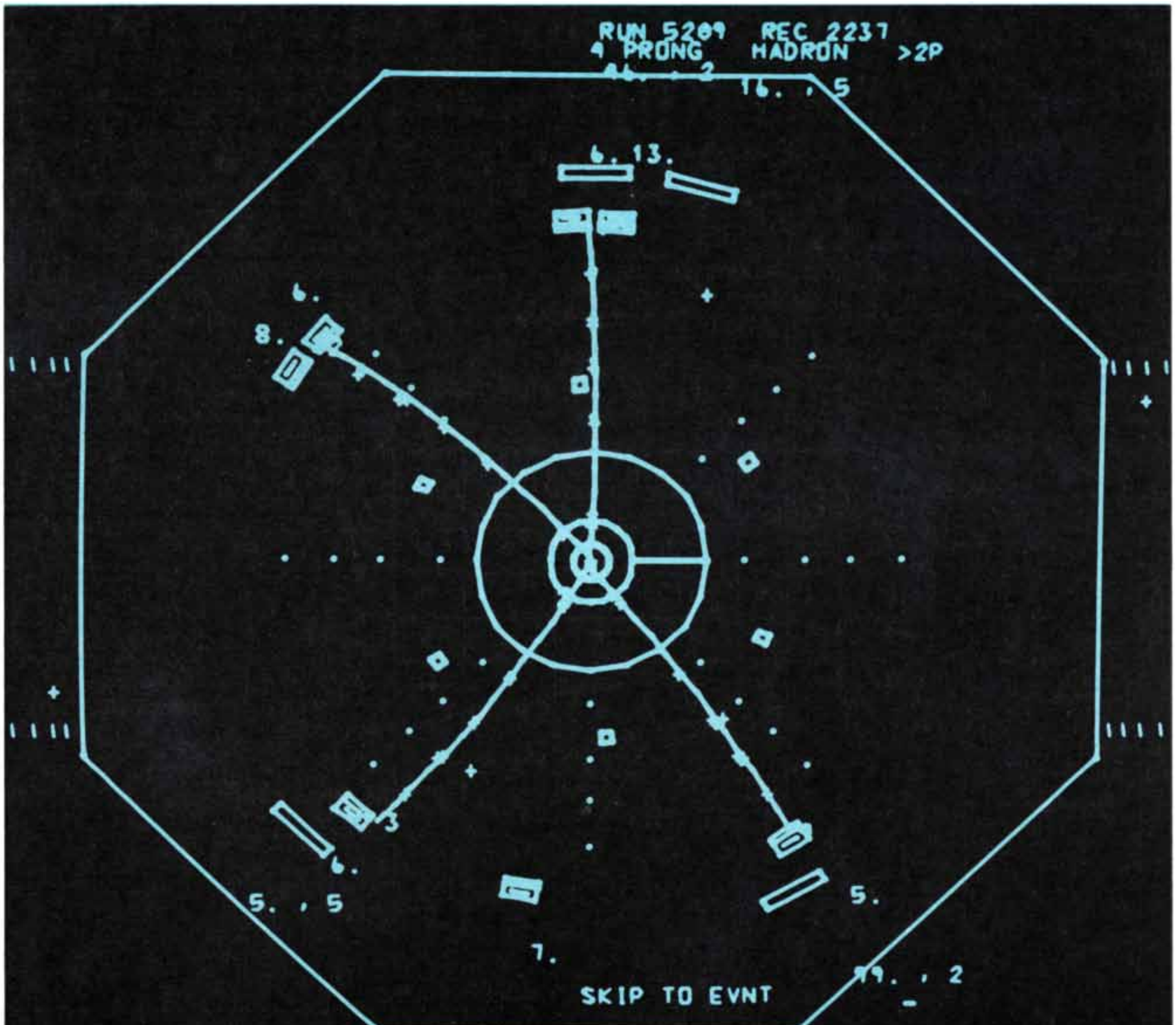
A theory that accounts for all the varieties of matter with just a few quarks and leptons has an appealing economy, but a significant disclaimer is necessary. Although the theory has gained widespread acceptance, there is no evidence at all that quarks exist in isolation. So far no one has been able to extract a quark from a hadron. Indeed, some theorists suggest that quarks may be so tightly entwined within hadrons that they may never be separated in the laboratory. The question of the existence of quarks will not be pursued here; instead the quarks will be considered as a means for interpreting relations between particles observed experimentally.

The quark hypothesis was proposed independently in 1963 by Murray Gell-Mann and George Zweig, both of the California Institute of Technology. In the original version of the concept there were three kinds of quark, given the labels *u*, *d* and *s*, for "up," "down" and "strange," and three corresponding antiquarks, designated $\bar{u}$, $\bar{d}$ and $\bar{s}$. Hadrons are formed by combining the quarks and antiquarks according to simple rules. One possibility is for a quark and an antiquark to bind together; the resulting particle is a member of the class of hadrons called mesons. An example is the positively charged pion (π^+), which is made up of a *u* quark and a $\bar{d}$ antiquark. Another allowed combination consists of three quarks in a bound system. Hadrons formed in this way are called baryons, and they include the proton (with the quark composition *uud*) and the neutron (*udd*). Finally, antibaryons can be formed from combinations of three antiquarks.

These are the only permissible ways of combining quarks to form hadrons. It is easy enough to imagine other combinations, such as particles made up of two quarks or of one quark and two antiquarks, but such hadrons do not exist.

The observed properties of hadrons are explained in a straightforward way by the properties assigned to their constituent quarks. With the exception of mass, all the properties of matter needed to identify a given elementary particle appear only in discrete units, or quanta, and they can therefore be measured in terms of the integers and simple fractions called quantum numbers. Spin angular momentum, for example, is observed only in integer and half-integer quantities (when it is measured in natural units); intermediate values are not possible. In all observed particles electric charge exists only in integer units of the electron's charge.

Most quantum numbers of a hadron



DECAY OF A CHARMED PARTICLE is reconstructed in a computer-generated display on a cathode-ray tube. The display includes a schematic cross section of the particle detector in which the event took place. The octagon is the basic form of the detector; the rows of four dots give the radial positions of four cylindrical spark chambers; the innermost circle at the center of the image represents the beam pipe in which electrons and their antiparticles, positrons, collide to produce other particles. The trajectories of particles are plotted from the triggering of discharges in the spark chambers and from two concentric rings of scintillation counters, the inner ones being represented by small boxes and the outer ones by larger, flat boxes. In this event the particle track at the 12 o'clock position has been identified as that of a negatively charged *K* meson (K^-) and the track at the seven o'clock position is that of a positively charged pion (π^+). These two particles are thought to be the decay products of a particle, designated D^0, that bears the new property of matter called charm. The D^0 itself decays too quickly for it to be detected directly. In electron-positron collisions charmed particles can be created only in combination with charmed antiparticles, and in this case the antimatter companion of the D^0 is thought to be an excited state, the $\bar{D}^{*0}$, whose decay products have been identified only tentatively. The scintillation counter near the bottom of the display, which has no associated particle track, might have been triggered by a gamma ray emitted when the $\bar{D}^{*0}$ decayed to a $\bar{D}^0$. This particle might have decayed to yield a muon (μ^-) at the 10 o'clock position, a *K* meson (K^+) at the five o'clock position and a neutrino that would have escaped detection.

LEPTONS	MASS (GeV)	SPIN	ELECTRIC CHARGE
ELECTRON NEUTRINO ν_e	0	½	0
ELECTRON e⁻	.0005	½	−1
MUON NEUTRINO ν_μ	0	½	0
MUON μ^-	.105	½	−1

QUARKS	MASS (GeV)	SPIN	ELECTRIC CHARGE	STRANGENESS	CHARM
u	.1	½	+⅔	0	0
d	.1	½	−⅓	0	0
s	.4	½	−⅓	−1	0
c	1.5	½	+⅔	0	+1

LEPTONS AND QUARKS are the only major classes of particles now thought to be elementary. Both seem to be simple and pointlike entities, with no internal structure and no measurable size. There are four known leptons, arranged in pairs. In the initial formulation of the quark theory there were only three quarks: those labeled *u* and *d* formed a pair but the *s* quark had no companion. The charm hypothesis establishes a symmetry between leptons and quarks by adding a fourth quark, designated *c*. A special relation must exist between the *s* quark, which carries the property of matter called strangeness, and the *c* quark, which has the similar property charm. For each of the leptons and quarks there is an antiparticle with exactly opposite properties. Quarks have not been isolated, but they seem to be constituents of other particles.

	HADRON	SYMBOL	QUARK COMPOSITION	MASS (GeV)	SPIN	LIFETIME (SECONDS)	ELECTRIC CHARGE	STRANGENESS	CHARM
BARYONS	PROTON	p	uud	.938	½	STABLE	+1	0	0
BARYONS	NEUTRON	n	udd	.940	½	10^3	0	0	0
BARYONS	LAMBDA	Λ	uds	1.116	½	10^{-10}	0	−1	0
BARYONS	CHARMED LAMBDA	Λ_c	udc	2.260	½	?	0	0	+1
MESONS	PI MESONS	π^+	$u\bar{d}$	.140	0	10^{-8}	+1	0	0
MESONS	PI MESONS	π^-	$d\bar{u}$	.140	0	10^{-8}	−1	0	0
MESONS	PI MESONS	π^0	$u\bar{u} + d\bar{d}$	.135	0	10^{-16}	0	0	0
MESONS	K MESONS	K^+	$u\bar{s}$	.494	0	10^{-8}	+1	−1	0
MESONS	K MESONS	K^-	$s\bar{u}$	.494	0	10^{-8}	−1	+1	0
MESONS	PHI	ϕ	$s\bar{s}$	1.020	1	10^{-22}	0	0	0
MESONS	PSI	ψ	$c\bar{c}$	3.095	1	10^{-20}	0	0	0
MESONS	CHARMED MESONS	D^0	$c\bar{u}$	1.863	0	?	0	0	+1
MESONS	CHARMED MESONS	D^+	$c\bar{d}$	1.868	0	?	+1	0	+1
MESONS	CHARMED MESONS	F^+	$c\bar{s}$	?	0	?	+1	+1	+1

HADRONS are the particles constructed of quarks. Those called mesons are made up of a quark and an antiquark, the latter being represented by the symbol for a quark with a bar over it. Baryons consist of three quarks bound together; there are also antibaryons that combine three antiquarks. The properties of the hadrons (and of other particles) are accounted for in terms of the integers and simple fractions called quantum numbers. The first hadron that hinted at the existence of charm was the psi meson, with the quark constitution $c\bar{c}$, but since the charm quantum numbers of this quark and antiquark cancel, the net charm of the psi is zero. The lowest-mass hadrons that exhibit overt charm are the D^0, D^+ and F^+ mesons, and the first two of these have been observed. There is also suggestive evidence for a charmed baryon, named Λ_c.

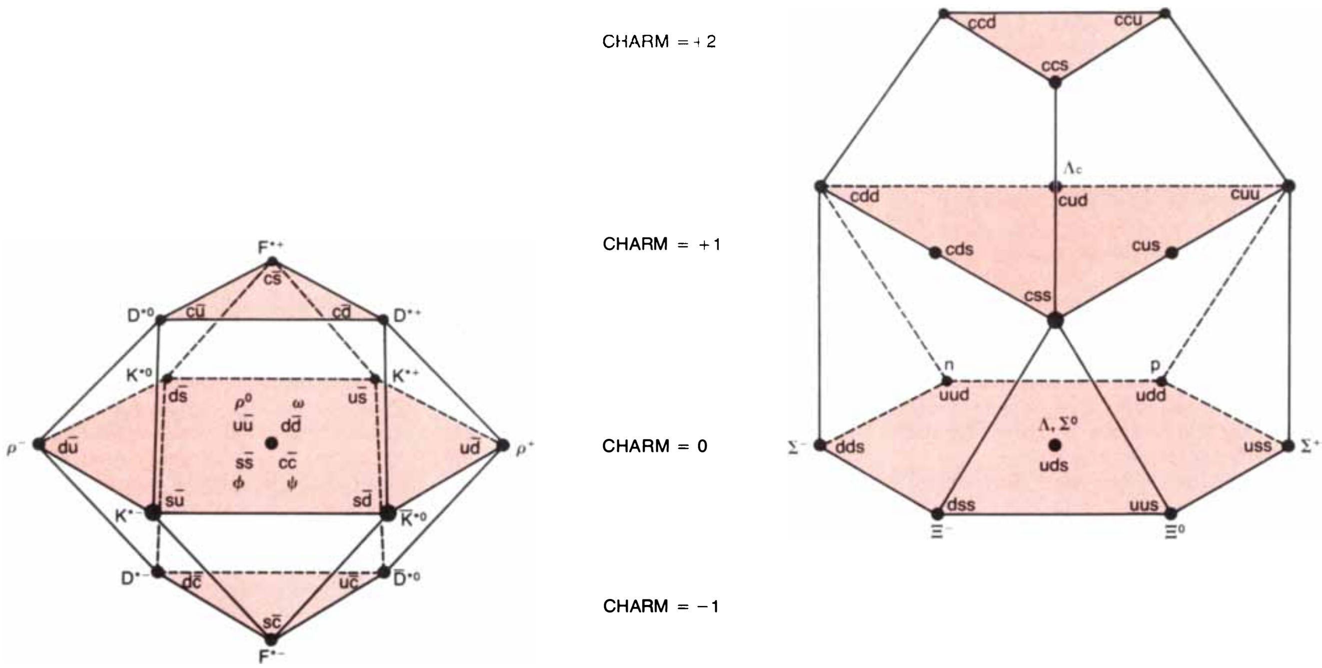
are determined by simply adding up the quantum numbers of the constituent quarks. In the case of electric charge this procedure requires that the quarks be assigned some rather bizarre quantum numbers: they must have fractional electric charge. The *u* quark has a charge of $+2/3$; the *d* and *s* quarks both have a charge of $-1/3$. Antiquarks have the opposite charges. Thus in the positively charged pion, made up of a *u* quark and a $\bar{d}$ antiquark, the charges of $+2/3$ and $+1/3$ add up to $+1$. The *uud* quarks of the proton, with charges of $+2/3$, $+2/3$ and $-1/3$, also give a sum of $+1$. In the neutron (*udd*) the charges of $+2/3$, $-1/3$ and $-1/3$ have a sum of zero. All these results are in agreement with the known electric charges of the hadrons.

Other quantum numbers can be treated in a similar way. Strangeness, for example, is a quantum number assigned to certain hadrons that have anomalously long lifetimes. In the quark model these particles are distinguished by the presence of an *s* quark or an $\bar{s}$ antiquark, which respectively carry strangeness quantum numbers of -1 and $+1$; the other quarks have zero strangeness. The strangeness of a hadron is then determined by adding up the strangeness quantum numbers of all its constituent quarks. The positively charged *K* meson (K^+), made up of a *u* quark and an $\bar{s}$ antiquark, has a strangeness of $+1$.

The concept of strangeness was introduced as an explanation for the slow decay of certain hadrons discovered in the 1950's. Massive hadrons generally decay through the strong force, which acts very quickly; in a sense the force is said to be strong precisely because it is so fast. A strongly decaying hadron exists for only 10^{-23} second before it breaks up into less massive hadrons. All the properties of matter must remain unchanged by this process; for example, the net electric charge of the decay products must be equal to the electric charge of the decaying particle. This requirement is expressed by saying that all quantum numbers must be conserved.

Strange hadrons can take part in such decays provided the strangeness quantum number can be conserved. In terms of the quark model the need to conserve strangeness implies that the strange *s* quark must be passed along intact to the decay products. The more massive strange particles do decay to lighter strange particles in this way, and their lifetimes are not substantially longer than those of other strongly decaying hadrons. Ultimately, however, there must be a set of lightest strange particles, which have no states of lower mass to which they can give the *s* quark. These are the *K* mesons and the lambda baryon (Λ).

The *K* and lambda particles cannot decay through the strong interaction. Indeed, they could not decay at all if it



FAMILIES OF HADRONS define all allowed combinations of quarks. There are separate families for each possible value of intrinsic spin angular momentum; those shown here are the mesons with one unit of spin and the baryons with 1/2 unit of spin. Until recently only the particles in the plane marked "Charm = 0" had been observed, but the discovery of charmed particles implies that all the others must also exist. The psi particle completes a group of four particles at the center of the array of spin-1 mesons. The D^0, D^+ and F^+ are members of another family of mesons, those with zero spin, but they are represented in the spin-1 group by excited states denoted D^{*0}, D^{*+} and F^{*+}. The D^{*0}, the D^{*+} and their antiparticles have been found. As yet, there is little evidence for the observation of charmed baryons.

were not for a remarkable feature of weak interactions: some conservation laws that are strictly obeyed in all strong and electromagnetic interactions can be ignored in weak interactions. Strangeness is among these approximately conserved quantum numbers; as a result weak interactions can convert an *s* quark into a *u* or a *d* quark and strange particles can decay through weak interactions into lighter nonstrange hadrons or into leptons. Particles that can decay only through the weak interaction, however, have much longer lifetimes than strongly decaying hadrons; the *K* and lambda particles have lifetimes ranging from 10^{-10} second to 10^{-8} second.

A few quantum numbers require a slightly more elaborate system of bookkeeping than electric charge and strangeness do. Most notable among them is spin angular momentum. All quarks and leptons are expected to have the same quantity of spin, namely 1/2 unit; rules of quantum mechanics require that such particles can have their spin axes aligned in only two possible directions (with respect to any arbitrary frame of reference). Thus the quark and antiquark in a meson can have their spins pointing in the same direction or in opposite directions. In the first case the net angular momentum of the meson is 1; in the second it is 0. From these rules it follows that mesons always have integer spin. In a baryon there are also two possible spin configurations: the spins of all three quarks can be parallel or the spin of one quark can be aligned in the opposite direction. The corresponding spins of the baryon are 1/2 and 3/2, and baryon spins are invariably half-integers. Further complicating the situation, a bound system of quarks can have orbital angular momentum in addition to spin: the quarks can revolve around each other or around their common center of mass as well as spin individually on their axes. Orbital angular momentum adds integer increments to the total angular momentum of the hadron. Because there are many possible spin states a single combination of quarks can give rise to numerous particles with various quantities of angular momentum and also with different energies, or masses. They are distinct states of matter even though they are made up of the same quarks.

By observing the rules for combining the quarks and for tallying their quantum numbers one can account for all the properties of hadrons. Every known hadron can be explained as a combination of either a quark and an antiquark or of three quarks. What is more, every allowed combination of quarks corresponds to a known hadron. There are no vacancies.

It was the very completeness of this scheme for classifying the hadrons that was challenged with the discovery of a new particle in 1974. The particle was a hadron, but it could not be formed by any allowed combination of the three quarks since all those combinations were accounted for.

The new particle was discovered independently and at about the same time by two groups of experimenters employing vastly different techniques. One group, whose members were from the Massachusetts Institute of Technology and the Brookhaven National Laboratory, found the particle during an experiment conducted at Brookhaven and gave it the name *J*. The other group, of which I was a member, was made up of physicists from the Stanford Linear Accelerator Center (SLAC) and the Lawrence Berkeley Laboratory. Our evidence for the new particle was obtained in an experiment at SLAC, and we chose to designate the particle by the Greek letter psi (ψ). Here I shall adopt the latter name. Last year the leaders of the two groups, Samuel C. C. Ting of M.I.T. and Burton Richter of SLAC, shared a Nobel prize for their discovery.

The new particle has a mass of about 3.1 billion electron volts (3.1 GeV). That is more than three times the mass of the proton, making the psi one of the heaviest particles known. Those of its quantum numbers that could be measured were found to be quite conventional. The psi has a spin angular momentum of 1, the integer unit indicating the particle must be a meson. It is electrically neutral, and it has zero strangeness. The products of its decay are familiar particles such as pions, electrons and muons. With these properties the psi might have passed for an ordinary hadron, quite unexceptional except for its mass. The problem was that all states of matter with the quantum numbers of the psi had long since been identified with other hadrons. In a world made of three quarks there was no need for the psi; there was not even room for it.

Experimentally the one distinctive trait of the psi is its lifetime, which is exceptionally long. Many hadrons with large masses are known, but almost all of them decay very rapidly through strong interactions; their lifetimes are generally on the order of 10^{-23} second. The psi particle also decays through the strong interaction, but it has a lifetime of about 10^{-20} second, 1,000 times longer than the lifetime of a typical hadron of comparable mass. Some explanation of its longevity is required.

A suggestive precedent for the long lifetime of the psi is found in another particle discovered some years earlier: the phi meson (ϕ). The phi meson is made up of a strange quark bound to a strange antiquark, and since the strangeness quantum numbers of the quark and antiquark cancel, it has a net strangeness of zero. The phi can decay through the strong interaction; all that is required is that the quark and the antiquark annihilate each other. This process conserves all quantum numbers, but it nonetheless seems to be inhibited. The lifetime of the phi is roughly 10 times longer than might have been expected.

The simplest way to provide a place among the hadrons for the psi particle is to assume it is a meson made up of a new, massive quark bound to the corresponding antiquark. This structure automatically accounts for the quantum numbers of the psi. The long lifetime might be explained by a mechanism similar to that which retards the decay of the phi meson: strong decays would be allowed but they might be slowed somewhat.

A bound state of a new quark and a new antiquark was the leading explanation for the psi particle from the time of its discovery. In large measure this hypothesis was credible only because the existence of a fourth quark had been proposed long before and on grounds that had nothing to do with the discovery of new hadrons.

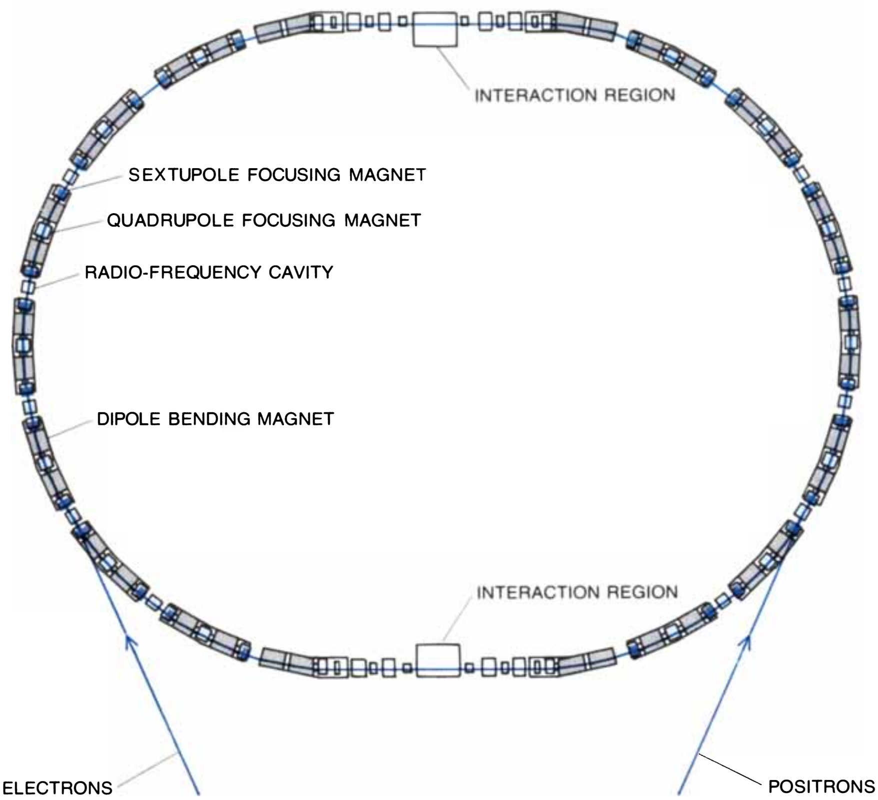
The new quark was suggested by a number of theorists as a natural extension of the model formulated by Gell-Mann and Zweig. Initially it was supported by an esthetic argument based on the notion that there may be a deep connection between leptons and quarks. Since there are four known leptons, it was argued, the spectrum of elementary particles would be far more attractive if there were also four quarks. The leptons come in pairs, the electron being associated with one neutrino and the muon with the other. The *u* and *d* quarks form a similar pair, but the *s* quark is without a companion. The new quark was designed to fill this gap. In order to do so it must have an electric charge of +2/3 and some new quantum number that would distinguish it from the known quarks. James D. Bjorken of SLAC and Sheldon Lee Glashow of Harvard University gave the quantum number the name charm. The charmed quark is designated *c*, the charmed antiquark $\bar{c}$.

Charm is a property much like strangeness: it must be conserved in all strong and electromagnetic interactions but not in weak ones. Hence the lightest charmed particles should decay only through the weak interaction, and they should have commensurately long lifetimes. The psi particle, however, is exempt from this rule because its net charm is zero. It consists of a *c* quark and a $\bar{c}$ antiquark whose charm quantum numbers cancel.

A substantial theoretical argument for the charmed quark was made in 1970 by Glashow in collaboration with John Iliopoulos and Luciano Maiani. It was based on observations that did not directly concern charmed particles at all but involved certain weak interactions of strange particles. Weak interactions can proceed either with or without a transfer of electric charge between the interacting particles. By 1970 a puzzling correlation had been established experimentally: with very rare exceptions strangeness is changed in weak interactions only when electric charge is transferred. The so-called neutral weak currents, which do not transfer charge, also seemed not to alter strangeness.

In the three-quark model there was no obvious connection between charge-transfer and strangeness-changing. Glashow, Iliopoulos and Maiani pointed out that the addition of a fourth, *c*, quark closely associated with the *s* quark could account for the suppression. How it does so involves a rather subtle quantum-mechanical argument. The fourth quark does not directly impede the interactions that change strangeness without transferring charge. Instead it provides an alternative channel for those interactions but in such a way that the effects of the two channels cancel.

By the time the psi was discovered charmed particles were already prominent items on a long list of entities that had been predicted to exist but had never been observed. The charm hypothesis could explain the properties of the psi, but so might other theories. On the other hand, the charm model predicts much more than a single new particle; if charmed quarks exist, there must be an



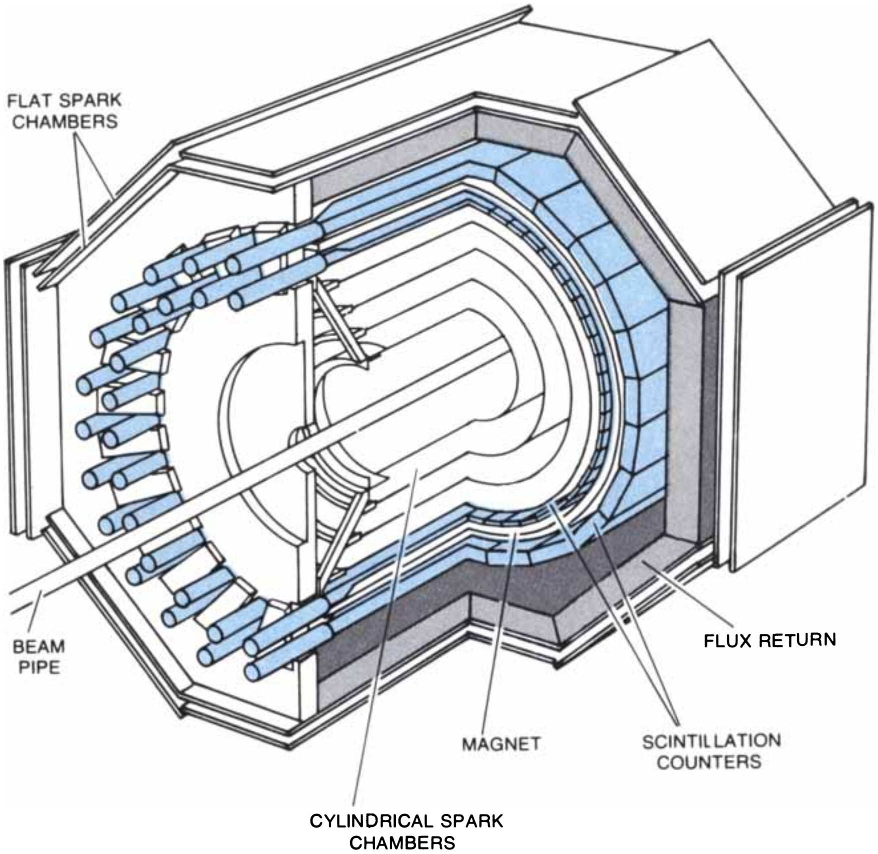
PARTICLE-STORAGE RING at the Stanford Linear Accelerator Center (SLAC) was one of the devices with which the psi particle was discovered and, more recently, where charmed mesons have been observed. The oval ring, which is called SPEAR, is supplied with electrons and positrons by the two-mile accelerator at SLAC. Counterrotating electrons and positrons are confined to the ring by magnets that bend the particles' trajectories and focus the beams into needle-shaped "bunches." Each bunch is made up of about 10^{11} particles, and the bunches are synchronized so that they pass through each other in the two straight sections of the ring. Energy lost by the particles is replenished by the radio-frequency cavities. The magnets confine particles moving in opposite directions since electrons and positrons have opposite charges.

entire spectrum of new states of matter. Some of these states are closely related to the psi: they consist of a c quark and a $\bar{c}$ antiquark, but they have different masses and different values of angular momentum. The charmed quark must also combine with the original three quarks to form dozens of new hadrons with overt charm. By searching for these particles we could determine whether or not the psi particle actually contains a new kind of quark and whether or not that quark possesses the proposed quantum number charm.

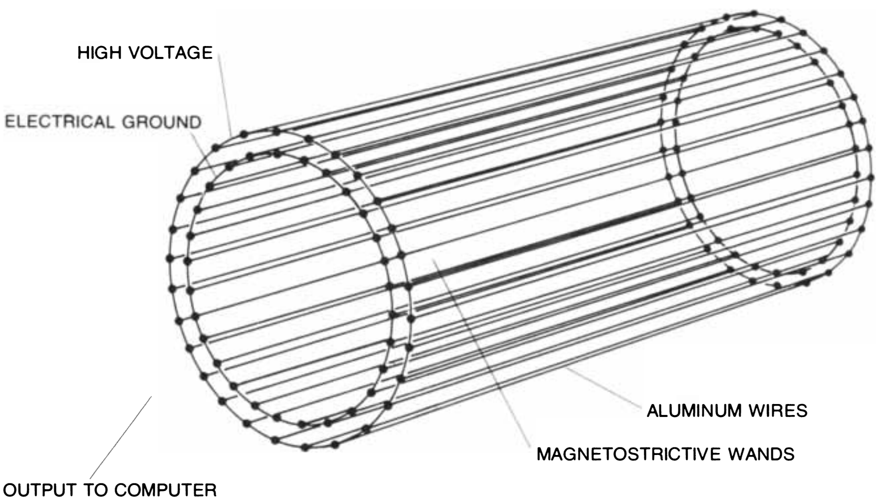
At SLAC the experiment in which the psi was discovered and subsequent searches for other particles containing charmed quarks were carried out by the same basic technique. Electrons and their antiparticles, positrons, are made to collide at high energy. The particles of interest are then found by sorting through the debris of the collisions.

The collisions take place in a device called a storage ring, in which electrons and positrons circulate in opposite directions at nearly the speed of light. The storage ring at SLAC is called SPEAR; it was built in 1971 and 1972 under the direction of Burton Richter and John Rees. At the heart of the SPEAR ring is a toroidal aluminum vacuum chamber with a mean diameter of 80 meters and a cross section of a few inches. When the ring is operating, about 10^{11} electrons circle it clockwise in a needle-shaped "bunch" a few centimeters long and less than a millimeter thick. A similar bunch of positrons circulates counterclockwise. The bunches cross twice during each revolution, and their orbits are timed so that the collisions take place in two short straight sections of the ring where particle detectors are emplaced. Elsewhere on the circumference of the ring the vacuum chamber is encased in large electromagnets, which bend the trajectories of the particles to the circular path and keep the bunches focused. There are also four cavities where radio-frequency energy is supplied to replenish the energy lost by the circulating electrons and positrons.

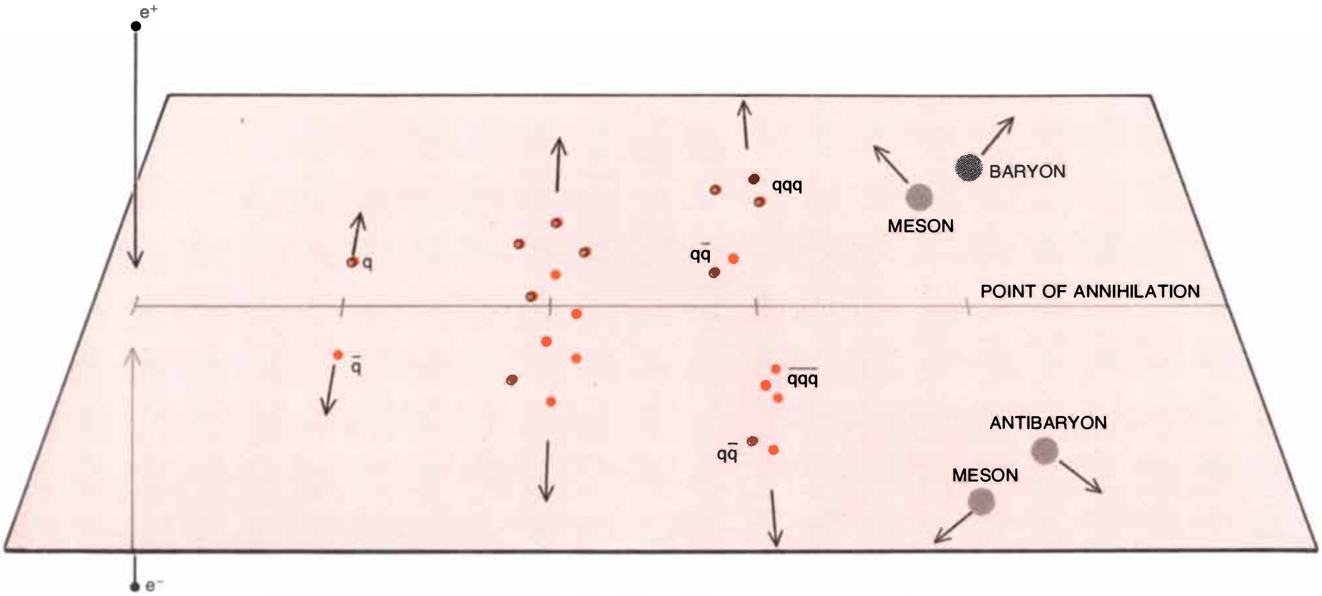
The electrons and positrons are generated by the two-mile-long linear accelerator at SLAC. At the far end of the accelerator electrons are boiled off a hot filament, much as they are in an ordinary vacuum tube. The electrons are injected into SPEAR merely by accelerating them and diverting them into the ring. Supplying positrons requires a slightly more elaborate procedure. First the electrons are propelled through a third the length of the accelerator, where they collide with a copper target, generating a shower of secondary particles that includes electrons, positrons and gamma rays. The remaining length of the accelerator is then adjusted to accelerate only the positrons. Replenishing the stored beams takes only a few



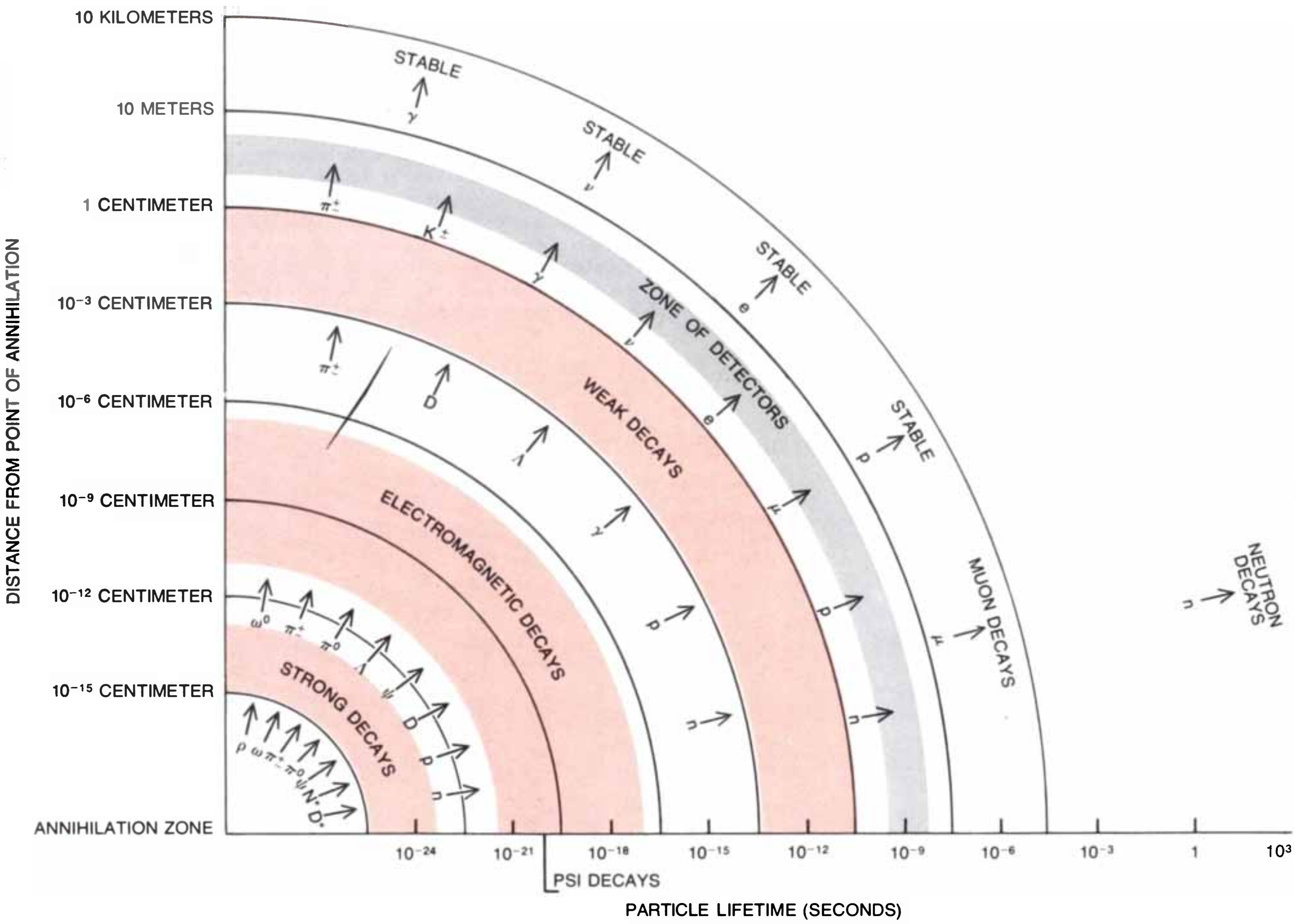
MAGNETIC DETECTOR at SPEAR was crucial to the discovery of the psi particle and of charmed mesons. The detector includes two kinds of device: spark chambers and scintillation counters. Both are sensitive to electrically charged particles emitted in electron-positron collisions at the center of the detector. A strong magnetic field bends the trajectories of the charged particles, and from the amount of bending their momentum can be determined. The scintillation counters are made of a plastic that emits a flash of light when a particle passes through it; they measure a particle's time of flight and hence its velocity. From the momentum and the velocity of a particle its rest mass can be estimated, providing an important clue to its identity.



SPARK CHAMBERS in the magnetic detector consist of concentric layers of fine aluminum wires; when the detector is triggered, a high voltage is applied across the layers. Only two layers are shown here; each chamber actually has four layers and there are four chambers, for a total of about 100,000 wires. When the spark chamber is in operation the wires are immersed in a noble gas that is readily ionized by the passage of an electrically charged particle. The ionized atoms left behind provide a conductive path for a spark between wires in the oppositely charged layers, thereby marking the position at which the particle passed through the cylindrical chamber. The positional information is recorded through a system of magnetostrictive "wands" laid down perpendicular to the aluminum wires at the end of the chamber. The surge of current that flows through the active wire when a spark is triggered gives rise to a pulse in the magnetostrictive wand, which propagates as a sound wave that can be detected and timed.



ANNIHILATION of an electron and a positron gives rise to equal numbers of particles and antiparticles. When hadrons are created as a result of a collision, the initial product is a single quark and an antiquark of the same type. From the energy imparted to this pair, other quarks and antiquarks materialize, and ultimately they coalesce in pairs and triplets to form hadrons. Only the hadrons can be detected.



SEQUENCE OF DECAYS following an electron-positron annihilation converts the great variety of particles created into a few stable or nearly stable species. Most hadrons that can decay through strong interactions have completed their decays within 10^{-23} second, when they have traveled only about 10^{-13} centimeter from the point of collision. The psi particle is an exception: it decays through the strong interaction but requires about 10^{-20} second. The products of the strong decays are longer-lived hadrons, some of which subsequently decay electromagnetically, yielding still other hadrons along with gamma rays (γ). Weak decays are the slowest, but most of them are completed before the particles have traveled more than a centimeter. After the weak decays only leptons, gamma rays and the longest-lived hadrons survive. Detectors cannot be placed closer than a few centimeters to the annihilation zone, and so they detect the products of a collision only after most decays have been completed. Nevertheless, collision products can often be identified. Gamma rays indicate that a particle has decayed electromagnetically, and leptons are evidence for a weak decay. *K* mesons signal the presence of a strange hadron.

minutes, after which they circulate stably for periods of up to several hours.

The energy of the stored beams can be adjusted between 1.2 GeV and 4 GeV. One of the principal motivations for employing storage rings is that all this energy is available for the creation of new particles. When an electron and a positron collide, both particles are annihilated and all their mass and kinetic energy is converted into a state of pure energy. Thus the machine has the potential for creating sets of particles with total masses of up to 8 GeV.

The energy liberated by the annihilation is confined to a small region of space, with dimensions of 10^{-15} to 10^{-14} centimeter. In an unmeasurably brief interval—the time required for light to traverse such a region—the energy is transformed into material particles, which then fly away from the point of annihilation. There are few constraints on the kinds of particle that can be created from this dense bundle of energy. One obvious requirement is that energy must be conserved: the rest masses of the particles created cannot exceed the sum of the electron and positron energies. Quantum numbers must also be conserved, but this requirement is less restrictive than it might seem at first.

The annihilation of an electron and a positron is mediated by the electromagnetic force and the immediate product is a photon, a quantum of electromagnetic energy. The photon decays so quickly it can never be detected, even in principle (it is called a virtual particle), but it nonetheless determines the properties of all subsequent states of the system. If all the decay products of the virtual photon are detected and their quantum numbers added up, the totals must be equal to the quantum numbers of the photon. The psi particle can be created in these events because its quantum numbers are precisely those of the photon: it has one unit of angular momentum but all its additive quantum numbers—those determined by simply adding up the quantum numbers of the constituent quarks—are zero. Why this should be so becomes apparent when the quark composition of the psi is considered. It consists of a quark bound to its own antiquark, and all quantum numbers except angular momentum exactly cancel.

In this context what had seemed a restrictive demand that the properties of the photon be conserved is seen to confer almost complete freedom for the creation of any particle as long as it is accompanied by its own antiparticle. In the psi the particle-antiparticle pair is internal, but it need not be so. For example, a charmed meson could be produced in combination with another meson having the opposite charm quantum number. Given enough beam energy almost any state of matter (and antimatter) could be formed. Collisions at SPEAR are energetic enough to yield up to 50 pions (half π^+ and half π^-) or four protons together with four antiprotons or even a complete atom of helium and an atom of antihelium.

As particles emerge from the annihilation region at speeds near the speed of light they commonly decay through several generations to lighter and longer-lived daughter particles. Most of the massive hadrons have undergone strong decays by the time they have traveled 10^{-13} centimeter, which is approximately equal to the diameter of a hadron. The products of these decays are lighter hadrons, such as pions, *K* mesons and protons and neutrons. Hadrons that decay electromagnetically generally travel less than 10^{-6} centimeter before they emit a photon and are thereby converted into lower-mass hadrons. Most weakly decaying particles, such as the strange lambda baryon, travel a few centimeters before they are transformed into the lowest-mass hadrons, such as protons, neutrons and pions, along with leptons. Ten centimeters from the point of interaction most of the surviving particles are protons and neutrons, electrically charged pions and *K* mesons, gamma rays and leptons. Within a few meters the pions and *K* mesons decay to leptons, leaving only protons, neutrons, gamma rays, electrons, muons and neutrinos. Much farther along the neutrons and the muons decay.

It is the goal of the experimenter to detect as many of these particles as possible and to extract a maximum amount of information from them. In practice the closest a detector can approach the interaction region is a few centimeters, and so there is no hope of seeing particles before both strong and electromagnetic decays are completed. At the other end of the scale detectors larger than a few meters in diameter become too costly and unwieldy to be practical. Even within this limited range of distances it is generally only the charged particles and in some cases the gamma rays that can be detected. Nevertheless, it is often possible from the particles detected to reconstruct the chain of events that gave rise to them. For example, the presence of *K* mesons among the decay products can be read as a sign that strange particles were created. Gamma rays indicate that an electromagnetic decay has taken place and leptons are a clue to the passage of a weakly decaying particle.

The detector in which the psi was discovered surrounds one of the interaction regions at SPEAR. It was built by physicists from SLAC and from the Lawrence Berkeley Laboratory while SPEAR was under construction. Various individuals and small groups from both laboratories had responsibility for components of the detector and for its computer programs.

The core of the detector is a 150-ton solenoid magnet whose useful field volume is a cylinder three meters in diameter and three meters long. Within the field are four cylindrical shells of spark chambers, which record the trajectories of charged particles as they pass through the detector. The spark chambers contain a total of 100,000 aluminum wires, which sense the momentary ionization of an inert gas caused by the passage of a particle. The magnetic field deflects charged particles, and by measuring the curvature of their trajectories one can determine the momentum of the particles.

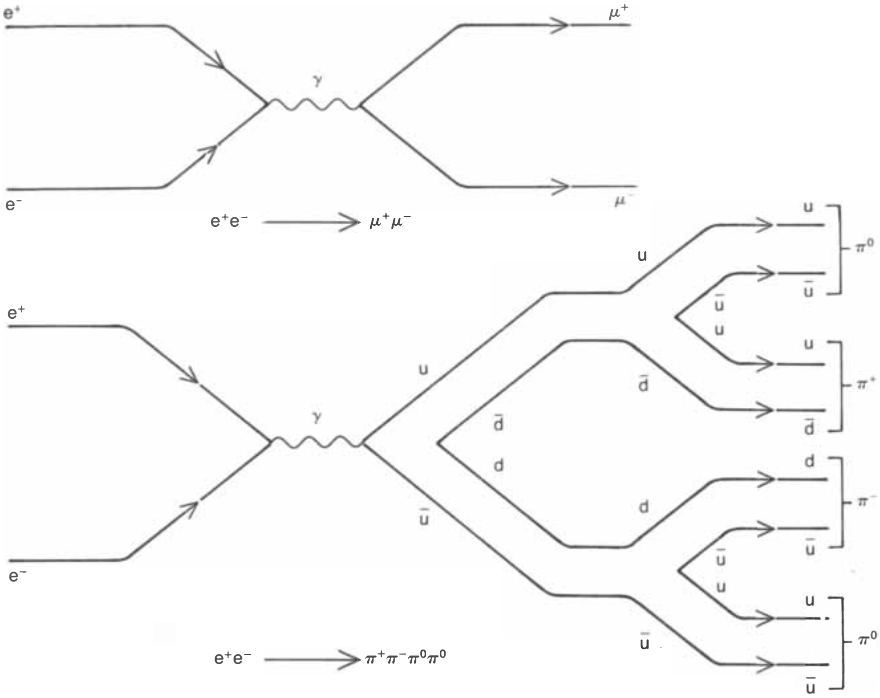
Beyond the spark chambers are scintillation counters, where brief flashes of light emitted when a particle penetrates a strip of phosphorescing plastic are detected photoelectrically. The scintillations can be timed precisely, and so the time of flight and the velocity of the particles can be calculated. Knowing the velocity of a particle and its momentum, one can deduce its rest mass, an important clue to its identity. The comparison of velocity and momentum is our principal means for distinguishing between the three charged hadrons that survive long enough to be detected, namely pions, *K* mesons and protons.

A final peripheral array of scintillation counters and spark chambers is shielded from the interior of the detector by filtering layers of lead and iron. The differential abilities of various particles to penetrate these metals allow us to discriminate between hadrons and leptons and, among the leptons, between electrons and muons.

Most of the time the electron and positron beams pass through each other without interacting at all. Even though each bunch contains 100 billion particles, the leptons are so small—indeed, their size has so far proved unmeasurable—that the probability of a collision is low. On the other hand, the beams move so rapidly, circling the ring more than a million times a second, that practical rates of data collection can be achieved.

The detector is triggered whenever two or more charged particles coming from the vicinity of the interaction region strike scintillation counters simultaneously with the passage of the beams through the detector. When a triggering event is recorded, all the information from the spark chambers and the scintillation counters is transferred to a computer, where it is recorded on magnetic tape. The detector is triggered about once a second, but only a few percent of these events derive from electron-positron annihilations. The rest are caused by cosmic rays and by interactions between the particle beams and residual gas molecules in the vacuum chamber.

The data collected on tape are analyzed with the aid of another computer. Background events are identified and discarded and for the more interesting events the trajectories of the particles



ANNIHILATION EVENTS can be divided into two broad categories: those that give rise to a pair of leptons, such as a positive and a negative muon, and those that lead to the creation of a quark and an antiquark, which eventually appear as a cluster of hadrons. If both leptons and quarks have a pointlike distribution of electric charge, then the rates of production can be calculated; the rates are determined by the number of kinds of leptons and quarks and by the squares of their charges. Measurements of the rate of hadron production therefore test various quark models, discriminating, for example, between those that include charmed quarks and those that do not. In practice the absolute production rates are not measured; instead the ratio of hadron production to muon production is determined, a ratio designated by the letter *R*.

are reconstructed. In a typical event involving hadrons four or five particles may be detected; their momenta and velocities are measured, and if possible the particles are identified.

At any given time approximately 35 physicists are involved with the operation of the detector and data analysis. Although our number varies as students and visitors come and go, without the dedication and talents of a great many individuals in our collaboration, this large detection and analysis apparatus could not function.

When the psi particle was discovered in 1974 we were not searching for charmed particles but were engaged in a much broader inquiry into the structure of hadrons. We were attempting to test the assumption that hadrons are made up of electrically charged, pointlike particles, such as the quarks are supposed to be.

Quantum electrodynamics, a theory that has been verified with great precision, is able to predict the rates at which particles are created in electron-positron annihilations provided the particles have a pointlike distribution of electric charge. The rate is proportional to the sum of the squares of the particles' charges. This prediction has been tested for electrons and for muons, with good agreement between experiment and theory. Quarks are thought to behave exactly like leptons in electromagnetic interactions, and so the same rule should predict their rate of production also.

Of course, we cannot directly observe the quarks created in the aftermath of an electron-positron collision, but it turns out this is no impediment to measuring their rate of production. When hadrons appear among the final products of an annihilation, it is assumed they all derive ultimately from a single quark-antiquark pair. Hence the rate of quark production is given simply by the number of events in which any hadrons appear. In practice the absolute rate is not measured. It is more convenient to measure the ratio of hadron production to muon production, a ratio that by convention is designated by the letter *R*. If the quark hypothesis is valid, then *R* should be equal to the sum of the squares of the quark charges, and it should be approximately constant, independent of the collision energy.

The experimental measurement of *R* is a straightforward procedure. At the energies of the SPEAR storage ring, events in which hadrons appear almost always have three or more particles in the final state, whereas muons are produced strictly in pairs, which leave the interaction with equal but opposite momentum. An excellent estimate of *R* can therefore be obtained simply by adding up the number of events in which more than two particles are detected and dividing by the number of two-body, back-to-back events.

The ratio *R* is an important test of quark theories. Because *R* is determined by the sum of the squares of the quark charges it can discriminate between variants of the quark model having different numbers of quarks or different charge assignments.

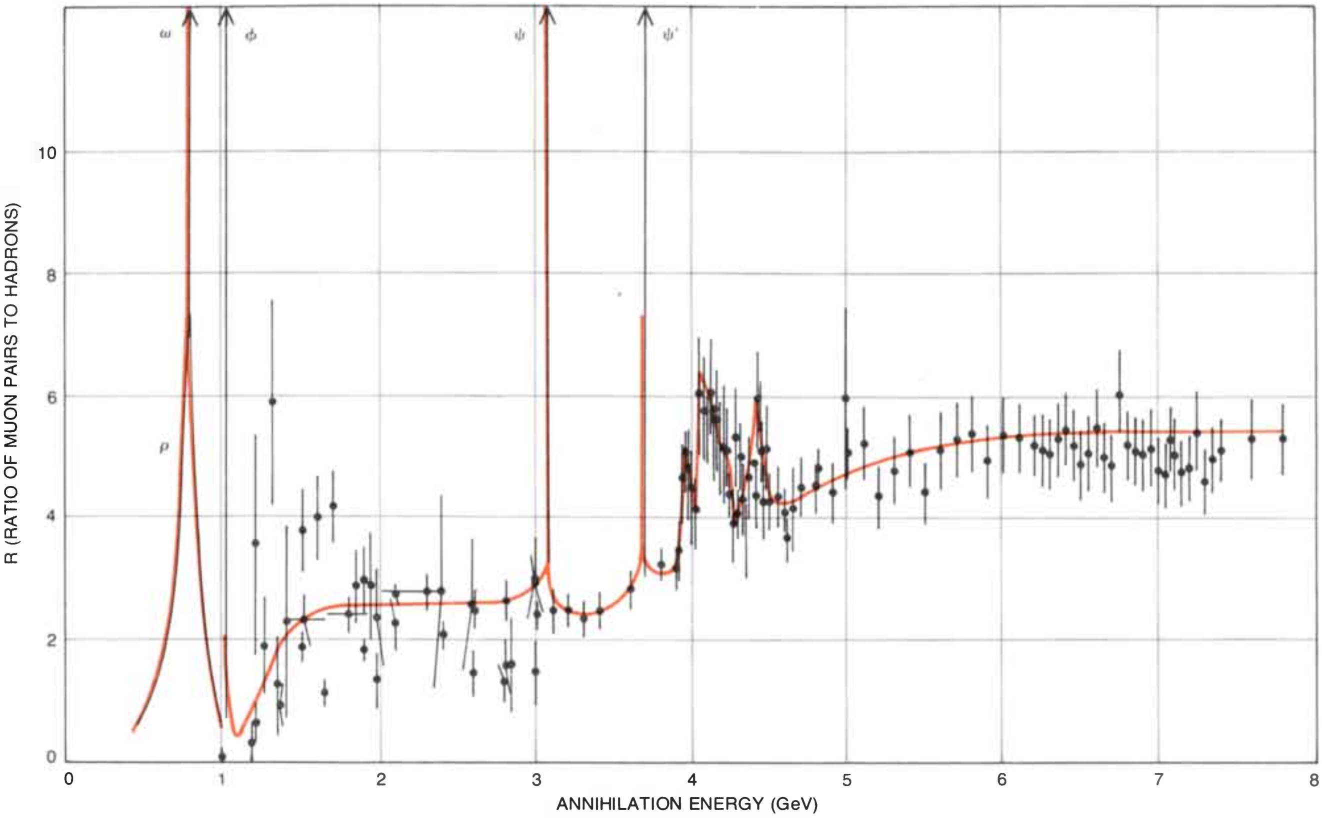
Experimentally determined values of *R* have an interesting history: they have consistently been too large for the fashionable theories of the day. In the original model of Gell-Mann and Zweig, with three quarks, *R* should have been equal to $(2/3)^2 + (1/3)^2 + (1/3)^2$, which adds up to 2/3. That value of *R* implies that for every two events involving hadrons there should have been three muon pairs produced. The first measurements of *R* were made with the storage ring called ADONE at Frascati in Italy. These measurements, all made at energies below 3 GeV, gave values not of 2/3 but of approximately 2; in other words, there were twice as many hadronic events as muon pairs.

Shortly thereafter theory briefly caught up with experiment. It was proposed, and it is now widely accepted, that each of the quarks must come in three varieties in order to accommodate a new quantum number, generally called color. Unless a quark can be isolated the color quantum number will remain permanently hidden from view. Color has proved to be an exceptionally fruitful concept in explaining the interactions of quarks, but it is a purely "internal" quantum number and it predicts no additional hadrons. Indeed, one of the few ways in which it impinges on experimental results is in the value of *R*. By tripling the number of quarks, *R* is also tripled—to a value of 2.

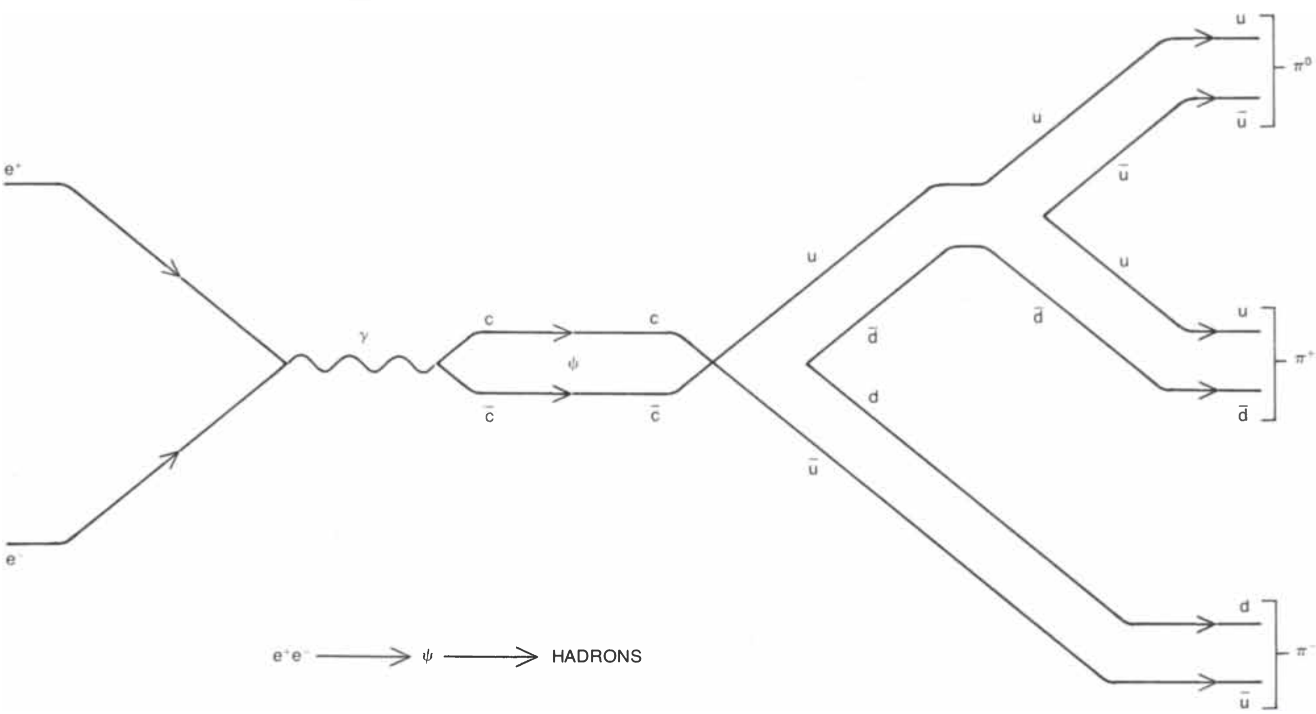
There was little time to gloat over this result. In the early 1970's two more tantalizing measurements of *R* were made at higher energy with a storage ring at the Cambridge Electron Accelerator. They showed that at 4 GeV *R* is 4, and at 5 GeV it rises almost to 6.

The Cambridge Electron Accelerator was shut down shortly after these measurements were made, but the initial program of experiments at SPEAR, beginning in the fall of 1973, confirmed and extended the results. *R* was measured at several energies separated by fairly wide intervals. Below about 3 GeV the value is 2.5, reasonably close to the value of 2 expected if there are three quarks and each comes in three colors. Above about 4 GeV, however, the ratio rises steeply, then above 5 GeV it levels off at a value of about 5.5. Clearly some new phenomenon must be making itself known. Even charm may not be sufficient to explain the behavior of the curve. Adding the charmed quark to the model (in three colors) raises the expected value of *R* only to 3.3.

These disconcerting results aroused



RATIO *R* **rises from about 2.5 at energies below 3 billion electron volts (3 GeV) to about 5.5 at energies above 5 GeV. The most plausible interpretation of the increase is that a threshold for the creation of a new kind of quark has been crossed in this interval. What is more remarkable than the overall increase in the ratio is that the curve has a series of tall and extraordinarily narrow peaks. The three peaks clustered near 1 GeV represent production of the rho (ρ), omega (ω) and phi (φ) mesons, which have quark compositions involving combinations of $u\bar{u}$, $d\bar{d}$ and $s\bar{s}$. The spike at 3.1 GeV is the psi meson (ψ), made up of a $c\bar{c}$ quark-antiquark pair. The psi was discovered in 1974 during a measurement of** ***R;*** **the psi prime (ψ′), which is an excited state of the same $c\bar{c}$ system, was found 10 days later. The broader peak at 4 GeV represents several short-lived particles and is now recognized as the threshold for the creation of a pair of charmed mesons.**



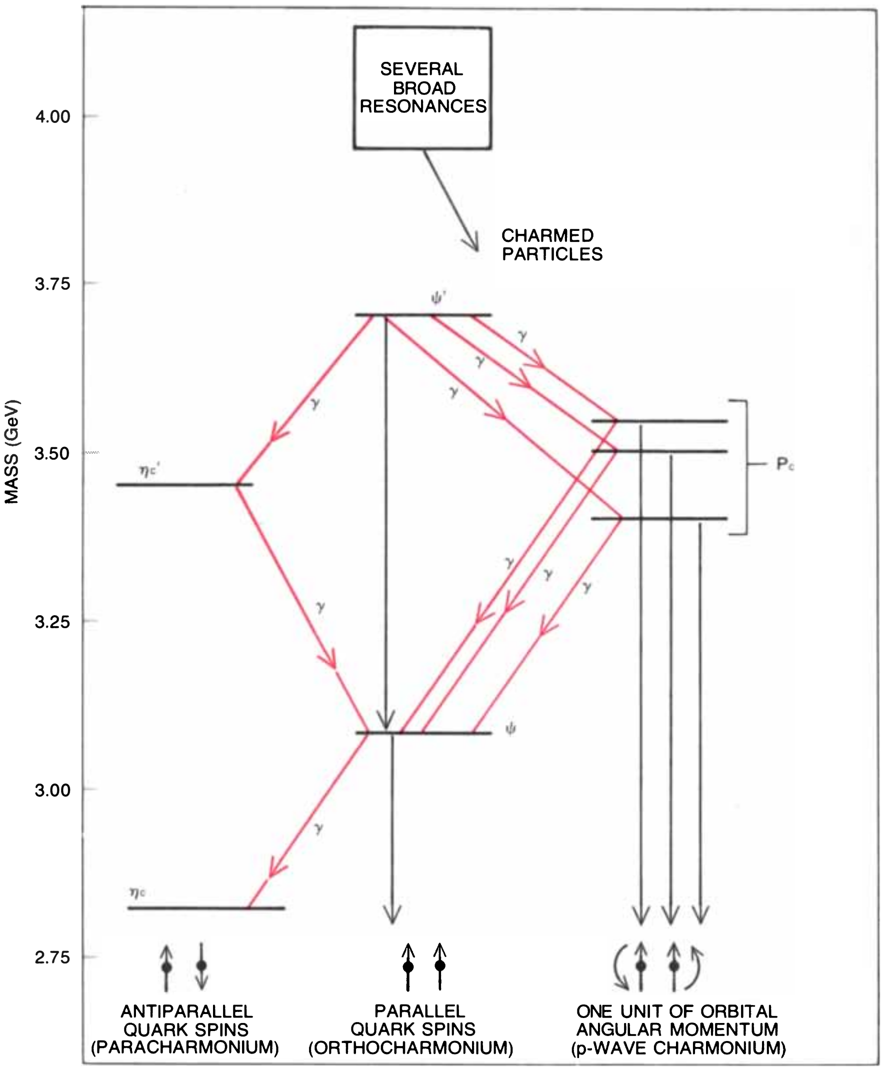
PSI PARTICLE can be created singly in electron-positron annihilations because it consists of a quark and an antiquark of the same type. The immediate product of the electron-positron collision is always a photon, but the photon invariably decays to other particles before it can be observed. The photon has one unit of spin angular momentum but all its other quantum numbers are zero; any particle created from the photon must have the same quantum numbers. The psi is one such particle; the rho, omega and phi mesons are others. The psi decays through the annihilation of the charmed quark and antiquark, giving rise to ordinary hadrons such as the three pions shown here.

much interest, but they were soon overshadowed by another feature discovered in the *R* curve: the peak that signaled the presence of the psi particle. In a graph of *R* the psi appears as a narrow spike of enormous height at an energy of 3.095 GeV. If the spike could be perfectly resolved, it would be seen to rise above the background level by a factor of 3,000; in other words, at 3.095 GeV hadrons are 3,000 times more likely to be created than they are at neighboring energies. Such an enhancement in hadron production is called a resonance. When the collision energy of the electron and positron exactly matches the mass of the psi particle, the probability is high that a psi will be created. The subsequent decay of the psi yields the hadrons that are detected as a resonance. At energies even slightly higher or lower than that of the resonance, energy and momentum can no longer be balanced to produce only the psi and the rate of hadron production declines.

Even more remarkable than the height of the psi resonance is its extreme narrowness, a property related to the lifetime of the particle. In fact, it is only from measurements of the width of the resonance that the lifetime can be deduced. Part of the observed width results from the finite resolution of the experimental apparatus. Even with perfect apparatus, however, the peak could not be confined to a single uniquely defined energy. The range of energies over which the peak is distributed can be interpreted in quantum mechanics as representing a residual uncertainty in the actual energy of the particle, and that uncertainty is related to the lifetime. The shorter the lifetime, the greater the uncertainty. It is as if there were not enough time before the particle decayed to make an accurate measurement of its energy. Most resonances in this mass region are very broad enhancements hundreds of millions of electron volts wide; the width of the psi resonance is 67,000 electron volts, or about .002 percent of its total energy.

The unexpected discovery of such a narrow peak in the graph of the ratio *R* prompted a change in the mode of operating storage rings. Instead of sampling hadron production at a few widely separated energies, high-resolution surveys were made with the smallest possible increment between sampled energies. By far the most successful high-resolution survey made at SPEAR was the very first one: 10 days after the discovery of the psi it disclosed a second, closely related particle with a mass of 3.684 GeV. This second particle was given the name psi prime (ψ'). Like the psi itself, psi prime is electrically neutral and has one unit of spin; it is not quite as long-lived as the psi, but the resonance width of 228,000 electron volts is still much narrower than most.

Further exploration of the *R* curve revealed a third feature in the vicinity of 4.1 GeV. This peak is much less distinct; it is an enhancement with a width of about 200 MeV. How it should be interpreted was not immediately apparent.



SPECTRUM OF PARTICLES that includes the psi consists of at least seven states. All of them are made up of a charmed quark and a charmed antiquark, but they are distinguished by various combinations of spin and orbital angular momentum. Only the psi and the psi prime can be created directly from electron-positron annihilations because only those states have the same quantum numbers as the photon. The quark spins in the psi are oriented in the same direction and they add to give the meson one unit of spin. Two states of somewhat lower energy, designated η_c and η'_c, are formed when the quark spins are antiparallel. Three states of intermediate mass, called *p*-wave excitations, are formed when a pair of quarks with parallel spins acquires a unit of orbital angular momentum. The bound system of a charmed quark and a charmed antiquark in many ways resembles a simple atom, and the name "charmonium" has been applied to all the states of this system. As in the spectrum of energy levels in an atom, one state can be transformed into another of lower energy through the emission of a photon.

A bound system of a charmed quark and a charmed antiquark has much in common with simple atoms such as the hydrogen atom. Just before the discovery of the psi, Thomas Appelquist and H. David Politzer of Harvard proposed the name "charmonium" for this bound system. The name implies an analogy with another exotic atomlike species, positronium, which is a bound state of an electron and a positron.

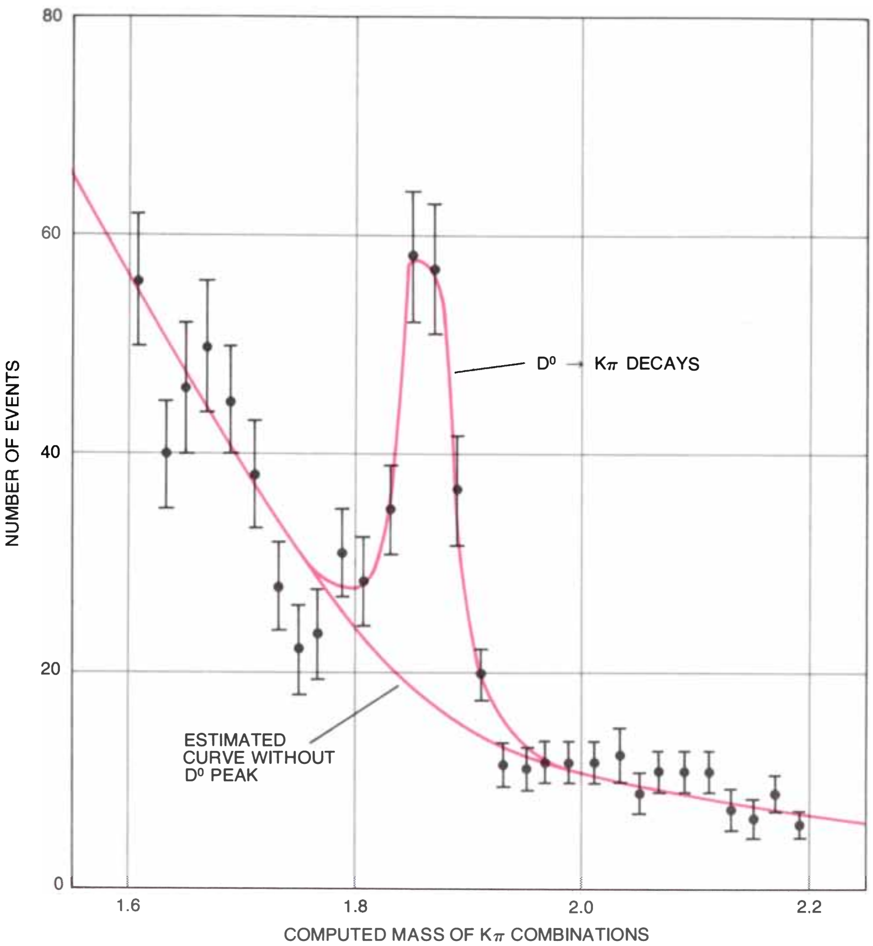
According to the charmonium model, the quark and the antiquark in the psi particle have their spins aligned in parallel, but they have no orbital angular momentum. The psi prime is an excited state of this system, analogous to one of the excited states of a hydrogen atom. In this state too the quark spins are parallel and there is no orbital angular momentum. The parallel quark spins combine to give each meson one unit of spin. This arrangement of spins is known in atomic physics as the "ortho" configuration, and the psi and psi prime particles can together be called orthocharmonium.

Continuing to argue from the principles of atomic physics, Appelquist and Politzer were able to predict additional forms of charmonium. For example, there should be states without orbital angular momentum and with the quark spins antiparallel. The oppositely directed spins are subtracted, with the result that the particle as a whole has zero angular momentum. This configuration is denoted by the prefix "para," and it has a lower overall energy than the ortho configuration. Thus there should be a paracharmonium state with a mass slightly less than that of the psi and another with a mass slightly less than that of the psi prime. The quark and the antiquark should also combine with parallel spins and with one unit of orbital angular momentum. Three such states are expected, known collectively as *p*-wave charmonium (another term borrowed from atomic physics). Their masses should all lie between that of the psi and that of the psi prime.

Because the quantum numbers of the photon must be conserved, electron-positron annihilations can give rise to quark-antiquark pairs only in the ortho configuration, and so neither paracharmonium nor *p*-wave charmonium can be observed as peaks in the ratio *R*. It should be possible to produce them, however, through electromagnetic transitions from the psi and psi prime states. Once again these transitions are analogous to events in atomic physics, where an atom can move from an excited state to a state of lesser energy by emitting a photon. In the atomic transitions the photons emitted are typically in the visible portion of the electromagnetic spectrum; in the transitions between charmonium states the photons are high-energy gamma rays.

The most straightforward approach to finding these states is to produce psi prime particles copiously, then search among their decay products for telltale gamma rays. Such experiments were undertaken at SPEAR and at the storage ring called DORIS associated with the German Electron Synchrotron near Hamburg.

The main difficulty in recognizing the gamma rays is in distinguishing them from numerous other photons emitted by unrelated processes. One aid to discrimination is that all the photons emitted during transitions between two given states have the same energy, so that they should stand out above the background events, which have a wide range of energies. In the initial experiments the background was further reduced by selecting for examination only those events that seemed to conform to a particular sequence of decays. We decided to study events in which a psi prime particle emits a photon and decays to some intermediate state, which then emits another photon and yields a psi particle. We were thus able to exclude from consideration all events except those in which the presence of a psi could be deduced and in which exactly two gamma rays were present. Such procedures for



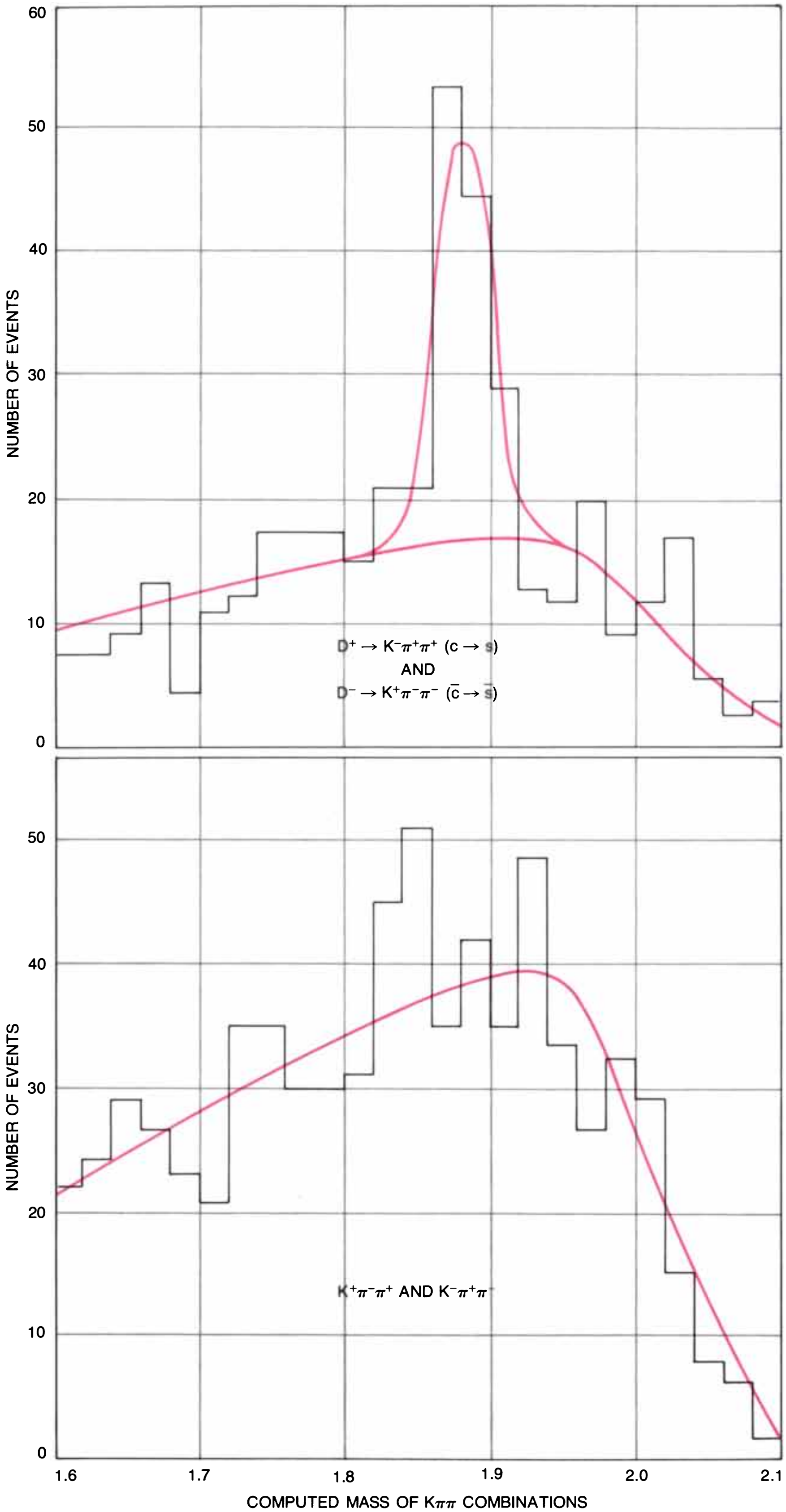
FIRST CHARMED PARTICLE to be discovered was the D^0, composed of a *c* quark and a $\bar{u}$ antiquark. Unlike the psi, the D^0 has overt charm: the quantum numbers of the *c* quark are not canceled by those of a $\bar{c}$ antiquark in the same hadron. As the lightest charmed meson, the D^0 can decay only through the weak interaction, which can transform the charmed quark into a quark of another type. Because of the special relation that exists between strangeness and charm, the charmed quark is usually converted into a "strange" quark; consequently strange hadrons (such as *K* mesons) are expected among the decay products. The D^0 was discovered as a bump in the energy distribution of *K* mesons and pions that have a net electric charge of zero ($K^+\pi^-$ or $K^-\pi^+$). Such combinations can be produced by many unrelated processes, but their number should decline smoothly as the mass of the pair increases. The presence of a bump at a given energy indicates that a particle with that mass is decaying into the $K\pi$ combination.

"cutting" data are common in high-energy physics; they can reduce hundreds of thousands of recorded interactions to a handful of interesting events.

The first results were reported by a group of experimenters working at DORIS; they found 14 events that met the criteria. Shortly thereafter our group at SPEAR contributed several more candidates. These few events retained one important item of information that was not a mere artifact of the selection process, namely the energies of the two photons. In most of the events one of the gamma rays had an energy near 170 MeV; by subtracting this energy from the mass of the psi prime we deduced that the psi prime was decaying at least part of the time into an intermediate state with a mass near 3.5 GeV. The intermediate state subsequently decays to yield a psi particle. The 3.5-GeV particle has turned out to be one of the states of *p*-wave charmonium, those with a unit of orbital angular momentum. The group at DORIS named it P_c. Since then more examples of this so-called cascade decay have been collected at both SPEAR and DORIS and the two related *p*-wave states, which have slightly different masses, have been found.

A second major search for gamma-ray transitions was undertaken by our group at SPEAR. We sought evidence for psi prime decays in which the final state included a particular ensemble of hadrons and a single gamma ray. These decay products would be the signature of an intermediate charmonium state that decayed not to a psi particle but to ordinary hadrons. We were gratified to discover that the photons again clustered at certain discrete energies, two of which coincided with energies observed in the cascade decays.

Within a year of the discovery of the psi particle a spectrum of closely related



D^+ MESON provides the strongest evidence that the new particles being observed have the property of charm. The charmed quark in the D^+ must decay into a strange *s* quark, which is a constituent of the K^- meson. The antiparticle of the D^+, the D^-, contains a charmed antiquark that must give rise to a strange antiquark, $\bar{s}$, found in the K^+ meson. Combinations of *K* mesons and pions consistent with this decay mode (*top*) show a distinct bump at an energy near 1.87 GeV. What is equally important, combinations that would arise from the forbidden transition from charmed quark to strange antiquark show no enhancement at that energy (*bottom*).

states had emerged. These included not only the *p*-wave excitations but also the paracharmonium state just below the psi prime. Two groups of experimenters working at DORIS also obtained evidence for the paracharmonium state below the psi itself; this is the lowest-mass state of charmonium. Finally, in 1976, most of the peaks in the gamma-ray spectrum of the psi prime were observed in a single measurement, made by physicists from five institutions: the University of Maryland, the University of California at San Diego, the University of Pavia, Princeton University and Stanford University. This gamma-ray spectrum is like the line spectrum of an ordinary atom, with one spectacular difference: the energy of a typical photon in an atomic transition is usually a few electron volts, whereas the gamma rays emitted in charmonium transitions have energies of several hundred million electron volts.

The spectrum of charmonium states represents strong evidence that the psi and its relatives consist of a particle bound up with its own antiparticle, but there is no assurance that the quantum number carried by those particles is charm. The property itself is hidden in all the charmonium states. It becomes manifest only in hadrons that include a charmed quark or antiquark in combination with uncharmed quarks.

In electron-positron collisions such charmed hadrons can be created only in pairs. Hence the lowest annihilation energy at which charm could appear is equal to twice the mass of the lightest charmed particle. This threshold energy can be estimated from an examination of the ratio *R*. A lower bound is the mass of the psi prime: if the threshold were below this mass, then the psi prime itself would decay rapidly to yield a pair of charmed particles and the resonance would be much broader than it is. Just above the psi prime mass a significant increase in *R* with several broad peaks suggests the creation of highly excited states of charmonium that do decay strongly to particles with nonzero charm. Thus the threshold appears to lie between 3.7 GeV and roughly 4 GeV, implying that the lightest charmed particle has a mass between 1.85 GeV and about 2 GeV.

Like massive strange particles, massive charmed ones could decay strongly while conserving charm by passing on the *c* quark or *c* antiquark to a lighter charmed state. The lightest charmed particles, however, should decay only through weak interactions, and they should therefore be comparatively long-lived. The lifetimes might be on the order of 10^{-13} second.

The lightest charmed states were expected to comprise six mesons, three of them made by combining a charmed *c* quark with a $\bar{u}$, a $\bar{d}$ or an $\bar{s}$ antiquark, the other three by combining a $\bar{c}$ antiquark with a u, a d or an s quark. The mesons containing a charmed quark have been named respectively D^+, D^0 and F^+; the corresponding antiparticles are D^-, $\bar{D}^0$ and F^-.

How could charmed particles be recognized when they were found? Even if long-lived mesons with the appropriate masses were discovered, how could it be established that they are distinguished by charm rather than by some other new property of matter? The answers to these questions lie in the special relation that must exist between the charmed quark and the strange quark if the suppression of strangeness-changing neutral weak currents is to be explained. Because of this relation weakly decaying charmed particles should usually include strange particles among their decay products. K mesons, the strange particles of lowest mass, serve as a distinctive signature of events involving the decay of charmed particles.

These properties of charmed-particle decays suggest an experimental procedure for finding charm. That procedure is to examine electron-positron collisions above the psi prime mass for multiparticle events (which usually contain hadrons) in which there are long-lived particles that include K mesons among their decay products.

The usual method for conducting such searches is to hypothesize a particular mode of decay, then to select from a large sample of events all those that might be examples of the presumed mode. For example, a reasonable hypothesis is that some charmed particles might sometimes decay into a K meson and two pions, and so all events that included these three particles would be selected. From measurements of the momentum of the decay products in each event the mass of the parent particle can be calculated. If the particles detected are in fact the products of a single mode of decay, then the majority of the computed masses will be clustered at a single value. If the selected particles are present through mere coincidence, on the other hand, then the computed masses will have a broad and featureless distribution. This procedure, which is quite common in high-energy physics, is called "bump hunting."

Our first attempts at bump-hunting at annihilation energies near 4 GeV were unsuccessful. The absence of a perceptible signal was discouraging, but it was not fatal to the charm hypothesis. Charmed mesons might decay through many different modes, so that no one mode is represented by a large number of events.

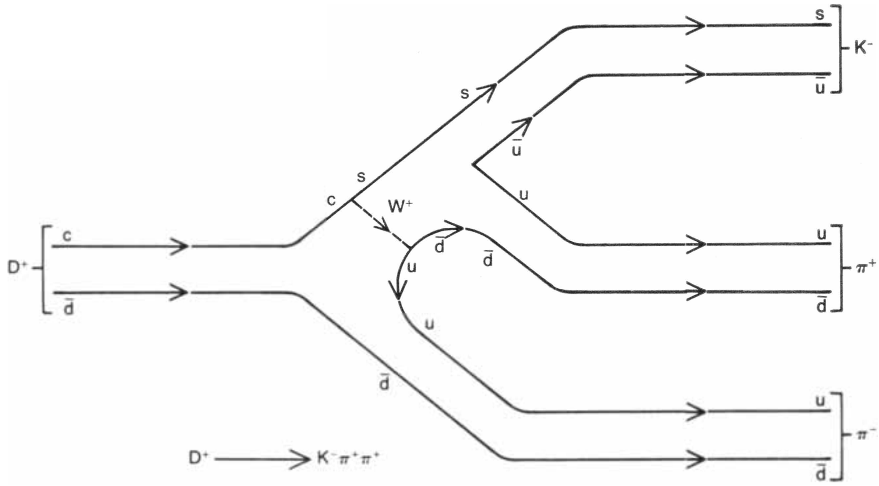
Meanwhile circumstantial evidence of charmed-particle production began to appear in experiments involving quite different techniques. In these experiments events were initiated by high-energy collisions between neutrinos and nuclear matter; because the neutrino interactions are necessarily mediated by the weak force, charmed particles might be created singly rather than in pairs. At Brookhaven and at the European Organization for Nuclear Research (CERN) a few bubble-chamber photographs showed events that could be interpreted as decays of charmed particles. At the Fermi National Accelerator Laboratory (Fermilab) an array of particle counters registered numerous candidate events distinguished by the presence of two leptons in the final state. The pairs of leptons were strong but nevertheless inconclusive evidence for the decay of a charmed particle with a mass in the vicinity of 2 GeV. With the exception of a single bubble-chamber photograph made at Brookhaven, none of these events provided enough information to precisely determine the mass of the parent particle.

At SPEAR we continued to accumulate data while studying the broad features in R at energies near 4 GeV. In the spring of 1976, with improved techniques for identifying strange particles and with a much larger sample of events, a bump was finally recognized. Gerson Goldhaber and François Pierre of the Lawrence Berkeley Laboratory found small but significant peaks at an energy of 1.863 GeV in two decay modes leading to the emission of K mesons and pions.

One class of events had in the final state a single K meson and a single pion with opposite charges, that is, $K^+\pi^-$ or $K^-\pi^+$. The other class of events had a K meson accompanied by three pions with the charge assignments $K^+\pi^-\pi^+\pi^-$ or $K^-\pi^+\pi^+\pi^-$. All these combinations of particles have a net electric charge of zero, and they all seem to derive from the decay of the same object. The most plausible candidate for that object is the electrically neutral D^0 meson, made up of a c quark and a $\bar{u}$ antiquark.

Additional bumps have since been discovered. The most important of them results from the decay of the charged relative of the D^0, the D^+ meson. The evidence for this state is an enhancement, at an energy 1.868 GeV, in the number of events yielding the combination of particles $K^-\pi^+\pi^+$. Significantly, the decay mode $K^+\pi^-\pi^+$, which has the same net electric charge of +1, has no corresponding peak at the same energy. This subtle distinction between charge states has a simple and elegant explanation in the charm theory, and it represents the best available evidence for the existence of charm. The D^+ meson consists of a c quark bound to a $\bar{d}$ antiquark. The relation between charm and strangeness specifies that a c quark can decay through the weak interaction to yield a strange s quark, but it cannot be converted into an $\bar{s}$ antiquark. The K^- meson contains an s quark, and so it can be produced in the decay of the D^+; the K^+ meson, on the other hand, contains an $\bar{s}$ antiquark, and as theory requires it is not observed in D^+ decays. No competing model of the new particles offers such definitive predictions, and none is so consistently in agreement with experimental findings.

In electron-positron annihilations charmed particles are invariably created in pairs, but the bumps by which they are identified represent the decay products of only one member of the pair. It is possible, however, to deduce some of the properties of the missing partner. Because the detected particle is recoiling from the one that escapes detection, both states must have the same momentum. From the known momentum and the known total energy of the system of particles (which is equal to the annihilation energy of the electron and



DECAY OF THE D^+ MESON is mediated by a weak interaction between quarks. Whereas the strong force can only create or annihilate matched pairs of quarks and antiquarks, the weak force can change one kind of quark into another. In the decay of the D^+ the weak force is transmitted by a particle called the W^+. The c quark emits a W^+ and is transformed into an s quark; the W^+ then decomposes into a u quark and a $\bar{d}$ antiquark. Analysis of the interaction shows that neither charm nor strangeness is conserved, although electric charge is conserved.

positron) the rest mass of the escaping particle can be determined.

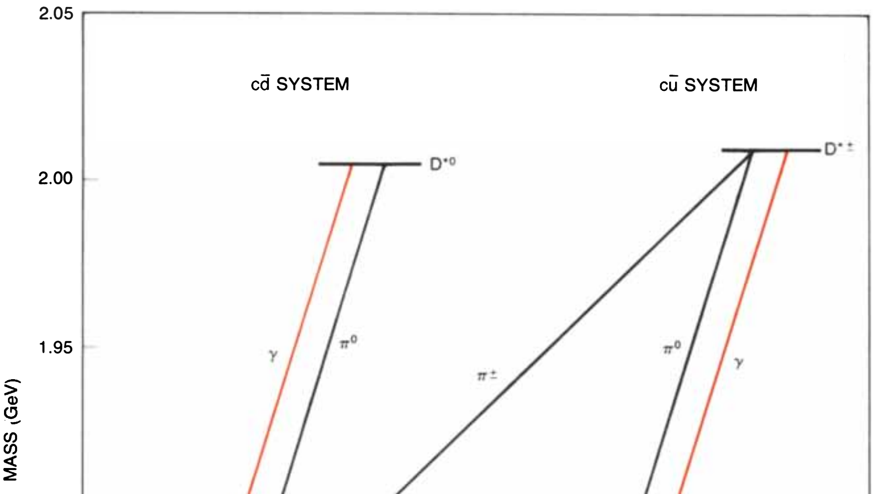
When the events involving D^0 and D^+ mesons were analyzed in this way, we found that in some cases the recoiling objects were simple antiparticles, $\overline{D}^0$ and D^-, with masses identical to those of the detected particles, about 1.87 GeV. A more prominent peak in the recoiling-mass spectrum was observed near 2.01 GeV. The peak is interpreted as evidence for excited states of the mesons, labeled D^{*0} and D^{*+}. If this interpretation is correct, the lowest-energy states (D^0 and D^+) are made up of quarks with their spins oriented antiparallel, so that the meson itself has zero angular momentum, whereas in the excited states parallel quark spins give the mesons one unit of angular momentum. Various angular correlations between the particles and their decay products depend on the spins, so that this hypothesis can be tested. Preliminary measurements of those correlations are in agreement with the model.

Once the mass of the psi had been established, psi particles could be produced at will and in large numbers. The more recent experimental successes have not been won as easily. States of matter with overt charm have proved remarkably difficult to observe. Even after three years of intensive searching only a few hundred events have been recorded in which all the decay products of a charmed particle are identified. All but a handful of those events are decays of charmed mesons. There must also be an extensive family of charmed baryons, made up of three quarks. A few events offering glimpses of charmed baryons have been recorded at Fermilab, but most properties of the particles remain to be determined.

The most elusive charmed particle is another meson, the F^+. It is made up of a c quark and an $\bar{s}$ antiquark, and so it is both charmed and strange. Several searches for the F^+ are under way. Although there is no conclusive evidence for its existence neither is there much doubt it will eventually be discovered.

Even if all the evidence in the case for charm has not yet been presented, there can no longer be any doubt about the verdict. The charmed quark began with an abstract symmetry principle relating quarks and leptons, a principle that can almost be reduced to a longing for tidiness. At the outset the hypothesis was proposed without the support of a single experimental observation; today a great many experiments point to a conclusion that is all but incontestable. An entire spectrum of particles testifies to the existence of the charmed quark: the psi and the other states of charmonium and the charmed mesons D^0 and D^+. The most compelling evidence is also the subtlest: it is the special relation between strangeness and charm manifested in the decay of the D^+ to the three mesons $K^-\pi^+\pi^+$ but not to $K^+\pi^-\pi^+$. If this is not charm, it is a superlative counterfeit.

Perhaps the most significant result of the discovery of charm is an improved understanding of ordinary matter. It has often been said that high-energy physics needs an equivalent of the hydrogen atom, the comparatively simple system of bound particles that served as a constant point of reference for the development of quantum mechanics 50 years ago. In a sense, of course, the quark model makes simple atomlike systems of all hadrons, but that structure was never very obvious in experiments until the discovery of the psi particle. Now the sharp transitions between charmonium states, with the emission of photons, provide an exact analogy to atomic spectroscopy, but at energies a million times greater. The discovery of charm marks the beginning of quark chemistry.



SPECTRUM OF CHARMED MESONS that have been observed in electron-positron collisions includes four states and their antiparticles. The states of lowest mass have antiparallel quark spins and a total spin of zero; the excited D^* states are made up of the same quarks but have the quark spins parallel, giving the mesons a total spin of 1. The one missing charmed meson, which would also have excited states, is the F^+. This meson is made up of a c quark and an $\bar{s}$ antiquark and thus is simultaneously charmed and strange. There is some evidence that the F^+ may have been seen in a European experiment but the evidence is not conclusive.

The case for charmed quarks may by now have been proved, but it would be entirely misleading to suggest that it ought to be closed. One conspicuous problem is the value of the ratio R. The most plausible quark model predicts a value of 3.3, but the observed ratio is nearer 5.5. Once again R seems to be much too large. Furthermore, there is a growing number of anomalous events, observed both at SPEAR and at DORIS, that are not readily explained by present versions of the quark model, even with the inclusion of charm. The events are characterized by the presence of two charged particles, often leptons, and undetected neutral particles, which may be neutrinos.

At the moment the hypothesis that seems best able to account for the anomalous events and for most of the excess in R is that a fifth member of the lepton family is appearing. It would have a mass of roughly 2 GeV (or 4,000 times the electron mass) and might be accompanied by still another new lepton, a neutrino. Of course, if there are additional leptons, the appealing symmetry between leptons and quarks would be destroyed. The obvious solution is to postulate more quarks, and preliminary evidence for particles containing a fifth kind of quark has now been reported in high-energy neutrino interactions at Fermilab. Further experiments will be needed to confirm that discovery, but it appears that the future of elementary-particle physics is at least as bright as the recent past.

If it takes high-speed movies to explain what's going on talk to some of the following:

Cordin
2230 South 3270 West
Salt Lake City, Utah 84119
(801) 972-5272/Telex 38-8395

Instrumentation Marketing Corporation
820 South Mariposa Street
Burbank, Calif. 91506
(213) 849-6251/Telex 67-3205

LW International
6416 Variel Avenue
Woodland Hills, Calif. 91367
(213) 348-8614

Marco Scientific, Inc.
1031H E. Duane Avenue
Sunnyvale, Calif. 94086
(408) 739-9418/Telex 35-7454

Mitchell Camera Corporation
11630 Tuxford Street
Sun Valley, Calif. 91352
(213) 768-6400

Photo-Kinetics, Inc.
10 Commerce Street
Poughkeepsie, N.Y. 12603
(914) 473-3840

Photo-Sonics, Inc.
820 South Mariposa Street
Burbank, Calif. 91506
(213) 849-6251/Telex 67-3205

Redlake Corporation
1711 Dell Avenue
Campbell, Calif. 95008
(408) 866-1010/
TWX: 910-590-2594

Teledyne Camera Systems
131 North 5th Avenue
Arcadia, Calif. 91006
(213) 359-6691

Visual Data Systems
5617 West 63rd Place
Chicago, Ill. 60638
(312) 585-3060

Visual Instrumentation Corporation
903 North Victory Blvd.
Burbank, Calif. 91501
(213) 849-7333/ Telex 68-6432

We can supply the right film.

Kodak

Early Schlumberger loggingoperation in Pechelbronn, France.

**50 years ago,
Conrad and Marcel Schlumberger
started answering
the energy problems of today.**

SCIENCE AND THE CITIZEN

Proliferating Quarks

At first there were three kinds of quark, called "up," "down" and "strange." They were proposed in 1963 as the hypothetical constituents of protons and neutrons and of the many related particles known collectively as hadrons. A year later a fourth kind of quark was added to bear the proposed new property of matter called charm. Perhaps the most remarkable development in high-energy physics in the past few years has been the gradual accumulation of evidence that the charmed quark does exist (see "Fundamental Particles with Charm," by Roy F. Schwitters, page 56). The study of the hadrons formed by these two pairs of quarks—up and down, strange and charmed—has hardly begun, but already an experimental result suggests there may be still more quarks. The result consists of evidence for a new hadron, by far the most massive subnuclear particle yet observed.

The new hadron was found in an experiment with a beam of high-energy protons at the Fermi National Accelerator Laboratory (Fermilab). The experiment was conducted by a group of 16 physicists from Columbia University, the State University of New York at Stony Brook and Fermilab, led by Leon M. Lederman of Columbia. They report their results in *Physical Review Letters*.

In the experiment protons accelerated to an energy of 400 billion electron volts (400 GeV) were allowed to collide with atomic nuclei in a target made of copper or platinum. The debris resulting from the collisions was analyzed in an apparatus called a double-arm spectrometer. Particles emitted in a narrow range of angles on each side of the proton beam were detected, and by measuring their deflection in a magnetic field their momenta were determined.

Lederman and his colleagues examined a specific class of events: those in which the products of the collision included a pair of muons with opposite electric charge and with an effective combined mass of between 5 GeV and 14 GeV. (The muon is a particle that is not made up of quarks but is closely related to the electron.) In some 10^{16} proton-nucleus collisions about 9,000 muon pairs in the selected energy range were detected.

At high energy, muon pairs can be created by a great variety of processes, but no single energy should be favored in their production. If the number of muon pairs is plotted as a function of the muons' effective mass, a smooth curve should result. When Lederman and his colleagues constructed such a graph for the events they had detected, the curve was smooth, as had been expected, with the exception of one distinctive feature: a bump at a mass of about 9.5 GeV. The bump represents strong evidence for the existence of a particle with that mass which sometimes decays into a pair of muons with opposite charges. The particle has been assigned the name upsilon. Its mass of 9.5 GeV is roughly nine times the mass of the proton.

The nature of the upsilon is still quite uncertain, but an interpretation that appeals to many physicists explains the particle as a bound system of a new quark and a corresponding antiquark. Since the more familiar quarks come in pairs, it is expected that there are actually two new quarks, even though the upsilon represents evidence for only one of them. The new quarks will apparently be called "top" and "bottom," the names being meant to suggest properties surpassing those of the up and down quarks found in ordinary matter. If the two new quarks do exist, there must also be two new properties of matter, which some physicists have taken to calling "truth" and "beauty."

In the original Fermilab experiment about 350 muon pairs were detected in excess of the number expected if no new particle existed. The number of events has since been increased to about 1,000. Subsequent analysis of the bump in the mass spectrum suggests strongly that it represents not one particle but two, with masses too close for the apparatus to resolve them clearly. Both are thought to consist of a bottom quark bound to a bottom antiquark, the particle of higher mass being an excited state of the lower-mass one. There is no sign yet of particles containing top quarks.

It is notable that the quark model was introduced in order to simplify the observed spectrum of hadrons, which seemed to be growing without limit. Now it is the quarks that seem to be proliferating.

Gossamer Wings

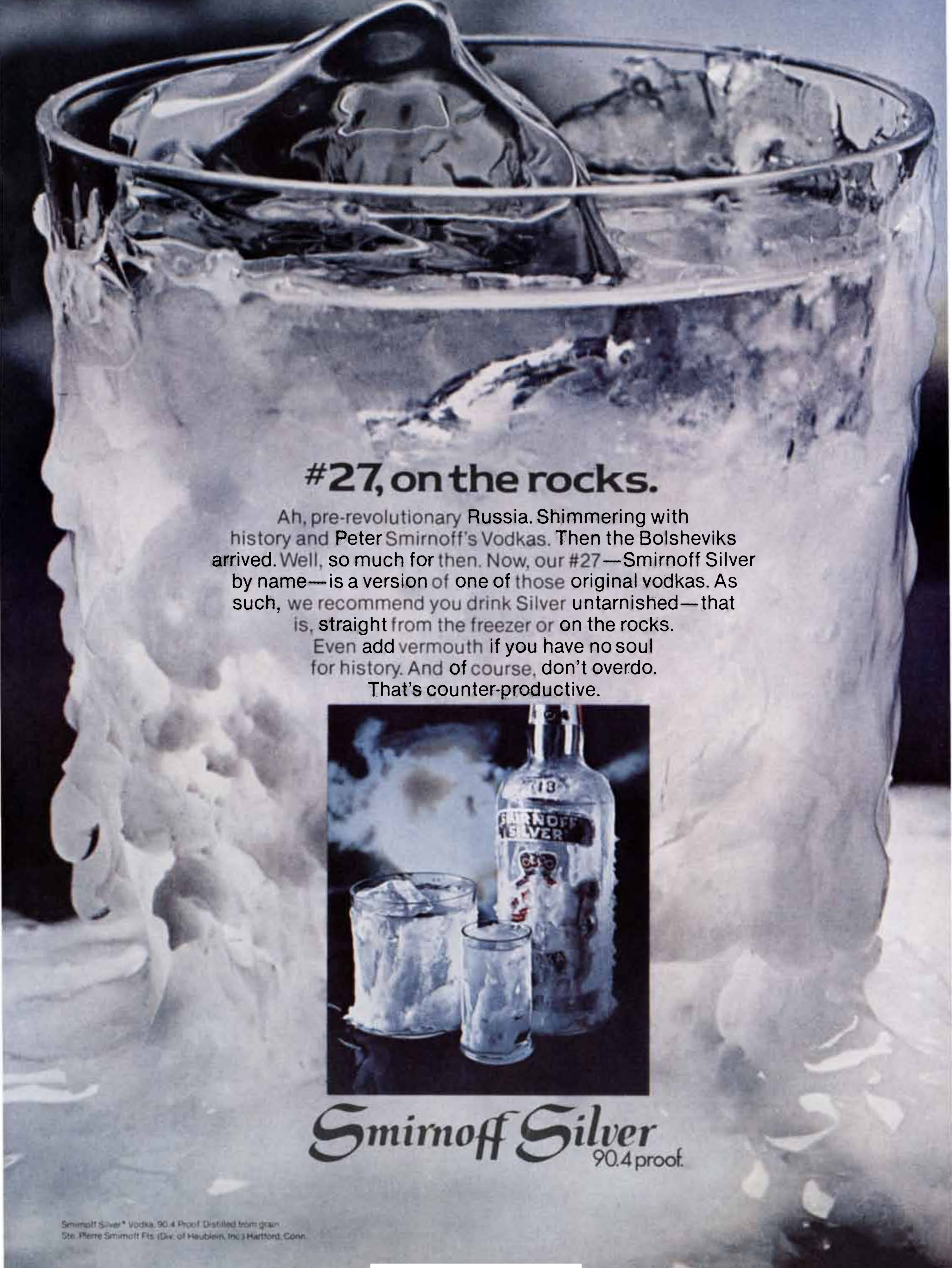
For a man to fly like a bird under his own power is the oldest dream of aeronautics, going back at least to the Greek myth of the inventor Daedalus, who escaped from the Labyrinth on wings covered with feathers held on by wax. The modern era of man-powered flight began in 1935, when two British engineers developed a propeller-driven "flying bicycle" that traveled through the air for short distances. In the 1950's a British industrialist, Henry Kremer, offered under the auspices of the Royal Aeronautical Society a substantial prize (now worth about $86,000) for the first man-powered flight to meet certain demanding conditions: the craft had to take off unassisted from level ground in a wind of 10 miles per hour or less, fly in a figure-eight pattern around two pylons half a mile apart and pass over a 10-foot hurdle at the start and finish. The use of stored power or buoyant gases was forbidden, as were any stops or assistance along the way.

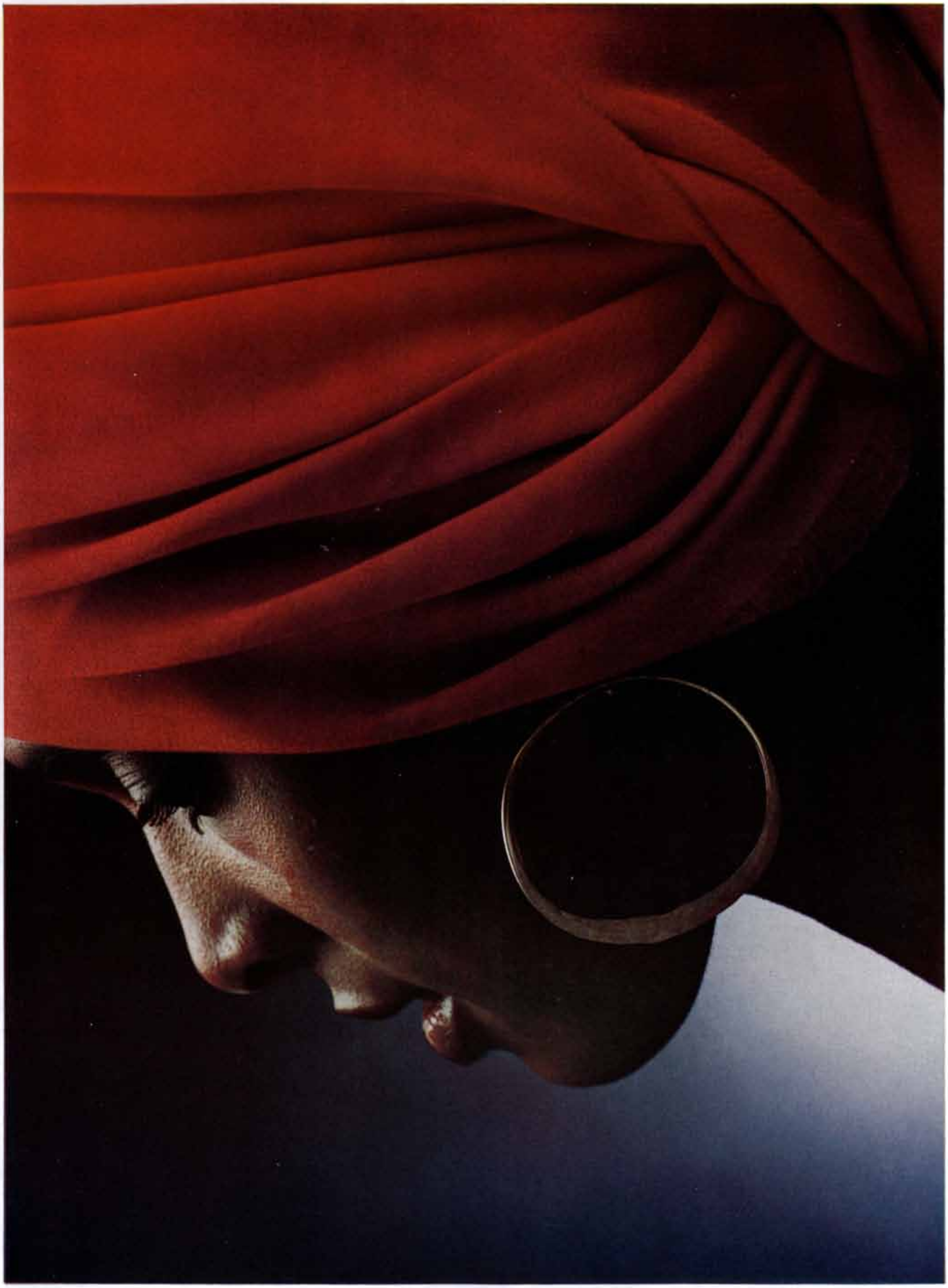
Over the years Kremer's challenge gave rise to a series of delicate, broad-winged flying machines. Several were able to make extended straight flights, but none could negotiate the course; the turns proved too difficult from the standpoint both of maneuverability and of power. Now a new type of aircraft developed by Paul B. MacCready, president of AeroVironment, Inc., of Pasadena, Calif., has finally succeeded where others had failed.

MacCready's craft, the *Gossamer Condor,* is aptly named: it is 30 feet long and eight feet high, has a 96-foot wingspan and a total lifting area of 835 square feet and yet weighs only 70 pounds. The disproportionate size of the wing is necessary to get maximum lift at low speed. The craft is propelled by pedaling a bicycle chain connected to a propeller behind the cockpit, which hangs from the center of the wing. A stabilizer mounted on a thin boom extending to the front of the wing and fitted with two small ailerons helps to control pitch and yaw and dispenses with the need for a vertical control surface. The craft is maneuvered into a turn by directly twisting the wing with the aid of a lever, which can be set at a predetermined turn radius. Once the twist of the wing is set, the stabilizer can still be manipulated to maintain control during the turn.

The main difference between the *Gossamer Condor* and preceding man-powered aircraft is that it was inspired by hang-glider technology. These lightweight gliders gave MacCready the idea of using an aluminum-tube skeleton braced with piano wire. Says MacCready: "The fact that the basic structure was very easy to build, modify and repair made an integrated development program possible. We could build a new wing in two weeks, and a new stabilizer in three days." The craft's design, he says, "is a mixture of the simplest and the most advanced aeronautics technology." Although the airfoil and the propeller were designed with the aid of a computer by AeroVironment's vice-president, Peter P. S. Lissaman, no plans for the craft were ever drawn up. Instead the design evolved over a period of a year, during which 12 successive models were built and improved on.

No custom-made materials were necessary. The fuselage was built from balsa wood, corrugated cardboard (which also forms the leading edge of the wing),


#27, on the rocks.
Ah, pre-revolutionary Russia. Shimmering with history and Peter Smirnoff's Vodkas. Then the Bolsheviks arrived. Well, so much for then. Now, our #27—Smirnoff Silver by name—is a version of one of those original vodkas. As such, we recommend you drink Silver untarnished—that is, straight from the freezer or on the rocks. Even add vermouth if you have no soul for history. And of course, don't overdo. That's counter-productive.
Smirnoff Silver
90.4 proof.
Smirnoff Silver* Vodka. 90.4 Proof. Distilled from grain.
Ste. Pierre Smirnoff Fls. (Div. of Heublein, Inc.) Hartford, Conn.



Last year, Technics gave you everything you wanted in direct drive. This year we're giving you less.

Introducing three new Technics turntables: The SL-1600 automatic, the SL-1700 semi-automatic and the SL-1800 manual. All with the Technics direct-drive system. The system FM stations use and discos abuse. And all with performance specs even better than last year's.

Less Acoustic Feedback.

This year we did more to give you less acoustic feedback. With Technics double-isolated suspension system. A floating system that not only damps out external vibrations at the base but also from the platter and tonearm. Combined with our new sensitive gimbal-suspension tonearm, it dramatically reduces feedback. Even at high music levels.

Less Wow & Flutter.

Amplified, even the smallest fluctuations in platter speed can be your biggest problem. So this year with a B-FG speed stabilizer built into our 321-element one-chip IC, you get infinitesimal wow and flutter, equal to Technics professional level. An incredible 0.025% WRMS.

Less Rumble.

The less noise your turntable makes, the more music you hear. That's why professionals loved Technics last year. This year, we're even more lovable, because there's even less rumble, −73 dB (DIN B) to be exact.

For Less Money.

Even with features like oil-damped cueing. A computer-analyzed S-shaped universal tonearm with anti-skate control. Pitch controls variable by 10%. And detachable dust cover. Technics turntables cost less this year than last.

Technics. We're giving you less. And that means you're getting more.

Technics
by Panasonic

sheets of Styrofoam and a thin skin of Mylar. Obviously every attempt was made to keep weight to a minimum. The project called for an investment of about $25,000 and thousands of man-hours in unpaid time. MacCready and his collaborators first built small-scale models to check the bracing and construction techniques. Then they went to a version with an 88-foot wingspan but no propeller. Finally they built a full-sized version complete with a propeller and a bicycle seat. To pilot the craft MacCready hired Bryan Allen, a 24-year-old championship-level bicycle racer who is also an experienced hang-glider pilot.

The craft was tested in the Mojave Desert. At that time it was discovered that the ends of the wings were too broad to allow for reasonable maneuverability at the craft's low speed (seven to 11 m.p.h.). When the ends of the wings were made narrower, it became necessary to further reduce drag by enclosing the pilot in a streamlined Mylar cockpit. With the pilot in place the finished craft weighs 207 pounds and requires about .35 horsepower to fly in a straight line. Says MacCready: "It is the absolute minimum airplane."

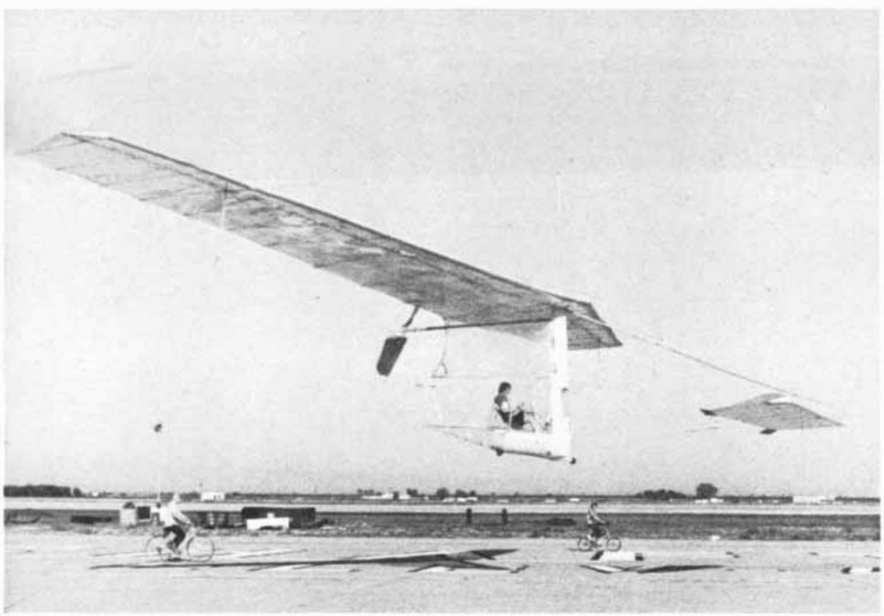
In February, MacCready and his collaborators moved the plane to the Kern County Airport in Shafter, Calif. The airport was chosen because it is relatively deserted, has a large hangar and has lighter winds than most airports in southern California. On August 23, after some final modifications on the fuselage, MacCready decided to try for the Kremer prize. His confidence was bolstered by a "confluence of beneficial factors": Allen was trained for his best effort, the wind conditions were satisfactory and Bob Richardson of the Kern County Department of Airports, who had been approved by the Royal Aeronautical Society to verify a successful flight, was present.

As Allen pedaled powerfully the *Gossamer Condor* accelerated down the runway for 30 feet and took slowly to the air. Moving at about 11 m.p.h., it traveled the specified figure-eight pattern around two markers half a mile apart, although because of the large wingspan and low speed of the craft the turns had to be wide and flat. When the successful flight was completed, the craft had covered from takeoff to landing a distance of 1.3 miles in seven minutes 22.5 seconds, and had won MacCready and his collaborators the Kremer prize.

Now that the competition is over MacCready wants to take a vacation from an activity that has absorbed nearly every hour of his spare time for the past year. The *Gossamer Condor,* he says, was "a crude vehicle with one purpose only—to win the Kremer prize. It was a compromise between a sophisticated design and a finite amount of time, money and effort." MacCready believes, however, that his basic design could be refined to reduce its weight from 70 pounds to 55 and make its wing surface smoother. Then the craft would require only .29 horsepower to stay aloft, making it possible for an athletic bicyclist to fly it for several hours. As for the victorious craft, it is slated for retirement. MacCready hopes it will find a place in the Smithsonian Institution.

Claudius Ptolemy: Fraud

Claudius Ptolemy, the Greek scholar who flourished in Alexandria in the second century, has long been considered the greatest astronomer of antiquity. His astronomical treatise *'E Mathematike Syntaxis,* written in A.D. 142, has been called the most important work of



At a Shafter, Calif., airport, the **Gossamer Condor** *makes the first sustained man-powered flight* mathematical astronomy until *De Revolutionibus Orbium Coelestium* was published by Copernicus 1,400 years later. In the *Syntaxis* Ptolemy expounded a mathematical system for predicting the future positions of the planets, in which the earth is at rest and the sun, the moon, the planets and the stars move around it in combinations of circular epicycles. The Ptolemaic planetary theory so influenced the development of Islamic astronomy that Arab astronomers gave the *Syntaxis* the name *Almagest,* meaning "the greatest." Moreover, the Ptolemaic theory molded the development of European cosmology until the work of Copernicus and Kepler. More recently the *Syntaxis* has been important to historians as well as astronomers because Ptolemy cited astronomical theories and observations that he credits to his Greek and Babylonian predecessors, and in some cases the book is the sole surviving source of information about early Greek culture and Babylonian chronology. Now, according to Robert R. Newton of Johns Hopkins University in a book titled *The Crime of Claudius Ptolemy,* soon to be published by the Johns Hopkins University Press, Ptolemy is revealed as the most successful fraud in the history of science.

The principal evidence for the fraud exists in the *Syntaxis* itself. According to statistical analyses by Newton, the accuracy of certain observations of the planets and the stars that Ptolemy claims to have made is so great that in some cases the odds are a billion to one against his having made them with the instruments he described. In other cases Ptolemy's alleged observations of the equinoxes and solstices are off by more than a day, an error so unreasonably large that it is 10 times greater than the error in similar observations made by his predecessors. In Ptolemy's "updated" version of the star catalogue of his predecessor Hipparchus the positions of the 1,000 or so stars that Ptolemy allegedly observed were obtained by shifting each position by an identical amount. In a number of examples Ptolemy's conclusions based on the alleged observations are accurate to the last digit even when his calculations are in error. All these oddities have led Newton to conclude that Ptolemy derived his observations from his tables instead of the other way around. Writes Newton: "He developed certain astronomical theories and discovered they were not consistent with observation. Instead of abandoning the theories, he deliberately fabricated observations from the theories so that he could claim that the observations prove the validity of his theories."

Newton has also discovered that a large fraction of the observations Ptolemy attributed to other astronomers is not material he has preserved from the past but material he has forged. "At a guess," Newton estimates, "the realization of Ptolemy's fraud destroys half of what we have been accepting as Greek astronomy." Moreover, Ptolemy's forgery may have extended to inventing the length of reigns of Babylonian kings. Since much modern reconstruction of Babylonian chronology has been based on a list of kings that Ptolemy used to pinpoint the dates of alleged Babylonian observations, according to Newton "all relevant chronology must now be reviewed and all dependence upon Ptolemy's [king] list must be removed."

Newton has examined the possibility that an assistant of Ptolemy's provided Ptolemy with false observations out of laziness, but he believes that it is unlikely that such an assistant would also have deliberately falsified history. Newton also mentions as a slight possibility that the *Syntaxis* could have been written by someone other than Ptolemy. Newton has not yet found any evidence for the motivation of the fraud, although he suggests that the most likely one is simply that Ptolemy wanted to be known as a great astronomer. "He may have found, early in his career, that he did not have the qualifications, so he turned to the only remaining way of satisfying his ambition, which was to replace ability by fraud."

To Newton the most difficult problem presented by the *Syntaxis* is: Why has it been accepted as a great work for 18 centuries? During that time a few astronomers and historians discovered certain inconsistencies in it, but apparently the only scholar who seriously questioned the value of the work was J. B. J. Delambre, a French astronomer of the early 19th century.

"It is clear," Newton concludes his book, "that no statement made by Ptolemy can be accepted unless it is confirmed by writers who are totally independent of Ptolemy on the matters in question. All research in either history or astronomy that has been based upon the *Syntaxis* must now be done again."

Cull of the Wild

Hybridization is the chief means of improving plant productivity. Similar improvements in strains of domestic animals, however, have virtually eliminated natural variability and have considerably reduced the heritability of desired characteristics. Why not infuse new genes from wild animal species? The suggestion is made by R. V. Short, director of the British Medical Research Council Unit of Reproductive Biology at the University of Edinburgh. At a symposium celebrating the 150th anniversary of the establishment of the Zoological Society of London, Short pointed out that many animal species from the Tropics breed all year round, and species from polar and high-altitude habitats are notable for their large size and rapid maturation. All these characteristics are much desired in domestic animals.

Short suggests that hybridizing the domestic goose (*Anser anser*) with such wild stock as the red-breasted goose (*Branta ruficollis*), the bar-headed goose (*A. indicus*) and the nene (*B. sandvicensis*) could produce birds that grow faster, breed over an extended period and do not develop wasteful deposits of fat. Moreover, unlike chickens and turkeys, geese thrive on an exclusively vegetarian diet that requires no animal-protein supplement. Some 33 wild species and subspecies of geese, all mutually fertile with *A. anser*, offer a virtually inexhaustible gene pool for experimentation.

The domestic duck (*Anas platyrhynchos*) might similarly be improved by hybridization, Short observes. More than 50 wild species are available as gene donors, and crosses with most of them are known to be fertile. A particularly attractive candidate is a duck of the tropical Pacific, the Laysan teal (*A. laysanensis*). The bird not only has a long laying season but also almost never swims, thereby minimizing the growth of subcutaneous fat. The protein requirement for ducks, unlike that for geese, totals nearly 20 percent of their total caloric intake but less than 10 percent of the protein need be of animal origin.

Short also notes that the largest of all wild sheep is a subspecies popularly known as Marco Polo's sheep (*Ovis ammon poli*). A ram of this subspecies is about the size of a red deer: it is more than three feet high at the shoulder and can weigh nearly 300 pounds. Russian experimenters have successfully crossed another subspecies of *O. ammon* with domestic sheep. Observing that red deer survive in the highlands of Scotland under conditions that are too severe for the small Scottish blackface sheep, Short proposes that a hybrid of a domestic sheep and *O. ammon poli* might also fit this "bare, sodden, windswept" ecological niche.

Dodo Ecology

"The Dodo," wrote the humorist Will Cuppy, "seems to have been invented for the sole purpose of becoming extinct, and that was all he was good for." Not so. That large (25 to 50 pounds), ill-favored, flightless *Raphus cucullatus* of the island of Mauritius in the Indian Ocean, also known as *Didus ineptus* and dubbed *doudo* (silly) by the Portuguese sailors who discovered it in 1507, that slow-moving member of the pigeon family whose name became a byword for stupidity and failure to adapt (the last dodo had succumbed to man and his domestic animals by 1681), was apparently indeed good for something else. It played an essential role in the life cycle of a large tree, *Calvaria major*. Absent the dodo, *Calvaria* seeds have apparently not germinated for nearly three centuries, and the tree is now nearly extinct—going the way of the dodo.

That is the conclusion of Stanley A. Temple, an ecologist at the University of Wisconsin, who reports on the interaction of *R. cucullatus* and *C. major* in *Science*. The tree was once common in the forests of Mauritius, Temple points out, but by 1973 only 13 old and dying specimens were known to have survived, each of them estimated by foresters to be more than 300 years old. The survivors produce well-formed fruits, but the seeds do not germinate even under nursery conditions. The reason appears to be that the seed is enclosed in an extraordinarily thick (up to 15 millimeters) endocarp, or pit, that keeps the embryo from expanding. Noting "the temporal coincidence between the extinction of the dodo 300 years ago and the last evidence of natural germination of *Calvaria* seeds," Temple hypothesized that a period of coevolution had led to a "mutualistic" relation between the tree and the bird. He proposed that the thick endocarp was evolved by *Calvaria* in response to the destruction of its seeds by the fruit-eating dodo; the tough pit resisted destruction by the large stones grinding away in the dodo's powerful gizzard, but the gizzard did abrade and scarify the pit enough so that when it was excreted, the seed could break through and germinate.

Temple calculated the forces likely to have been generated in dodo's gizzard. He found that intact pits could indeed have withstood such forces without being crushed; the pits would not have been crushed until after they had been abraded to a size at which they would be likely to be excreted. He tested his hypothesis by feeding fresh *Calvaria* pits to turkeys. Of 17 ingested pits, seven were eventually crushed in the turkey's gizzard and 10 were regurgitated or excreted after being reduced in size by abrasion. When he planted the 10 pits, three of them germinated. They "may well have been the first *Calvaria* seeds to germinate in more than 300 years." Temple points out that many instances of plant-animal mutualism are known in which the elimination of a plant species has affected an associated animal population. This appears to be the first documented example of the reverse.

Side-looking Airborne Radar

A radar antenna mounted on the belly of an aircraft and aimed to the side can record microwave images of terrain in striking detail, regardless of the weather or the time of day or night

by Homer Jensen, L. C. Graham, Leonard J. Porcello and Emmett N. Leith

The first aerial images of the earth's surface, made as an aid to topographic mapping and military reconnaissance, were crude photographs taken from balloons in the middle of the 19th century. As cameras, films and airplanes evolved, aerial imagery was employed to satisfy an ever increasing variety of needs. In the 1920's vertical photographs made from airplanes became the basis for aerial photogrammetry. Today nearly all geological and topographical maps are based on images of the earth's surface made from aircraft and artificial satellites.

In the 1950's imaging devices were invented whose sensitivity extended beyond the visible region of the spectrum, particularly into the infrared wavelengths. With the advent of infrared sensors the region of the electromagnetic spectrum at which the earth's surface could be observed ranged between .4 micrometer (the shortest visible-light wavelengths) and 10 micrometers (in the thermal-infrared wavelengths). All of the devices detected energy that was either sunlight reflected from the terrain and man-made structures or was radiation naturally emitted by them as a function of their temperature and emissivity.

At visible and infrared wavelengths, however, the atmosphere absorbs a significant amount of radiation even when the air is clear. When the weather is cloudy or rainy, the performance of the sensors is seriously impaired. Even with such a broad region of the spectrum at the aerial surveyor's disposal his longstanding dream of some kind of magic camera that could be operated from the air to obtain pictures of the earth's surface regardless of illumination and weather remained unfulfilled.

The requirements for an all-weather, day or night imaging system are basically simple. In order to be independent of sunlight the system needs its own source of illumination. In order to penetrate clouds the system must employ a kind of radiation that is not attenuated or dispersed by water vapor. There is an imaging system that fulfills both of these requirements: a radar system working at the microwave wavelengths of between one centimeter and 30 centimeters.

Radar systems provide their own illumination, and microwaves of certain frequencies are little affected by the atmosphere and the weather. Radar systems working at those wavelengths currently monitor the weather from the ground and from satellites, assist in the control of air traffic and have measured the distance to the moon and the planets. Although microwaves are about 100,000 times longer than the waves of visible light, they are still short enough to resolve detail fine enough for most geological and geographical purposes. In fact, a radar system operating at those wavelengths carried aboard an aircraft obtains images of the earth's surface that display fine detail and spectacular relief. In this article we shall confine our attention to airborne radar systems that form images of the adjacent terrain by looking to the side of the airplane.

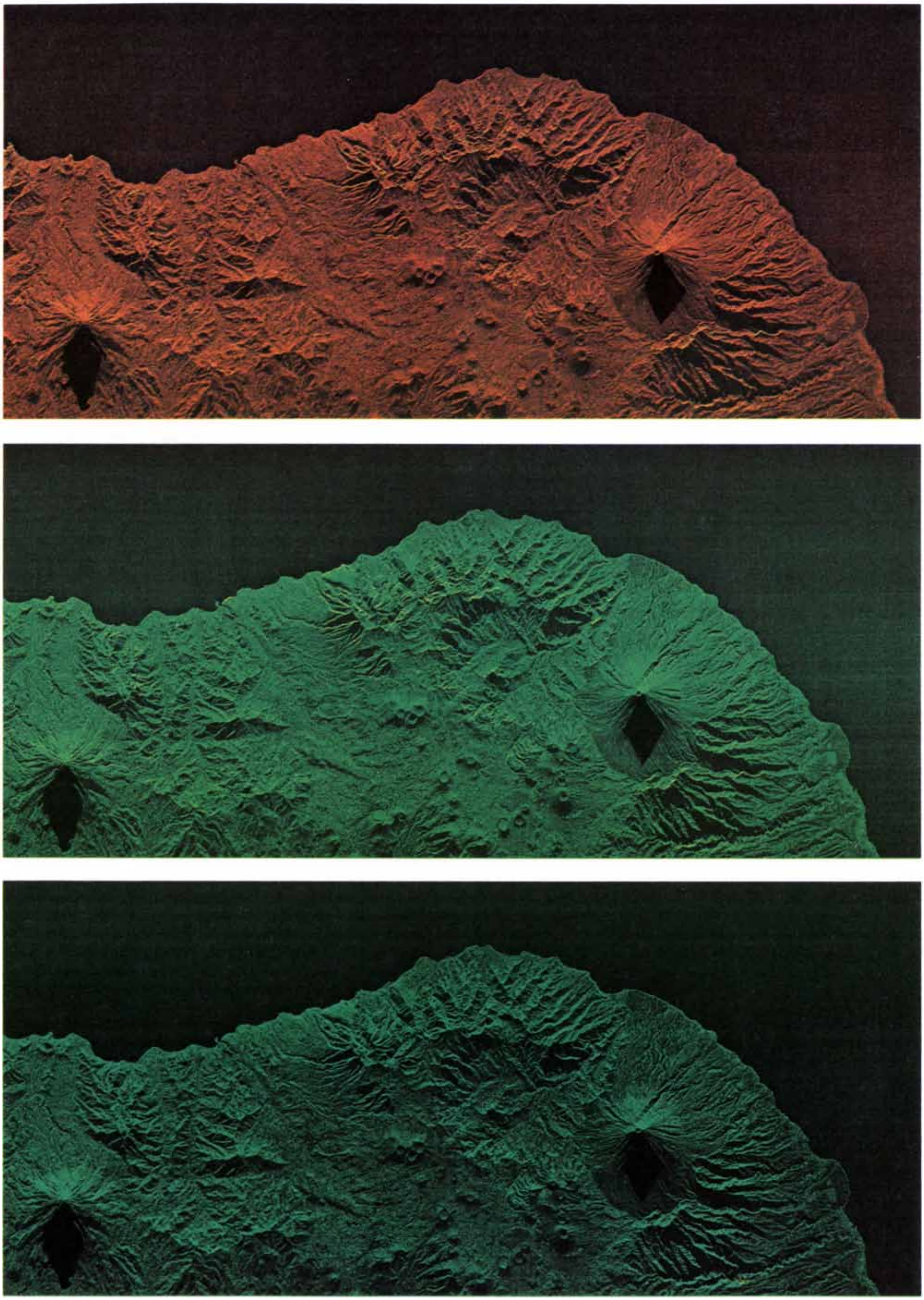
Resolution of the System

An obvious question is: Why should the radar antenna be pointed to the side? Almost everyone is familiar with circular-scan radar systems, which present on a cathode-ray tube an image obtained by a rotating antenna. Television weather reports, for example, employ such a system to show precipitation in the area of the weather-scan antenna. In a circular-scan radar system the raster of the cathode-ray tube rotates synchronously with the antenna, and the varying intensity of the reflected energy is expressed as the varying brightness of the scanning spot. The image of each scan is retained by the cathode-ray tube so that a picture is continuously maintained.

Experience has shown that most of the images made by circular-scan radar systems aboard airplanes are poorly defined. The poor definition results from a fundamental reason: Most airborne circular-scan radar antennas are rather small, and fine angular resolution can be obtained only with an imaging system that has a large aperture with respect to the wavelength of radiation received. In other words, the resolution of a large-aperture lens or antenna is finer than the resolution of a small-aperture lens or antenna. The limiting angular resolution is proportional to the ratio between the wavelength received and the size of the aperture. On an aircraft it is difficult to mount a rotating antenna that is very large, but it is easy to mount a fixed antenna as long as five meters on the belly of the aircraft, pointing to the side.

A side-looking airborne radar system generates images that record the reflective properties of the terrain at microwave wavelengths. The images, preserved on photographic film, resemble an ordinary photograph made from the air but they also show some fundamental differences. Unlike an aerial photograph made by the light of the sun, the radar antenna is the terrain's source of illumination, and the image it constructs relies on its own reflected energy. Since radar waves travel in straight lines, areas obscured by hills or other vertical objects are not illuminated; therefore they neither receive nor return the microwave pulses. The resulting shadows

SIDE-LOOKING RADAR IMAGES display in sharp relief the terrain of an island in Indonesia. As is indicated by the shadows, the aircraft making the images was flying on a course parallel to the images' long dimension. Area shown is 40 kilometers of southern coastline of the island of Flores. The peak at the left is an active volcano 7,000 feet high; the peak at the right is a dormant volcano of the same height. The grazing angle of the side-looking radar imagery also picks out the smaller volcanic craters in the middle of the picture and the sharp ridges and the valleys near the sea. All three pictures are derived from the same radar image of the terrain. The colors of the pictures are different because each picture is generated and displayed using a different wavelength of monochromatic light. Each of the images shows more than one color, however, because in certain portions of the picture the monochromatic light was sufficiently intense to excite a second color-emulsion layer in the color film. Displaying the radar images in color can convey more information than displaying them in black and white.



on the radar image of the terrain are dark voids; they are not areas weakly illuminated by atmospheric scattering, as are similar shadows of rough terrain photographed by the light of a rising or setting sun. On the radar image the detailed character of the reflections is determined by the wavelength and the polarization of the incident signal and by the geometrical and electrical properties of each reflecting surface on the terrain. Let us consider only the wavelength of the signal and the geometrical properties of the terrain.

Geometry and Reflection

The length of an electromagnetic wave with respect to the size of features of the terrain determines whether a surface appears rough or smooth at that wavelength. Thus a surface that is rough at a visible wavelength of .5 micrometer may be quite smooth at a microwave wavelength of three centimeters. A rough surface scatters the incident energy in all directions, diffusing it and returning only a small portion of it to the antenna. A smooth surface reflects the incident energy in one direction, acting like a mirror. If the smooth surface is at right angles to the incident radar beam, the energy returned to the antenna is intense; if the surface is at any other angle to the radar beam, none of the energy is returned. Some features, for example a field of corn, are diffuse reflectors in both the visible and the microwave regions of the spectrum. Other features, for example a concrete parking field, are diffuse reflectors in the visible region and are specular, or mirrorlike, reflectors in the microwave region. The net result is that there are more specular reflectors in microwave imagery than there are in photographs made with visible light.

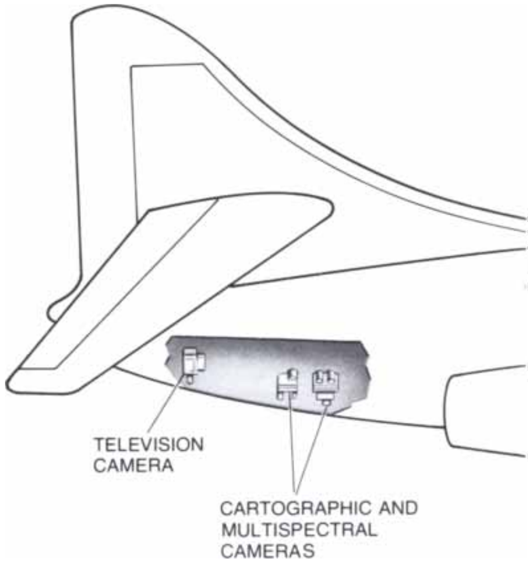
Smooth surfaces of water are specular reflectors par excellence. Since they are not usually viewed at right angles to the radar beam they specularly reflect all the microwave energy off into space and return none of it to the radar antenna. As a result in side-looking radar imagery rivers and lakes usually appear completely black. Conversely, related horizontal and vertical surfaces (such as a building next to a road) may work together to form a corner reflector, returning a large part of the energy directly back to the antenna. Such surfaces may therefore appear to the radar receiver thousands of times brighter than diffusing surfaces of a similar size. Cities and towns have many such multiple reflectors at microwave wavelengths.

The geometry of a photographic image is determined by the fact that the angular relation of rays entering a lens from an object or scene are preserved by the rays transmitted through the lens and exposing the film, as if all the rays had passed through a common pinhole. The size of the image of a distant object is inversely proportional to the distance of the object from the lens. Geometrical relations between objects in the object plane in front of the lens will be preserved on a parallel image plane in back of the lens; the relations between tilted surfaces on the object plane will follow straightforward trigonometric relations on the image plane.

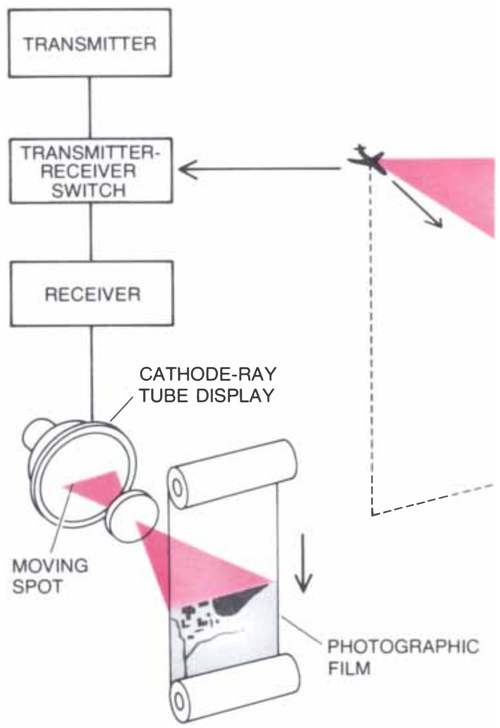
A side-looking airborne radar system operates with a different set of geometrical considerations. One coordinate of the image formed by such a system is established by the ranging of radar pulses across the track; the other coordinate is established by the motion of the radar-bearing aircraft parallel to the track. In the ranging process a short pulse of microwave energy is directed from a high-powered transmitter along a vertical fan-shaped beam by the long antenna in the belly of the aircraft, which is pointed at right angles to the flight path. The pulse travels with the speed of light; the time required for it to reach an object and be reflected back to the aircraft is proportional to the range, or distance, of the object at right angles to the flight path.

The intensity of the returned signal controls the intensity of a bright spot moving across a cathode-ray tube at a synchronized proportional velocity. Successive positions of the spot correspond to successive ranges across the track. In this way a line of varying intensity is swept across the cathode-ray tube with each pulse from the radar transmitter. An overlapping succession of such lines is traced on a strip of photographic film moving at right angles to the direction of the scan lines with a velocity proportional to the velocity of the aircraft. Thus a continuous image of the terrain is created on the film line by line.

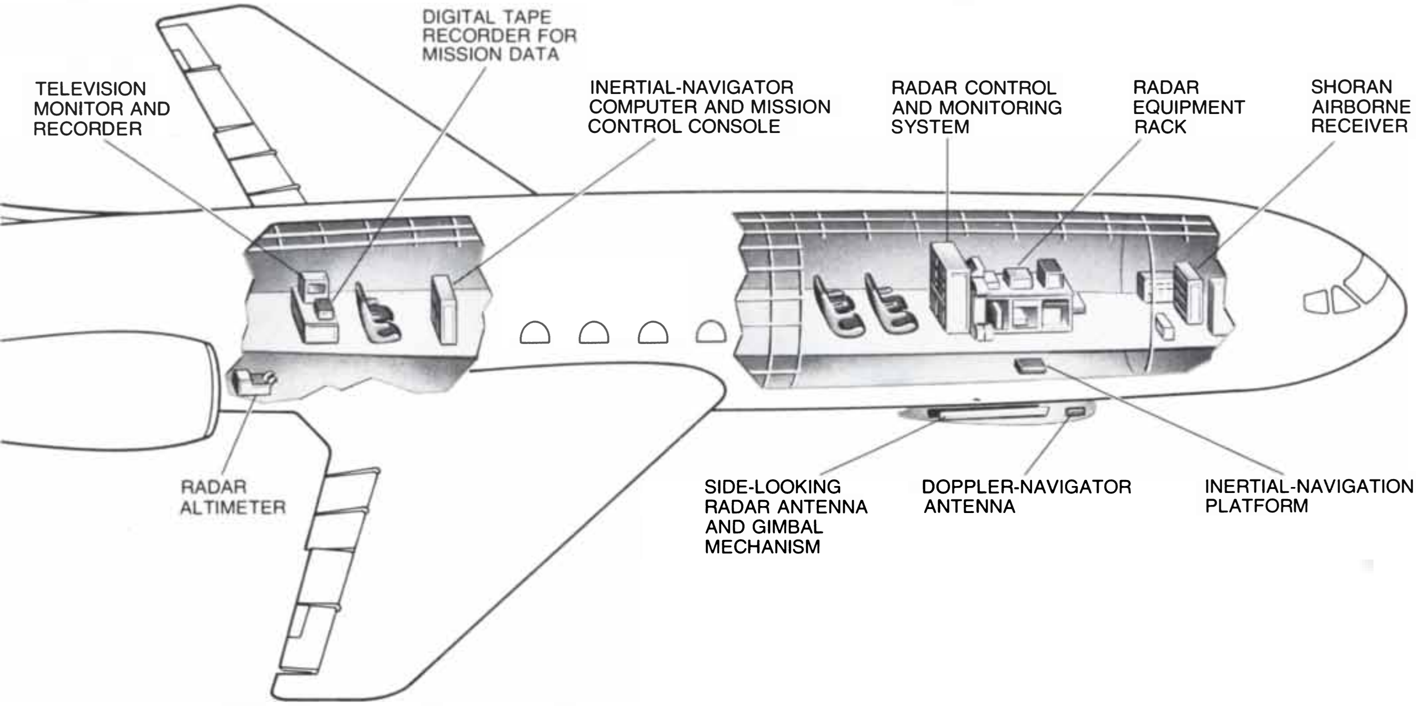
Relief in the terrain displaces details in the image from the positions they would have in the across-track dimension if they were viewed from directly overhead. The reason is that the portion of the microwave pulse reflected from elevated areas closer to the aircraft returns to the antenna sooner than the energy reflected from the surrounding lower terrain. Thus images of elevated areas on the terrain closer to the aircraft are displaced toward the aircraft's line of flight. This elevation displacement appears in all radar imagery of elevated terrain. It is somewhat analogous to the perspective distortion of conventional photography. In photography all objects with the same angular coordinates as seen from the lens appear to coincide on the film. In radar imagery all objects on a single scan at the same range as seen from the antenna appear to coincide on the image. A radar image is thus roughly analogous to a photographic image made by a hypothetical camera situated



SIDE-LOOKING RADAR SYSTEM built by the Goodyear Aerospace Corporation is installed on a Caravelle aircraft owned by the

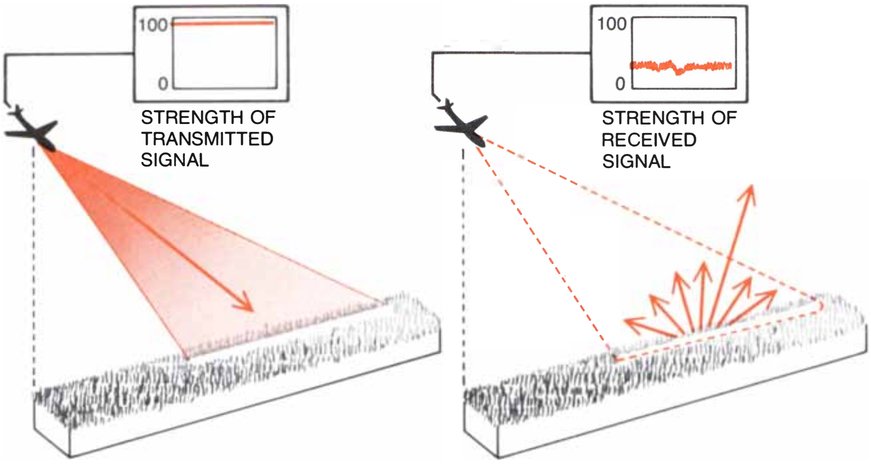


GEOMETRY of side-looking radar is shown for a hypothetical example of an aircraft surveying mountainous terrain. The antenna on the belly of the aircraft directs short pulses of microwave energy in a narrow fan-shaped



Aero Service Division of the Western Geophysical Company of America. The aircraft is kept accurately on course by a specially programmed inertial navigation system, which also controls many aspects of the side-looking system. Other cameras aboard the aircraft (television, cartographic and multispectral cameras) provide complementary coverage of the terrain at visible and infrared wavelengths.

beam at right angles to the flight path of the aircraft. As each pulse impinges on the terrain it is scattered, and a portion of it is reflected back to the aircraft. There it is received by the same antenna and sent to a sensitive radio receiver. In the receiver a signal is created the amplitude of which depends on the strength of the microwave energy received at any instant. That signal controls the brightness of a moving spot of light on a cathode-ray tube, which is recorded on a moving strip of photographic film. Since microwaves travel in straight lines, areas such as the far side of the mountain neither receive nor return the pulses, and they appear on the radar imagery as black voids.



DIFFUSE REFLECTOR, such as a field of corn, scatters the energy from the microwave pulses in all directions, returning only a small portion of the energy back to the radar antenna.

on a line at right angles to a line connecting the object to the radar antenna.

Along the track the aircraft is moving very slowly with respect to the speed of the microwave pulse, and the fan-shaped beam of microwave radiation is always pointing at right angles to the aircraft's line of flight. The result is that whereas the across-track dimension has a normal radar perspective, in which images of objects are displaced toward the aircraft as a function of their height, the along-track dimension is totally lacking in perspective. Thus whereas the scale and appearance of the across-track dimension of the radar image are determined by the speed of light, the scale and appearance of the along-track dimension of the image are determined by the speed of the aircraft. Reconciling the two independent scales and equalizing them on the final image is one of the more elegant and demanding problems in making maplike radar surveys of the earth's surface.

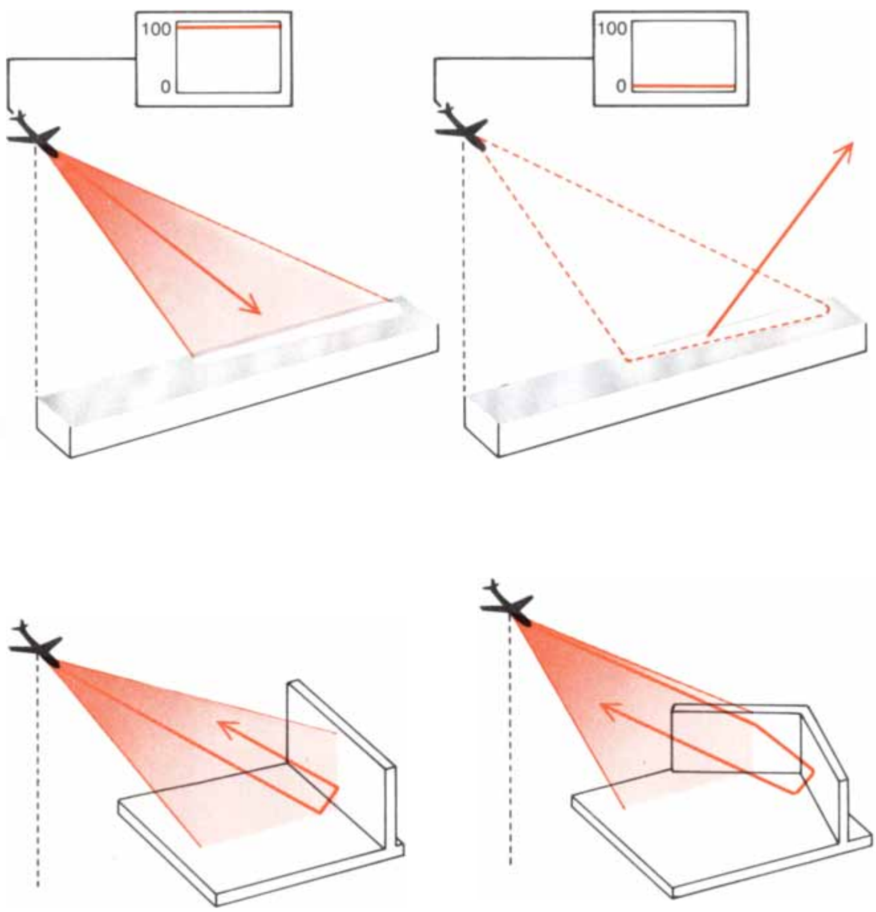
In practice this problem and many other complex problems involved in generating a satisfactory radar image have been solved by installing an inertial navigator in the aircraft. The inertial navigator employs a flexible computer that is programmed to provide outputs that serve the wide variety of information and control requirements of a radar survey. It guides the aircraft at an appropriate operating altitude along a series of great-circle paths as long as 1,000 kilometers, each path almost precisely parallel to the previous one so that a mosaic of the terrain composed of the series of strips can be built up. The inertial navigator guides the aircraft above clouds or through them during the day and at night, without correction from visual or radio navigation aids.

By providing signals to the airplane's autopilot the inertial navigator keeps the plane on course in a disciplined mode that avoids erratic maneuvers that would distort the image. Moreover, the inertial navigator's angular sensors control the attitude of the radar antenna in three angular coordinates (yaw, roll and pitch), making sure that the antenna is always pointed toward the ground at a constant angle with respect to the aircraft's line of flight. The velocity output of the inertial system meters both the rate at which the radar pulses are emitted and the velocity at which the image-recording film is advanced. The speed at which the moving spot sweeps across the cathode-ray tube screen controls the across-track scale of the image, and the velocity of the image-recording film controls the along-track scale.

Aperture Synthesis

Since the side-looking airborne radar system measures the distance to objects on the terrain by means of pulses of microwave energy, the resolution of the image in the across-track coordinate is proportional to the length of the microwave pulse. The pulses are very short, and the signals reflected from two distinct objects can be resolved if their respective ranges are separated by at least half the pulse length. For example, a pulse lasting 10^{-7} second is 30 meters long; thus it yields a resolution of 15 meters. If finer resolution is required, the length of the pulse can be reduced.

The resolution of the image in the along-track coordinate, however, is proportional to the width of the beam of microwaves sent out by the antenna. Two objects at the same range separated



SPECULAR REFLECTOR, such as a smooth body of water or a paved surface, acts as a mirror and reflects the energy from the microwave pulses in one well-defined direction. If the surface is not at right angles to the antenna, no energy returns to the antenna (*top*), and on the radar image the surface will appear black. Related surfaces such as streets and buildings, however, act together to form corner reflectors (*bottom*) that return a strong signal to the antenna.

in the along-track direction can be resolved only if their separation is larger than the antenna's beam width at their range. Otherwise they are in the radar beam simultaneously, the antenna will receive their reflections of the microwave pulse at the same time and they will appear on the image as a single strong return. Therefore the angular beam width of the antenna itself determines the resolution along the track of the airplane. Since that angle remains essentially constant across the range, the width of the beam increases with range and the resolution in the along-track dimension deteriorates with range.

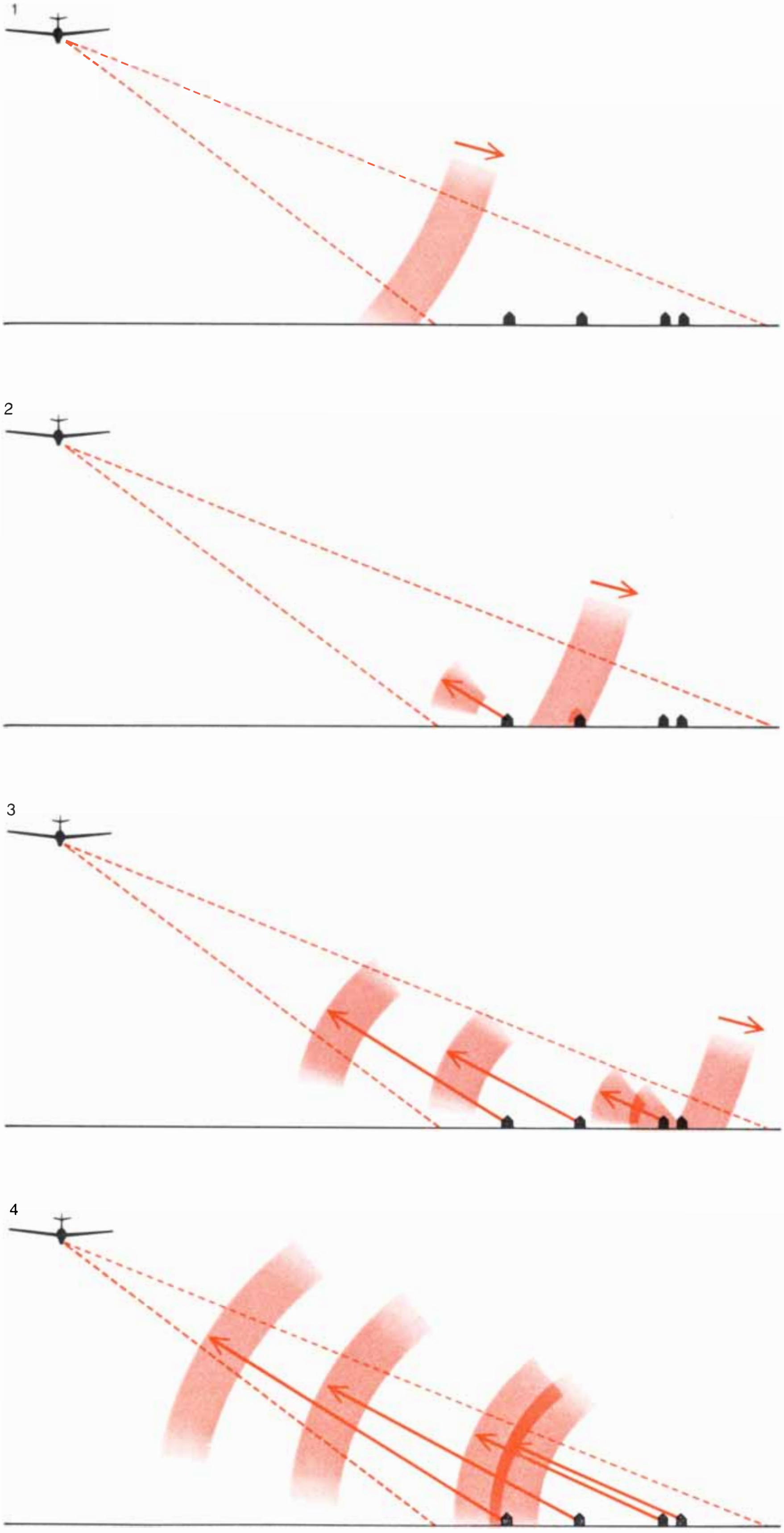
The angular width of the antenna's beam is inversely proportional to the length of the antenna, so that the resolution could be much improved by employing a larger antenna with a narrower beam width. Since the physical length of an antenna that can be mounted on an airplane is limited by the size of the plane, even with the longest practical antenna a conventional side-looking radar system has far coarser resolution along the track than across it, except at very short ranges.

It is possible, however, to make a relatively short antenna with a wide beam "synthesize," or behave like, a very long antenna with a narrow beam by taking advantage of the motion of the aircraft. This is accomplished in the following way. The short real antenna transmits pulses at regular intervals along the flight path. As the airplane flies past a feature of the terrain below, the feature enters the antenna's beam, moves through the beam and finally leaves it; during the entire time it is in the beam it reflects the series of microwave pulses. Now, the greater the range of the feature, the longer it remains in the beam of the antenna. In other words, seen from features of the terrain the synthetic antenna is longer for more distant features and shorter for closer features. In fact, the effective length of the synthetic antenna is directly proportional to the range to the feature. Since resolution is proportional to the length of the antenna but inversely proportional to the range, for synthetic-aperture radar the two effects compensate for each other, and the resolution of the image in the along-track dimension remains the same at all ranges. Synthetic-aperture radar thus makes it possible to obtain high-resolution images of terrain many miles away.

Radar Holography

An interesting analogy exists between a synthetic-aperture radar system and a visible-light holographic system. Holography is the imaging process in which light waves from a source of coherent light such as a laser illuminate an object or a scene. Each point on the object reflects the light, the reflected waves are

RADAR IMAGE OF THE OHIO RIVER in eastern Ohio and Kentucky shows the different action of diffuse and specular reflectors. On the picture the town of Portsmouth is the cluster of specular reflectors on the north bank of the river at the top left. To the west of Portsmouth the Scioto River joins the Ohio. Note that the water of both rivers appears black. The road running along the base of the hills of the western side of the river in Kentucky is U.S. Highway No. 23. The land to the east of the river is the edge of Wayne National Forest. This radar image shows the hilly terrain to particular advantage and clearly delineates the river's flat floodplain. The two bridges crossing the northern part of the river are brilliant specular reflectors of radar waves, as are the locks and the dam in the middle of the page. The elongated white flecks visible in the middle of the river are barges. Straight cuts across wooded hills are clearance paths for power lines. The terrain in the image is 10 miles by 15 miles. Image is a small section of a mosaic of images covering about 40,000 square miles of portions of Ohio, Kentucky, West Virginia and Virginia, made for a study of geological features associated with reserves of natural gas.



RANGE OF OBJECTS from the aircraft, or their distance across the track, is determined by the length of time required for the pulses of microwaves traveling at the speed of light to reach the object and be reflected back to the antenna. In the illustration signals returned from the two houses closer to the antenna will arrive sooner than the signals returned from the two houses farther away (*1 and 2*). The resolution across the track is equal to half the length of each pulse. Two objects, such as the third and fourth houses, separated across the track by less than half the radar-pulse length will be perceived by the antenna as one broad object (*3 and 4*).

made to interfere with a beam of reference light waves and the resulting multitude of circular interference patterns are recorded on fine-grained photographic film. When the developed film is illuminated with a beam of coherent light, the multitude of interference patterns reconstructs the original set of reflected waves, which form a three-dimensional image of the original scene having a startling degree of detail and fidelity.

Synthetic-aperture radar is an imaging process in which coherent microwaves illuminate the terrain below and to the side of the airplane. The radar receiver detects the radar waves scattered from each point on the terrain. In the electronic equipment aboard the aircraft the signals generated in the receiver by the train of reflected microwave pulses are combined with a train of reference signals to create interference patterns. Line by line each interference pattern is displayed on a cathode-ray tube and photographed on a moving strip of data film, the velocity of which is proportional to the velocity of the aircraft. The data film can be regarded as a radar hologram.

In synthetic-aperture radar the interference pattern from each scatterer on the terrain is recorded not as a circular pattern but as a narrow broken line parallel to the edge of the data film. The broken line can be considered, however, as a cross section through a circular interference pattern; the width of the line corresponds to the length of the radar pulses. The range of the scatterer is recorded as the distance from the near edge of the data film. The broken lines from two scatterers at the same range and sufficiently close together to be simultaneously illuminated by the radar beam will give rise to two overlapping but slightly displaced sequences of broken lines. As in visible-light holograms the two sequences of interference patterns can be superimposed without destroying their ability to form separate images.

When the data film is illuminated with a beam of coherent light, the interference pattern from each scatterer on the film re-forms a small portion of the light transmitted through the film into a diverging wave that can be brought to a point focus by a lens. Many overlapping signals from many scatterers simultaneously in the radar beam are transformed into a corresponding set of well-resolved images. In this way a miniature visible-light replica of the radar waves received by the antenna is formed, recreating a miniature image of the original terrain. Since the radar hologram possesses all of the data collected from every point of each feature of the terrain for the entire length of the synthetic antenna, the resolution of the holographic image corresponds to the resolution of the long synthetic-aperture antenna rather than to the resolution of the short real-aperture antenna. Thus the image formed by the hologram has the fine resolution we seek.

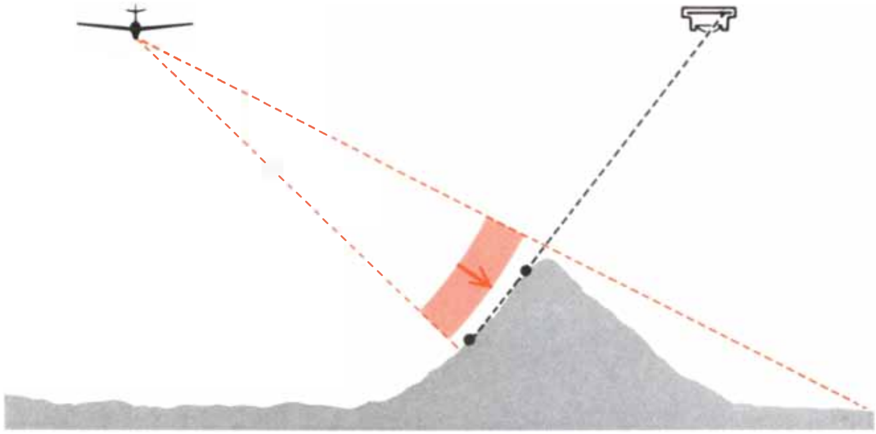
A convenient way to record the image from the hologram is to advance the data film through the beam of coherent light, moving the image-recording film synchronously with the moving image. In this way one generates a continuous picture of the terrain for the entire length of the aircraft's flight. Someone familiar with holography might suppose that the optical system for obtaining the final image film is a simple one, since the image from the hologram is formed simply by illuminating the hologram with a beam of coherent light. With the radar hologram, however, the task is not so simple.

There are two practical problems in recording the image. One is that the holographic aspect of the data record exists only in the along-track direction of the film. The signals are already resolved according to their respective ranges across the width of the film. As a result the process of focusing the signals in the along-track direction defocuses the signals in the across-track direction. This problem is corrected by introducing a cylindrical lens between the data film and the image film; the focusing power of the cylindrical lens is exerted only in the across-track direction. With such optical compensation it is possible to form images that are in sharp focus in both directions.

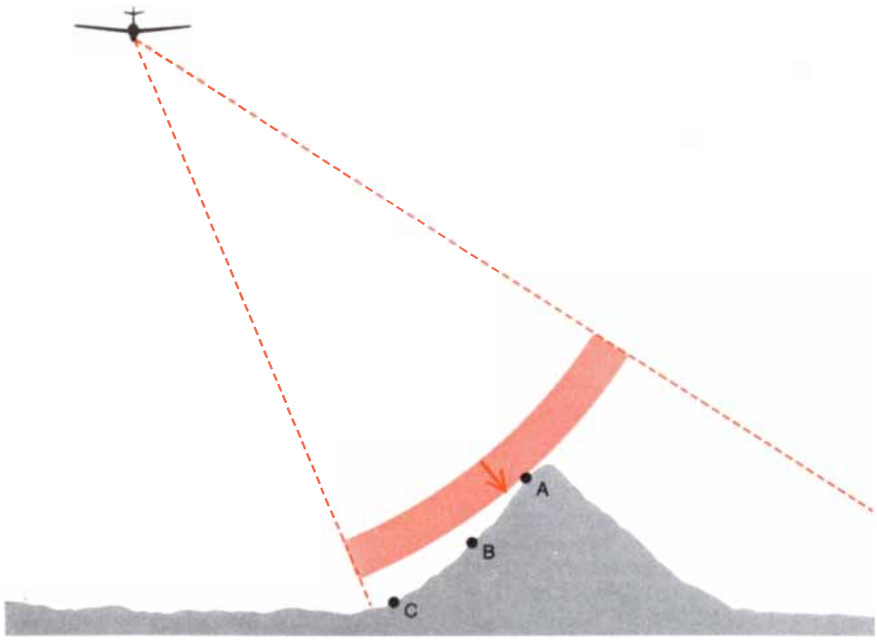
The other problem is that when the data film is illuminated, signals from objects at long range produce wave fronts that have a larger radius of curvature than signals from objects at short range. The signals from the opposite edges of the data film thus focus light at different distances from the film. The distance of the focus from the film depends on the range of the object. The image thus generated is tilted, and a camera recording the image could not simultaneously focus on the entire width of the swath. The problem is actually to be expected from elementary considerations; after all, the terrain being mapped lies at an extremely oblique position with respect to the radar antenna on the aircraft, so that the resulting image is similarly tilted. The problem can be corrected by placing a conical lens in front of the data film. The optical systems for reconstructing the radar holograms and the corrective lenses employed for capturing the image on film are some of the strangest ever devised, but they are quite practical in field applications.

Applications of the System

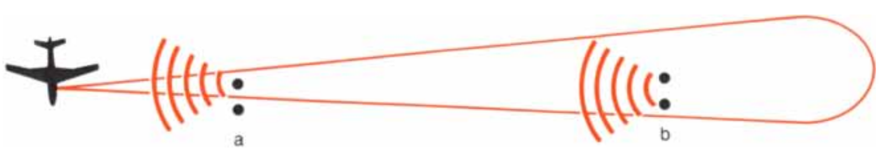
The first large-scale project for mapping terrain with side-looking airborne radar was a complete survey of the province of Darien, which connects Panama and South America. In 1968 the Westinghouse Electric Corporation,



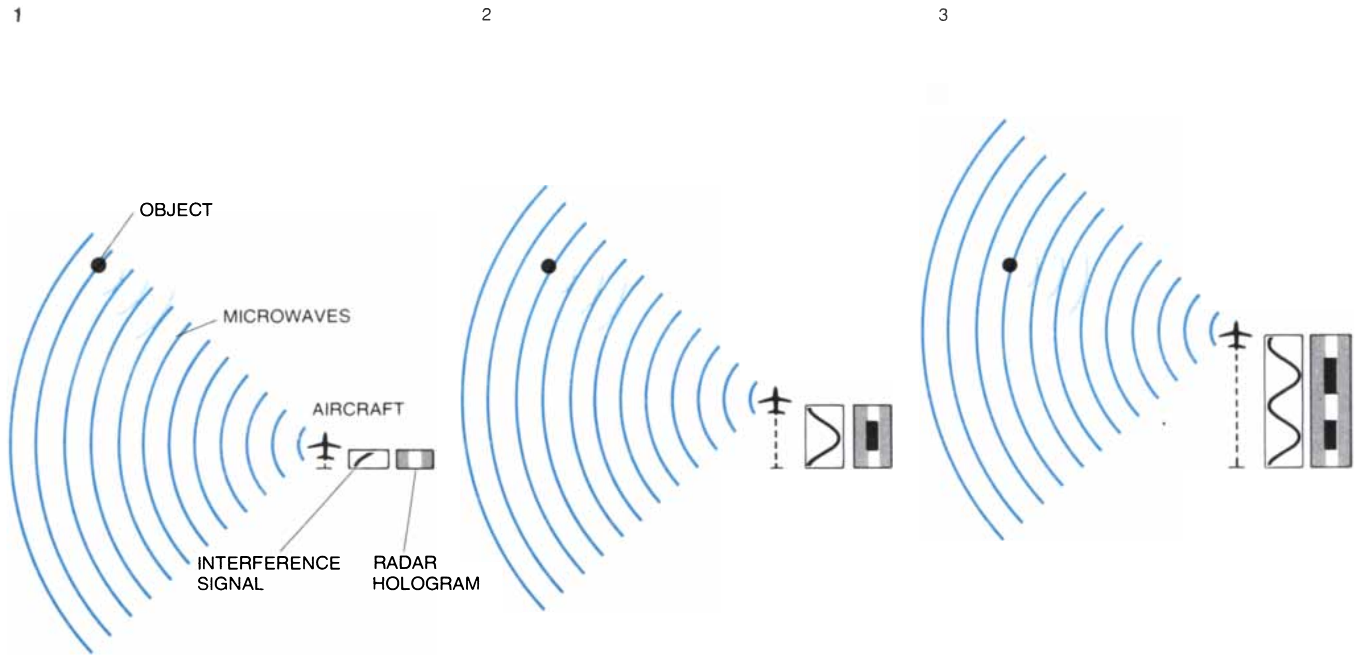
PERSPECTIVE IN RADAR IMAGERY is somewhat different from the perspective in ordinary aerial photographs made with visible light. In the radar imagery two objects closer together than the imaging beam width will appear to coincide if they are both at the same range. The reason is that the microwave pulses will reach them both at the same time and will return to the antenna at the same time, and so both objects will be seen as one. In photography two objects appear to coincide if they have the same angular coordinates as they are seen from the lens. Thus the perspective in the radar imagery is roughly analogous to perspective of a camera in the same vertical plane with the aircraft and objects but photographing the terrain at right angles.



ELEVATED OBJECTS ARE DISPLACED toward the aircraft's line of flight in radar images of the terrain, making the near side of mountains seem steeper than they are. The reason is that the peak of a mountain (*A*) is closer to the aircraft than the side (*B*) or base (*C*), and so pulses from the peak return to antenna first. Amount of displacement increases with elevation.



RESOLUTION DETERIORATES WITH RANGE in the along-track direction because the radar beam of a real-aperture side-looking radar system fans out across the track and is wider at a longer range than it is at a shorter one. Two objects close together will be seen as two objects at short range (*a*) but will be seen as one object at long range (*b*). A longer antenna would have better resolution, but there is a limit to length of antenna that can be mounted on an airplane.



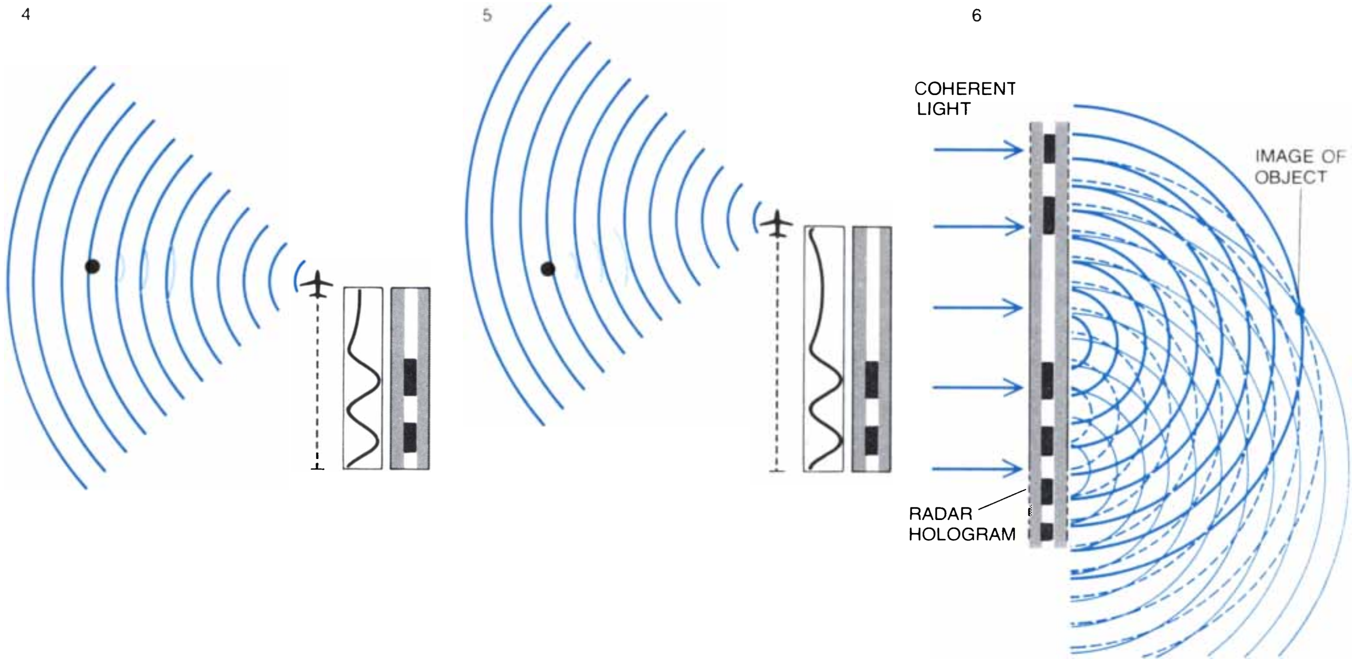
LONG ANTENNA CAN BE SYNTHESIZED by a short antenna by taking advantage of the airplane's motion. As the airplane flies along a straight line a short real antenna mounted on its belly sends out a series of pulses at regular intervals. Each pulse consists of a train of coherent microwaves (*dark color curves*). Although the length of the pulses determines the resolution across the track, it is the wavelength of the microwave radiation that determines the resolution along the track. As an object (*black dot*) enters the antenna's beam (*left*) it reflects the portion of the pulses it receives back toward the antenna. At some points in the aircraft's path the object will be an integral number of microwave wavelengths away; between those points it will not be. In the illustration the object is first 11 wavelengths away (*1*), then 10 (*2*), then 9 (*3*), then $8^1/_2$ (*4*), at which point the object is at right angles to the antenna. From then on the airplane is increasing its distance from the object (*5*). The antenna receives the series of reflected waves (*curves in light color*) and electronically combines them with a train of reference wavelengths (not shown), causing the two series of oscillations to interfere. The interference signal emerges as a voltage, which controls the brightness of a spot scanning across the screen of a cathode-ray tube. At the times that a returned pulse coincides with

in cooperation with the Raytheon Company, employed a real-aperture side-looking radar system that Westinghouse had developed for the U.S. Army and successfully made images for a mosaic of an area of 20,000 square kilometers. The area had never been seen or mapped in its entirety because of an almost perpetual cloud cover. Other surveys were conducted by Westinghouse and later by Grumman Ecosystems with a radar system developed by Motorola Inc., also for the Army. The Motorola equipment, with some changes, is now being employed for surveys by Motorola Aerial Remote Sensing, Inc.

In 1970 the synthetic-aperture radar systems were released from military security classification. A group from the Goodyear Aerospace Corporation and the Aero Service Division of the Western Geophysical Company of America, a subsidiary of Litton Industries, adapted for civilian surveys a synthetic-aperture radar system built by Goodyear and installed it in a Caravelle jet aircraft belonging to Aero Service. The two companies then jointly offered radar surveying services with the system. Similar radar surveying systems were constructed and operated by a group of investigators at the University of Michigan, whose programs are being continued by the Environmental Research Institute of Michigan. Several other organizations have also made substantial contributions to the field.

In 1971 the rain forests of South America, mostly in the Amazon Basin, remained the last great land area in the world that was still poorly and inexactly mapped. Venezuela and Brazil, embarking on plans for developing their rain-forest area, arranged for radar surveys of major regions of the two countries. Several requirements had to be met. The images of all the land surveyed had to be consistent, so that one region could be easily compared with another. Moreover, the radar imagery had to be consistent with the imagery made by the Earth Resources Technology Satellite (now called *Landsat*), which was then just beginning to survey the earth. With those considerations in mind the Goodyear/Aero Service group directed all its survey flights north and south, with the radar antenna illuminating the land toward the west. In that way the appearance of the land and the shadows cast would closely match the appearance of the images from *Landsat*, which was set to photograph all the land at a local time of about 9:00 A.M., when the ground was illuminated by the slanting rays of the morning sun.

Since the radar imagery was to be used to compile maps, the highest practical level of positional accuracy was sought. The inertial navigator, completely adequate for controlling the aircraft's flight, was not sufficiently accurate for determining exact reference points for maps. Therefore an array of some 30 points was established by the then new Transit Satellite Positioning System throughout the Amazon Basin. These points were surveyed to an accuracy of about 10 meters. At each one a transponder was set up for a shoran radio-positioning system. Equipment aboard the aircraft could detect signals from the transponder as far as 500 kilometers away. By simultaneously receiving the broadcasts from two such points the position of the aircraft was determined to an accuracy of about 75 meters and was checked every six seconds. During the survey the plane's position was recorded on magnetic tape along with its altitude and other significant data. With this system in less than a year images of the entire Brazilian Amazon Basin—an area of some four million square kilometers, about half the size of the U.S.—were recorded and assembled



a reference pulse the interference is constructive; the voltage will be high and the moving spot will be bright. At the times that the phase of the returned wavelength does not coincide with the phase of the reference frequency the interference is destructive; the voltage will be low and the moving spot will be dim. The moving spot thus traces out a series of light and dark dashes of unequal length, which are recorded on a strip of data film moving at a velocity proportional to the velocity of the airplane. The series of opaque and transparent dashes on the film is actually a one-dimensional interference pattern; the film on which they are recorded is a radar hologram. When the developed hologram is illuminated by a source of coherent light (*6*), each transparent dash will act as a separate source of coherent light. Behind the hologram there will be a single point where the resulting light waves all constructively interfere. There the 11th wavelength of light (*thin color curves*) from transparent dash created by 11th microwave will meet 10th wavelength of light (*broken color curves*) from transparent dash created by 10th microwave, and both will meet 9th wavelength of light (*thick color curves*) from transparent dash created by 9th microwave. At that one point light from the entire length of interference pattern is focused to form miniature image of original object.

into mosaics. The completed survey images were accurate to about 500 meters.

To support the radar survey and make full use of its imagery Brazil set up Projecto Radam (for radar of the Amazon) based at Belem. Projecto Radam involved scores of geologists, soil scientists and other technically trained workers. The mosaics formed the base maps for all their studies. Geological analysis led to detailed mapping and extensive fieldwork in certain areas, employing the imagery as a guide. Structural information apparent in newly disclosed surface features provided clues for revising and expanding existing geological knowledge of the Amazon Basin. On the basis of the mosaics tentative routes were selected for sections of the Trans-Amazonas Highway. Large deposits of important minerals were discovered, and new geographical features were revealed, including previously unknown volcanic cones and even large rivers.

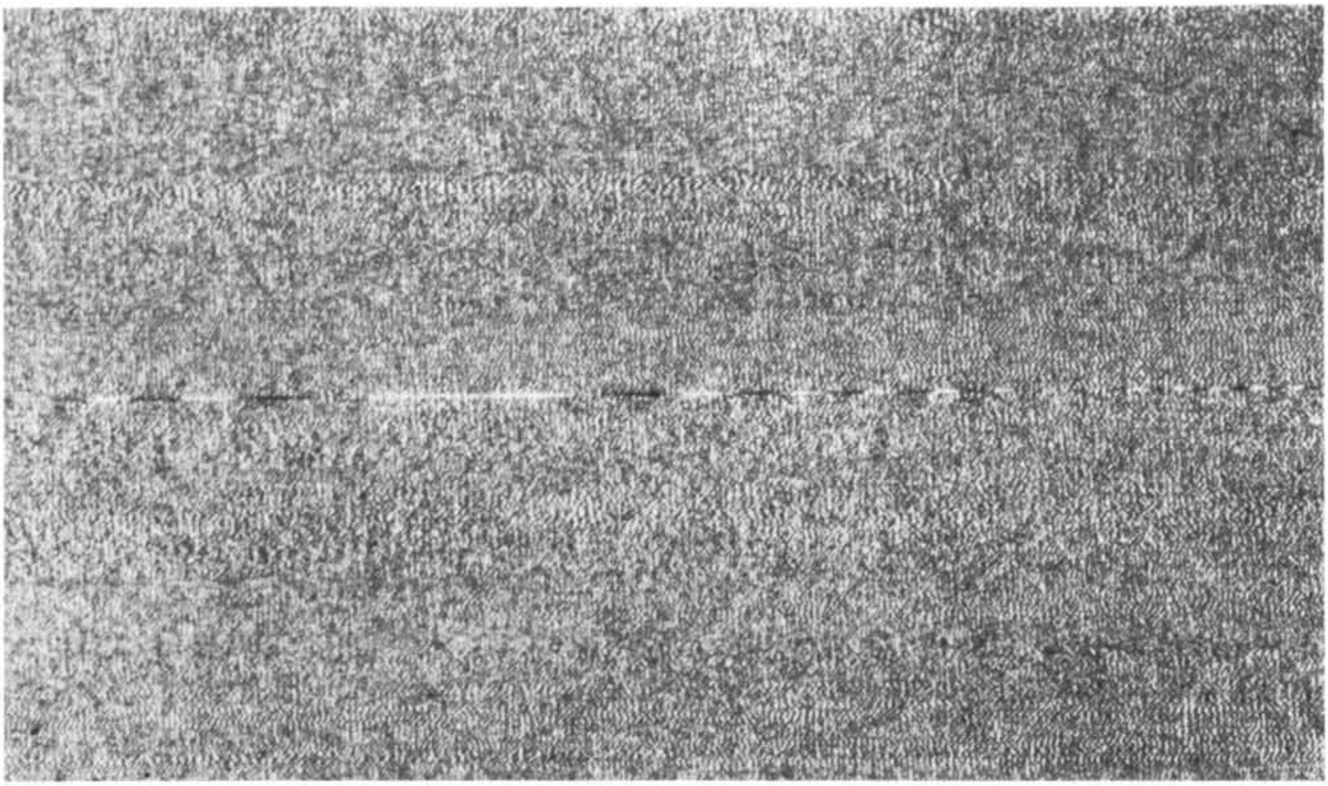
So far the Goodyear/Aero Service group has surveyed the rain forests of the Amazon in Brazil, Venezuela, eastern Colombia, Peru and Bolivia. The surveys have pretty well erased the unknown areas of South America. No other tool could have done the job. The success of the surveys has led to continued work in areas already mapped by other means. And in March, 1976, the last series of flights completed a radar survey of all of Brazil, a mapping project of unprecedented scope.

Other Applications

Side-looking airborne radar has been a tool of unique value in other applications. Studies of continental drift have shown that it is important to recognize certain types of faults, lineaments and other tectonic features. In many cases the features are evident from the surface but too subtly or on too large a scale to be easily recognized by most means. The acute grazing angle of the microwave illumination of side-looking radar emphasizes the form of the land, and the large areas that can be surveyed under constant conditions favor the recognition of extensive features. In view of the importance of avoiding active geologic faults in the placement of nuclear power plants, radar imagery plays an important role in analyzing the sites selected. Moreover, because of the complete radar blackness of unruffled bodies of water, rivers and drainage patterns are clearly shown on the images, as are the limits of flood inundation.

What might be some future applications of side-looking airborne radar? Among those working in the field it is generally accepted that the ability to generate radar imagery has outstripped the ability to extract information from it. Therefore more attention is now being given to displaying the information.

Some fairly obvious improvements in the display techniques have dramatically increased the ability to extract visual information from the images. One method is the use of color coding to display the great brightness range of radar energy, a range that often far exceeds the capability of black-and-white film. A second method is the use of stereographic techniques to display the elevation data that are intrinsic in suitably generated pairs of radar images. A third method is to capitalize on the fact that two remote sensors operated in combination are usually far more effective than either one might be alone. Investigators in the U.S. Geological Survey and at Goodyear Aerospace have worked together on combining images made by side-looking airborne radar with images made by *Landsat,* a combination that preserves the distinctive capabilities of both systems. In this combination the diagnostic color presenta-

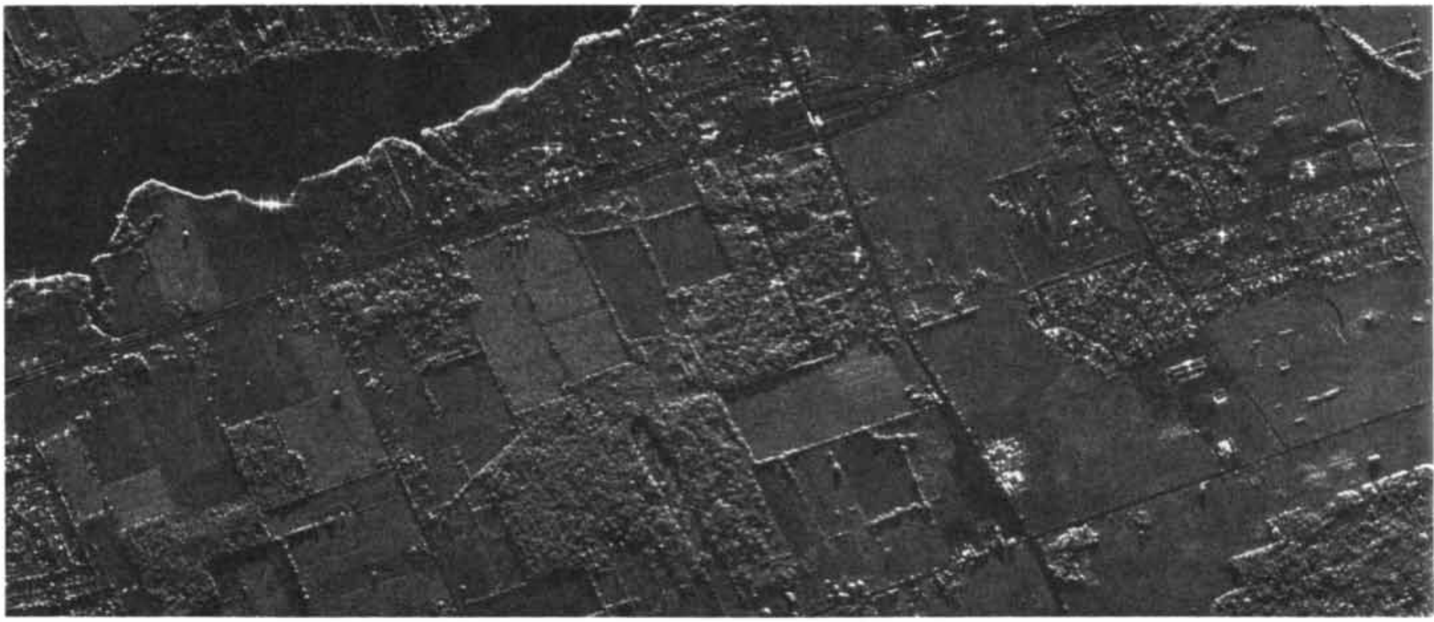


SECTION OF DATA FILM shows the interference patterns produced by the returning trains of microwaves in the synthetic-aperture side-looking radar. Radar hologram is enlarged 10 times to show the detail. The long horizontal line of black and white dashes in the center is the interference pattern of one strong return from an unusually large specular reflector such as one of the power-line towers in the radar image on the cover. The smaller and less organized variations are less intense returns from hundreds of smaller targets.

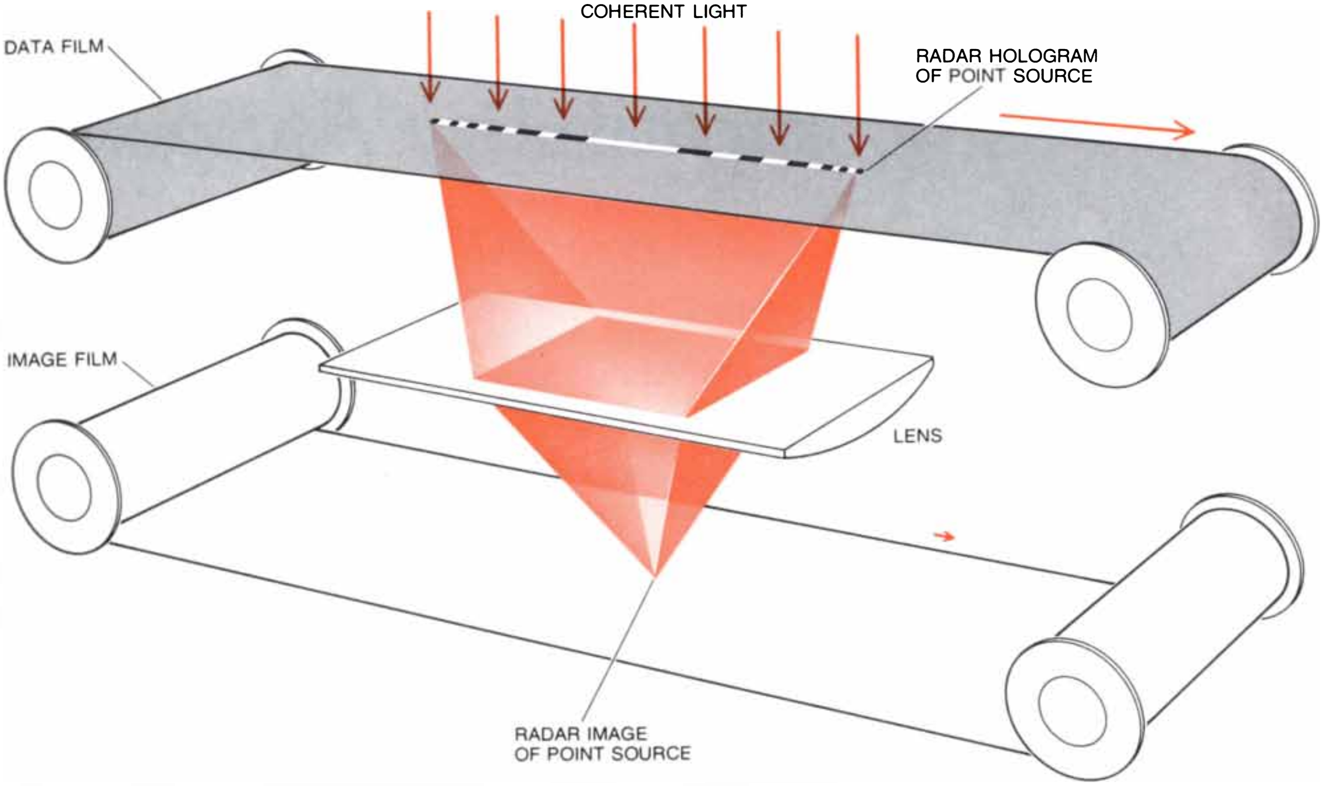
tions of the *Landsat* imagery become more significant in the context of the vivid presentation of land forms and the rich detail of the radar imagery.

At a more sophisticated level it may become possible to extract a great deal of information from multispectral ensembles of radar images generated at different wavelengths and with microwave signals polarized in different ways. By studying how the reflecting properties of a surface vary with both wavelength and polarization it may be possible to improve the microwave perception of surface differences and distinguish between, say, different soils and different crops.

Side-looking radar may also be able to make an important contribution to the exploration of the planets. Pictures of the surface of the moon and of Venus

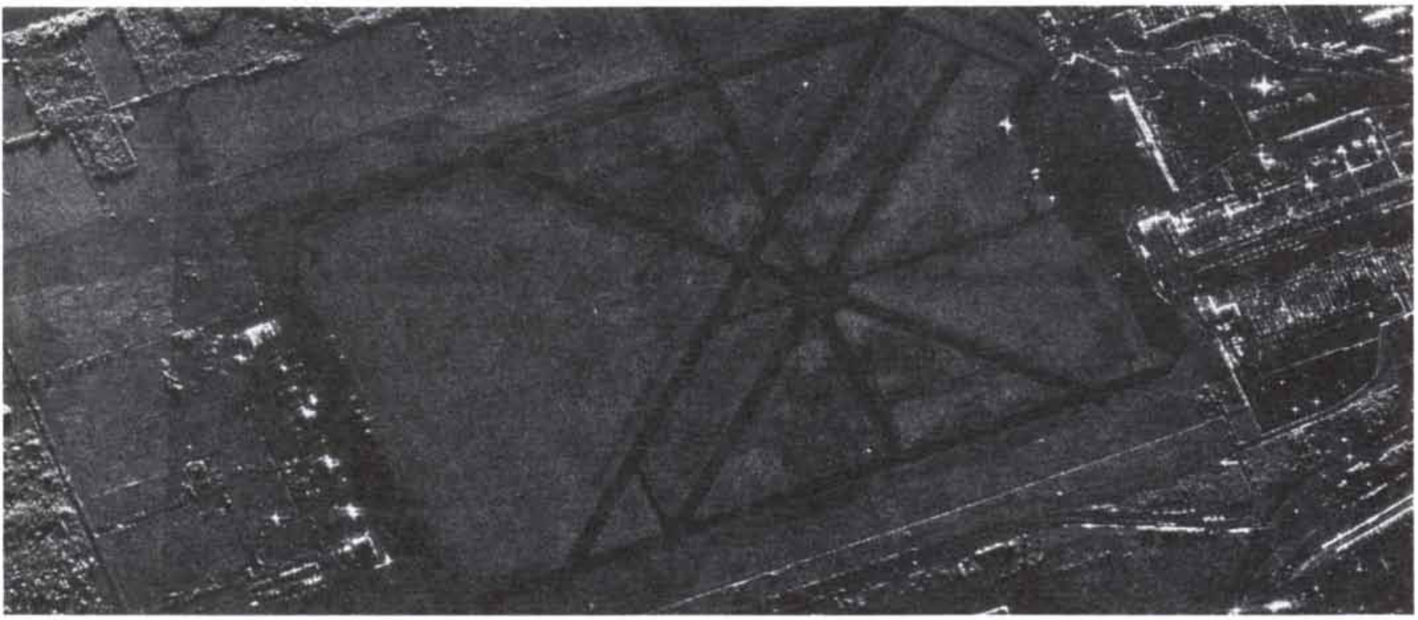


SECTION OF IMAGE FILM recorded by side-looking radar by the Environmental Research Institute of Michigan shows a swath of terrain near Detroit and Ann Arbor. North is to the bottom of the image. In this orientation the aircraft was flying in the same direction as the long dimension of the photograph. The water of the Belleville Lake (a dammed-up portion of the Huron River) and the runways of the



HOLOGRAPHIC IMAGE IS RECONSTRUCTED by shining the coherent light of a laser through the data film. As the data film is advanced through the beam of light the reconstructed image is recorded on another moving strip of film. Because the data film is holographic only in the along-track coordinate the images in the across-track coordinate must be focused with a cylindrical lens. Wealth of detail on image arises in part from the fact that length of interference pattern on data film is 72 times longer than image it reconstructs.

have already been made by large radar systems based on the earth. The *Apollo 17* spacecraft carried a three-wavelength synthetic-aperture radar system that made pictures of the moon's surface from an altitude of 100 kilometers with a resolution of 10 meters at the shortest wavelength employed (two meters). The surface of the moon, however, is never obscured by clouds, whereas the surface of Venus always is. Over the past few years enthusiasm has grown for putting an imaging-radar system on a satellite in orbit around Venus and having it transmit radar pictures back to the earth. Such a system would be capable of displaying the terrain of the shrouded planet in fine detail over limited areas, complementing the broad-coverage, limited-resolution radar pictures made from the earth.



Willow Run Airport are black voids where the radar pulses were specularly reflected off into space. The trees and bushes are diffuse reflectors. Buildings and pavements acted together to form corner reflectors and direct a brilliant specular reflection back toward the antenna. The image is somewhat distorted, causing squares to be seen as rhombuses, because the velocity of the data film was metered inexactly.

The Structure and Function of Histocompatibility Antigens

Histocompatibility antigens are proteins of the cell surface that cause the rejection of tissue grafted from one person to another. They may also help defend the body against infection and cancer

by Bruce A. Cunningham

When a patch of skin or an organ is transplanted from one person to another, the graft does not usually survive. After a week or so the transplanted tissue becomes puffy and inflamed, and soon the graft dies and drops away. Premodern medicine blamed such failures on sepsis, but it is now known that the immune system of the person receiving the transplant recognizes certain molecules on the grafted cells as being foreign, and attacks and destroys the cells as though they were bacteria or other invading agents. The molecules that are recognized as being foreign are called histocompatibility antigens, and they give the tissues of each individual their own chemical identity.

Although early interest in these cell-surface molecules stemmed from their role in graft rejection, this cannot be their normal biological function because in nature tissues are not usually transplanted between individuals. Moreover, the ability to reject foreign grafts can be traced back along the evolutionary tree to the most primitive vertebrates, suggesting that the histocompatibility antigens confer some important survival advantage on the host. Indeed, recent experiments have provided evidence that these antigens are part of an immunological defense system that protects the body from its own cells when they become infected with viruses or turn cancerous.

Histocompatibility antigens are defined as any molecules that differ from individual to individual and are recognized in graft rejections. Nearly 30 such antigens have been detected in the mouse, and each of them can vary in non-inbred populations. This genetic diversity provides biological uniqueness for every individual in a normal population. Studies of histocompatibility antigens suggest that the number of possible combinations is so large that (with the exception of identical twins) exact matches are very rare.

Histocompatibility antigens differ in their ability to elicit an immunological response in a graft recipient. In mice one set of antigens known as the *H*-2 system provides by far the strongest barrier to transplantation. If skin grafts are exchanged between mice having different *H*-2 antigens on their cells, the grafts will invariably be rejected within 14 days. Grafts differing in "weak" histocompatibility antigens outside the *H*-2 system, however, can take up to 200 days to be rejected. The *H*-2 antigens were first described by Peter Gorer and his colleagues at the Lister Institute of Preventive Medicine in 1937 and were intensively investigated by him and by George Snell and his colleagues at the Jackson Laboratories in Bar Harbor, Me., over the next 20 years.

The strong *H*-2 antigens of the mouse have their counterpart in the *HLA* antigens of man, which were first detected on the surface of white blood cells in the early 1950's. Although these antigens were for genetic reasons more difficult to define, the overall features of the *HLA* system were discerned by correlating data from a large number of laboratories in a series of international workshops on tissue typing held during the late 1950's and early 1960's. Work on these antigens has already had a direct clinical application in the matching of tissues and organs for transplantation surgery so as to reduce the immunological reaction of the recipient against the graft. In addition, the realization that diseases such as multiple sclerosis, Hodgkin's disease (cancer of the lymph nodes) and chronic hepatitis occur with a higher frequency in people with certain *HLA* types has raised significant questions about the genetic origin of these diseases and their connection to the function of the histocompatibility antigens.

Because the *H*-2 and *HLA* antigens have so many genetic variants, it would be extremely difficult to investigate their molecular nature in outbred populations such as those of man. Fundamental questions about such antigens can most easily be explored in mice, which can be made genetically homogeneous by intensive inbreeding, usually by many brother-to-sister matings. Moreover, it is possible to make many functional tests with mice that cannot be made with human beings. For these reasons the bulk of this discussion will focus on the *H*-2 antigens.

The *H*-2 and *HLA* antigens are protein molecules firmly attached to the surface of nearly all cells. Part of the molecule is probably embedded in or passes through the cell's outer membrane. Like other protein molecules, the antigens consist of chains of amino acids, but in addition they contain a small amount of carbohydrate; hence they are termed glycoproteins. They are detected and classified by their ability to bind to specific antibodies, which are manufactured when cells from one animal are injected into a genetically different animal of the same species. Some of the molecules on the injected cells are identical to those of the recipient; others, because of the genetic differences, are not. The recipient typically makes antibodies against the molecules that are different, and those antibodies define the antigenic specificities of the injected cells. A given antigenic specificity may reflect the presence of an antigen molecule on the donor's cells that is not present on the recipient's cells, or it may reflect structural differences between variants of a histocompatibility antigen that is present on both the donor's cells and the recipient's.

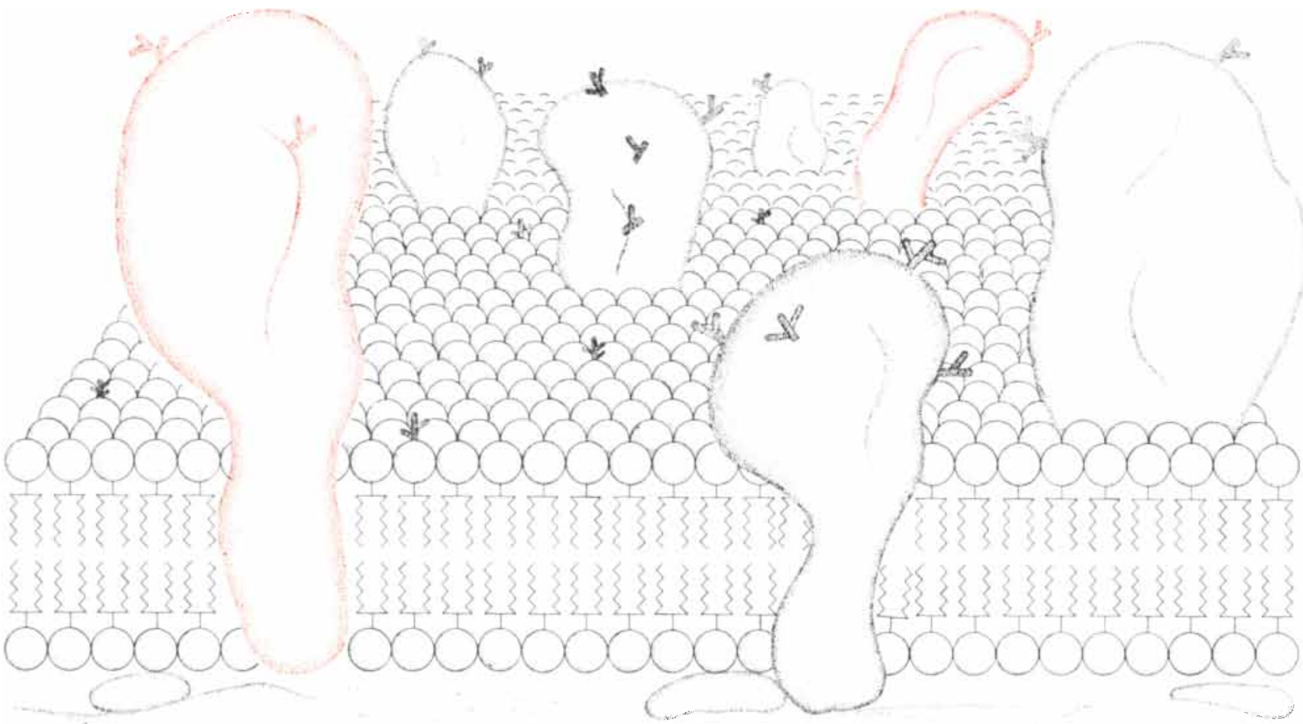
Antibodies to specific *H*-2 antigens can be made by injecting mice from one inbred strain with cells from mice of another inbred strain. Through careful crossbreeding separate strains can be developed that differ in only a small part of their genetic makeup. Such strains can then be utilized to make antibodies that are specific for only one type of *H*-2 antigen. The systematic application of such specific antibodies, together with genetic mapping and population-genetics studies, have enabled immunogeneticists to describe the genes that code for the *H*-2 antigens in considerable detail. These investigations have revealed that the *H*-2 genes have many variants, and also that they are closely associated with other genes that play an important role in the functioning of the immune system and in the development of the embryo into the adult.

The region of the genetic material of the mouse that codes for the *H*-2 antigens is a complex of genes located on a relatively short segment of the mouse's chromosome No. 17, representing only a three-thousandth of the mouse's total complement of DNA. This region is nonetheless long enough to code for up to 200 average-sized proteins. Two of these proteins are the *H*-2 antigens, which are coded for by genes located at either end of the *H*-2 gene complex; they are designated *H*-2*K* and *H*-2*D*. A mouse's body cells contain two No. 17 chromosomes, one inherited from each parent. In outbred populations the parental genes will be different, giving rise to four different *H*-2 antigens: two *K* and two *D* antigens.

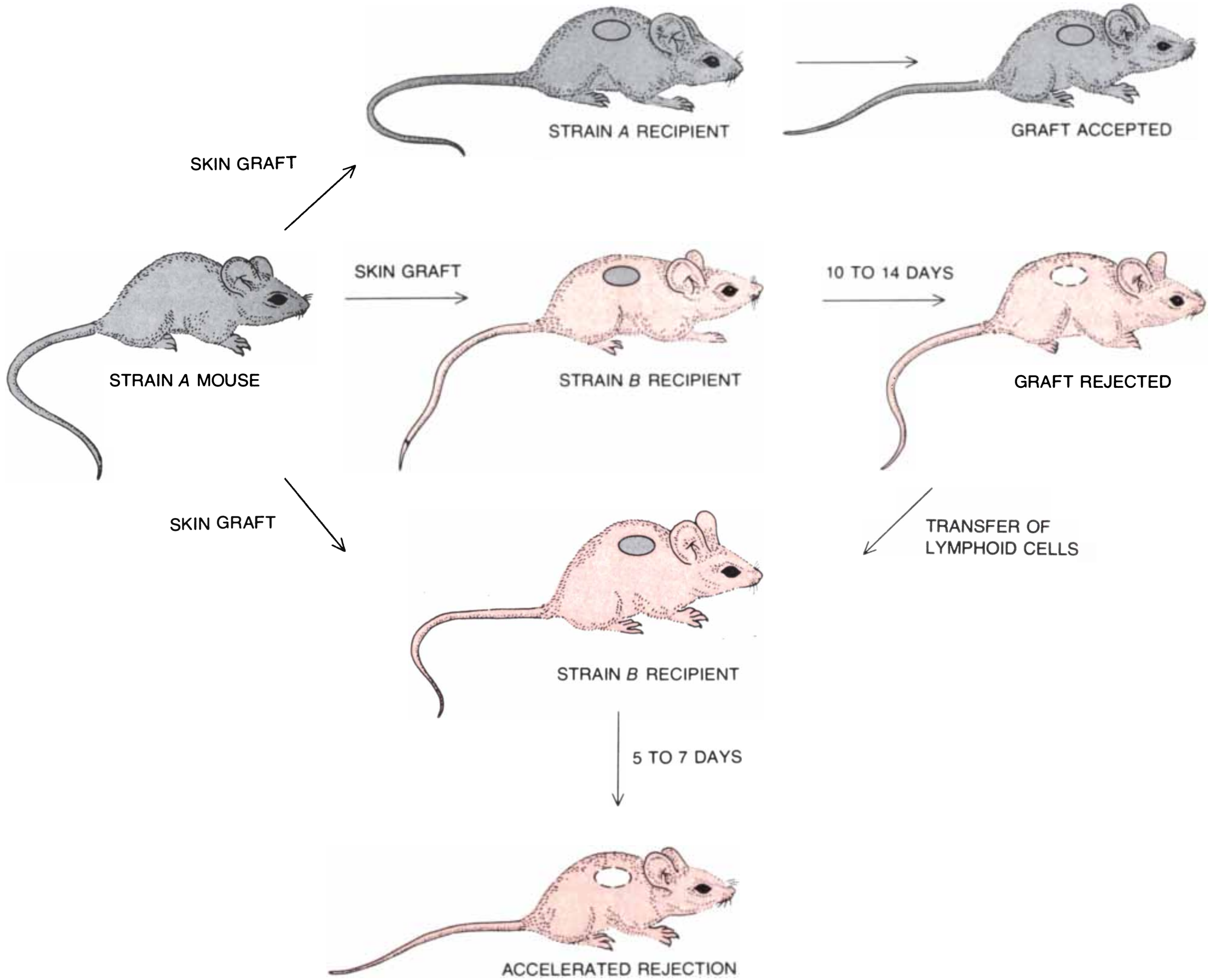
At least 10 variants of both the *H*-2*K* and *H*-2*D* genes have been described in inbred strains of mice, and studies by Jan Klein of the University of Texas Southwestern Medical School in Dallas on wild outbred strains of mice suggest that the number of variant forms of each gene may exceed 100. The antigens specified by the *H*-2*K* and *H*-2*D* genes are immunologically distinct, but the two genes have many features in common. Donald C. Shreffler and his colleagues at the University of Michigan Medical School have suggested that the two genes may have evolved from duplicate copies of an ancestral gene that through mutation gradually came to specify different proteins.

Other parts of the *H*-2 gene complex that have been identified include the *I* region and the *S* region, both of which are closely related to the immune system. The importance of the *I* region was first recognized when Baruj Benacerraf, working at the National Institute of Allergy and Infectious Diseases, Hugh O. McDevitt of the Stanford University School of Medicine and Michael Sela of the Weizmann Institute of Science showed that certain inbred mouse strains could make antibodies to a particular antigen whereas other strains could not. They subsequently found that the ability to manufacture antibodies is controlled by a series of genes within the *I* region known as immune-response (*Ir*) genes. Although the products of the *Ir* genes have not been identified, certain proteins designated *Ia* (*I*-region-associated) antigens have been found on the surface of lymphocytes, the white blood cells involved in immunological reactions. It is not yet known, however, whether the *Ia* antigens are coded for by the *Ir* genes. The *S* region, which was first detected by Shreffler and his colleagues, has been shown to control the manufacture of a component of serum complement, the system of proteins in the blood serum that acts to destroy antigenic cells such as bacteria and grafted cells once they have reacted with specific antibodies.

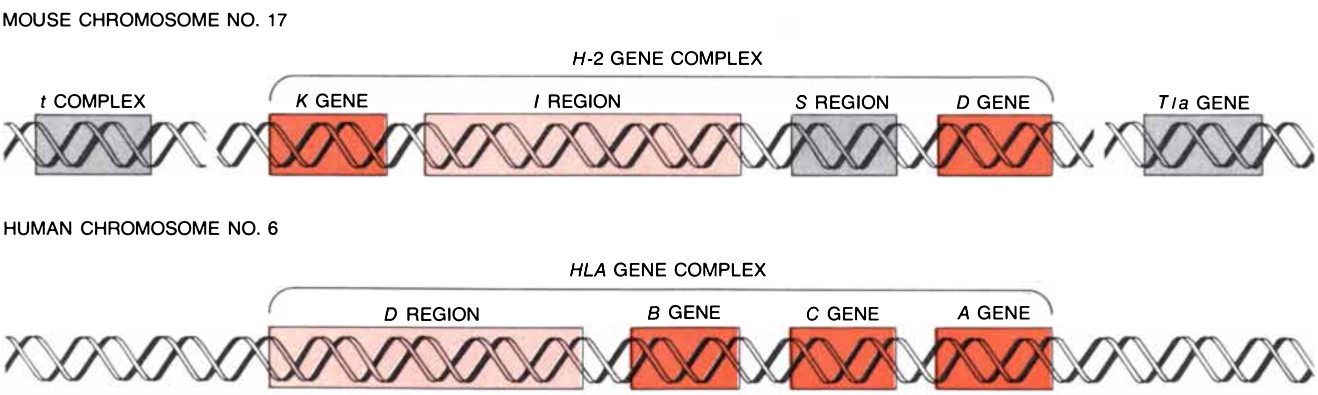
In short, the *H*-2 complex is a unique chromosomal region carrying a variety of closely related genes, all of which have some connection with the immune system. Outside the complex but also on chromosome No. 17 are at least two other regions that appear to be related to the *H*-2 genes. One of them is the *Tla* gene, which codes for the thymus-leukemia (*TL*) antigens defined by Edward A. Boyse, Lloyd J. Old and Elisabeth Stockert of the Sloan-Kettering Institute for Cancer Research [see "Cancer Immunology," by Lloyd J. Old; SCIENTIFIC AMERICAN, May]. The *TL* antigens, which appear on the surface of normal thymus cells in certain mouse strains



MAJOR HISTOCOMPATIBILITY ANTIGENS, here shown schematically (*color*), are among the vast array of proteins embedded in the bilayer of lipid molecules that constitutes the outer membrane of the cell. One part of the antigen molecule extends into the interior of the cell, where it may interact with cytoplasmic proteins; another part juts out from the cell surface and is available for interactions with other cells. Because of their genetic variability the histocompatibility antigens provide the tissues of each individual with a unique chemical identity. When skin or an organ is transplanted from one individual to an unrelated individual, the killer lymphocytes of the recipient's immune system recognize the foreign histocompatibility antigens on the transplanted cells. They then attack and destroy the transplant.



GENETICS OF TRANSPLANTATION are best investigated in mice, which can be made genetically homogeneous by intensive inbreeding. If skin is transplanted from a mouse of an inbred strain (*gray*) to a second mouse of the same strain, the graft is retained indefinitely. If, however, the skin is transplanted to a member of a genetically different strain (*color*), the graft will be rejected within two weeks. When lymphocytes from a mouse that has already rejected a graft are injected into a genetically identical mouse at the same time it receives a transplant from the unrelated donor, the graft will be rejected twice as rapidly. This accelerated rejection occurs because the transferred lymphocytes have already been sensitized to the foreign histocompatibility antigens on the surface of the transplanted cells.



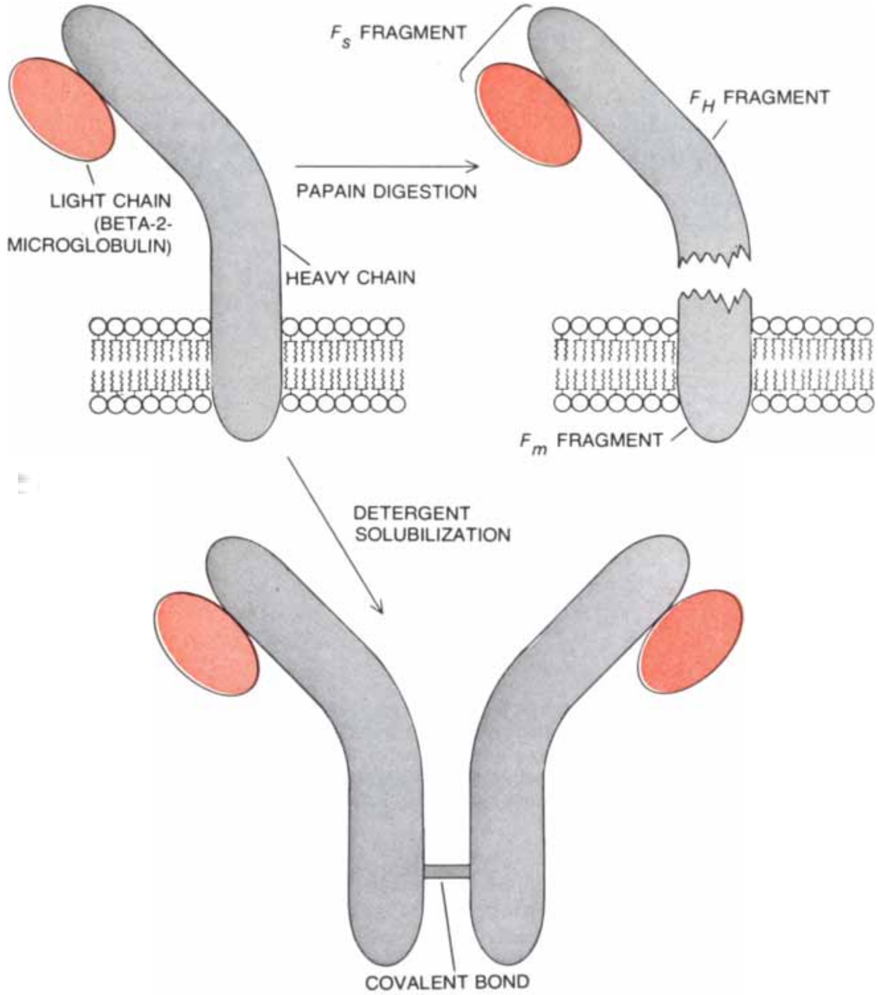
SPECIFIC CHROMOSOMAL REGIONS code for major histocompatibility antigens of mouse (the *H*-2 antigens) and man (the HLA antigens). *H*-2 gene complex (*top*) is located on chromosome No. 17 of the mouse. Located at both ends of complex are genes specifying "strong" antigens *H*-2*K* and *H*-2*D*. Between these genes are genetic regions associated with the function of immune system: the *I* region, which controls manufacture of antibodies to certain foreign antigens, and the *S* region, which controls manufacture of a component of serum complement, a system of proteins that destroys antigenic cells once they have reacted with antibodies. Regions designated *t* and *Tla* are located outside *H*-2 complex but may be functionally related to *H*-2 genes. *HLA* gene complex (*bottom*) is located on human chromosome No. 6. It has three genes coding for histocompatibility antigens (*HLA-A, HLA-B, HLA-C*) and a *D* region which is comparable to the *I* region of the mouse. Note that topography of the two gene complexes is different (*homologous regions are the same color*).

and also on the leukemia cells of nearly all strains, have many structural features in common with the *H-2K* and *H-2D* histocompatibility antigens. Moreover, the expression of the thymus-leukemia antigens on the surface of cells is closely related to the expression of *H-2* antigens. The second *H-2*-related region, designated the *t* complex, resembles the *H-2* complex in that it has many genetic variants. L. C. Dunn and Dorothea Bennett of Columbia University showed some time ago that the *t* complex plays a critical role in the embryonic development of mice. Mutations in this region can cause visible changes in the size and shape of the mouse's tail, and certain mutations in these genes are lethal.

A similar but less detailed genetic map has been obtained for the human *HLA* gene complex. Our relative lack of knowledge stems from the fact that human antibodies for detecting *HLA* antigens are more difficult to obtain in a systematic way than mouse antibodies to *H-2* antigens are. *HLA* antibodies are usually found in the serum of women who have had multiple pregnancies and have made antibodies against the paternal *HLA* antigens expressed on the cells of the fetus. Antibodies for experiments can also be obtained from people who have had multiple blood transfusions and who have accordingly been exposed to foreign *HLA* antigens on the white blood cells of the donor, or from volunteers who have been deliberately immunized with white blood cells from another person.

The *HLA* gene complex is located on human chromosome No. 6, and it has many features in common with the *H-2* complex. Three genes in the complex code for histocompatibility antigens; they are designated *HLA-A, HLA-B* and *HLA-C*. The *HLA-A* and *HLA-B* genes appear to be closely related to the *H-2D* and *H-2K* genes of the mouse. The topography of the *HLA* complex differs from that of the *H-2* complex, however, in that *HLA-D* (an area comparable to the *I* region of the mouse) is outside the *A* and *B* genes, whereas the *I* region is between the *H-2D* and *H-2K* genes.

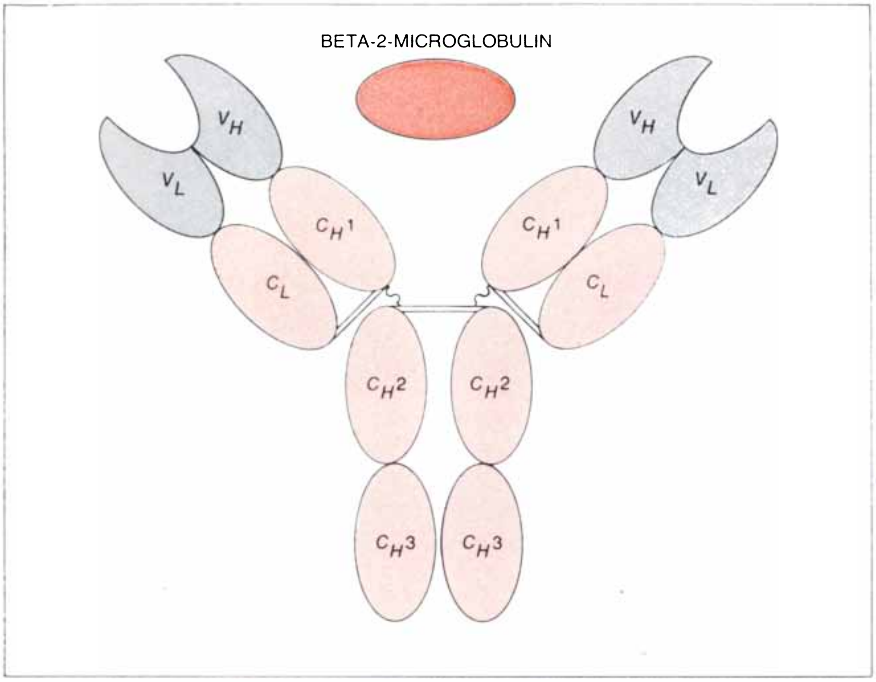
Little is known about the molecular structure of the histocompatibility antigens, primarily because, like other cell-surface proteins, they present major experimental problems to the protein chemist: they are not readily obtained in large amounts, they are not soluble in water or simple salt solution and they are intermingled with many other molecules on the cell surface. In spite of these limitations the *H-2* antigens are more amenable to structural analysis than most other cell-surface glycoproteins, for a variety of reasons. First, the numerous genetic and immunological studies of the *H-2* antigens provide an excellent background for chemical investigations. Second, antibodies can be used to isolate specific *H-2* antigens from complex mixtures of proteins. Third, the assays for *H-2* antigens, although not ideal, are well established and reliable. As a result the analysis of the molecular structure of *H-2* antigens, in addition to its value in understanding the histocompatibility system, is becoming the paradigm for future analyses of other cell-surface glycoproteins. Together with Gerald M. Edelman, Roland Henning, Robert J. Milner and Konrad Reske of Rockefeller University I have been employing a variety of methods for characterizing proteins to obtain a more precise model for the structure of *H-2* antigens, both in solution and on the cell surface.



STRUCTURE OF *H-2* ANTIGENS differs depending on the manner in which they are removed from the cell membrane. When the *H-2* molecules are extracted with detergent, they appear to be a complex of two "heavy" (long) and two "light" (short) polypeptide chains. The two heavy chains are linked by a covalent bond, but the heavy and light chains are held together only by weak, noncovalent interactions. When the *H-2* molecules are snipped from the cell membrane with the protein-cleaving enzyme papain, one obtains a single fragment designated F_s, which consists of part of the heavy chain (F_H) and an entire light chain. The shorter fragment of the heavy chain, designated F_m, probably remains associated with the cell membrane. On the cell surface the covalent bond between two heavy chains found in detergent solution is not present. Thus the structure of *H-2* antigens may be a complex of one light and one heavy chain, a complex of two light and two heavy chains or a mixture of the two forms.

The *H-2* antigen molecule is a complex of polypeptide chains, each consisting of a linear sequence of amino acid subunits linked by peptide bonds. Two kinds of constituent polypeptide chains have been identified: "heavy" (long) chains and "light" (short) ones. As has been shown by Stanley Nathenson and his colleagues at Albert Einstein College of Medicine in New York, *H-2* molecules can be obtained for structural studies either by disrupting the cell membrane with a detergent or by using the protein-cleaving enzyme papain to chemically snip the molecule away from the cell surface.

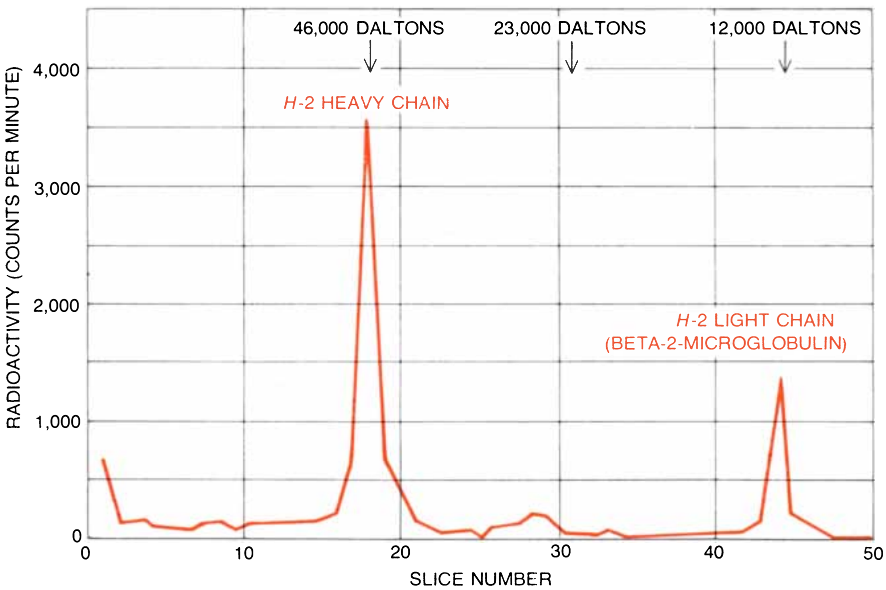
When the *H-2* molecule is extracted with detergent, it will stay in solution only if the detergent is still present. We have found that under these conditions the *H-2* molecule consists of two identical heavy chains and two identical light ones. At least one covalent chemical bond holds the two heavy chains together. The light chains are probably held to



STRUCTURAL SIMILARITY between beta-2-microglobulin (the light-chain component of *H*-2 and *HLA* antigens) and the constant regions of antibody molecules (*light color*) is suggested by the homologies in their amino acid sequences. This finding implies that the histocompatibility antigens are genetically related to antibodies and raises the question of whether the *H*-2 molecule is folded into functional domains similar to those of the antibody molecule.

the heavy chains only by weaker, noncovalent forces such as hydrogen bonds and electrostatic forces. When the molecule is extracted with papain, a water-soluble fragment of the *H*-2 molecule, designated F_s, is released. A similar fragment is released when the detergent-solubilized material is treated with papain. In both instances the F_s fragment consists of one complete light chain in association with a fragment (F_H) of the heavy chain. A smaller portion of the heavy chain, designated F_m, probably remains associated with the cell membrane after the heavy chain has been cut by papain. This smaller fragment has not been isolated, but it apparently includes the region of the heavy chain that interacts with the membrane, together with the amino acid subunits that form the covalent bonds between two heavy chains in detergent solution.

The *H*-2 molecules on cells differ from those in detergent solution in that they lack any covalent bonds between the heavy chains. This finding suggests the following possibilities for the structure of the *H*-2 molecules on the cell surface: (1) a complex consisting of one light chain and one heavy chain, (2) a complex consisting of two light chains and two heavy chains or (3) a combination of both. These possibilities are functionally relevant because they imply that *H*-2 antigens on cells may be available to interact with other molecules as they interact with one another in detergent solution.



SEPARATION of the heavy and light polypeptide chains of *H*-2 antigens is achieved by electrophoresis. Radioactively labeled *H*-2 molecules are applied to the top of a cylindrical gel of polyacrylamide, and an electric field is placed across the gel. The heavy and light chains migrate toward the bottom of the gel at different rates according to their molecular weight. Gel is then sliced into thin sections, and the radioactivity of each section is measured. Two peaks of radioactivity, corresponding to the bands of purified *H*-2 heavy and light chains, are obtained.

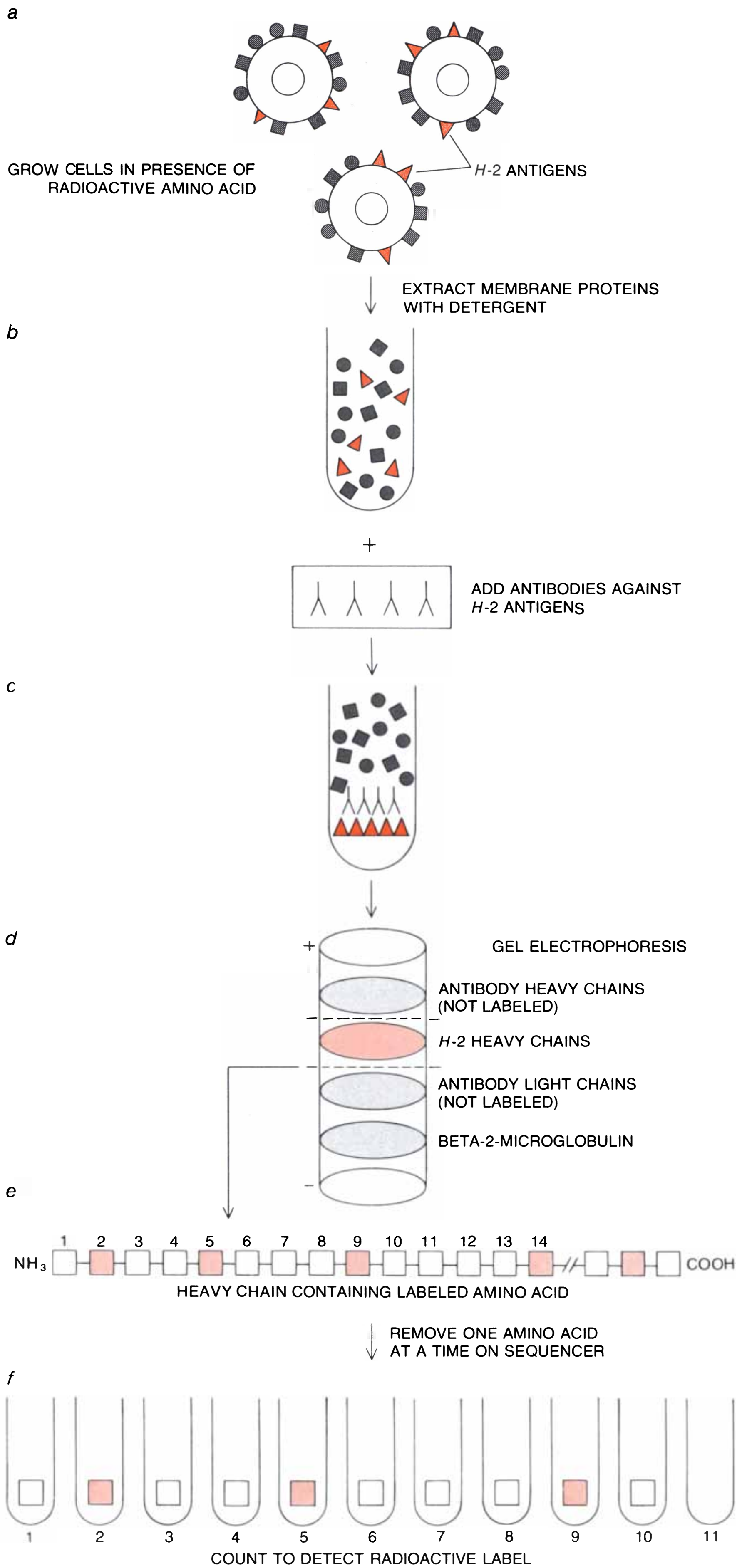
Recent investigations in a number of laboratories have identified the light polypeptide chain of *H*-2, *HLA* and *TL* antigens as beta-2-microglobulin, a protein already well known to biochemists. This molecule incorporates no carbohydrate and appears to carry none of the antigenic determinants characteristic of a given *H*-2 or *HLA* type. Hence all the carbohydrate and the antigenic specificities must be carried by the heavy chains.

Human beta-2-microglobulin was first discovered by Ingemar Berggård and Alexander G. Bearn at the Rockefeller University in 1968, in the urine of patients with kidney malfunctions. Since then the protein has been characterized in detail and has been shown to be manufactured by nearly every cell in the body. It has also been described in other species, including the mouse, the dog, the rabbit, the guinea pig and the rat. In contrast to the variability of the heavy chains of histocompatibility antigens, no variant of beta-2-microglobulin has been detected within a species by either immunological or chemical techniques. Indeed, the amino acid sequences of beta-2-microglobulins obtained from several species have been partially determined, and they show only a few differences, about the number one would expect for functionally similar proteins from different species. This similarity suggests that the structure of beta-2-microglobulin in man is representative of the structure of the molecule in all other vertebrate species, including the mouse.

The most striking feature of beta-2-microglobulin is its resemblance to portions of antibody molecules. It has been known for some time that the light and heavy polypeptide chains that make up antibody molecules have two types of regions: variable regions that differ in amino acid sequence from one molecule to another, and constant regions that are invariant in amino acid sequence for any class of heavy chains or light chains. Three segments in the constant region of the antibody heavy chain resemble one another and the constant region of the light chain in size, overall structure and amino acid sequence. On the basis of these observations Edelman proposed that both the heavy and light polypeptide chains of antibodies are folded into compact domains corresponding to the variable and constant regions. This hypothesis has now been verified by structural analysis utilizing the technique of X-ray crystallography [see "The Antibody Combining Site," by J. Donald Capra and Allen B. Edmundson; SCIENTIFIC AMERICAN, January]. In addition, these data, together with genetic evidence, suggest that the genes coding for antibody molecules arose in the course of evolution through the duplication of an ancestral gene large enough to code for a single antibody domain, a polypeptide some 100 amino acids long.

The polypeptide chain of beta-2-microglobulin is also some 100 amino acids long, and it resembles the constant domains of antibodies in many of its structural features. Studies by Oliver Smithies at the University of Wisconsin, in collaboration with Dave Poulik of Wayne State University, and our own determination of the complete amino acid sequence of human beta-2-microglobulin, have shown that it is nearly as closely related to the constant domains of antibodies as they are to one another. On the basis of this evidence we have surmised that the gene for beta-2-microglobulin arose from the same ancestral gene that gave rise to the genes for antibody molecules.

The fact that beta-2-microglobulin is associated with both *HLA* and *H*-2 antigens suggests a possible evolutionary relation between antibody molecules and histocompatibility antigens. In addition,

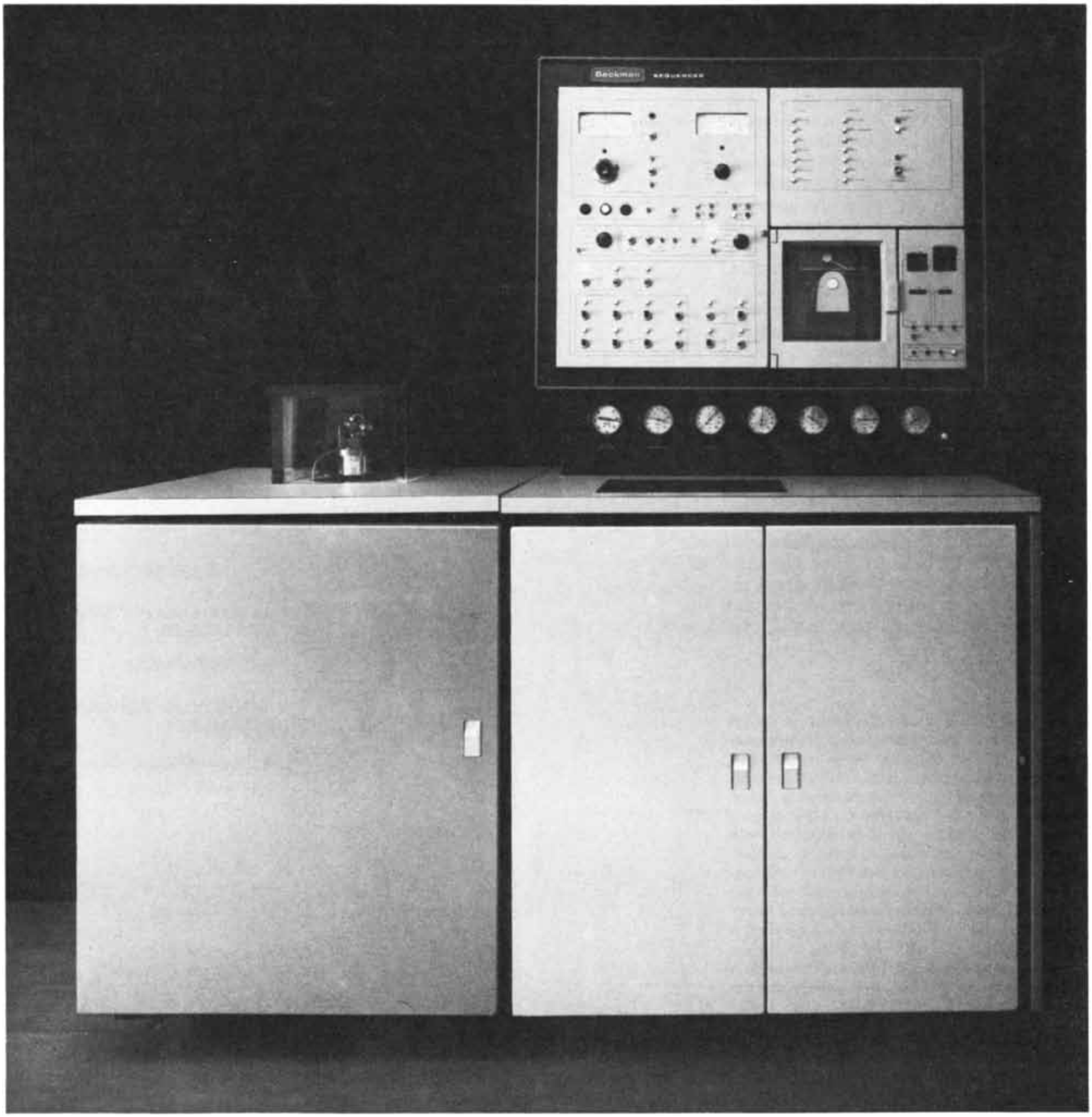


SEQUENCING of the *H*-2 heavy chain requires a new microchemical analytic technique. Mouse cells are cultured in the presence of a radioactive amino acid that is incorporated into all newly manufactured proteins, including the *H*-2 antigens (*a*). The labeled membrane proteins are then extracted with detergent (*b*), and the *H*-2 antigens are isolated by precipitating them with specific antibodies (*c*). Next the antigen molecules are resolved into their component light and heavy chains by electrophoresis across a polyacrylamide gel (*d*). The purified heavy chains are eluted from the gel, and the positions at which the radioactive amino acid appears along the chain are determined by removing one amino acid at a time from the protein with the aid of an automatic machine called a sequencer (*e*). Finally the radioactivity of the sequentially removed amino acids is measured to detect the presence or absence of the radioactive label (*f*). The entire procedure must be repeated many times to determine locations along the polypeptide chain of each of the 20 amino acids from which protein molecules are built.

it raises the question of whether or not the heavy and light polypeptide chains of *HLA* and *H*-2 antigens are folded into functional domains similar to those seen in antibodies. The answers to these questions should be provided by the determination of the amino acid sequence of the *H*-2 heavy chains and by the X-ray crystallographic analysis of the three-dimensional structure of beta-2-microglobulin, both of which are now under way in our laboratory.

The analysis of the amino acid sequence of *H*-2 heavy chains is far more difficult than that of the amino acid sequence of beta-2-microglobulin. For one thing, the intact *H*-2 heavy chain must be extracted from cells with detergent and kept in detergent solution. For another, in order to get enough *H*-2 heavy chains to determine their amino acid sequence by conventional sequencing techniques such vast numbers of cells are required that this approach has been too costly to be practical. We and others in the field have therefore developed a microchemical approach for determining the amino acid sequence of a protein when only minute amounts of it are available.

The method involves incorporating radioactively labeled molecules of a given amino acid into *H*-2 heavy chains so as to identify all the positions along the polypeptide chain where that amino acid is found. Specifically we grow mouse spleen cells or mouse tumor cells in a medium containing an amino acid, say tyrosine, that has been labeled by replacing the hydrogen atoms in the molecule by atoms of tritium, the radioactive isotope of hydrogen. The tritium-labeled tyrosine is then taken up by the cells and incorporated into all newly synthesized proteins including the *H*-2 antigens, which are constantly being broken down and resynthesized. Next the cells are disrupted, and the mem-



AUTOMATIC SEQUENCER in the author's laboratory at Rockefeller University can determine the amino acid sequence of a polypeptide chain for up to 50 consecutive amino acids. The machine makes use of carefully controlled conditions to carry out a reaction that sequentially removes one amino acid at a time from the chain. The entire reaction sequence is controlled by a paper-tape program.

CLOSEUP VIEW shows the reaction chamber of the sequencer. The glass cup spins throughout the reaction sequence and is kept at a constant temperature by means of a heating element. Plastic inlet tubes on the delivery head (*lifted up for viewing*) allow the introduction of solutions and nitrogen into the chamber. At the start of the reaction the polypeptide sample is introduced as a film on the inner surface of the spinning cup. The reagent phenyl isothiocyanate is then allowed to couple to the end of the polypeptide chains, and a single amino acid is removed from the chains by exposing the material to acidic conditions. Finally the released amino acid derivative is extracted from the reaction mixture with an organic solvent, and it is sucked out of the spinning cup through the *L*-shaped tube on the side of the delivery head (*top*). Reaction sequence then begins anew to remove next amino acid from the chain. Sequentially removed amino acids are later identified with a variety of chromatographic techniques.

brane proteins are solubilized with detergent. Specific antibodies to the *H*-2 antigens are added to the mixture of solubilized proteins so as to remove the *H*-2 molecules as a precipitate.

This precipitate of *H*-2 antigens bound to specific antibodies is concentrated by spinning the solution in a centrifuge. It is resolubilized in detergent and separated into its component molecules by high-voltage electrophoresis across a gel of polyacrylamide. In this process the *H*-2 heavy and light chains separate and migrate at different rates through the gel, separating into narrow bands on the basis of size. The region of the gel containing the band of radioactively labeled *H*-2 heavy chains is isolated, and the molecules are eluted from it. Thus the end products of the purification are *H*-2 heavy chains containing a radioactive tyrosine subunit at each position where tyrosine normally appears in the amino acid sequence. The positions of other amino acids along the chain can be determined by incubating the cells in a medium containing tritium-labeled molecules of each amino acid.

The exact positions at which the labeled amino acid appears are determined with the aid of an automatic machine called a sequencer, which chemically removes one amino acid at a time from the polypeptide chain beginning at the amino terminal (NH_2) end. The chemical reaction required was developed by the late Pehr Edman at the University of Lund, who in collaboration with Geoffrey Begg at St. Vincent's Hospital for Medical Research in Melbourne, Australia, subsequently devised an automatic system to precisely control the reaction conditions. The yield for the removal of each amino acid from the chain exceeds 95 percent, and amino acid sequences can be determined for 30 to 50 consecutive amino acid subunits, which is much farther along a polypeptide chain than can be attained with manual analytic methods. Usually the derivatives of amino acids released by the Edman reaction are identified by a variety of chromatographic separation techniques. Because of the minute quantities of protein dealt with in our studies, however, we can identify each amino acid only by the presence of its radioactive label.

With this microchemical approach we can detect amino acids in the *H*-2 heavy chain at the level of about 10^8 molecules, whereas conventional sequencing techniques call for perhaps a million times more (10^{14}) molecules for accurate detection. The fact remains that microchemical sequencing is particularly time-consuming because accurate identification of each labeled amino acid requires that we work on only one or two amino acids in each experiment. A separate preparation of purified *H*-2 heavy chains must therefore be made to determine the locations along the chain of each of the 20 amino acids of which proteins are built. Moreover, some amino acids cannot be detected because they are incorporated into the protein in insufficient amounts, either because they are not efficiently transported into the cells or because the radioactively labeled molecules are diluted in a large pool of similar unlabeled molecules within the cells. Our sequences therefore have a number of gaps.

In spite of these limitations, over the past two years the sequences of the first 27 amino acids in several *H*-2 heavy chains have been obtained by us, by Nathenson and his colleagues, by Jack Silver and Leroy E. Hood at the California Institute of Technology, and by a research group headed by Jonathan W. Uhr at the University of Texas Southwestern Medical School. Although the data are limited, they already provide valuable information on the kind of ge-

AMINO ACID POSITION ALONG HEAVY CHAIN

	MAJOR ANTIGENS	GENETIC VARIANTS					5					10					15					20					25		
MOUSE	*H-2K*	K^b	■	Pro	His		Leu	Arg	Tyr	Phe	Val	Thr	Ala	Val		Arg	Pro		Leu			Pro	Arg	Tyr				Leu	Tyr
		K^d	Met	■	His		■	Arg	Tyr	■	■	Thr		■		Arg	Pro		■			Pro	Arg	Phe					Tyr
		K^k	Met	Pro	His		Leu	Arg	Tyr	Phe	His		Ala	Val		Ile	Pro		Leu		Lys	Pro	Phe	Ala					Tyr
	H-2D	D^b		Pro				Arg	Tyr				Ala	Val		Arg	Pro		Leu			Pro	Arg	Tyr					Tyr
		D^d	Met	■	His		Leu	Arg	Tyr	Phe	Val	Thr	Ala	Val	Thr	Arg	Pro		■			Pro	Arg	Tyr					Tyr
MAN	*HLA-A*	*A*2	Gly	Ser		Ser	Met	Arg	Tyr	Phe	Phe	Thr	Ser	Val	Ser	Arg	Pro	Gly		Gly	Glu			Phe	Ile	Ala	Val		
		*A*1,2		Ser		Ser	Met	Arg	Tyr	Phe	Phe	Thr	Ser	Val	Ala	Arg	Pro	Gly											
	HLA-B	*B*7	Gly	Ser		Ser	Met	Arg	Tyr	Phe	Tyr	Thr	Ser	Val	Ser	Arg	Pro	Gly		Gly	Glu			Phe	Ile	Ala	Val		
		*B*8,13		Ser		Ser	Met	Arg	Tyr	Tyr	Tyr	Ser	Ala	Val	Ser	Arg	Pro	Gly											

Ala ALANINE	Ile ISOLEUCINE	Pro PROLINE
Arg ARGININE	Leu LEUCINE	Ser SERINE
Glu GLUTAMIC ACID	Lys LYSINE	Thr THREONINE
Gly GLYCINE	Met METHIONINE	Tyr TYROSINE
His HISTIDINE	Phe PHENYLALANINE	Val VALINE

- □ AMINO ACIDS AT THESE POSITIONS NOT YET DETERMINED
- AMINO ACIDS SHARED BY BOTH *H*-2 and *HLA* ANTIGENS
- AMINO ACIDS SHARED BY *H-2K* AND *H-2D* ANTIGENS
- AMINO ACIDS SHARED BY *HLA-A* AND *HLA-B* ANTIGENS
- ■ AMINO ACIDS NOT DETERMINED BUT KNOWN NOT TO BE AN AMINO ACID SEEN IN OTHER CHAINS AT THIS POSITION

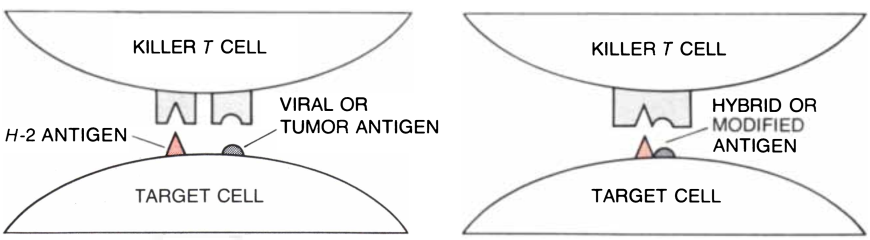
PARTIAL AMINO ACID SEQUENCES of the heavy polypeptide chains of *H*-2 and *HLA* antigens reveal broad regions of similarity along with distinct differences that may reflect the overall genetic variability of these molecules. The differences in sequence are well distributed over the length of the chain segments, suggesting that different antigenic specificities are determined by large regions of the molecule. This is in contrast to the antibody molecule, where the variable amino acid sequences are confined to a segment of each chain.

netic variation that is found in heavy chains, together with clues to the origin of the *H*-2*D* and *H*-2*K* genes.

The amino acid sequences of all the *H*-2 antigens examined so far are quite similar. Products of the *H*-2*K* genes so closely resemble products of the *H*-2*D* genes that there is no feature in the amino acid sequence that distinguishes one type of antigen from another. These results support Shreffler and Klein's hypothesis that the *H*-2*K* and *H*-2*D* genes are derived from a common ancestral gene and also the hypothesis that all the genes in the *H*-2 complex are closely related.

The amino acid sequences in the genetic variants of the *H*-2 heavy chain are similar, but distinct differences can be recognized. For example, *H*-2 antigens can have different amino acids at positions 1, 2, 8, 9, 14, 17, 22 and 23. We have not yet been able to correlate such differences in amino acid sequence with the different antigenic specificities of the *H*-2 variants, but the two properties are undoubtedly related. The differences in amino acid sequence appear to be well distributed along the sequence and are not confined to a specific short segment. Thus the different antigenic specificities that reflect the genetic variability of the *H*-2 antigens appear to be generated by amino acid substitutions involving a large portion of the *H*-2 heavy chain.

An important question raised by the study of beta-2-microglobulin is whether or not the *H*-2 heavy chain is similar in amino acid sequence to the variable or constant regions of antibody molecules. We have compared our sequences of *H*-2 heavy chains with those of various antibody molecules, but we have found no convincing similarities. The data on the *H*-2 heavy chains are limited, however, and confined to a short region near the amino terminal end of the molecule, so that we shall have to wait for more complete sequence determinations before we can obtain a definitive answer.



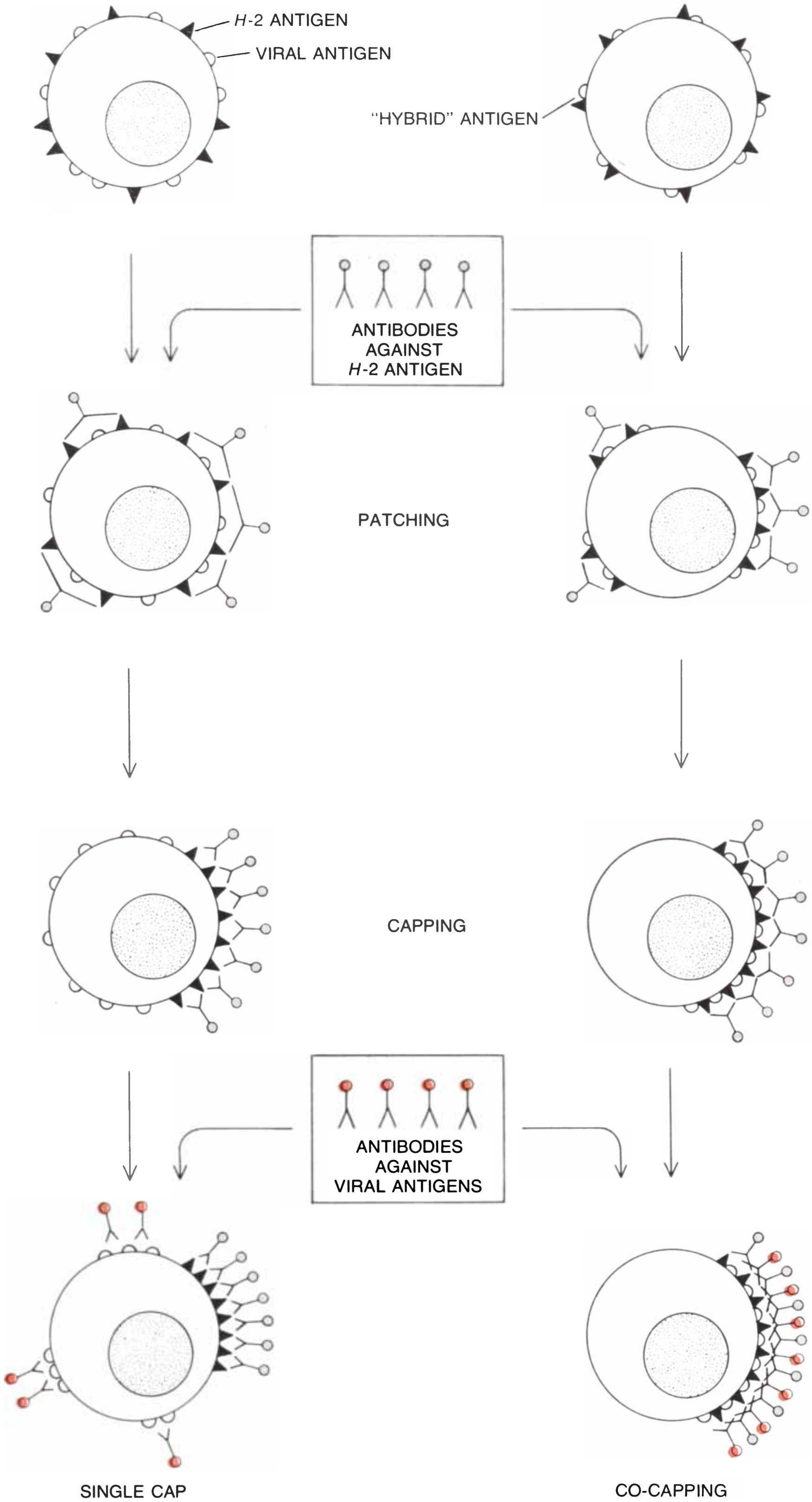
ALTERNATIVE MODELS explain the observation that lymphocytes called killer *T* cells, which can selectively destroy virally infected or cancerous body cells, have to detect at least one *H*-2 antigen molecule on the surface of the target cell before killing can occur. The recognition process may involve two separate binding molecules, or receptors, on the surface of the killer *T* cell (one for the *H*-2 antigen and one for the viral antigen). Alternatively a single receptor on the killer cell recognizes a hybrid structure formed by the association of the two antigens.

Paralleling the work on *H*-2 heavy chains is an equally intensive effort to determine the structure of *HLA* heavy chains. Jack L. Strominger and his colleagues at Harvard University and a group of English investigators from the laboratories of John Bridgen of the University of Cambridge, Walter Bodmer of the University of Oxford and Michael Crumpton of the National Institute for Medical Research at Mill Hill, have approached the problem by growing in culture large quantities of human cells from lines that have an unusually large amount of *HLA* antigens on their surface. Both of these research groups have managed to obtain sufficient quantities of *HLA* heavy chains to determine partial amino acid sequences by conventional methods, without the need for incorporating radioactive amino acids into the polypeptide chains.

Comparison of the *H*-2 and *HLA* sequences that have been obtained so far satisfy the expectation that they would be very similar. This finding is significant for two reasons. First, it supports the hypothesis that the major histocompatibility antigens of man and mouse are closely related and probably serve the same biological function. Investigations of the mouse *H*-2 system will therefore be directly applicable to the *HLA* system. Second, the similarities in the *H*-2 and *HLA* sequences validate the microchemical technique we have employed for the sequence analysis of the *H*-2 antigens, since the *HLA* sequences were determined by an entirely different method. This verification is important because it shows that the method is suitable for studying any cell-surface protein to which antibodies can be prepared. The challenge remains, however, to extend this approach so as to determine the amino acid sequences of larger regions of the *H*-2 molecule. By determining the sequence of a large portion of an *H*-2 heavy chain we hope to distinguish those features of the *H*-2 molecule that have implications for understanding its biological function.

As I indicated at the beginning of this article, histocompatibility antigens cannot have evolved simply to thwart the efforts of the transplantation surgeon. What, then, can be said about their biological function in the host? The most promising explanations come from a consideration of the possible role of *H*-2 antigens in defending the body against abnormalities among its own cells. According to the theory of immunological surveillance, deleterious mutations, such as those that cause cancer and other cellular derangements, are constantly arising within the body. Such mutations cause changes in the structure of the cell surface that are subsequently recognized by the immune system, specifically a class of wandering lymphocytes that constantly scan the body's own cells and tissues. If these lymphocytes encounter a defective cell, they carry the alarm back to the lymph nodes, alerting a class of lymphocytes known as killer *T* cells to attack and destroy the defective cells. The killer *T* cells then destroy these cells in much the same way that they kill foreign cells during graft rejection.

There are, however, some important differences between graft rejection and the postulated mechanism of immunological surveillance. In graft rejection



CAPPING EXPERIMENT makes it possible to determine whether *H*-2 antigens and viral antigens are closely associated on the surface of tumor cells. First, antibodies specific for *H*-2 antigens and those specific for viral antigens are labeled with the dyes fluorescein and rhodamine, which glow respectively green and red under the fluorescence microscope. Then the antibodies against the *H*-2 antigens are added to the cells, linking the *H*-2 antigens on the cell surface to form patches and finally a cap at one end of the cell. The cells are then fixed, so that no new caps can form, and the antibodies against the viral antigens are added. If the two antigens are independent, the *H*-2 antigens will form a cap but the viral antigens will remain spread over the cell surface (*a*). If the two antigens are associated, they will form a cap together (*b*).

the targets of the killer *T* cells are clearly the foreign *H*-2 antigens of the donor, whereas in the case of the body's own abnormal cells the *H*-2 antigens on the *T* cells and the target cells are identical. In attacking the host's own abnormal cells the *T* cells must therefore recognize some feature such as a component of a virus (a viral antigen) or a component unique to a tumor cell (a tumor-specific antigen) that allows them to distinguish the altered cells from normal cells.

Until recently no direct evidence was available to indicate that *H*-2 antigens might play a role in this recognition process. Then Rolf Zinkernagel and Peter Dougherty of the Australian National University discovered that killer *T* cells could not destroy cells infected with virus unless the killer cell and the target cell had at least one *H*-2 antigen in common. John Schrader, Edelman and I found the same to be true of the killing of cancer cells by killer *T* cells. When antibodies that bind to the shared *H*-2 antigens were added to the incubation medium, the killer lymphocytes were unable to destroy the cancer cells, presumably because they could no longer recognize the *H*-2 antigens. Additional experiments showed that it was the *H*-2 antigens on the target cells and not those on the killer *T* cells that had to be blocked by antibodies in order to prevent the destruction of the target cells.

Two general mechanisms have been proposed to explain the role of *H*-2 antigens in the destruction of abnormal cells by killer *T* cells. One set of hypotheses suggests that the killer lymphocytes carry two kinds of binding molecules, or receptors: one receptor that binds to *H*-2 antigens and a second receptor that binds to the abnormal (viral or tumor) antigen. On this hypothesis both receptors must bind if the abnormal cell is to be killed. Another set of hypotheses suggests that the killer cell has only one kind of receptor, but that the receptor selectively binds to a hybrid antigen consisting of an *H*-2 antigen bound to an abnormal antigen. Such a close interaction between two types of antigens is theoretically possible because it is now known that the cell membrane is not a rigid structure but rather a semifluid one, with antigens and other proteins resembling icebergs afloat in a sea of membrane lipids. The *H*-2 and abnormal antigens are thus in constant motion over the cell surface, and they could easily come in contact through random interactions.

One way to determine whether *H*-2 antigens and abnormal antigens really do find each other and specifically associate on the cell surface is to perform what is known as a capping experiment. If antibodies that bind to cell-surface molecules such as *H*-2 antigens are added to a suspension of cells, the antibodies link the antigen molecules together, aggregating them into patches on the cell surface. Eventually all the patches fuse at one end of the cell to form a cap. The patches and caps are detected with the aid of antibodies labeled with a fluorescent dye that glows with a distinctive color in the fluorescence microscope. Dyes of different colors (such as fluorescein, which is green, and rhodamine, which is red) can serve to label antibodies with different specificities, so as to determine the distribution of different types of antigens on the same cell.

In order to find out if two different antigens are closely associated on the cell surface one first causes one type of antigen to form a cap by adding antibodies with a green label and then fixes the cells so that no new caps can form. Next one treats the cells with antibodies specific to the second type of antigen and bearing a red label. If in the fluorescence microscope the red and green labels are both associated with the cap, one can assume that the two kinds of antigen first physically associated and then formed the cap together. If, on the other hand, the green label appears as a cap and the red label is spread diffusely over the cell surface, one can assume that the antigens are independent.

We have performed this kind of test with *H*-2 antigens and viral antigens on the surface of tumor cells that react with antibodies to the Rauscher leukemia virus. The two types of antigen do form a cap in the same place on the cell surface, suggesting that they are closely associated. Although the experiment is not definitive because it depends on the specificity of the antibodies involved, we and others have confirmed the observation by somewhat different methods. We therefore favor the hypothesis that *H*-2 and foreign antigens can form hybrid structures that are recognized preferentially by killer *T* cells. A corollary of this hypothesis is that *H*-2 antigens have special structural features that enable them to interact on the cell surface with viral or tumor antigens. On a more speculative note, the genetic variability of the *H*-2 system could conceivably affect both the interaction of *H*-2 antigens with abnormal antigens and their detection by killer *T* cells.

It is still a matter of debate whether the one-receptor or two-receptor model of killer *T*-cell recognition is the correct one, but in any event it is likely that our thinking will continue to focus on the role of *H*-2 antigens in mechanisms for immunological surveillance. Whatever the detailed function of the histocompatibility antigens turns out to be, that function will surely be of great importance to immunology and may have profound practical implications for tissue transplantation and the control of infectious and malignant disease. In the modern era medicine and molecular analysis have become companions in both understanding and practice.

The Solution of the Four-Color-Map Problem

Four colors suffice to color any planar map so that no two adjacent countries are the same color. This famous conjecture has been proved true by a new kind of proof, one that relies on high-speed computers

by Kenneth Appel and Wolfgang Haken

In 1852, a few months after he had completed his studies at University College London, Francis Guthrie wrote a letter to his brother Frederick, who was still at the college as a student of the mathematician Augustus De Morgan. Francis pointed out to Frederick that it seemed that every map drawn on a sheet of paper can be colored with only four colors in such a way that countries sharing a common border have different colors. He asked if there was any way to prove this mathematically. Frederick did not know, and he asked De Morgan, who did not know either. For the next 124 years Guthrie's four-color problem, the problem of proving that every map requires at most four colors or of drawing a map that requires five colors, intrigued professional mathematicians, amateur mathematicians and high school students who felt that all unsolved problems remain so only because of the incompetence of the older generation.

In 1976 we solved the four-color problem. Guthrie's conjecture was proved mathematically, but in a way quite different from what he might have expected. Even among present-day mathematicians those who were not aware of the developments leading to the proof are rather dismayed by the result because the proof made unprecedented use of computers; the computations of the proof make it longer than has traditionally been considered acceptable. In fact, the correctness of the proof cannot be checked without the aid of a computer. Moreover, some of the crucial ideas of the proof were perfected by computer experiments. Of course, a short proof of the four-color theorem may some day be found, perhaps by one of those bright high school students. It is also conceivable that no such proof is possible. In this case a new and interesting type of theorem has appeared, one which has no proof of the traditional type.

In spite of the novel aspects of the proof, both the four-color problem and the basic method of proof have deep roots in mathematics. We will begin to examine them by returning to the initial formulation of the problem in Francis Guthrie's letter. By "neighboring" countries Guthrie must have meant countries adjacent along a borderline rather than at a single point; otherwise a map whose countries look like the wedges of a pie would require as many colors as there are countries. By "country" he certainly must have meant a connected region, because if a country is allowed to consist of more than one region and the regions are separate, it is not at all hard to construct an example of a map with five countries each of which is adjacent to each of the other four [*see top illustration on page 111*].

Guthrie and De Morgan certainly realized that a map with four countries can be drawn in which each country is adjacent to the other three [*see left side of lower illustration on page 111*]. Such a map requires four colors and therefore a three-color conjecture is false. Three colors will not suffice to color all maps.

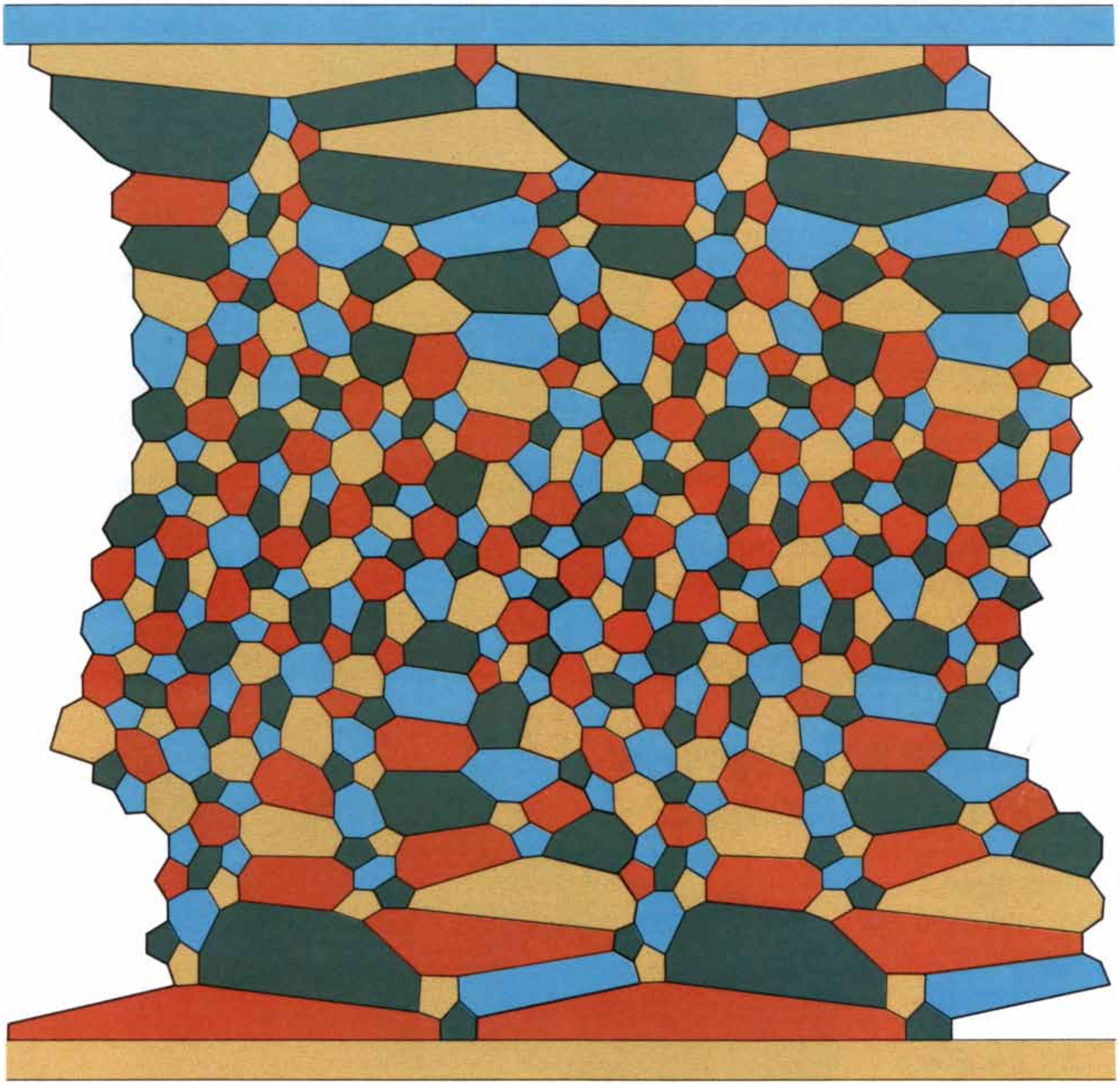
De Morgan proved that it is not possible for five countries to be in a position such that each of them is adjacent to the other four. This led him to believe that five colors would never be needed and thus that the four-color conjecture was true. Proving that five mutually adjacent countries cannot exist in a map does not prove the four-color conjecture, however [*see lower illustration on page 111*]. Many amateur mathematicians, not understanding this fact, have independently discovered proofs of De Morgan's result and have then thought that they had proved the four-color conjecture.

In 1878 the mathematician Arthur Cayley, unable to prove or disprove the four-color conjecture, presented the problem to the London Mathematical Society. Less than a year later Alfred Bray Kempe, a barrister and member of the society, published a paper purporting to show that the conjecture is true. Kempe's argument was extremely clever, and although his proof turned out to be incomplete, it contained most of the basic ideas that led to the correct proof a century later. Kempe tried to prove the conjecture true by the classical method of *reductio ad absurdum;* he assumed that the conjecture is false (that is, that there is at least one map which requires five colors) and then proceeded to show that this assumption leads to a contradiction. Reaching a contradiction shows that the original assumption (some maps require five colors) is wrong and therefore the four-color conjecture (four colors are always enough) is true.

Kempe began by defining normal maps: A map is normal if none of its countries enclose other countries and if no more than three countries meet at any point [*see illustration on page 112*]. He then showed that if there were a map that required five colors, a "five-chromatic" map, then there would have to be a normal five-chromatic map. Thus to prove the four-color conjecture it is sufficient to prove that a normal five-chromatic map is not possible. Kempe noted that if there were a normal five-chromatic map, then there would have to be such a map with a smallest number of countries, a "minimal normal five-chromatic" map. (In other words, any map with fewer countries than the minimal five-chromatic map can be colored with four or fewer colors.) Therefore to prove the four-color conjecture it is sufficient to prove that a minimal normal five-chromatic map is impossible, that is, that postulating the existence of a minimal normal map requiring five colors leads to a contradiction.

Kempe approached a contradiction as follows. He proved that any normal map must have some country with five or fewer neighbors. He then argued that if a minimal normal five-chromatic map has a country with fewer than six neighbors (which, as he had just shown, every normal map must have), then there would have to be a normal map with fewer countries that is also five-chromatic. If this argument had been totally correct, a contradiction would have been reached. By assuming the existence of a minimal five-chromatic map Kempe would have proved the existence of a smaller five-chromatic map. That would contradict the definition of a minimal five-chromatic map, and so no such map would be possible. Since that implies that there can be no five-chromatic map at all, the proof would have been complete. Kempe proved correctly that a country with two, three or four neighbors existing in a minimal five-chromatic map leads to a contradiction, but his proof of the case of five neighbors was faulty. In our proof of the four-color theorem we corrected Kempe's flawed analysis of the last case by examining some 1,500 arrangements of countries. Our methods were basically extensions of valid parts of Kempe's proof that have been the object of great attention and refinement by mathematicians over the past 100 years.

Kempe had shown that in every normal map there is at least one country with two, three, four or five neighbors. (In other words, there are no normal maps on a plane in which every country has six or more neighbors.) This may be expressed by the statement that the set of "configurations" consisting of a country with two neighbors, a country with three neighbors, a country with four neighbors and a country with five neigh-



MAP OF 846 COUNTRIES created by Edward F. Moore of the University of Wisconsin is colored with four colors in order to illustrate the four-color theorem. Although some maps (including this one) are fairly difficult to color with four colors, no one has ever created a map that requires five or more colors. Until last year, however, all known efforts to prove that four colors are enough to color any map drawn on a sphere or plane had failed. The difficulty of coloring a map depends on the way in which its countries border on one another. The "configurations," or arrangements of neighboring countries, within Moore's map helped the authors estimate the computational difficulties of implementing their ultimately successful approach to proving the four-color theorem. Only part of Moore's map is shown in the illustration. The complete map is a cylindrical projection with octagons (that is, countries with eight neighbors) at the north and south poles. Moore's map is composed of 54 octagons, 228 heptagons, 96 hexagons and 408 pentagons. It was colored by Tom Burket.

The Kodak Ektasound movie camera makes it so easy to take sound movies. Just drop in a cartridge of film, such as Kodak Ektachrome 160 sound movie film, and push the shutter release. It's that simple.

And the microphone? It's right on the handle! So, you're free to move around and shoot wherever you want without worrying about cords or microphone placement.

You don't need movie lights, either. Indoors or outdoors, in room light or sunlight, the Kodak Ektasound movie camera takes colorful sound movies without movie lights.

There's also an automatic exposure control that adjusts the amount of light reaching the film.

Kodak's unique on-camera microphone. Sound made simple.

And, there's a sound monitor so you can hear what the camera is hearing—both before and during use.

Ask your photo dealer for a demonstration. There are several models to choose from including the new model 250 with power zoom and the new model 260 with power zoom and coupled rangefinder.

Kodak Ektasound movie cameras

bors [*see top illustration on page 114*] is "unavoidable," that is, every normal map must contain at least one of these four configurations. Unavoidability is one of the two important ideas that are basic to our proof of the four-color theorem.

The second important idea is reducibility. A configuration is intuitively reducible if there is a way of showing, solely by examining the configuration and the way in which chains of countries can be aligned, that the configuration cannot possibly appear in a minimal five-chromatic map. Kempe showed that three of his four configurations are reducible but failed to show that a country with five neighbors is a reducible configuration. The methods of proving configurations reducible grew out of Kempe's proof that a country with four neighbors cannot occur in a minimal five-chromatic map. The use of the word reducible stems from the form of Kempe's argument; he proved that if a minimal five-chromatic map contains a country with, say, four neighbors, then there is a five-chromatic map with a reduced number of countries [*see bottom illustration on page 114 and illustration on page 116*].

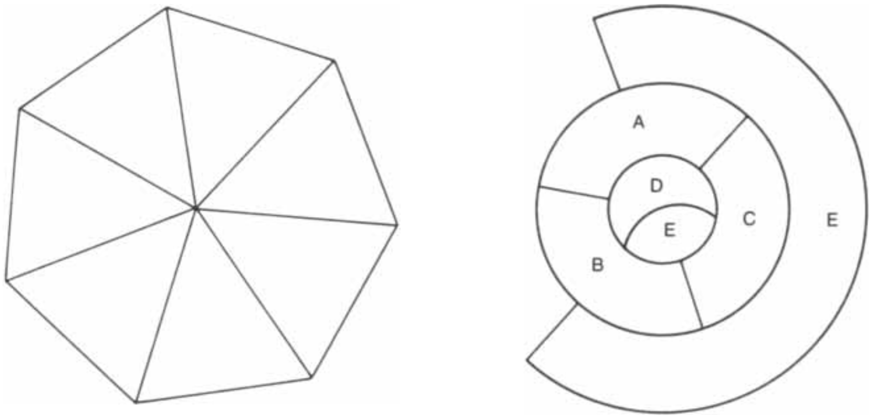
We can describe Kempe's attack on the four-color conjecture as an attempt to find an unavoidable set of reducible configurations. Finding such a set is sufficient for proving the four-color conjecture because it shows that every map contains a configuration that cannot be part of a minimal five-chromatic map. Therefore there can be no minimal five-chromatic map, and as Kempe proved correctly this implies that there can be no five-chromatic map at all. Like Kempe, we attacked the four-color problem by constructing an unavoidable set of reducible configurations. Instead of four simple configurations, however, our set consisted of some 1,500 complex figures.

In 1890, 11 years after Kempe published his proof, Percy John Heawood pointed out that Kempe's argument that no minimal five-chromatic map could contain a country with five neighbors was flawed, and that the error did not appear easy to repair. In his attack on the problem Heawood investigated a generalization of the original four-color conjecture. The maps studied by Guthrie and Kempe were maps in a plane or on a sphere. Heawood, considering maps on more complicated surfaces, was able to obtain an elegant argument that provided an upper bound for the number of colors required to color maps on these surfaces. If the method he used had been applicable to the plane, it would have provided a proof of the four-color conjecture.

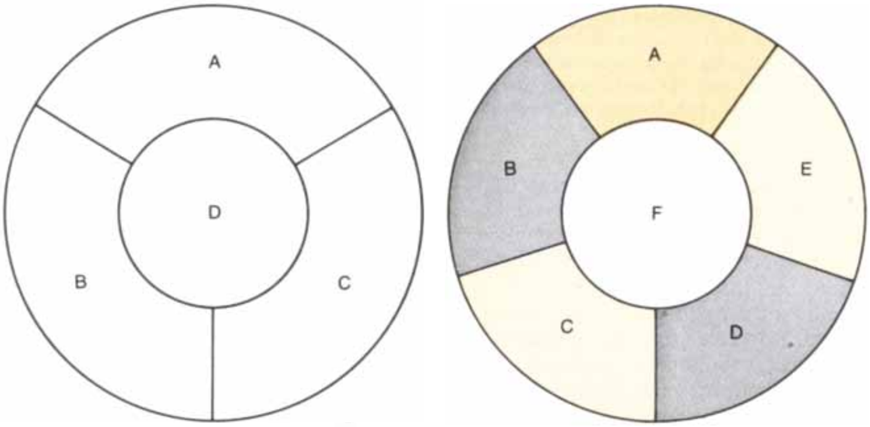
Heawood continued to work on the problem for no less than 60 years. During that time many other eminent mathematicians devoted a great deal of effort to the four-color conjecture. One may wonder why so many mathematicians would spend so much time on what appeared to be a question of so little practical significance. The explanation lies in the discoveries made about the nature of pure mathematics over the past century and in the effect of those discoveries on the faith of mathematicians in the power of mathematics.

Toward the end of the 19th century mathematicians were able to build powerful theories that enabled them to settle many difficult questions. The feeling grew that any question that could be reasonably posed in the language of mathematics could be answered by the use of sufficiently powerful mathematical ideas. Moreover, it was believed that these questions could be answered in such a way that a competent mathematician could check the correctness of the answer in a reasonable length of time. The four-color conjecture certainly appeared to be such a problem. If mathematicians could not solve the problem, it seemed clear that they had simply not yet developed the appropriate mathematical tools.

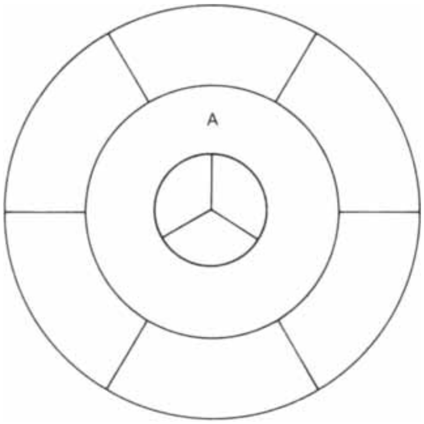
In the 1930's, however, new discoveries challenged the 19th-century faith in the completeness of mathematics. Kurt Gödel and Alonzo Church obtained new and disturbing results in formal logic, the branch of mathematics in which the concept of proof is stated most precisely. It was proved that in what seems to be the most natural logical system there are true statements that cannot be proved true (or false) in the system. Moreover, there are theorems in the system with relatively short statements whose shortest proofs are too long to be written down in any reasonable length of time. In the 1950's further investigation showed that the same difficulties affect branches of mathematics other than logic. Some mathematicians began to think that the four-color conjecture



FOUR-COLOR THEOREM states that four colors are always sufficient to color a planar map so that no two neighboring countries are the same color. For the theorem to make sense a map must consist of contiguous countries. Neighboring countries must be adjacent along a line, because if countries adjacent at single points were considered neighbors, then map at the left would require a different color for each of its seven countries. Country must be a single connected region, otherwise map at right, with country *E* in two pieces, would require five colors.



THREE COLORS ARE NOT ENOUGH to color all maps, as is shown by the map of four countries at the left. Each country is adjacent to the other three, and so the map obviously requires four colors. On the other hand, it is not correct to assume that the number of colors needed to color a map is the same as the highest number of mutually adjacent countries in the map. In the map at the right, for example, no more than three countries are mutually adjacent, and yet four colors are needed: three for outer ring of countries and one for country in the center.



"NORMAL" MAPS were defined by Alfred Bray Kempe, who published a faulty proof of the four-color theorem in 1879. A map is normal if no country in it completely surrounds another country or countries and if no more than three countries meet at any point. Map shown here is not normal, since country *A* encloses three other countries. Kempe proved correctly that if the four-color theorem could be proved for normal maps, it would be true for all maps, and thus the authors considered only normal maps in their work on theorem.

might be one of those statements that can neither be proved nor disproved; after all, the conjecture had been studied without success for 80 years. Other workers felt that if a proof existed, it would be too long to write down. Many believed, however, that the disease of unsolvability could not spread to this area of mathematics and that an elegant mathematical argument would be found to decide the truth or falsehood of the four-color conjecture.

We now know that a proof can be found. But we do not yet know (and may never know) whether there is any proof that is elegant, concise and completely comprehensible by a human mathematical mind. So many areas of mathematics have been involved in various attempts to prove the four-color conjecture that it would be impossible to discuss them all here. We shall restrict our attention to work that led directly to the proof.

In 1913 George D. Birkhoff of Harvard University improved on Kempe's reduction technique and was able to show that certain configurations larger than Kempe's were reducible. Philip Franklin of the Massachusetts Institute of Technology utilized some of Birkhoff's results to show that a map requiring five colors must have at least 22 countries, that is, that any map with fewer than 22 countries is four-colorable. Birkhoff's methods were improved by many mathematicians between 1913 and 1950. During this period reducible configurations were exploited primarily to improve Franklin's result; by 1950 it had been shown that any map with fewer than 36 countries is four-colorable.

Although the work in this area did show that many configurations are reducible, the set of all the configurations that had been proved reducible by 1950 did not even come close to forming an unavoidable set. Only a few mathematicians had constructed unavoidable sets of configurations, and there was little hope that their work would lead to an unavoidable set of reducible configurations.

Heinrich Heesch of the University of Hannover seems to have been the first mathematician after Kempe to publicly state that the four-color conjecture could be proved by finding an unavoidable set of reducible configurations. Heesch, who began his work on the conjecture in 1936, made several major contributions to the existing theory. In 1950 he estimated that the reducible configurations in the (then) hypothetical unavoidable set would be of certain restricted sizes and number about 10,000.

Configuration size must be a major consideration in this approach to the four-color problem. In the century since Kempe first introduced the idea of reducibility certain standard methods for examining configurations to determine whether or not they are reducible have been developed. Employing these methods to show that large configurations are reducible requires the examination of a large number of details and appears feasible only by computer. Every configuration is surrounded by a ring of neighbors; the ring size, or number of countries that form the ring around the configuration, has a direct bearing on the difficulty of proving the configuration reducible. When one is trying to construct an unavoidable set of reducible configurations and discovers that a particular configuration is not reducible, one can often replace it to good effect with one or more other configurations, usually configurations of a larger ring size. Replacing one configuration with another whose ring contains an additional vertex, however, greatly increases the difficulty of reducibility testing and the computer time needed for it. This is in part because the number of distinct proper colorings of the vertices of the new ring is about three times the number of colorings of the vertices in the original ring. Furthermore, programs to test reducibility consider each possible coloring several times. In 1950 the difficulties of computation seemed to rule out the possibility of producing an unavoidable set of configurations and proving that each of its members is reducible.

With the advent of high-speed digital computers, however, an attack on these problems became technically possible. Heesch formalized the known methods of proving configurations reducible and observed that at least one of them (a straightforward generalization of the method used by Kempe) was in principle a sufficiently mechanical procedure to be implemented by computer. Heesch's student Karl Dürre then wrote a computer program that used this procedure to prove configurations reducible. Whenever such a program succeeds in proving a configuration reducible, the configuration is certainly reducible. A negative result, however, shows only that the particular method of proving reducibility is not sufficient to prove the configuration reducible; it might be possible to prove it reducible by other methods. In some cases, when Dürre's program failed to prove a configuration reducible, Heesch succeeded. He was able to show the configurations reducible with data generated by the program and with further calculations using a stronger technique developed by Birkhoff.

Heesch described configurations in a convenient way. He began by transforming the original map into what mathematicians call a dual form: a planar graph in which each vertex of the graph represents a country and each line segment between vertices represents a border. To obtain a dual graph of a map mark the capital in each country in the map and then, whenever two countries are neighbors, join their capitals by a road across the common border [*see illustration on page 117*]. Now remove everything except the capitals (called vertices) and the roads (called edges) you have added. These vertices and edges form the dual graph of the original map. The edges of a graph divide the plane into regions that are called faces. If the original map is normal (and only normal maps must be considered in the proof of the four-color theorem), all of the faces are triangles. In this case the dual graph is called a triangulation. The number of edges that end at a particular vertex (in the dual graph) is called the degree of the vertex and is equal to the number of neighbors of the country (in the original map) that is represented by that vertex. A path of edges that starts and ends at the same vertex and does not cross itself separates the graph into two parts: its interior and its exterior. Such a path is called a circuit.

In the vocabulary of dual graphs a configuration is a part of a triangulation consisting of a set of vertices plus all of the edges joining them. The boundary circuit, consisting of those vertices adjacent to the configuration and the edges joining them, is called the ring of the configuration. (The ring in the graph corresponds to the ring of countries that bound the configuration in the original map.) Configurations are often described by the lengths of their rings: a six-ring configuration, for example, is one whose boundary circuit has exactly six vertices.

With Heesch's work the theory of reducible configurations seemed extremely well developed. Although certain improvements in the methods of proving reducibility have since been made, all of the ideas on reducibility that were needed for the proof of the four-color theorem were understood in the late 1960's. Comparable progress had not been made in finding unavoidable sets of configurations. Heesch introduced a method that was analogous to moving charge in an electrical network to find an unavoidable set of configurations (not necessarily reducible), but he had not treated the idea of unavoidability with the same enthusiasm as the idea of reducibility. This method of "discharging" that first appeared in rather rudimentary form in the work of Heesch has been crucial, however, in all later work on unavoidable sets. In a much more sophisticated form it became the central element in the proof of the four-color theorem; hence we will explain it in some detail.

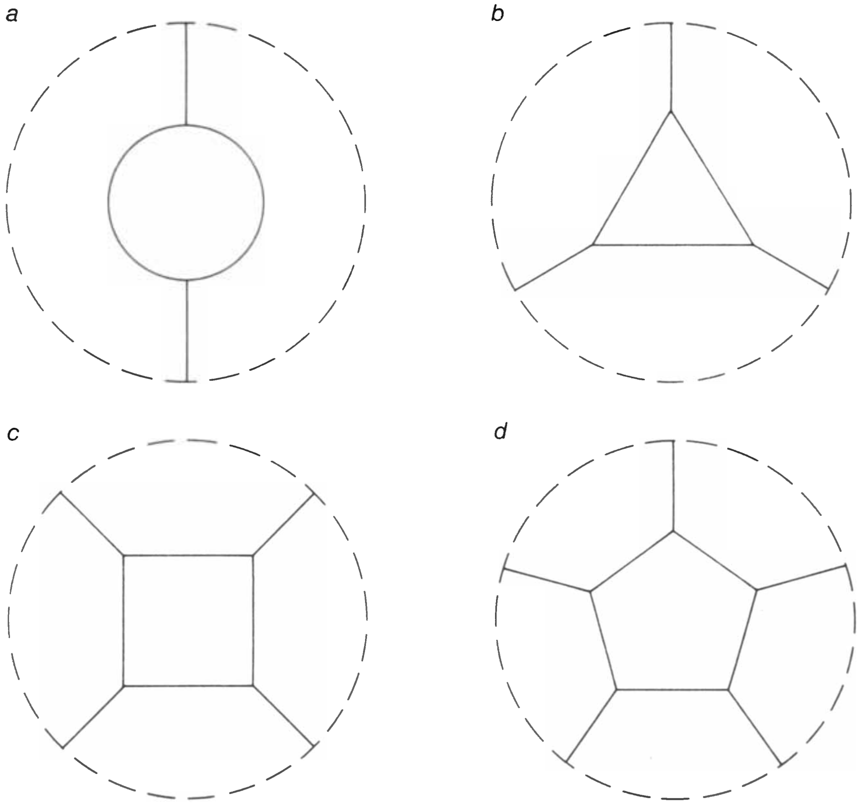
Kempe's work shows that a triangulation that represents a minimal five-chromatic map cannot have any vertices with fewer than five neighbors. Thus in what follows we will for convenience use the word triangulation to mean a triangulation with no vertices of degree less than five. If we assign the charge number 6-k to every vertex of degree k (that is, with k neighbors), then vertices of degree greater than six (major vertices) are assigned negative charge and only vertices of degree five are given positive charge. It follows from Kempe's work that the sum of the assigned numbers of any triangulation is exactly 12. This somewhat surprising result depends both on the fact that the graph is drawn in the plane and that it is a triangulation. The particular sum of 12 is not very important. What is extremely important is that for every planar triangulation this charge sum is positive.

Now suppose the charges in such a triangulation are redistributed, moved around without losing or gaining charge in the entire system. In particular suppose positive charge is moved from some of the positively charged (degree-five) vertices to some of the negatively charged (major) vertices. It is certainly not possible to change the (positive) sum of the charges by these operations, but the vertices having positive charge may change; for example, some degree-five vertices may lose all positive charge (become discharged), whereas some major vertices may gain so much charge that they end up with positive charge (become overcharged). Different vertices become discharged or overcharged according to the discharging, or redistribution, procedure chosen.

Given a specified discharging procedure on an arbitrary graph, however, it is possible to make a finite list of all the configurations that, after discharging is done, have vertices of positive charge. (Of course, each of these configurations can be repeated an unforeseeable number of times.) In other words, positive charge can only occur within this finite set of configurations. Since the total charge is always positive, there will always be some vertices of positive charge. Therefore since all possible receptacles of positive charge are included in the list of configurations, every planar triangulation must contain at least one of these configurations. This process—generating a list of specific configurations from an arbitrary map—works because it is possible to determine the layout of small pieces of the map without knowing the form of the complete map.

In the proof of the four-color theorem the purpose of this discharging of positive vertices is to find a procedure describing exactly how to move charge in such a way as to ensure that every vertex of positive charge in the resulting configuration must either belong to a reducible configuration or be adjacent to one. Since the configurations signaled by this procedure must form an unavoidable set, if they are also reducible, then the four-color conjecture is proved. Of course, if not all of the resulting configurations are reducible, then no real progress has been made. In fact, Kempe's unavoidable set can be considered to be the one resulting from the ineffective procedure of moving no charges at all.

MAP OF THE EASTERNMOST STATES of the U.S. is normal, but the map of the entire continental U.S. is not: Utah, Colorado, Arizona and New Mexico meet at a single point. (Note that the 48 contiguous states do not even make a proper map because Michigan is made up of two pieces that are not connected.) The maps in the lower illustration on page 111 are normal, but maps in top illustration are not.



SET OF FOUR CONFIGURATIONS was proved by Kempe to be "unavoidable," that is, every normal map includes as a part at least one member of this set. The configurations in the set are a country with two neighbors (*a*), a country with three neighbors (*b*), a country with four neighbors (*c*) and a country with five neighbors (*d*). Of course, no country can have zero or one neighbor, since neither an island nor an enclosed country is allowed in a normal map.

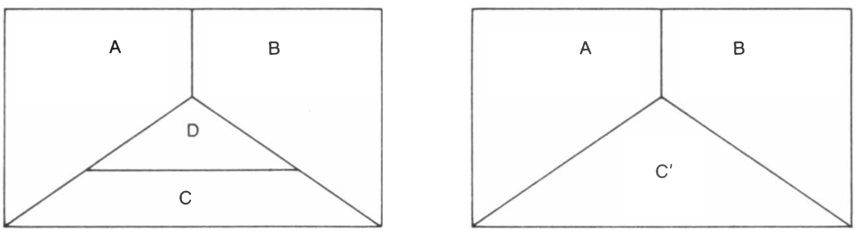
An example of a rather simple discharging procedure and the associated unavoidable set may make these ideas clearer. Consider the procedure that transfers 1/5 unit of charge from every degree-five vertex to each of its major neighbors [*see top illustration on page* 121]. The corresponding unavoidable set consists of two configurations. One is a pair of degree-five vertices joined by an edge and the other is a degree-five vertex joined by an edge to a degree-six vertex.

These configurations are obtained as follows. A degree-five vertex can only end up positive in this procedure if at least one of its neighbors is not major, so that the vertex is forced to retain positive charge; the vertex either has a degree-five neighbor (the situation corresponding to the first configuration in the set) or a degree-six neighbor (the second configuration).

A degree-six vertex starts with no charge and therefore cannot receive any. A degree-seven vertex can only become positive if it has at least six neighbors that are degree-five vertices; if it has at least six such neighbors, two of them are joined by an edge (the first configuration of the unavoidable set). A vertex of degree eight or higher cannot become positive even if all of its neighbors are degree-five vertices. This can be seen by examining a degree-eight vertex. Its charge is minus two and the maximum positive charge it can receive is eight times 1/5, or $1\frac{3}{5}$. Thus the two (nonreducible) configurations form an unavoidable set; that is, since these calculations apply to any planar triangulation (with no vertices of degree less than five) a member of the two-configuration set will be found in every such triangulation of the plane.

In 1970 one of us (Haken) noticed certain methods of improving discharging procedures and began to hope that such improvements might lead to a proof of the four-color conjecture. The difficulties, however, still appeared to be formidable. One was that it was believed that very large configurations (with rings of neighbors containing as many as 18 vertices) would be included in any unavoidable set of reducible configurations. This meant that the problem might be beyond the capabilities of existing computers because although testing configurations of small ring size (up to about 11 vertices) for reducibility was reasonably simple on a computer, the computer time required increased by a factor of four for every unit increase in ring size. To make things worse, the computer storage requirements increased just as quickly. When Dürre's program was applied to a particularly difficult 14-ring configuration, it took 26 hours to prove that the configuration did not satisfy even the most mechanical definition of reducibility (the definition stated with the fewest machine instructions). Even if the average time required to examine a 14-ring configuration were only 25 minutes, testing an average 18-ring configuration would require $4^4 \times 25$ minutes, or more than 100 hours, of computer time and more storage than was available on any existing computer.

Another major difficulty was that no one really knew how many reducible configurations would be needed to form an unavoidable set. It seemed likely that the number of configurations would be in the thousands, but no upper limit had been established. Suppose that on a computer with sufficient storage it takes 100 hours to show that an 18-ring configuration is reducible. If there are 1,000



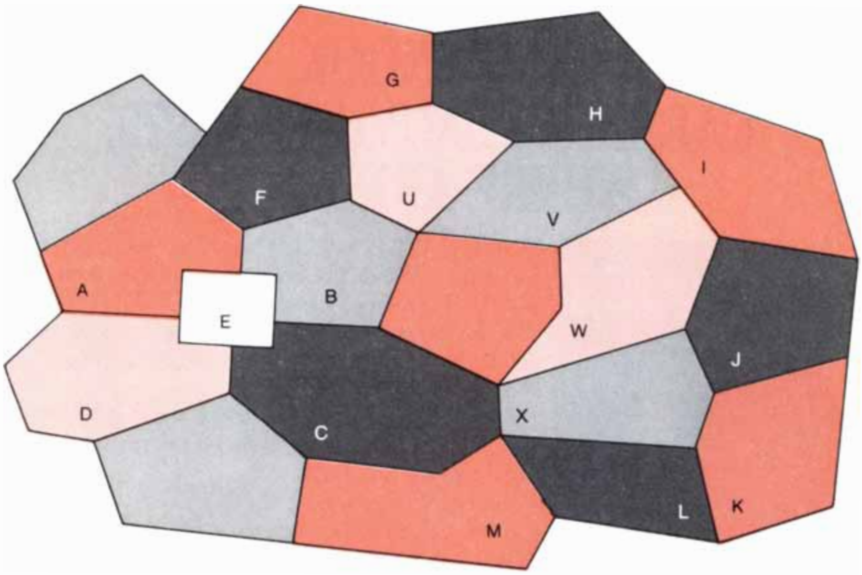
KEMPE SHOWED that to prove the four-color theorem it is sufficient to prove that no minimal five-chromatic map can exist. A minimal five-chromatic map is a map with the fewest countries that requires five colors, that is, any smaller map (any map with fewer countries) can be colored with four colors or fewer. Kempe tried to prove that no minimal five-chromatic map is possible by showing that none of the configurations in his unavoidable set can exist in a minimal five-chromatic map. Each planar map contains one of these configurations, so that if all Kempe's arguments had been correct, he would have proved the theorem. Kempe's correct proof that a country, *D*, with three neighbors (*left*) cannot be part of a minimal five-chromatic map is as follows. Combine *D* with one of its neighbors, *C*, to form a new country, *C'* (*right*). The new map has fewer countries than the original minimal five-chromatic map and so can be colored with four colors. Give all countries of the original map except *D* colors they would have in four-coloring of the smaller map. *D* can be colored with color not used for *A, B* or *C*. Therefore original map can be colored with four colors: it is not a minimal five-chromatic map. Similar argument shows that no country with two neighbors can be part of minimal five-chromatic map. Proof for country with four neighbors is described in illustration on page 116.

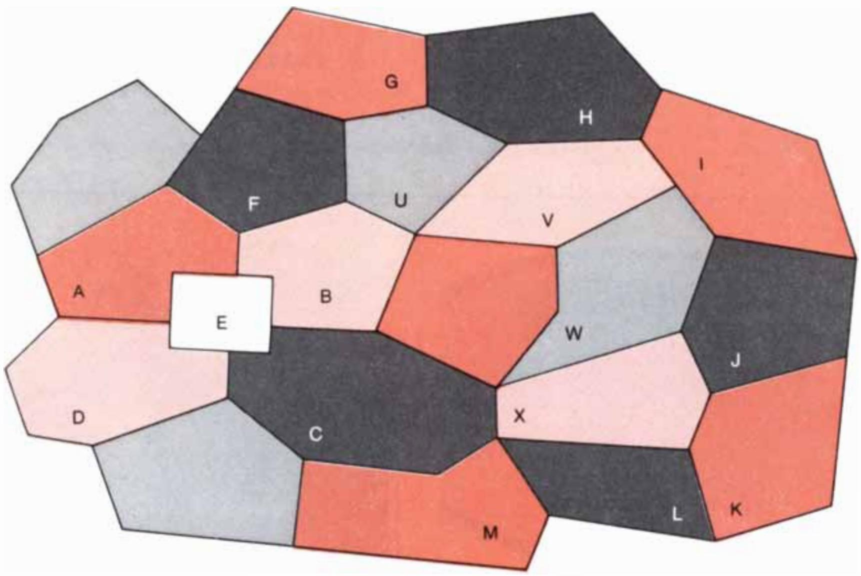
18-ring configurations in the set, it would take 100,000 hours, or more than 11 years, to prove them reducible on a very large computer. For all practical purposes if the set had been that large, the proof would have had to wait for computers much faster than those currently available.

Even if the theorem could be proved by finding an unavoidable set of reducible configurations, the proof would not satisfy those who demand mathematical elegance. What would be even more upsetting to many mathematicians, no one would be able to check the reducibility of all the configurations in the set by hand. By 1970, however, many experts on the four-color problem were quite pessimistic about finding a short proof. The problem had received a great deal of attention since its formulation more than 100 years earlier. Many approaches to the problem had been tried, but although some had led to important results in other areas of mathematics, none had ever led to a proof of the four-color theorem.

When we began our work on the problem in 1972, we felt certain that the techniques available to us would not lead to a noncomputer proof. We were even quite doubtful that they could lead to any proof at all before much more powerful computers were developed. Therefore our first step in attacking the problem of finding an unavoidable set of reducible configurations was to determine whether or not there was any hope of finding such a set with configurations of ring size sufficiently small that the computer time for the proofs of reducibility would be within reason. By the very nature of this question it was clear that we should not begin by examining the reducibility of all the configurations considered; otherwise the time spent in making the estimate would exceed the expected time needed for the entire task.

Here a thought of Heesch's proved extremely useful. While he was testing configurations for reducibility he observed a number of distinctive phenomena that provide clues to the likelihood of successful reduction. For example, there are certain conditions involving the neighbors of vertices of a configuration under which no reducible configuration had ever been found. No reducible configuration had ever been found that contained, for instance, at least two vertices, a vertex adjacent to four vertices of the ring and no smaller reducible configurations. Although no proof is known that reducible configurations with these reduction obstacles could not exist, it seemed prudent to assume that if one wanted reducible configurations, one should avoid such configurations. Heesch found three major reduction obstacles, including the one described above, that could be easily described [see bottom illustration on page 118]. No configuration containing one of them





REDUCIBLE CONFIGURATION is an arrangement of countries that cannot be part of a minimal five-chromatic map. To prove the four-color theorem Kempe tried to show that the four configurations in his unavoidable set are reducible. He failed because he could not prove that a country with five neighbors is a reducible configuration. The authors succeeded in proving the theorem by producing an unavoidable set of some 1,500 reducible configurations. The concept of reducibility is derived from Kempe's correct proof that a country with four neighbors, say *E*, cannot be part of a minimal five-chromatic map. As in the case of a country with three neighbors (*see bottom illustration on page 114*), *E* can be combined with a neighbor to establish a four-coloring for all the other countries in the map (*top*). If the four neighbors of *E* are colored with fewer than four colors, an unused color can be assigned to *E* to show that the map is four-colorable. Otherwise consider the colors of a pair of countries on opposite sides of *E*. Either there is a path of countries colored with these two colors leading from one country to the other or there is no such path. In this example *A* and *C* are connected by a path of dark-color and dark-gray countries, but no path of light-gray and light-color countries connects *B* and *D*. A theorem states that if both pairs of countries are joined by such paths, then the paths have a country in common. This is clearly impossible, since the common country would have to be colored with two distinct colors. Therefore at least one pair of countries (*here B and D*) will not be linked by a path of countries the colors of the pair. The light-gray and light-color countries that form a path from *B* are *U, V, W* and *X*. Reverse colors of the countries in this group. In this case make the light-gray countries light color and the light-color countries light gray (*bottom*). Now uncolored country has neighbors colored with only three colors, so that it can be given the fourth color (*light gray here*) to produce a four-coloring for the minimal five-chromatic map. In other words, a country with four neighbors cannot be part of a minimal five-chromatic map.

has ever been proved to be reducible.

It is easy to determine whether or not a configuration contains a reduction obstacle, and configurations without reduction obstacles have a very good chance of being reducible. If there was a manageable unavoidable set of configurations free of reduction obstacles, we felt there would have to be an unavoidable set of roughly the same size containing only reducible configurations.

We therefore decided to first study certain kinds of discharging procedure in order to determine the types of sets of obstacle-free configurations that might arise. To gain an understanding of what was needed, even for this study, we restricted the already restricted problem to what are called geographically good configurations: configurations that do not contain the first two of Heesch's three obstacles.

In the fall of 1972 we wrote a computer program that would carry out the particular type of discharging procedure that seemed most reasonable to us. We were not ready to prove the theorem, so that the output of our program was not an unavoidable set but rather a list of the configurations that resulted from the most important situations. Although a computer program could not be expected to proceed as cleverly as a human being, the immense speed of the computer made it possible to accept certain inefficiencies. In any event the program was written in such a way that its output could be easily checked by hand.

The first runs of the computer program in late 1972 gave us a great deal of valuable information. First, it appeared that geographically good configurations of reasonable size (with a ring size of no more than 16) would be found close to most vertices of ultimately positive charge. Second, the same configurations occurred frequently enough for it to appear that the list of configurations might be of manageable length. Third, as the procedure was originally organized the computer output was too large to handle; similar cases repeated the same argument too often. Fourth, there were clearly some flaws in the procedure, since there were vertices of ultimately positive charge in whose neighborhoods no geographically good configurations could be guaranteed. Finally, the program generated a tremendous amount of information in only a few hours of computer time, so that we knew it would be possible to experiment often.

The program and certain aspects of the discharging procedure had to be modified to overcome the problems indicated by the first computer runs. Since we could preserve the basic program structure, the changes were not too difficult to make, and a month later we made a second set of runs. Now that the major problems had been corrected, we could perceive subtler problems. After some study we found solutions to these problems and again modified the program.

The man-machine dialogue continued for another six months until we felt that our procedure would obtain an unavoidable set of geographically good configurations in a reasonable amount of time. At this point we decided to prove formally that the procedure would provide a finite unavoidable set of geographically good configurations. We had to put aside the experimental approach and describe the total procedure. It was necessary to prove that all cases had been covered and that those cases that were not handled by the computer program were as simple as they appeared to be.

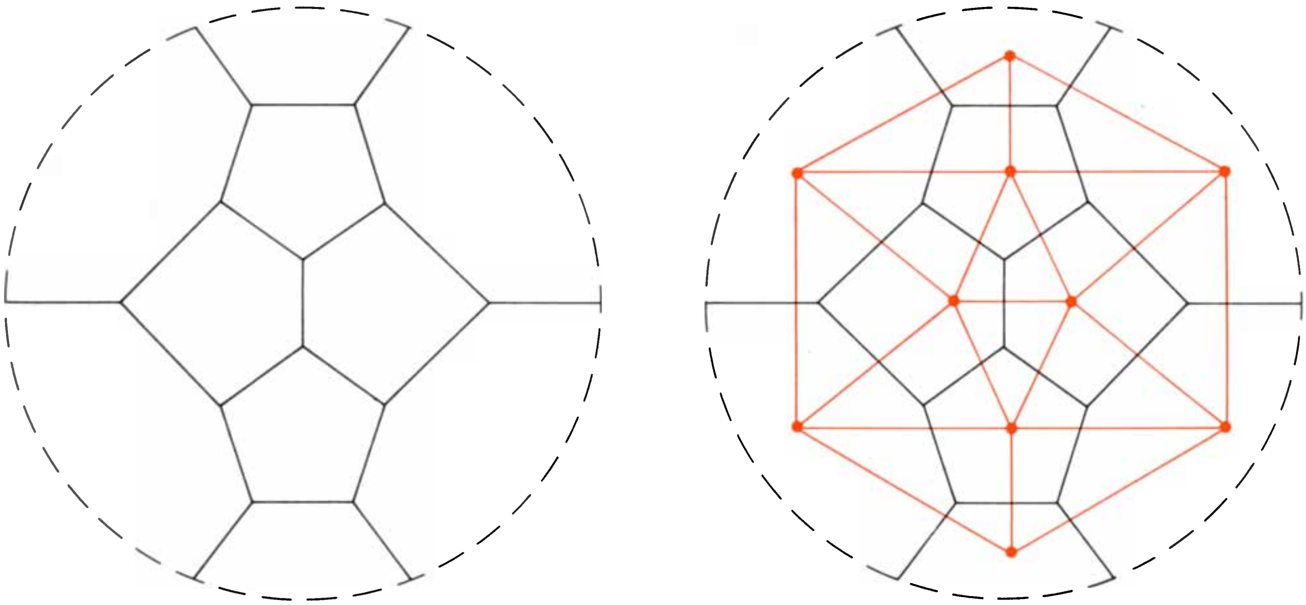
Much to our surprise this task proved extremely difficult and took more than a year. It was necessary to formulate general definitions of terms and to prove abstract statements about them. Special cases, even those that were not likely to arise in practice, had to be examined in detail and often required complicated analyses. Finally, by the fall of 1974, we had a lengthy proof that a finite unavoidable set of geographically good configurations does exist, and we had a procedure for constructing such a set with precise, although much larger than desirable, bounds on the size of the configurations in the set. The procedure we had designed was extremely important to us because we intended to apply it in the proof of the four-color theorem. (A short time later Walter Stromquist, a major contributor to reduction theory who was then a graduate student at Harvard, devised an elegant proof of the existence of unavoidable sets of geographically good configurations. Since Stromquist's proof did not provide a method of actually constructing the configurations in the set, however, it appeared unlikely that it would be immediately applicable to the four-color conjecture itself.)

When we had proved that our procedure would work for geographically good configurations, we still did not know how complicated carrying out the procedure would be. We decided to try it on the restricted problem of triangulations having no pairs of adjacent degree-five vertices. This, of course, is a strong restriction, but the unavoidable set of geographically good configurations we obtained was quite small: only 47 configurations and none of them of a ring size larger than 16. We guessed that the solution to the unrestricted problem might be 50 times more unwieldy (this turned out to be a bit optimistic), and we decided that there was good reason to continue. Early in 1975 we modified the program so that it produced obstacle-free configurations rather than geographically good ones and forced it to search for sets in which more of the configurations had small ring size. When we ran the modified program, certain flaws became evident but there was also a very pleasant surprise: Replacing geographically good configurations with obstacle-free ones only doubled the size of the unavoidable set.

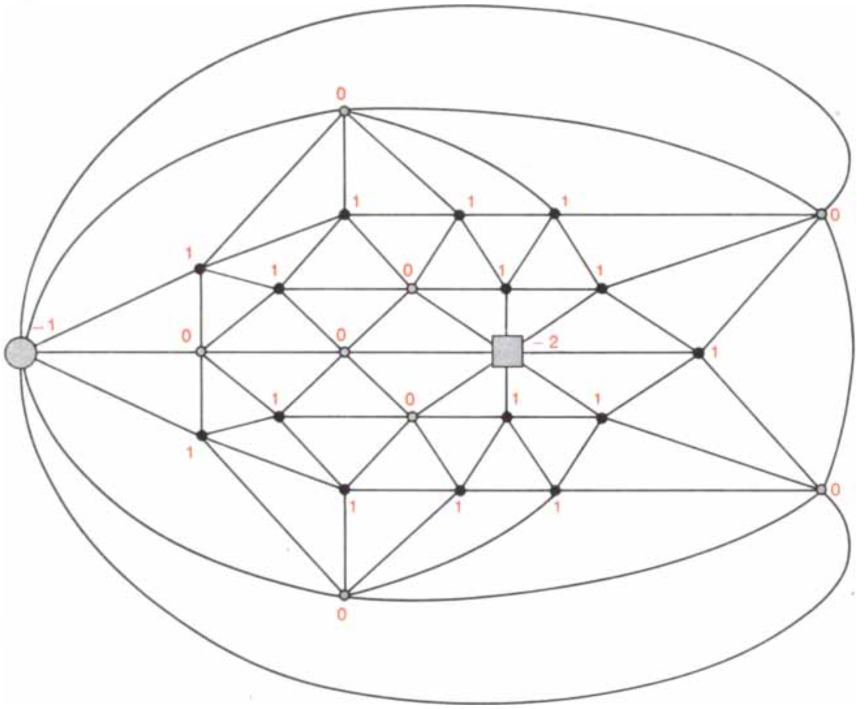
At this point the program, which had been absorbing our ideas and improvements for two years, began to surprise us. When we had hand-checked the analyses produced by the early versions of the program, we were always able to predict their course, but now the computer was acting like a chess-playing machine. It was working out compound strategies based on all the tricks it had been taught, and the new approaches were often much cleverer than those we would have tried. In a sense the program was demonstrating superiority not only in the mechanical parts of the task but in some intellectual areas as well.

By the summer of 1975 we believed that there was a good chance that we could find an unavoidable set of configurations each of which would be obstacle-free and likely to be reducible. Such a set would certainly contain some irreducible configurations, but we felt that by some small change of procedure we would be able to replace them with reducible configurations. Now for the first time we would have to test configurations for reducibility.

We began by writing a program to test for the most mechanical reducibility, working with the assembler language for the University of Illinois IBM 360 computer. We had been joined the preceding fall by John Koch, a graduate student in computer science who chose to write a dissertation on the reducibility of configurations of small ring size. (Frank Allaire of the University of Manitoba and Edward Swart of the University of Rhodesia were doing somewhat similar work of which we were unaware.) By the fall of 1975 Koch had written programs to check for the most mechanical reducibility in configurations of a ring size up to 11 and had gone on to more general investigations.



A DUAL GRAPH (*color*) displays the countries and borders of the original map (*left*). The dual graph is constructed by marking a vertex in each country of the original map and then connecting the vertices of each pair of neighboring countries with an edge across their common border. These edges can always be drawn as straight lines, so that they divide the plane into polygonal faces. When the original map is normal, these faces are triangles and the graph is called a triangulation of the plane. The number of edges meeting at a vertex is called the degree of the vertex and is equal to the number of neighbors of the country (in the original map) represented by that vertex.



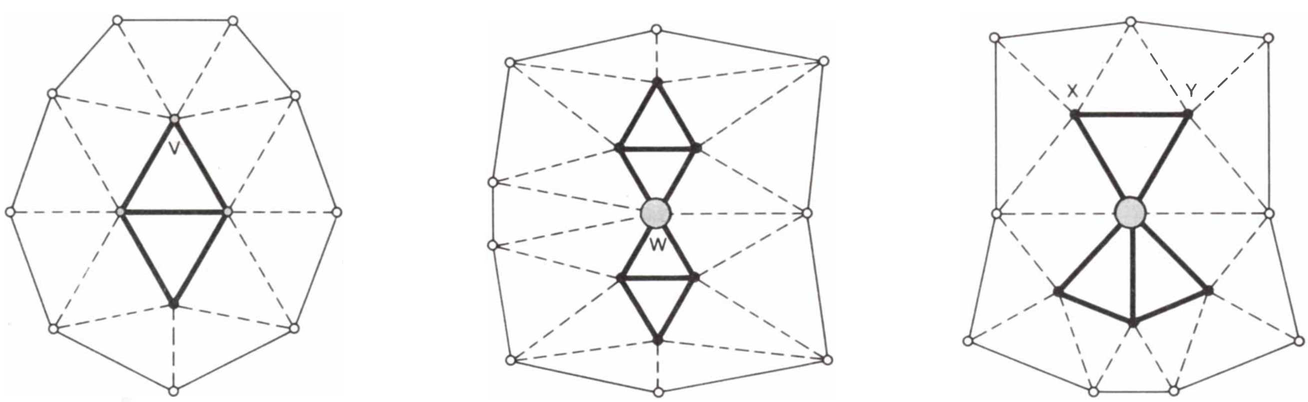
EXAMPLE OF A PLANAR TRIANGULATION includes one degree-eight vertex (*gray square*), one degree-seven vertex (*large gray circle*), eight degree-six vertices (*small gray circles*) and 15 degree-five vertices (*small black circles*). Kempe proved that vertices of degree two, three and four are reducible, that is, they cannot occur in a minimal five-chromatic map. Therefore in order to disprove the existence of a minimal five-chromatic map the authors needed to consider only those triangulations with vertices of degree five or more. The "charge" (*numerals in color*) of a vertex is defined as six minus the degree of that vertex. Since the triangulations under consideration have no vertices of degree four or less, the only vertices with positive charge are those of degree five. It is not difficult to prove that the total charge in any normal map equals 12. This fact implies that there will be positively charged (that is, degree-five) vertices in every triangulation of the plane involved in the proof of four-color theorem.

Over the next few months, with Koch's aid, we modified his work on 11-ring configurations to produce programs for checking the reducibility of 12-, 13- and 14-ring configurations. Finally we further modified these programs to apply a more general procedure for reduction that had been developed by Birkhoff.

At this point our work on the discharging procedure reached an impasse. Structural changes, not just adjustments of parameters and small additions, were required to improve the program, and each change would mean a major modification of it. We therefore decided to discard the program and implement the discharging procedure by hand. This would ensure greater flexibility and allow the procedure to be modified locally whenever necessary. In December, 1975, we made an encouraging discovery. One of the rules that defined our discharging procedure was too rigid. Relaxing the rule made the procedure considerably more efficient.

With the new discharging procedure it looked as though we could build an unavoidable set whose reducible configurations would be smaller than those produced by earlier procedures. The computer time required for the proof would probably be less than previous estimates indicated. It was still evident, however, that a considerable computational effort would be required to produce even the best unavoidable set of reducible configurations. Edward F. Moore of the University of Wisconsin had developed a powerful method for constructing maps that contained no small reducible configurations. For example, he had created a map in which the ring size of the smallest reducible configuration is 12 [*see illustration on page 109*]. Therefore any unavoidable set of reducible configurations must have at least one configuration of ring size 12. Moore's work provides a lower bound on ring size, but we now feel sure that a ring size of 13 is necessary, and it is quite likely that a ring size of 14 is also necessary. (Our proof demonstrates that no larger configurations are needed.)

In January, 1976, we began the construction of an unavoidable set of reducible configurations by means of our new discharging procedure. The final version of the procedure had one further advantage for ensuring the reducibility of the final configurations. The proce-



THREE REDUCTION OBSTACLES were observed by Heinrich Heesch of the University of Hannover. These arrangements of vertices seem to occur only in configurations that cannot be proved reducible. Thus the authors could use the obstacles to identify potential problem areas in their proof. In the graphs shown here the configurations are the arrangements of vertices joined by heavy solid lines. Thinner solid lines connect the vertices of the ring, or outer boundary, of the configuration. Broken lines connect configuration vertices to ring vertices. Each configuration contains one of Heesch's reduction obstacles: vertex *V* (*left*) has four neighbors on the ring of the configuration; vertex *W* (*middle*) has three nonconsecutive neighbors on the ring of the configuration; vertices *X* and *Y* (*right*) form what is called a hanging pair, that is, these vertices are joined to each other and to ring vertices but have only one other neighbor in the configuration.

dure was essentially self-modifying; the program was designed to identify critical, or problem, areas—configurations that looked as though they would resist reduction efforts—and to modify itself to move positive charge to a different place. Since we were doing the discharging procedure without the computer, we knew that the procedure could be modified as we encountered critical areas.

We began by making an approximation, a preliminary run of our discharging procedure. We considered each possible case in which a vertex was forced to be positive, and in each case the neighborhood was examined to find an obstacle-free configuration. If none was found, the neighborhood was called critical, which meant that the discharging would have to be modified to avoid this problem. Even when an obstacle-free configuration was found, however, we could not guarantee a reducible configuration. The new reduction programs were used to try to find some obstacle-free configuration in the neighborhood that was also reducible. If none was found, the neighborhood was also called critical.

This method of developing an unavoidable set of reducible configurations (as we perfected a discharging procedure) was only possible by another dialogue with the computer. To determine which neighborhoods were critical it was necessary to check for reducibility quickly, in terms of computer time and in terms of real time. Fortunately it was seldom necessary to wait more than a few days for results, even though often a considerable amount of computer time was needed. Since this intensive man-machine interaction was indispensable to our work, we should explain the circumstances that made it possible.

Although our arrangement for computer usage seemed quite natural to us at the time, we have since discovered that we were indeed fortunate to be working at the University of Illinois, where a combination of a large computing establishment and an enlightened policy toward the research use of computers gave us an opportunity that seems to be unavailable at almost any other university or research establishment. Although we could not guarantee that our work would lead to a proof of the four-color theorem, we were given well over 1,000 hours of computer time in what we feel was an admirable display of faith in the value of pure mathematical research. We were informed by the computer center that because the university's computers were not fully utilized at all times by class work and other kinds of research we could be included in a small group of computer users who were allowed to share the surplus computer time. We now know that this policy was essential to our success.

In June, 1976, we completed our construction of an unavoidable set of re-

"I would like to present to you the 1978 Cordoba, and to tell you about this extraordinary new automobile in greater detail than is possible in a television commercial."

Ricardo Montalbán

My first look at the 1978 Cordoba occurred here at the Chrysler Styling Center in Highland Park, Michigan.

Naturally, it was the handsome styling changes that I noted first. Refreshingly contemporary . . . unlittered by unnatural lines. Yet the classic look remains a strength . . . a look that will continue to stand the test of time.

In the realm of engineering, Chrysler has been a leader . . . and in the realm of electronics a true pioneer.

Electronic Ignition, for example, was an integral part of Chrysler Corporation cars years before it became the standard for the entire industry.

Likewise Cordoba's ingenious Electronic Lean Burn System. This year, it includes an even more sophisticated Spark Control Computer for improved ignition performance.

Cordoba's suspension, too, has benefited from subtle refinements.

Cornering and braking are . . . as they have been . . . remarkable.

Yet this year, engineers have made refinements which have increased rear suspension travel, and enhanced the already low sound levels in the interior.

The interiors. Magnificent. Newly styled. Available with new fabric blends, and Corinthian leathers in new colors.

For seating I prefer two options: either an individually adjustable 60/40 split design (most impressive when ordered with a supple new velour textured like corduroy) . . . or a new bucket design, pillowed with leather in a most indulgent fashion. Beautiful . . . and extravagantly comfortable.

In all, this new Cordoba is an extraordinary automobile, the very ultimate personal car from Chrysler. I am pleased that it has managed to exceed my expectations.

And I am pleased to inform you that you can examine, Buy or Lease one today at your Chrysler-Plymouth dealer.

Chrysler Cordoba

CHRYSLER

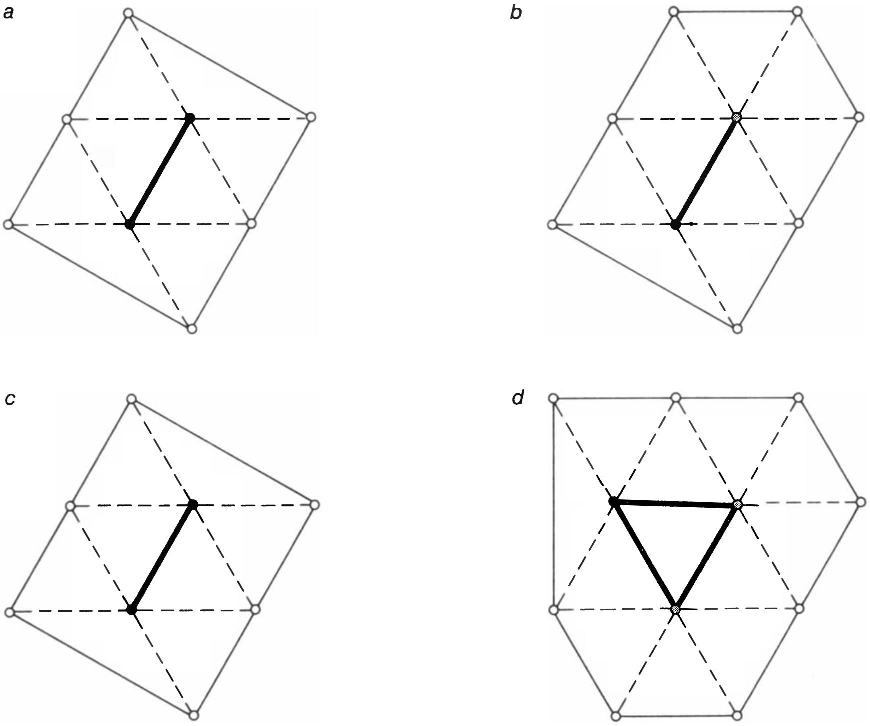
A PRODUCT OF CHRYSLER CORPORATION

ducible configurations. The four-color theorem was proved. We had used 1,200 hours of time on three different computers. The final discharging procedure differed from our first approximation by some 500 modifications resulting from the discovery of critical neighborhoods. The development of the procedure required the hand analysis of some 10,000 neighborhoods of vertices of positive charge and the machine analysis of more than 2,000 configurations. A considerable part of this material, including the reduction of 1,482 configurations, was used in the final proof. Although the discharging procedure (without the reductions) can be checked by hand in a couple of months, it would be virtually impossible to verify the reduction computations in this way. In fact, when we submitted our paper on the proof to *Illinois Journal of Mathematics,* its referees checked the discharging procedure from our complete notes but checked the reducibility computations by running an independent computer program.

Many mathematicians, particularly those educated before the development of high-speed computers, resist treating the computer as a standard mathematical tool. They feel that an argument is weak when all or part of it cannot be reviewed by hand computation. From this point of view the verification of results such as ours by independent computer programs is not as convincing as the checking of proofs by hand. Traditional proofs of mathematical theorems are reasonably short and highly theoretical—the more powerful the theory, the more elegant the proof—and reviewing them by hand is usually the best method. But even when hand-checking is possible, if proofs are long and highly computational, it is hard to believe that hand-checking will exhaust all the possibilities of error. Furthermore, when computations are sufficiently routine, as they are in our proof, it is probably more efficient to check machine programs than to check hand computations.

If many mathematicians are disturbed by long proofs, it may be because until quite recently they only employed the kinds of methods that produce short proofs. We still do not know whether or not a short proof of the four-color theorem can be found. Several new proofs of moderate length have been announced, but none of them has survived expert scrutiny. Although it is conceivable that one of these proofs is valid, it is also conceivable that the only correct proofs will be based on unavoidable sets of reducible configurations and that they will therefore require computations that cannot be done by hand. We believe that there are theorems of great mathematical interest that can only be proved by computer methods. Even if the four-color theorem is not such a problem, it provides a good example of what might be done to prove one. There is no reason to believe that there is not a large number of problems calling for this different kind of analysis.

Our proof of the four-color theorem suggests that there are limits to what can be achieved in mathematics by theoretical methods alone. It also implies that in the past the need for computational methods in mathematical proofs has been underestimated. It is of great practical value to mathematicians to determine the powers and limitations of their methods. We hope that our work will facilitate progress in this direction and that this expansion of acceptable proof techniques justifies the immense effort devoted over the past century to proving the four-color theorem.



DISCHARGING PROCEDURE generates an unavoidable set of reducible configurations by redistributing the positive charge of an arbitrary triangulation (with no vertices of degree less than five) so that positive charge occurs only in reducible configurations. Since there is positive charge in every map, an unavoidable set can be formed by choosing configurations occurring in every type of neighborhood of positive charge. If each configuration in the set is reducible, then there can be no minimal five-chromatic map and the four-color theorem is proved. An example of a simple discharging procedure is the transfer of 1/5 unit of charge from each positively charged vertex to each of its negatively charged neighbors. In this process a degree-five vertex (*black circles*) ends with positive charge only if it has (*a*) a degree-five neighbor or (*b*) a degree-six neighbor (*gray circles*), so that it is forced to retain charge. A degree-six vertex never becomes positive because, having an initial charge of zero, it never receives positive charge. A degree-seven vertex finishes with positive charge only if it has at least six degree-five neighbors; then at least two of them will be adjacent (*a*). Vertices of degree greater than seven can never receive enough positive charge to overcome their initial negative charge. The unavoidable set generated by this discharging procedure consists of a degree-five vertex joined by an edge to another degree-five vertex (*a*) and a degree-five vertex joined by an edge to a degree-six vertex (*b*). These configurations are not reducible. If procedure is modified to transfer 1/3 unit of charge from each positively charged vertex to each of its negatively charged neighbors, slightly better set (*c, d*) is generated. If 1/2 unit of charge is transferred, resulting set is close to one produced by early version of authors' discharging procedure.

POSTAGE METER STAMP is used by department of mathematics at University of Illinois at Urbana-Champaign to honor the solution of four-color-map problem by two of its members.

How the Iron Age Began

Until almost the end of the second millennium B.C. bronze was the utilitarian metal of the Mediterranean world. Within a few centuries it was replaced by a new kind of metal: "steeled" iron

by Robert Maddin, James D. Muhly and Tamara S. Wheeler

It is an old adage that necessity is the mother of invention. The many unanticipated inventions of modern times would seem to be counterexamples, but over the span of human history the adage probably has a certain rough justice. A current investigation in the history of technology is providing a case in point: the appearance some 3,000 years ago of a kind of steel as a substitute for bronze, which until then had been the dominant metal of the civilized world.

For some two millenniums, up to about 1200 B.C., the civilizations of the Old World had satisfied their needs for a utilitarian metal—for tools, weapons, armor and many other durable articles—with various types and alloys of copper, including bronze. Toward the end of that era, known loosely as the Bronze Age, the civilizations of the eastern Mediterranean suffered a series of disturbances. Exactly what caused these disturbances is not known, but one fact about the ensuing centuries has become clear: the use of bronze rapidly diminished and the use of iron—specifically "steeled," or carburized, iron—increased even more rapidly.

Various items of evidence, necessarily fragmentary but diverse enough to inspire a degree of confidence, suggest the speed of the transition. Iron was known as a workable metal during most if not all of the Bronze Age. Nevertheless, Bronze Age sites in the eastern Mediterranean and Southwest Asia representative of a 2,000-year span of history have yielded a total of fewer than 500 iron artifacts, most of them ornamental. The bronze artifacts recovered from the same sites are numbered in the scores of thousands.

Contrast this picture with the number of knives and weapons made from both metals that have been found at sites in Greece representative of the period between 1050 and 900 B.C. A. Snodgrass of the University of Cambridge tabulated the relative abundance of bronze and iron as follows: bronze knives none, iron knives more than 15; bronze swords one, iron swords more than 20; bronze spearheads eight, iron spearheads more than 30.

This distribution is remarkable on two counts. First, bronze articles of this kind can be made by casting, a quick and easy production method, whereas similar iron articles must be individually shaped by forging, a comparatively arduous process. Second, bronze, like its principal constituent, copper, is remarkably durable and can be scrapped and recast repeatedly. Why had the people at these sites substituted a complex process for a simple one? And what had become of the bronze articles they had made in earlier years?

Future archaeological investigation may reveal why bronze went into a decline. Among the causes could have been a breakdown of the trade in tin, the principal alloying ingredient in the articles of the later Bronze Age. With respect to the rapid rise of iron as a substitute metal the picture is a good deal clearer. Both archaeological investigations and studies of ancient inscriptions have in recent years substantiated much that was conjectural and revealed much that was entirely unknown. To present this information in context let us briefly describe copper and iron metallurgy as it was known in the last millenniums before the Christian Era.

Copper is found both as a metal (in lumps of "native" copper) and as an ore that must be heated to yield the metal. For the purposes of this discussion native copper can be ignored. The oldest copper-smelting sites known are at locations in Iran and Israel and date from the fifth and fourth millenniums B.C. Copper ores are found in modest abundance on Cyprus, in various parts of Turkey, in Iran and in Israel. The commonest ore is chalcopyrite, a complex copper iron sulfide that must be roasted in the open air to remove the sulfur before it can be smelted. Certain copper ores contain arsenic; when they are smelted they yield an arsenical copper alloy that can be characterized as "natural bronze."

The early metalworkers of the eastern Mediterranean commonly smelted copper by filling a stone furnace with alternating layers of charcoal and ore combined with a flux. The flux served the purpose of removing from the copper ore the constituents the metalworker did not want in his finished product. These constituents are collectively known as gangue (a word originally derived from the Greek for a vein of ore). In the hot furnace the flux tended to combine with the gangue and remove it from the metal. In many ores of the eastern Mediterranean the gangue was silica: any one of a number of silicon oxides. The appropriate flux for these ores was the iron oxide hematite; the heat of the furnace combined the silicon and iron oxides to form an iron silicate.

When the metalworker had filled his furnace, he ignited the charcoal. In some furnaces the heat of the fire was increased by the natural draft provided by a flue; in others air was forced into the top, sides or bottom of the furnace through clay pipes. As the charge in the furnace got hotter the charcoal was oxidized to carbon monoxide; the hot gas flowed upward through the mixture of ore and flux, chemically reducing both. Reduction took place at about 1,100 degrees Celsius, and the molten copper trickled down to form a puddle at the bottom of the furnace, leaving the gangue behind as a slag. The early metalworkers could not have accurately predicted or measured the furnace temperature; presumably if no puddle of copper appeared, they let the furnace cool off and started over again with a different charge or more draft.

If the ore happened to contain more than a few percent of arsenic what the metalworker found was a puddle of natural bronze. The bronze had the advantage of being harder than copper. Even if the puddle was merely soft pure copper, however, the cast metal could be made harder by hammering. Working the copper in this way made it possible for metalworkers to fashion reasonably durable copper articles during a pre-

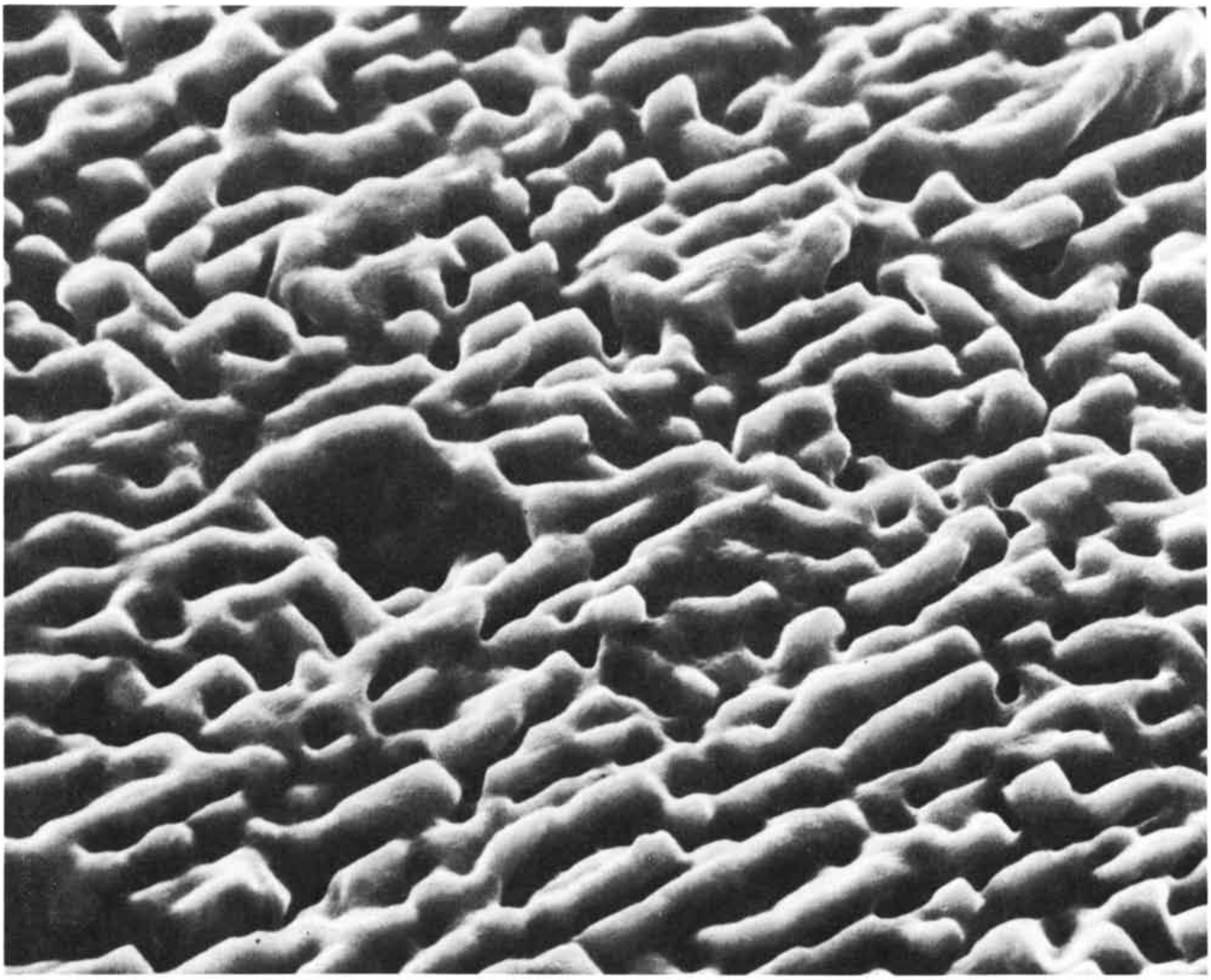
Bronze-Age period that varied in its duration in different parts of the eastern Mediterranean and Southwest Asia.

Arsenical copper ores eventually came to be widely smelted because of the greater hardness of natural bronze. It was later discovered that copper combined with tin instead of arsenic was also hard. It may be that the toxicity of arsenic led to the replacement of arsenical bronze by tin bronze. In the eastern Mediterranean and Southwest Asia tin bronze first appeared around the beginning of the third millennium B.C.; by the early years of the second millennium the production of tin bronze had surpassed that of arsenical.

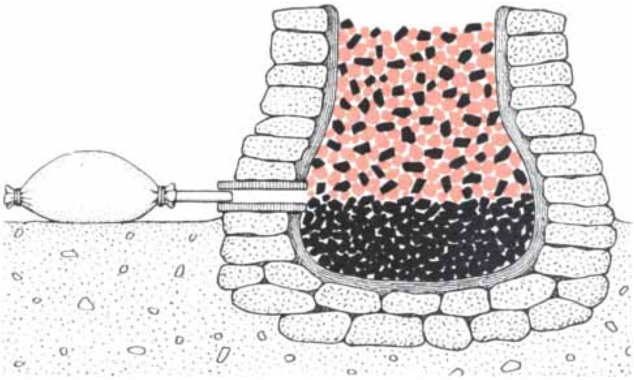
Late in the second millennium B.C. upheavals in the eastern Mediterranean, some of them attributed in Egyptian texts of the time to the activities of interlopers known collectively as the "Peoples of the Sea," led to the collapse of local authority in a number of areas. In the centuries that followed, described by students of the ancient world as a dark age, iron soon replaced bronze as the metal most commonly used for tools, weapons and other articles. Since bronze had been satisfactory for the same purposes for several thousand years and iron did not appear to be as useful, it must be inferred that iron was not suddenly adopted as a result of technical innovation but rather that bronze suddenly became scarce. The further inference is that the scarcity resulted from an interruption in the supply of tin and even of copper to the bronze smelters of the eastern Mediterranean. Where the tin had been coming from is not known. The source may have been such relatively nearby mining areas as the Balkans or the Eastern Desert of Egypt; it may even have been such distant areas as Cornwall or eastern Iran.

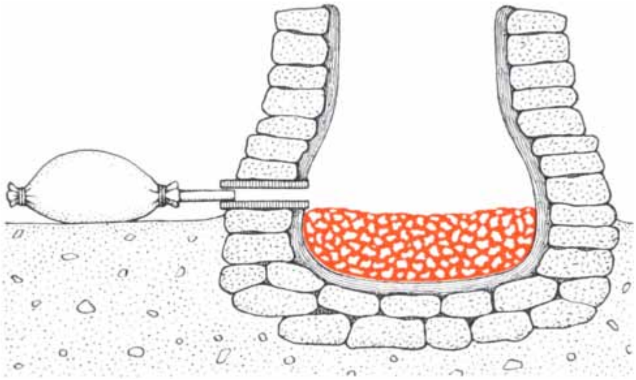
The early metalworkers produced iron from ores, mostly hematite and magnetite, by a smelting process much like the one used to produce copper. There was, however, an important difference. Iron does not melt at temperatures below 1,537 degrees C., and the highest temperature that could be reached in a primitive smelter appears to have been about 1,200 degrees. Smelting iron ore at that temperature yields not a puddle of metal but a spongy mass mixed with iron oxide and iron silicate. These nonmetallic substances, which collectively represent slag, arise from the combination of ferrous oxide and silica gangue in the reduction process.

The commonest of the nonmetallic substances is fayalite, which remains viscous at temperatures down to 1,177 degrees C. The metalworker therefore withdrew a mass of spongy iron from the furnace, reheated it in a forge and quite literally squeezed the fayalite out of it by hammering.



SPONGY STRUCTURE of bloomery iron is revealed in this scanning electron micrograph which shows a modern sample of iron that was smelted at a temperature below its melting point. To produce the sponge a very pure oxide of iron was reduced in an atmosphere of helium and carbon monoxide; the result is seen magnified 2,400 diameters. With a sample of ordinary iron ore the interstices of the sponge would be filled with slag, which could be expelled by hammering. The sample was prepared by Y. K. Rao of the University of Washington.





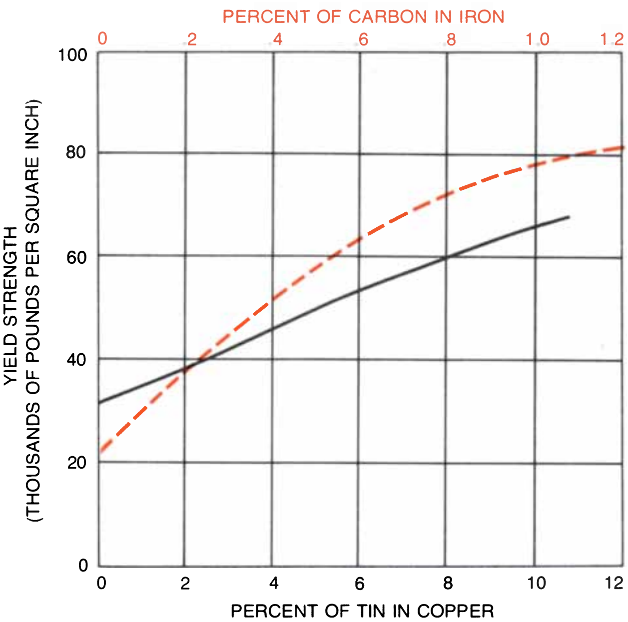
IRON SMELTER, seen here in a speculative reconstruction based on the remains of European Iron Age furnaces, was first filled with a mixed charge of ore, usually hematite or magnetite, and charcoal (*left*). The charge was ignited and the furnace temperature was raised to about 1,200 degrees Celsius by a draft. Because iron does not melt below 1,537 degrees C. the product of the smelting process (*right*) was the spongy mixture of nonmetallic wastes and iron (*color*) known as a bloom. The blacksmith reheated the bloom on a forge to above 1,170 degrees C., making the wastes viscous. The smith then removed the wastes from the iron by hammering. What remained was soft iron.

The hammering at the same time turned the porous iron "bloom" into a continuous network of iron grains interspersed with a few stringers of slag that had not been eliminated. The bloom was the blacksmith's raw material; the iron articles were made by heating and hammering the bloom further.

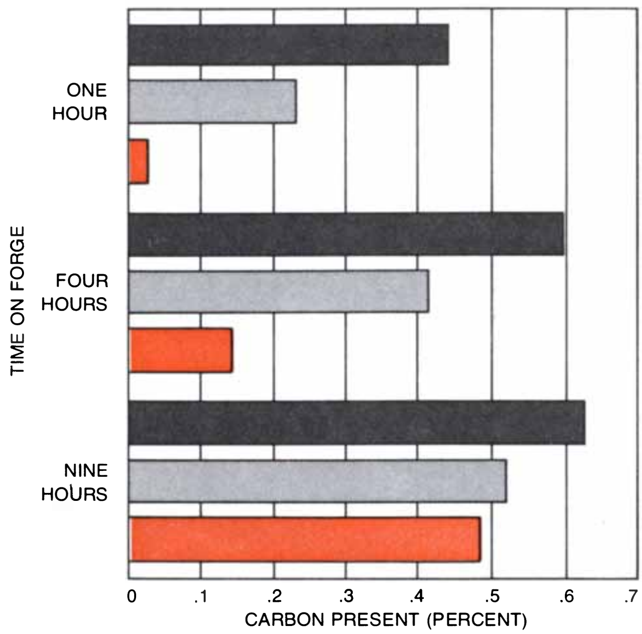
What the blacksmith had to work with was a poor substitute for bronze. Bloomery iron is a soft metal; its tensile strength is about 40,000 pounds per square inch, only slightly more than the strength of pure copper (about 32,000 p.s.i.). Work-hardening, that is, continued hammering, will bring the strength of iron up to almost 100,000 p.s.i. A bronze containing 11 percent tin, however, has a tensile strength after casting of some 60,000 p.s.i. and a strength after cold-working of as much as 120,000 p.s.i. Bronze was clearly a better material than bloomery iron for the manufacture of weapons and tools.

Bronze had other major advantages over iron. Since it melted at temperatures that the early metalworkers could attain, it was suited to casting. A bronze containing 11 percent tin begins to lose fluidity when it cools to 1,000 degrees C., and it is completely solid at 831 degrees. Since pure iron does not melt below 1,537 degrees, it could not be cast. When iron is alloyed with large amounts of carbon, say 4 percent, it can be made to melt at about 1,150 degrees. The resolidified metal, however, is very brittle. In any event iron was not cast before the middle of the first millennium B.C., when the process was pioneered by the Chinese in the Far East.

The casting methods used in the



EFFECT OF STEELING, the carburization of bloomery iron by alloying the metal with carbon, is to strengthen the metal until its yield strength is substantially greater than the yield strength of bronze. The addition of 1 percent of tin to copper produces a bronze alloy (*black*) with a yield strength of 30,000 pounds per square inch and the addition of 5 percent tin produces a bronze alloy with a yield strength of 50,000 p.s.i. The addition of .4 percent carbon to iron produces a carburized alloy (*color*) with a yield strength of more than 50,000 p.s.i. An iron that contains 1 percent carbon, in turn, is more than 10,000 p.s.i. stronger than a bronze that contains as much as 8 percent tin.



CARBURIZATION OF IRON results from the diffusion of carbon into the iron from a charcoal fire. The diffusion rate depends on the heat of the fire and the length of time the iron is in the forge. At a temperature of 920 degrees C. the amount of carbon diffused to a depth of half a millimeter below the surface of the iron in an hour (*top bars*) is nearly .5 percent by weight (*black*) but only .02 percent of the carbon has penetrated to a depth of 1.5 mm. (*color*). As the middle and bottom bars show, prolonged exposure increases the percentage of carbon diffused to depths of one mm. (*gray*) and 1.5 mm. below the surface of the iron without greatly increasing percentage at .5 mm.

THE CLASSIC GRAY'S ANATOMY 3 POUND, 1,257-PAGE COLLECTOR'S DELUXE EDITION

NOW ONLY $6.95!

AN EXTRAORDINARY PUBLISHING TRIUMPH—(ORIGINAL 1901 EDITION WORTH $200.00 OR MORE IF YOU CAN FIND ONE!)

For today's men and women truly interested in their own body and its functions. For students, artists and the medically curious—for everyone. The most fascinating book ever published!

IF you ever wanted to know how any part of your body really works, if you want to experience the same magic feeling a young medical student feels as the mysteries of the body are revealed for the first time, then this landmark edition of *Gray's Anatomy* is for you.

Considered one of the greatest texts of our time, *Gray's Anatomy* has been used by nearly every living physician and every physician has been exposed to it. But its appeal is not only to physicians, but to everyone, including students, artists, and collectors of classic books. Dr. John Crocco, Chief of Pulmonary Services, St. Vincent's Hospital and Medical Center of New York, writing in his new Introduction states:

"This stellar book represents the acme of anatomical description over the last century and will probably still be the premier text in anatomy over the next one hundred years."

Just what is Anatomy? Anatomy is the gateway to medicine and the queen of basic sciences. It is the parts and the whole of the human body and a prerequisite for entry into the field of human biology. In this, his masterpiece, Henry Gray unquestionably found the best method of imparting this knowledge to other minds.

To those interested, we say that this is an unprecedented opportunity to acquire *a magnificent classic* at the fantastically low price of only $6.95.

The book itself is divided into 16 main sections, with hundreds of major subdivisions.

The text includes:

The Skeleton. 181 pages covering all aspects of the 200 bones of the body.

The Articulations. 76 pages describing movements of joints, ligaments, cartilages, etc.

The Muscles. 158 pages covering every muscle from the tiniest in the eye to the largest in leg with every detail of their action and use.

The Blood-Vascular System. 167 pages describing circulation, arteries and veins. A veritable book within a book.

The Lymphatics. 14 pages on every lymph gland in the body from sex glands to neck and face glands.

The Nervous System. 171 pages on the brain, the extremities, the spinal cord, etc.

The Organs of Special Sense. 56 pages on the tongue, nose, eyes and ears.

The Organs of Digestion. 85 pages starting with the teeth and including the salivary glands, stomach, intestines, liver, etc.

The Organs of Voice and Respiration. 67 pages on the larynx, trachea, lungs, etc.

The Male Organs of Generation. 14 pages on the prostate glands, penis, testes, etc.

The Female Organs of Generation. 15 pages on the clitoris, vagina, ovaries, pregnancy, etc.

The Surgical Anatomy of Hernia. 20 pages including scrotal and femoral hernia.

General Anatomy or Histology. 73 pages on the animal cell, nutritive fluids, blood, the skin, secreting glands and much more.

Embryology. 66 pages on the embryo, fetus, ovum, and development of a baby.

The above list of contents is by the limitation of space just a mere summation. The actual content of *Gray's Anatomy* is so massive that the table of contents in the book needs 16 entire pages with 1,932 separate category entries. And the index of this masterwork covers 41 pages with 8,541 separate listings.

Again, as Dr. Crocco states, "There have been many imitations. there have been anatomical narratives, dissection manuals and pictorial atlases with exquisite photographs. However, there is only one *Gray's Anatomy*. *Gray's* is the book all students turn to."

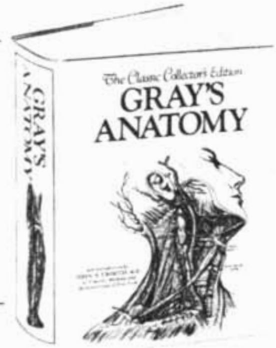
The medical contents alone makes this book invaluable, but the writing style makes it a true literary masterpiece. It combines both technical genius with a beautiful prose-like quality usually associated with a great novel. The 708 illustrations, including 172 original full color plates have never been surpassed for their authenticity and clarity.

Originally published more than 100 years ago, this classic edition is now available complete and in a new format that is outstanding for both its visual beauty and the unprecedented low price of only $6.95.

708 Illustrations with 172 in Full Color

Fig. 283—Surgical anatomy of the arteries of the neck, showing the carotid and subclavian arteries.

Here is an extraordinary opportunity to acquire one of the greatest books of all time. To order, simply fill out form below and enclose check or money order. You'll be extremely pleased that you did.



CANBRIDGE BOOKS, Dept. 81
6 Commercial Street, Hicksville, N.Y. 11801

Gentlemen. Please send me______copies of the Classic Collector's Edition of Gray's Anatomy at the introductory prices shown below. I enclose check or money order for $______

___Copies of Deluxe Heavy Laminated Soft Cover Edition @ only $6.95 each plus 75¢ for postage and handling.

___Copies of Special Hard Cover Library Edition @ only $8.95 each plus 75¢ postage and handling.

Name __________

Address __________

City __________

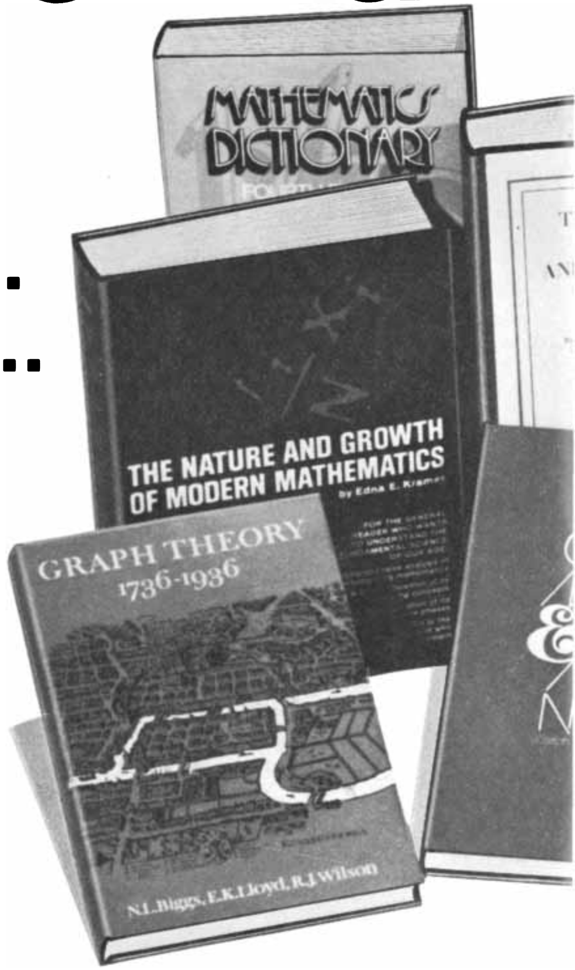
State ______ Zip ______

N.Y. Res. Add Applicable Sales Tax.

Announcing...

The Library of Math

the only book club devoted to mathematics.



Birkhoff...Knuth...Ulam... Magnus...von Neumann...

yes, you'll find all the people who have changed—and are changing—mathematics represented in this new book service. **The Library of Mathematical Sciences** is designed to bring you all the most useful, most important books in this fast-changing field—and at money-saving prices too. Here are the books you need to keep up on new mathematical tools and techniques... evolving interdisciplinary studies... emerging concepts and methods in computer sciences. Books that will spur your enjoyment of mathematics. Ideas that will excite your students. Problems to challenge you. Join **The Library of Mathematical Sciences** today, and see for yourself.

Take any 3 books all for only $3.95

(values to $66.95) if you will join for a trial period and agree to take 3 more books – at handsome discounts – within the next 12 months.

(Publishers' Prices shown)

81475. SURREAL NUMBERS. *Donald Knuth.* Mathematical creativity is explained and taught in this unique mathematical novelette. *Sofbound.* **$3.95**

61530. MATHEMATICS DICTIONARY. 4th Edition. *Edited by Robert James and Edwin Beckenbach.* More than 8,000 definitions. **$17.95**

39745. COMPANION TO CONCRETE MATH, Vol. 1. Mathematical Techniques and Various Applications. *Z. A. Melzak.* **$16.50**

61310. MATHEMATICAL IDEAS, MODELING AND APPLICATIONS. *Z. A. Melzak.* The accompanying volume to *Companion to Concrete Math, Vol. 1. Counts as 2 of your 3 books.* **$29.95**

69396. PLAYING WITH INFINITY/ARITHMETICAL EXCURSIONS. Two books to have fun with. First, Peter Rozsa concentrates on the concept of infinity, then Henry Bowers deals with little known facts concerning magic squares and other diversions. *Softbound. The 2 count as 1 book.* **$6.25**

61535. MATHEMATICS FOR ECONOMISTS. 2nd Edition. *Taro Yamane.* All the basic mathematical techniques currently employed by economists. **$14.95**

39540. COLLEGE MATHEMATICS WITH BUSINESS APPLICATIONS. 2nd Edition. *John Freund.* Basic business mathematics including elementary calculus, probability, linear programming, decision-making. **$14.95**

66061. ORDINARY DIFFERENTIAL EQUATIONS. *V. I. Arnold.* **$17.50**

50263. A GEOMETRIC INTRODUCTION TO TOPOLOGY. *G. T. Wall.* No knowledge of general topology assumed. Instead Wall works with point-sets in Euclidean space. **$11.95**

63860. THE NATURE AND GROWTH OF MODERN MATHEMATICS. *Edna Kramer.* From origins of digital computers in a Sumerian folk tale to future mathematics, a lively history. *Counts as 2 of your 3 books.* **$24.95**

32300. ADVENTURES OF A MATHEMATICIAN. *S. M. Ulam.* The birth and childhood of the Atomic Age written by a man who was intrumental in creating both the A- and the H-bombs. **$14.95**

83820. THEORY OF GAMES AND ECONOMIC BEHAVIOR. *John Von Neumann and Oskar Morganstern.* The classic exposition of the theory of games. *Counts as 2 of your 3 books.* **$25.00**

61300. MATHEMATICAL MAGIC SHOW. *Martin Gardner.* America's master of mathematical amusement presents his eighth array of puzzles, diversions, and arithmetical esoterica. Includes reader repartee and 133 drawings. **$8.95**

60760. THE MATHEMATICAL CARNIVAL. *Martin Gardner.* More brain ticklers, bizarre games, and wonderfully wacky sleight-of-mind tricks. **$8.95**

39837. THE COMPLEXITY OF COMPUTING. *John E. Savage.* An emerging discipline in applied mathematics. Spells out the trade-offs between computational resources. **$22.95**

60670. MATHEMATICS APPLIED TO DETERMINISTIC PROBLEMS IN THE NATURAL SCIENCES. *C. C. Lin and L.A. Segel.* Construction, analysis, and interpretation of the mathematical models used in science. **$16.95**

50051. GAME THEORY AND POLITICS/DECISION ANALYSIS. *Steven Brams* applies game theory to treat problems in political science, then *Howard Raiffa* shows how courses of action are selected under conditions of uncertainty. *Softbound.* **$14.90**

42010. THE DESIGN AND ANALYSIS OF COMPUTER ALGORITHMS. *Aho, Hopcroft, Ulman.* **$18.95**

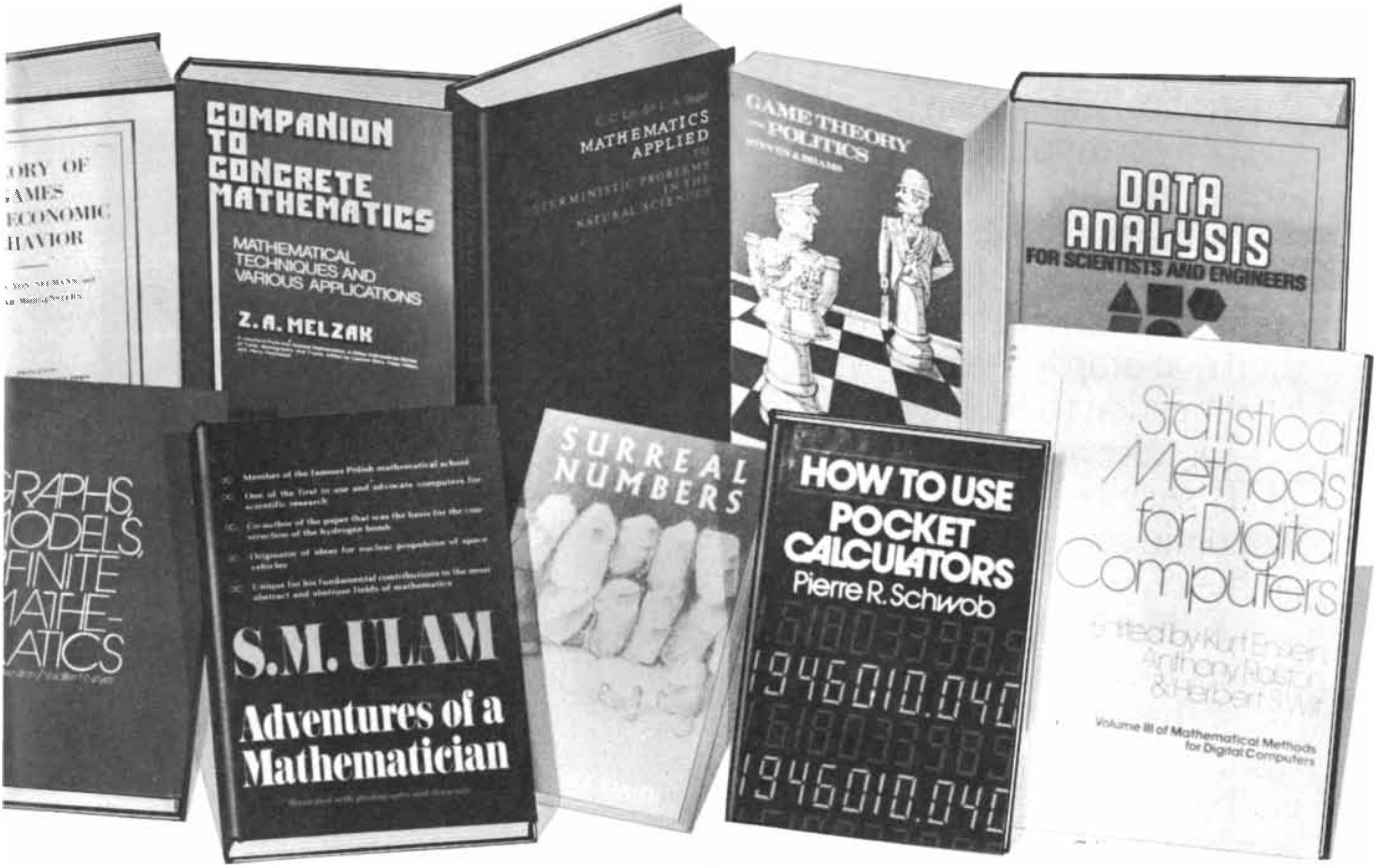
73191. QUEUEING SYSTEMS. Vol. II, Computer Applications. *Leonard Kleinrock.* How information-processing systems are used to solve real-world problems relating to computers and computer networks. *Counts as 2 of your 3 books.* **$25.00**

52292. A HANDBOOK OF NUMERICAL AND STATISTICAL TECHNIQUES. *J. H. Pollard. Counts as 2 of your 3 books.* **$24.95**

50560. GRAPH THEORY. *N. L. Biggs, E. K. Lloyd and R.J. Wilson.* An introduction to and history of graph theory. **$22.50**

53903. HOW TO USE POCKET CALCULATORS. *Pierre R. Schwob.* How to get the most out of a calculator. **$9.96**

ematical Sciences



00470. VAN NOSTRAND'S SCIENTIFIC ENCYCLOPEDIA. 5th Edition. Remarkable one-volume reference contains more than 2 million words. *Counts as 3 of your 3 books.* **$67.50**

39835. COMPLEX VARIABLES AND APPLICATIONS. 3rd Edition. *Ruel Churchill, James Brown, Roger Verhey.* Emphasis is on conformal mapping in boundary value problems and to aspects relating to important applications. **$11.95**

84396. TOPICS IN ALGEBRA. *I. N. Herstein.* Stimulating explorations of advanced algebra. **$17.50**

81490. A SURVEY OF MODERN ALGEBRA. 4th Edition. *Garrett Birkhoff and Saunders MacLane.* Linear algebra, abstract algebra, Boolean-algebra and lattice theory. **$13.50**

39585. COMBINATORIAL GROUP THEORY. *Wilhelm Magnus, Abraham Karrass, and Donald Soliter.* Factor groups, subgroups, Nielsen transformations, commutator calculus and more. **$6.00**

56260. INTRODUCTION TO FINITE MATHEMATICS. 3rd Edition. *John G. Kemeny, J. Laurie Snell, Gerald Thompson.* Hundreds of problems, exercises, charts, tables. **$14.95**

50576. GRAPHS, MODELS AND FINITE MATHEMATICS. *Joseph Malkevitch and Walter Meyer.* Delightful introduction to practical finite mathematics, using current examples (mail routes to predator-prey interaction). **$12.50**

70350. AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS, Vol I. 3rd Edition. *William Feller.* Broad array of practical problems and latest techniques for solving them. **$15.00**

70190. PRINCIPLES OF OPERATIONS RESEARCH. 2nd Edition. *Harvey Wagner.* Complete overview of all the fundamental concepts. **$21.50**

80530. STATISTICAL METHODS FOR DIGITAL COMPUTERS. *Edited by Kurt Enslein, Anthony Ralston, and Herbert Wilf.* 19 outstanding experts focus on multivariate methods. **$19.95**

41610. DATA ANALYSIS FOR SCIENTISTS AND ENGINEERS. *Stuart Meyer.* The statistical methods, concepts and techniques that experimental scientists need to deal with data. **$19.95**

MEMBERSHIP BENEFITS • You begin by choosing any three of these exciting books for only $3.95 as your introductory offer. • You keep saving substantially on the books you buy. Savings range up to 30% and occasionally even more. • If you continue your membership past the trial period, you will be eligible for our Bonus Book Plan—an important way to save more, at least 70% off publishers' prices. • At 3-4 week intervals (15 times per year) you will receive the Book Club **News,** describing the coming Main Selection and Alternate Selections, together with a dated reply card. In addition, up to 4 times a year, you may receive offers of special selections, always at substantial discounts. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. • You will have at least 10 days to decide. If, because of late mail delivery of the **News,** you should receive a book you do not want, we guarantee return postage.

The Library of Mathematical Sciences
Riverside, New Jersey 08075

Please accept my application for membership and send me the three volumes indicated, billing me only $3.95. I agree to purchase at least three additional books during the first 12 months I am a member, under the membership plan described in this ad. Savings range up to 30% and occasionally even more. My membership is cancelable any time after I buy these three books. A shipping and handling charge is added to all shipments.

3 books for only $3.95.
Indicate by number the 3 books you want.

A few expensive books (noted in book descriptions) count as 2 or 3 choices.

Name ______________________

Address ______________________

City ______ State ______ Zip ______

(Offer good in Continental U.S. and Canada only. Prices slightly higher in Canada.) Books purchased for professional purposes may be a tax-deductible expense.

M-8AA

Businessmen love Qantas.

Qantas 747B's are faster than Pan Am's SP to Australia. Faster than anybody's anything to Australia.

Take the Qantas Australian Express. From San Francisco to Honolulu, then non-stop to Sydney with a connection to Melbourne.

Or take our regular 747B service which calls at Honolulu and Nadi, Fiji, on the way to Sydney.

Either way, Qantas is fastest to Australia.

What's more, every Qantas flight is a 747B. This is a bigger, more comfortable airplane than Pan Am's SP.

Nobody beats us, so why not join us?

See your Transportation Manager or Travel Agent.

QANTAS
The Australian Airline.

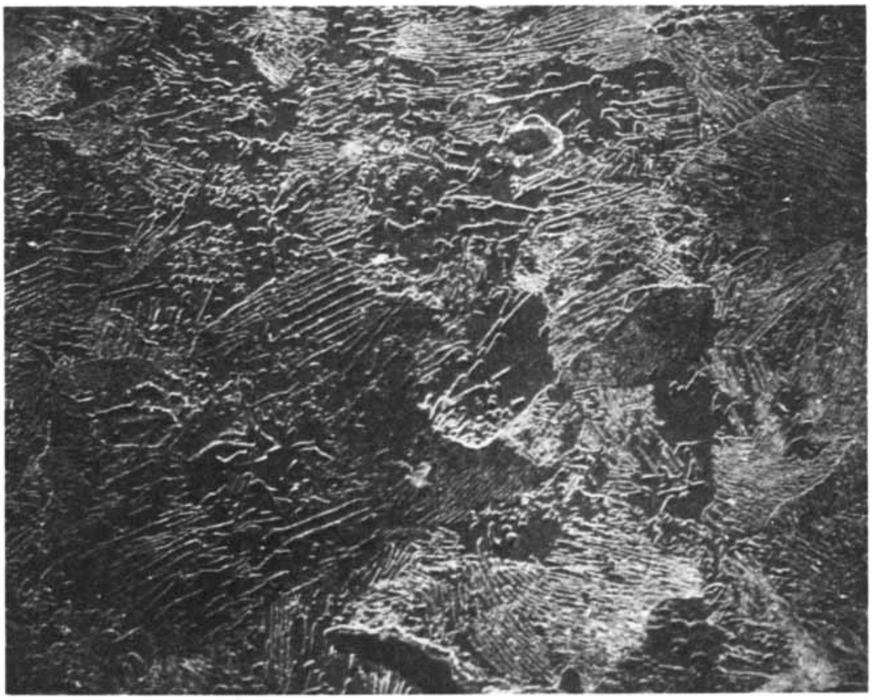
Because Qantas is fastest to Australia.

Bronze Age, as indicated by actual molds that have survived or as inferred from written references, were remarkably diverse. There were one-piece, two-piece and multipiece molds made of earth, clay and stone. Some molds had cores, and the lost-wax process made it possible to cast the bronze in complex shapes and with elaborate surface decoration. Iron, except for the brittle high-carbon metal, was useless for any of these processes, which turned out not only utilitarian articles but also a great number of purely ornamental objects such as figurines and jewelry. Bronze had one further advantage. It corrodes slowly, and the characteristic green patina is considered decorative. Iron corrodes rapidly, and in the process often suffers considerable damage.

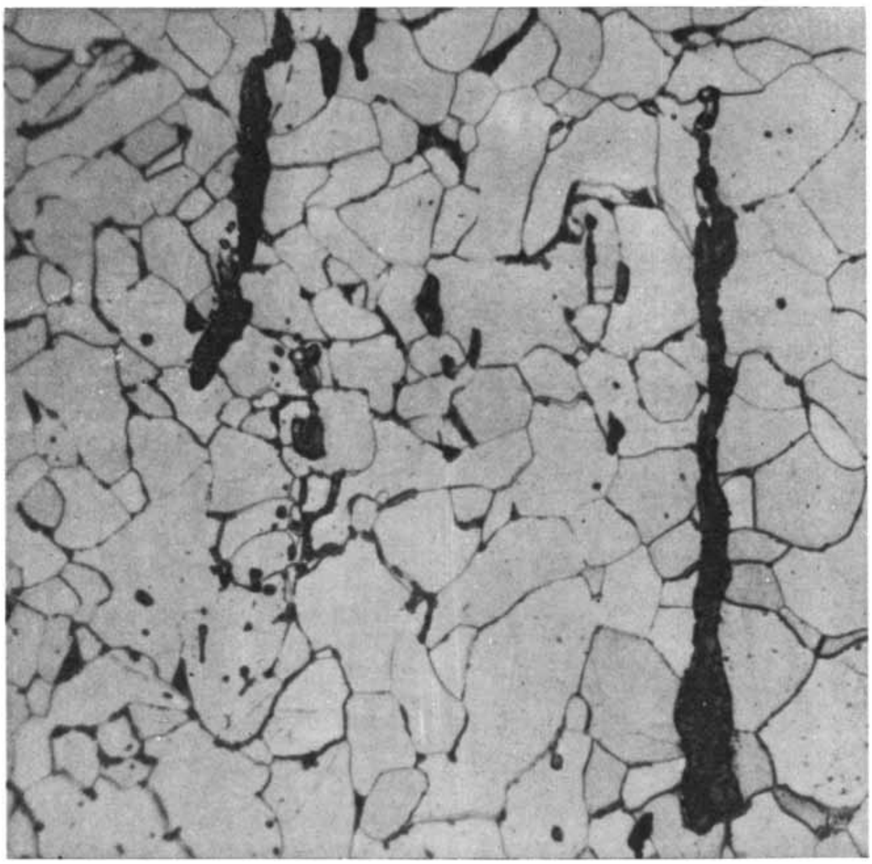
How, then, could iron become a satisfactory substitute for bronze as the second millennium B.C. drew to a close? The answer is that if bloomery iron is treated in a certain way, it can be transformed into an alloy that is for most purposes far superior to bronze. That treatment is steeling, and its initial discovery was probably accidental. What happened was as follows. When the blacksmith reheated the iron bloom in order to hammer out the slag, he did so with a charcoal fire in the forge. He needed to heat the bloom to 1,200 degrees C. to make the slag viscous, and he probably did not let the temperature go much below 800 degrees until the work was finished. The bloom was in direct contact with the white-hot charcoal and with the hot carbon monoxide evolved by its combustion. In that temperature range a small amount of carbon from both sources slowly diffused into the iron, in effect converting it into carbon steel down to a certain depth below the surface.

The time required for carbon to diffuse into iron follows simple physical laws. For example, if one plots carbon concentration and depth of penetration at a temperature of 950 degrees C., one finds that after nine hours the concentration at a depth of 1.5 millimeters below the surface is .5 percent. At higher temperatures the carbon atoms diffuse into the iron more rapidly; at 1,150 degrees C. after nine hours the concentration at the same depth can reach 2 percent.

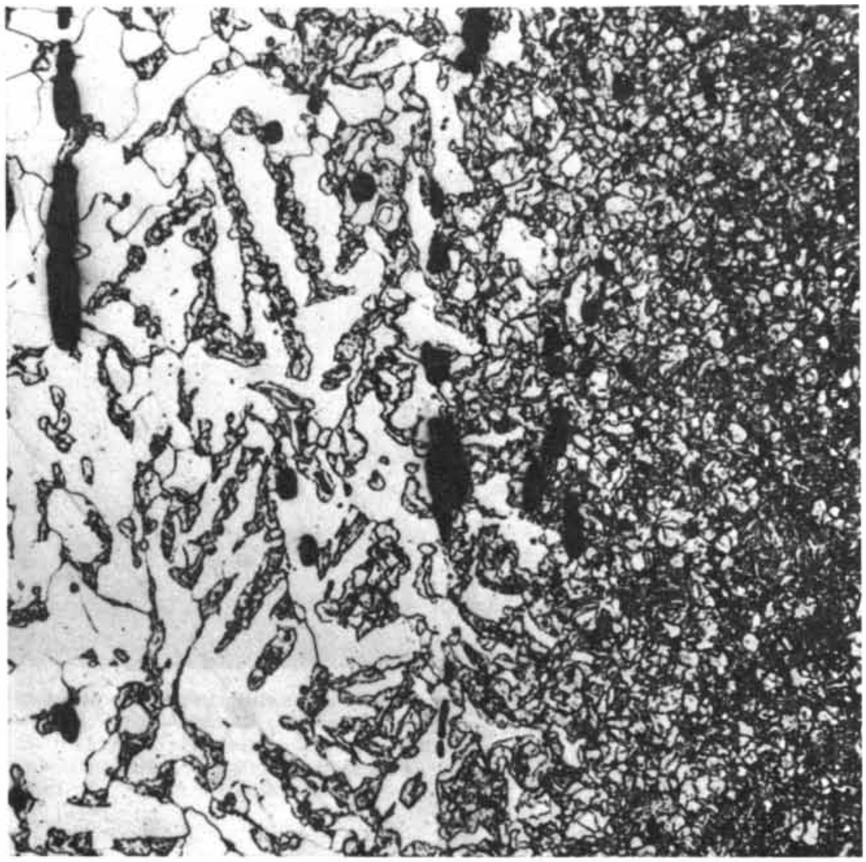
In modern metallurgical terms carburized iron at a temperature higher than 910 degrees C. has the microstructure of the form of steel known as austenite. When the temperature falls below 727 degrees C., the austenite breaks down into two components. One is ferrite, or pure iron. The other is the iron carbide known as cementite. Called the eutectoid reaction, this two-phase breakdown gives rise to the microstructure of the form of steel known as pearlite: alternating layers of ferrite and



PEARLITE, a characteristic microstructure of carbon steels, has a layered pattern at high magnifications. This scanning electron micrograph enlarges the sample 1,250 diameters. It shows a gold-covered replica of the corroded surface of a steeled-iron blade found at a site in Israel: Tel Fara South. Although the iron is oxidized a relic pearlite microstructure is preserved. The blade was made available for study by the Institute of Archaeology of the University of London.



IRON SPIT from a site in Greece is seen in longitudinal section at a magnification of 200 diameters. The narrow dark areas are stringers: fragments of slag in the original iron bloom that were deformed but not expelled by the hammering preparatory to the forging of the spit. The iron artifact was made available for study by the staff of the Numismatic Museum in Athens.



METALLURGICAL SKILL in the fourth century B.C. is indicated in this micrograph. Shown at a magnification of 300 diameters is a section of an adze blade unearthed at Al Mina, the ruins of a Greek trading colony on the coast of Turkey near Syria. The blacksmith who made the implement used a sheet of carburized iron for the working face of the adze but economized by using soft iron for the other face. The micrograph, showing the juncture between fine-grained steeled iron and coarse-grained uncarburized iron, is clear proof of the blacksmith's mastery.

cementite. Reflected in this microstructure is the proportion of carbon in the metal. If the iron is free of carbon, no pearlite will be present. If the alloy contains as much as .8 percent carbon, the microstructure that is formed will be 100 percent pearlite.

It is illuminating to compare the tensile strengths of carburized iron and bronze. A carbon content ranging from .2 to .3 percent gives the steeled iron a strength equal to that of unworked bronze: about 60,000 p.s.i. If the carbon content is raised to 1.2 percent, the steeled iron has a tensile strength of 140,000 p.s.i., which is somewhat greater than the strength of cold-worked bronze. If the blacksmith then cold-hammers the steeled iron, its tensile strength can be increased to 245,000 p.s.i., or more than twice the strength of cold-worked bronze.

The accidental discovery of steeling must have encouraged experimentation, because in due course the early blacksmiths could control the process well enough to develop properties in the metal appropriate to the function of the object they were making. It is possible that future metallurgical studies (and even some that are now in progress) will help to trace the progression from accidental steeling to purposeful steeling. For example, carbon can get into iron by pathways other than carburization. It can be trapped in the pores of the iron bloom; forging would then form the trapped carbon into stringers within the metal. In a polished metallographic section of such an accidentally steeled piece the stringers would appear as uneven streaks. Entirely accidental carburization through exposure to charcoal in the course of heating and forging, although more difficult to detect microscopically, should leave some trace. The carbon content would probably be low and the concentration would be uneven. It is incontrovertible proof of deliberate steeling, however, when one finds an iron object consisting of layers with dissimilar percentages of carbon. The blacksmith would have no reason to make such an object unless he understood the different properties of the different layers. The earliest-known object of this kind is an Egyptian iron knife that was probably made between 900 and 800 B.C.

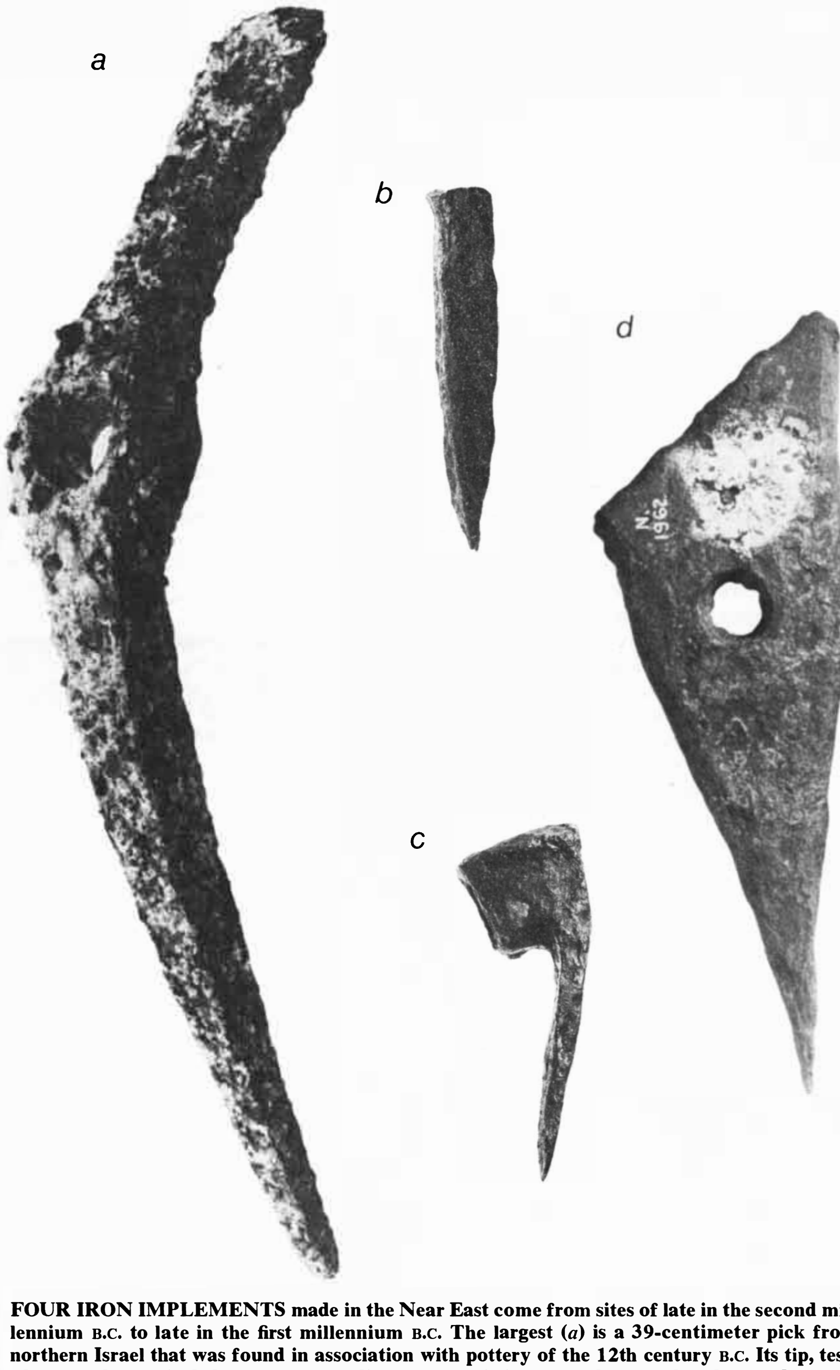
On the basis of our research and that of others it seems evident that by the beginning of the 10th century B.C. blacksmiths were intentionally steeling iron. Very few iron artifacts of the 12th century B.C. or earlier have been metallurgically analyzed, so that the prevalence of steeling in that period remains uncertain. Nevertheless, a 12th-century knife from Idalion, a site on Cyprus, was certainly carburized to improve its hardness. A site on Mount Adir in northern Israel has yielded an iron pick in association with 12th-century pottery. One would hesitate to remove a sample from the pick for analysis, but it has been possible to test the tip of it for hardness. The readings averaged 38 on the Rockwell "C" scale of hardness. This is a reading characteristic of a modern hardened steel.

Unlike the Mount Adir pick, many of the ancient iron artifacts that have survived in eastern Mediterranean sites are badly corroded, a condition that complicates metallurgical study. Several techniques, however, make it possible to detect pearlite even in objects that are totally oxidized. In one such technique a polished surface is prepared for study under the microscope at magnifications greater than 1,000 diameters. Since the ferrite layers of the pearlite differ from the cementite layers in their chemical composition, it is sometimes possible to detect them even in iron oxide. An even more promising technique makes use of the scanning electron microscope. Here it is not the polished surface that is examined but a gold-plated thin-film replica of it. The ferrite layers are softer than the cementite ones, even in a corroded artifact, and they stand out clearly when the replica is tilted to accentuate the different levels of the two constituents. The presence of pearlite in an iron artifact is a clear indication that the artifact has been carburized.

After 900 B.C. the production of iron implements rapidly increased. Even though at this time tin again became available in the eastern Mediterranean, bronze did not replace iron. Sites dating from the 10th century B.C. to the sixth have yielded great hoards of iron implements. At Hasanlu in northwestern Iran a University of Pennsylvania expedition unearthed iron weapons numbering in the thousands. At Gordion, the capital of ancient Phrygia, another University of Pennsylvania expedition discovered one of the largest collections of iron artifacts ever found in the eastern Mediterranean. At Nimrud in Iraq, the excavations of Max Mallowan of the University of Oxford uncovered another major iron assemblage.

The evidence for the growing popularity of iron goes beyond archaeological finds. Neo-Assyrian and Neo-Babylonian writings reflect a world technologically quite different from that of the period before 900 B.C. References are made to iron axes, iron hoes, iron picks, iron saws, iron arrowheads, iron scissors, iron fetters and even iron furniture and iron lamps. Iron also became the metal of choice for knives and daggers. A passage from the corpus of Babylonian "wisdom literature" strikes a sour note to this effect: "A woman is a pitfall. a hole, a ditch, a woman is a sharp iron dagger that cuts a man's throat."

We have not yet mentioned a second process that significantly enhances the quality of carburized iron. This is quenching: quickly cooling a hot piece of metal by plunging it into water. An article of steeled iron that is left to cool



FOUR IRON IMPLEMENTS made in the Near East come from sites of late in the second millennium B.C. to late in the first millennium B.C. The largest (*a*) is a 39-centimeter pick from northern Israel that was found in association with pottery of the 12th century B.C. Its tip, tested for hardness, yielded an average reading characteristic of a modern hardened steel. The chisel (*b*), from Al Mina, is almost 13 centimeters long; it was cleverly quenched to make its matrix hard without leaving its cutting edge brittle. The small adze (*c*), from Al Mina, is almost 12 centimeters long. It may have been made from the same bloom of iron as the chisel and perhaps by the same blacksmith. Its microstructure is shown in the illustration on page 126. The fourth tool (*d*), 25 centimeters long, comes from the ruins of Nimrud in northern Iraq. Under microscopic examination it shows the pearlite structure characteristic of carburized iron.

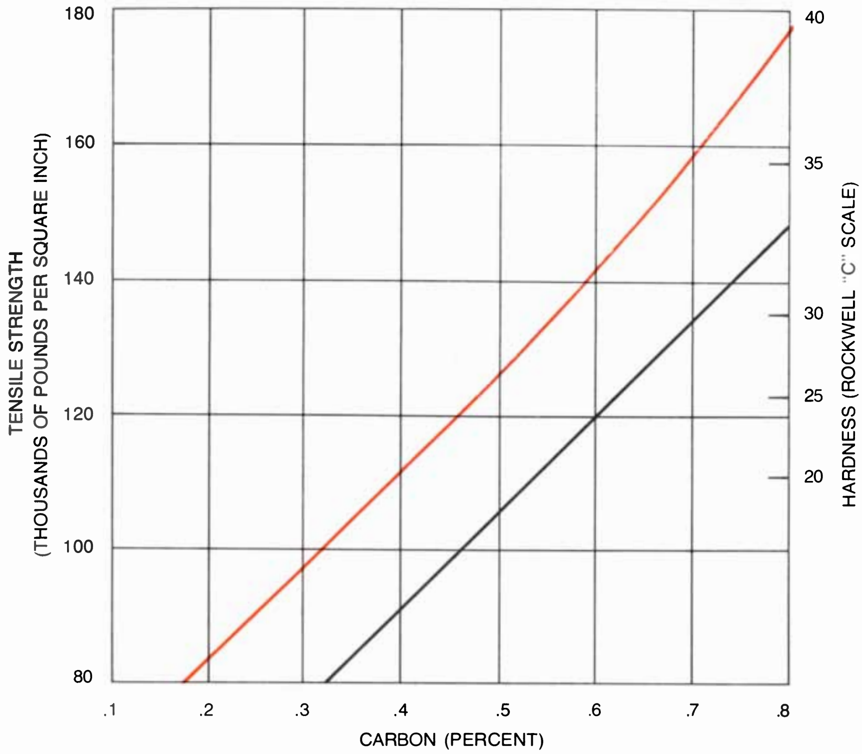
by itself in the open air develops a microstructure of coarse pearlite. If the blacksmith instead waves the finished article in the air, accelerating the process of cooling, the pearlite microstructure is much finer. Even faster cooling by quenching can suppress the development of pearlite altogether; the steeled material has a quite different microstructure and is known as martensite. Martensite is significantly harder than pearlite, although it is quite brittle.

When one displays the results of quenching graphically, it appears that different rates of cooling—from the temperature of the furnace or forge, about 1,200 degrees C., to the temperature of transformation, about 700 degrees—give rise to different microstructures. In an iron containing .8 percent carbon, if the cooling period is approximately 60 seconds, coarse pearlite will form. To produce fine pearlite the cooling period should be only two or three seconds. To produce martensite the temperature must be reduced from the forge level to below 220 degrees C. in less than a second; only quenching can cool the material so quickly. One may assume that for small objects such as arrowheads hardness would be important and brittleness could be tolerated. If such small objects were made from iron with a .8 percent carbon content quick quenching would make them martensite. The same could have been accomplished with the surface layers of larger iron objects where hardness was important and brittleness could be tolerated, even though the interior of the object, being slower to lose heat, would remain pearlite.

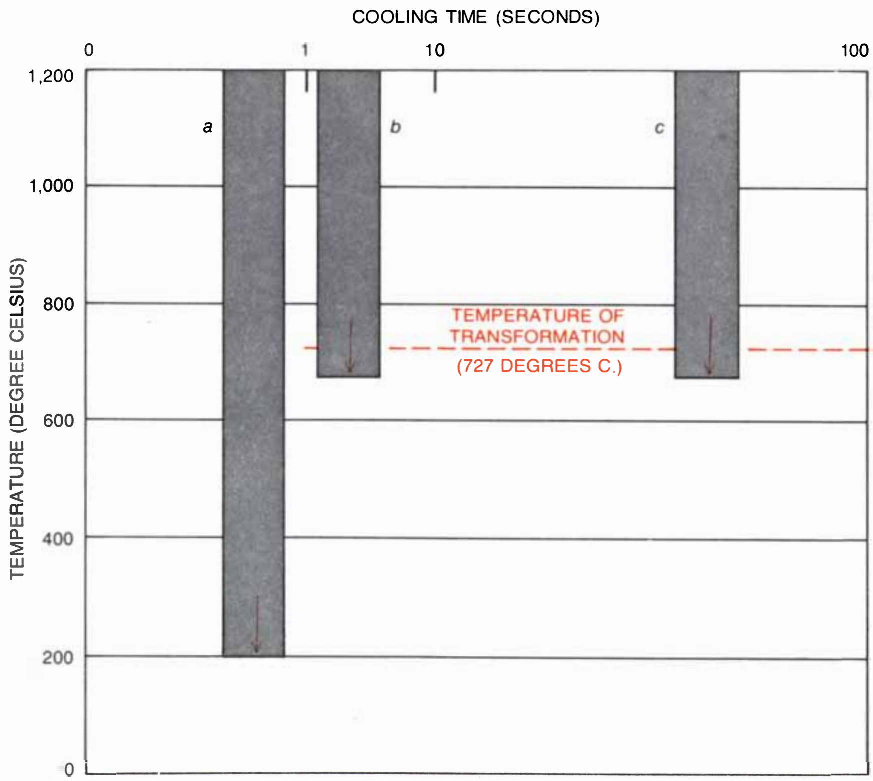
It is not possible to estimate when the quenching process was invented; like steeling, it could easily have been discovered accidentally. One item of literary evidence, however, clearly indicates that blacksmiths of the eastern Mediterranean were familiar with the process in the eighth or seventh century B.C. The passage is in the ninth book of the *Odyssey.* Trapped in the cave of Polyphemus, the one-eyed giant, Odysseus and his men manage to get the giant drunk. They decide to blind him by snatching a burning olive trunk out of a fire and thrusting it into his eye. Our translation follows Richmond Lattimore's:

"As when a man who works as a blacksmith plunges into cold water a great axe or adze which hisses aloud, 'doctoring' it, since this is the way that steel is made strong, even so Cyclops' eye sizzled about the beam of the olive."

The description could only have been written by someone who had watched a blacksmith quench hot iron and knew that the quenching was done to increase the hardness of the metal. It also suggests that quench-hardening was something of a novelty in the Greek world at that time. "'Doctoring' it" is a free rendering of the Greek word *pharmasso;* the



STRENGTH AND HARDNESS of carburized iron both increase as the percentage of carbon in the alloy rises. Degrees of hardness here are measured on the Rockwell "C" scale. If the speed of cooling is not rapid, the steel microstructure that develops is coarse pearlite (*black*). Regardless of its carbon content, coarse pearlite steel is less durable than fine pearlite steel (*color*).



EFFECT OF QUENCHING on the microstructure of carburized iron is shown on this graph in terms of the length of time required to lower the temperature of the iron from its forging heat (about 1,200 degrees C.). If cooling requires less than one second (*a*), the microstructure that forms is martensite, tough but somewhat brittle. If up to three seconds is required, the resulting microstructure is fine pearlite (*b*). Cooling as quick as this calls for quenching. Prolonged cooling, in excess of 10 seconds, forms a third kind of microstructure: coarse pearlite (*c*).

implication is that the iron was being treated in some magical way, as if by means of drug potions. Perhaps Homer was puzzled by the fact that whereas water softened or even dissolved many materials, it turned carburized iron into a metal harder than any that had been known before.

It is more difficult to obtain evidence of deliberate quenching than evidence of deliberate steeling. For one thing, water quenching could have been done simply to cool a forged object quickly, perhaps to be able to use it immediately or perhaps to avoid having too many hot objects around the smithy. For another, with an object of any size it is the outer layer that consists of martensite, and it is this layer that is most likely to have been removed by corrosion over the centuries. Nevertheless, some meager evidence of deliberate quenching exists. For example, so far 11 iron artifacts from Nimrud have been analyzed metallurgically. Although all are heavily corroded, five of them show possible indications of quenching. The rest show signs of some carburization but evidently were not quenched.

A third technique of ironworking arises directly from quenching. This is the process of tempering, a practice that reduces the brittleness induced by quenching. The early blacksmiths must have realized soon that quenching made their products brittle. The quenching would have left cracks along the edges of many articles, and users would have complained of breakage. Tempering, that is, reheating up to but not above the temperature of transformation (727 degrees C.), affects the iron carbide that quenching forces into the microstructure of iron so as to give rise to martensite. The carbide precipitates and then coalesces by diffusion. Both the exact temperature attained and the time that the article is held at that temperature determine the amount of iron carbide that coalesces and hence the final hardness and ductility of the metal. As ductility increases, hardness decreases.

For the blacksmiths of antiquity tempering was probably never done intentionally. There is no change in the color of the hot iron at the critical range of temperatures and thus no obvious way of gauging the correct heat in the forge. A method was developed early in the fourth century B.C., however, that surmounted the difficulty of achieving true tempering and at the same time produced carburized-iron tools that were both hard and durable. Our evidence for this statement is a stonecutter's chisel unearthed at the Greek trading colony of Al Mina, a site in coastal Turkey near the Syrian border. The chisel has a matrix composed largely of martensite but containing nodules of pearlite that increase in density toward the tip of the tool, rather than toward the interior as one would expect. Laboratory experiments suggest that this reverse distribution was achieved by covering the tip of the tool with some kind of insulating material, such as clay, heating the piece and then quenching it. Clay would crack off as soon as the tool hit the water, but it would nonetheless slow the cooling rate enough to reduce the amount of martensite formed at the cutting end of the tool. It was evidently by some cheap and simple means such as this one that the blacksmith of Al Mina created a chisel with a strong body and a durable edge.

Another example of innovative smithing was found at Al Mina. Analysis of the elements present in an adze uncovered with the chisel suggests that both tools were made from the same bloom of iron and thus perhaps even made by the same smith. The adze was formed by hot-hammering two sheets of iron together. One sheet was carburized and the other was not. The combined sheets were forged into shape, and the finished adze was air-cooled rather than quenched. The blacksmith, if he was the same man who made the chisel, evidently appreciated the fact that a woodworking adze need not have as hard an edge as a stoneworking chisel. He adjusted the manufacturing process accordingly, probably at a net saving in time and effort. The moment when smiths were first able to control the ironworking process so that the properties of the products were suited to their end uses is clearly a significant one in the history of technology. Such a moment may well be what we see in the finds at Al Mina.

To sum up, by the beginning of the seventh century B.C. at the latest the blacksmiths of the eastern Mediterranean had mastered two of the processes that make iron a useful material for tools and weapons: carburizing and quenching. And by the beginning of the fourth century B.C. at the latest a method had been found to overcome the disadvantages of brittle steel while preserving the advantage of its hardness. The smiths' quest had involved hundreds of years of experimentation and uncountable hours at the furnace and the forge. All the craftsmen of antiquity—potters, masons, stonecutters, weavers, carpenters and smiths of bronze, copper and precious metals—shared the blacksmith's empirical approach to their work, but none of them faced as large a challenge as he did. To be made strong and durable iron needs sophisticated treatments. The blacksmith could not have understood, at least at first, why these treatments improved the iron, but his tenacity, pragmatic knowledge and skill in the centuries following the end of the second millennium B.C. enabled the peoples of the eastern Mediterranean to take the momentous step from the Bronze Age to the Iron Age.

Hallucinations

These false perceptions, which can occur in any of the senses, turn out to be much alike from one person to another. Apparently they have their roots in excitations of the central nervous system

by Ronald K. Siegel

A motorist who drives alone at night in a state of extreme fatigue may well perceive things that are not there: people, animals, vehicles or strange forms. Such a percept is characterized as a hallucination. Although the definition of the word (which comes from the Latin *hallucinari,* meaning to prate, to dream or to wander in mind) is far from precise, one that is widely accepted in psychiatry is: "A false sensory perception in the absence of an actual external stimulus. May be induced by emotional and other factors such as drugs, alcohol and stress. May occur in any of the senses."

By this definition it is likely that everyone has had a hallucination at one time or another. Lonely explorers, isolated hunters in the Arctic and prisoners in dark cells have reported experiencing them. Some people seek out the experience by taking hallucinogenic drugs. Under the right social circumstances the perceptions may be regarded as valid. Joan of Arc became a saint because of her visions, and the flashes of light perceived by the astronauts were taken quite seriously. (They were actually caused by cosmic rays.) On the other hand, negative evaluations are applied to similar perceptions by inmates of correctional institutions. All such reports, however, are necessarily subjective. When one has a hallucination, one does so alone, in the privacy of one's mind.

Do the hallucinations of one person have anything in common with those of another? My colleagues and I in the Neuropsychiatric Institute of the University of California at Los Angeles have attempted to answer the question by means of experiment. We find that hallucinations do have a great deal in common. Moreover, the experiments point to underlying mechanisms in the central nervous system as the source of a universal phenomenology of hallucinations.

One of the earliest classifications of hallucinations was offered in 1853 by Brierre de Boismont of France. He found that hallucinations occurring in states of insanity, delirium tremens, drug intoxication, nervous disorders, nightmares, dreams, ecstasies and fevers were all characterized by excitation and the production of images from memory and the imagination. His countryman Jacques Moreau described hallucinations as being similar to dreams in which imagined visual, auditory and tactile stimuli seem to be real. Foretelling what future neurophysiological research would reveal, he maintained that hallucinations resulted from excitation of the brain. Moreau's technique, which he described in 1845, was to take hashish, which put him in a hallucinatory state while leaving him able to report his experiences. (Moreau also tried to persuade his medical colleagues and friends to take hashish. His colleagues were hesitant because they did not view the taking of hashish as an acceptable form of objective experimentation. The Bohemian artists and writers of 19th-century Paris were more receptive. One of them, the novelist Théophile Gautier, went on to organize the Club des Haschichins, whose members included Balzac, Baudelaire, the younger Dumas and Victor Hugo. Some of the club members' writings testify to the richness of the imagery induced by hashish.)

In Germany and the U.S. early students of hallucinations followed a similar course of self-experimentation and focused mainly on visual hallucinations. Using mescaline, a hallucinogenic alkaloid derived from the peyote cactus *Lophophora williamsii,* Heinrich Klüver began a series of investigations at the University of Chicago in 1926. He reported that mescaline-induced imagery could be observed with the eyes either closed or open and that with the eyes open it was impossible to look at a blank wall without seeing it as being covered with various forms.

Among these forms Klüver found four constant types. One he described with terms such as grating, lattice, fretwork, filigree, honeycomb and chessboard. A second type resembled cobwebs. A third was described with terms such as tunnel, funnel, alley, cone and vessel. The fourth type consisted of spirals. The form constants were further characterized by varied and saturated colors, intense brightness and symmetrical configurations. The visions seemed to be located at reading distance. They varied greatly in apparent size. In general they could not be consciously controlled.

Klüver made the crucial observation that these form constants appear in a wide variety of hallucinatory conditions. He listed a number of the conditions, and other investigators have added to his list, which now includes falling asleep, waking up, insulin hypoglycemia, the delirium of fever, epilepsy, psychotic episodes, advanced syphilis, sensory deprivation, photostimulation, electrical stimulation, crystal gazing, migraine headaches, dizziness and of course a variety of drug intoxications.

Most of the drugs that give rise to such imagery are classifed as hallucinogens. Other drugs and substances can give rise to similar effects, however, and so most psychoactive compounds (to the extent that they cause the mind or the attention to wander) can be regarded as hallucinogens. In this category are alcohol, carbon dioxide, cocaine, cortisol, digitalis, scopolamine and even tobacco with a high concentration of nicotine.

The form constants appear in the first of two stages of drug-induced imagery. The images of the second stage, which are more complex but can incorporate the simple constants, include landscapes, faces and familiar objects and places. The complex images, which are perhaps the most dramatic aspect of the hallucinatory experience, are usually regarded as an activation of images already recorded in the memory.

One would expect the forms and scenes of complex imagery to be almost infinitely diverse. Actually constants appear even at this stage. Indeed, a review of more than 500 hallucinations induced by lysergic acid diethylamide (LSD) revealed that whereas between 62 and 72 percent of the subjects experienced the simple form constants, more than 79 percent reported quite similar complex images. They included religious symbols and images (72 percent) and images of small animals and human beings (49 percent), most of them friendly and many in the nature of cartoons and caricatures.

Most of the investigators and the subjects did not describe the complex imagery in detail. Moreover, before Klüver's classic work little was said about the geometry of even simple hallucinatory images. Klüver attributed this omission to the novelty of the visions, expressing the view that many hallucinating people are so overwhelmed by the color or brightness of the images that they do not articulate the basic forms. It was this inarticulateness that challenged Klüver to describe the simple first stage of hallucinatory imagery. The apparent complexity of the second-stage images challenged us to do the same for them. We believed the study of such phenomena might point to a common visual imagery underlying hallucinations and so might help us understand the origin of these percepts and related ones.

Our first experiments were designed to see if the hallucinatory phenomena I have described appeared when hallucinations were induced in untrained subjects by drugs. Each subject was given either a standard dose of a hallucinogen (usually marihuana or its active principle, tetrahydrocannabinol) or an inactive placebo. (The subject did not know which substance he was receiving.) He was then asked to lie on a bed in a lightproof and soundproof chamber and to report his experiences. We recorded the reports on tape and analyzed them according to the frequency of different

HALLUCINATORY SHAPES AND COLORS are represented in this yarn painting made by a member of the Huichol Indian group in Mexico. The picture was made to show visions of the kind experienced in hallucination brought on by taking peyote. The Indian at the left is carrying a basket of freshly harvested peyote and viewing a vision that is exploding with color and streaks and flashes of light. The peyote cactus is represented at the right. The picture was made by putting beeswax on wood and then pressing yarn into the beeswax.

forms, colors, movements and complex images.

The results showed that normal imagery (that is, imagery not induced by drugs) is characterized by amorphous black-and-white forms (sometimes including lines and curves) that move about randomly in the visual field. Anyone who closes his eyes or goes into a dark room is likely to experience a baseline imagery of this kind. Indeed, one can induce similar imagery, some of it brightly colored and geometric, by gently rubbing one's closed eyelids.

With hallucinogenic drugs, however, the number of images reported by the subjects rose sharply. The consensus was that the imagery resembled what one would see in a motion picture or a slide show. A number of the subjects had difficulty in describing the imagery, but they agreed that there were many geometric forms in it. The imagery was characterized by a bright light in the center of the field of vision that obscured details but allowed images on the periphery to be observed.

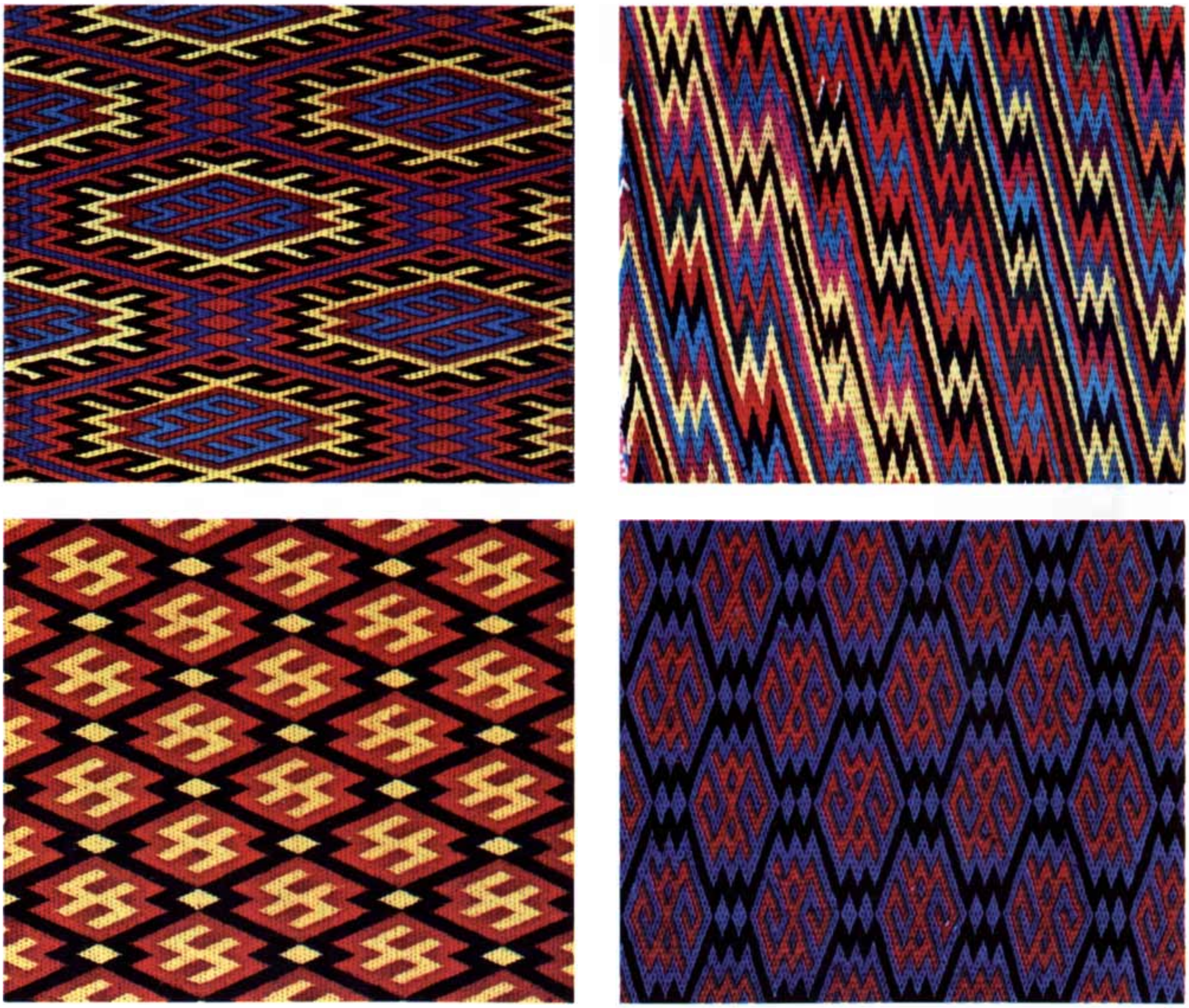
The location of this point of light created a tunnel-like perspective. The subjects reported viewing much of their imagery in relation to a tunnel. According to their reports, the images tended to pulsate, moving toward the center of the tunnel or away from the bright light and sometimes moving in both directions.

When the images appeared in color, all colors were reported, although the incidence of reports that the color was red increased as the dosage of the drug increased. Geometric forms frequently combined, duplicated and superimposed. At times the flow of imagery was so rapid that most subjects found it difficult to maintain a running commentary.

The geometric forms were soon replaced by complex imagery. The complex images reported included recognizable scenes, people and objects, many in cartoon or caricature form, with some degree of depth and symmetry. The images were often projected against a background of geometric forms.

In listening to these reports we encountered certain difficulties. The subjects differed widely in their choice of words. Moreover, the reports were riddled with idiosyncratic experiences. We therefore decided to facilitate the ease and accuracy of reporting by training our subjects to use a standard descriptive code.

Efforts along this line had previously been made at Harvard University by the psychologists Ogden R. Lindsley and Timothy Leary. They employed the



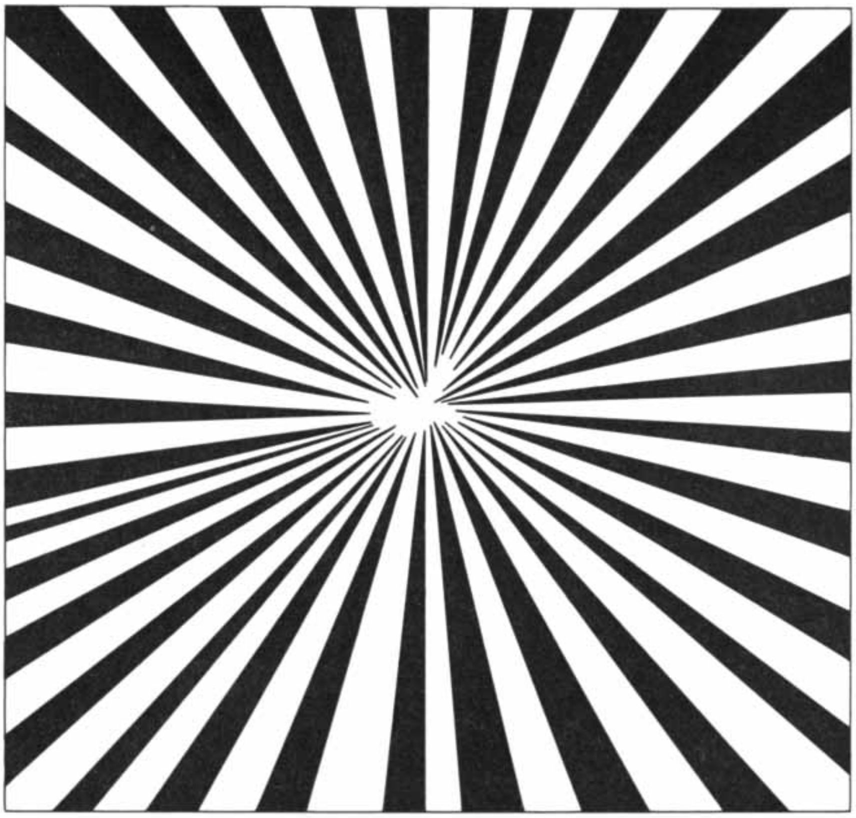
LATTICE FORMS, one of several form constants reported during drug-induced hallucinations, are depicted in four samples of Huichol Indian embroidery. The patterns illustrate designs commonly found in the hallucinatory visions induced by peyote. Some pictures made by schizophrenics exhibit a similar preoccupation with geometric designs, which are often distorted and repeated in symmetrical patterns.

operant key press as a reporting device; each key was equated with the occurrence of a subjective state. A hallucinating subject who wanted to report that he "saw" or "heard" something would press a given key for each condition. Lindsley and his colleagues had already demonstrated that the key-pressing technique provided a continuous and objective method for the study of gross behavior in altered states of consciousness.

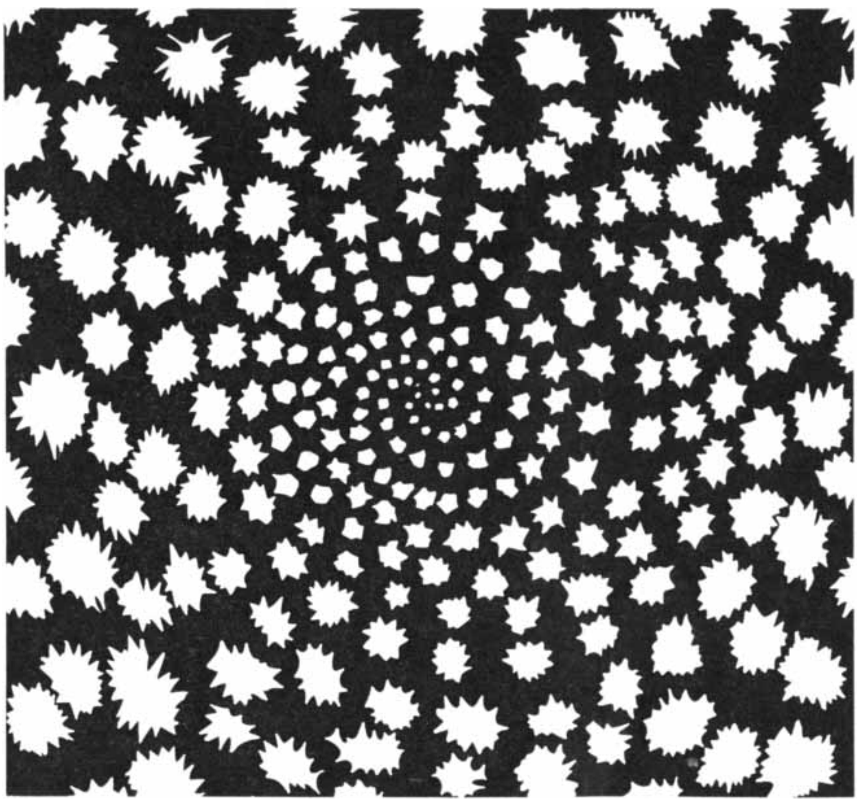
Nevertheless, a method for obtaining detailed information on subjective phenomena was not developed until Lindsley and Leary devised the "experiential typewriter." This apparatus consisted of a 20-key typewriter connected to a recording instrument. The keys served to code various subjective states, including modes of perception, internal images, external images, images seen with the eyes closed and with the eyes open, hallucinations, colors and so on. Training consisted of having the subjects memorize the categories and the corresponding keys. When a subject felt a bodily sensation such as "pain," he pressed a specific key to signify the event. A hallucination with oscillating colors could be signified by pressing a different key.

In tests with the hallucinogens LSD and dimethyltryptamine Leary found that the muscular discoordination associated with those drugs interfered with pressing the keys. The problem could have been avoided with lower doses, and the key press could have been replaced by a verbal report. (Leary had found that verbal reporting was not disrupted by the drugs.) Therefore we decided to develop a verbal code for reporting drug-induced hallucinations.

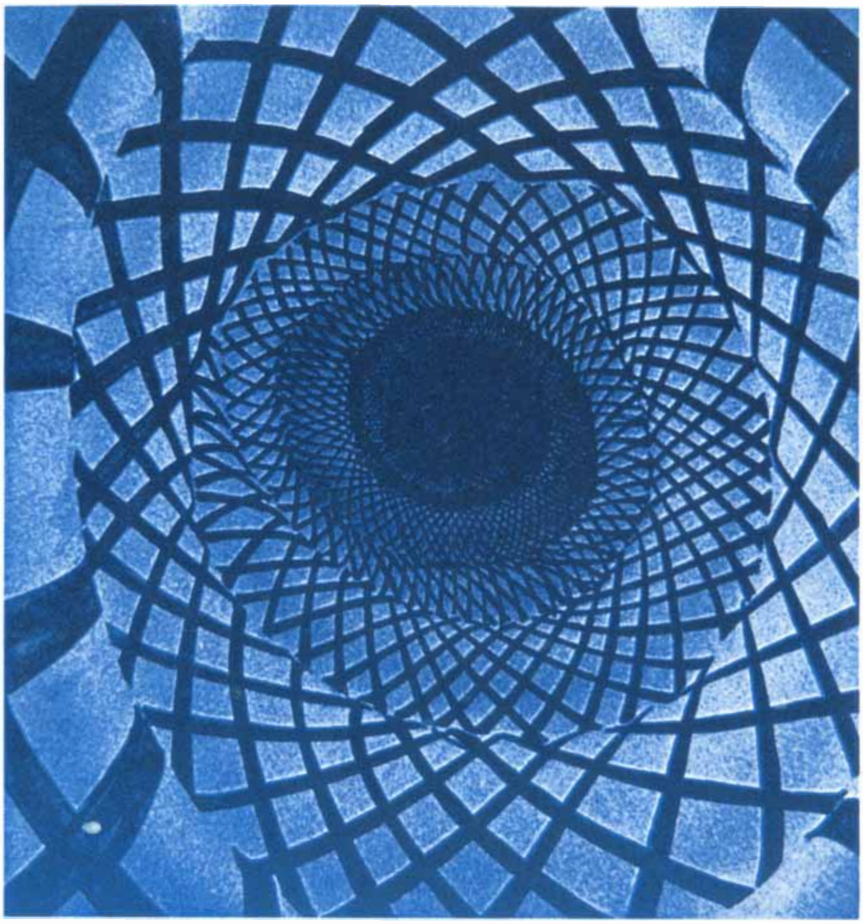
On the basis of the previous results we constructed a list of eight forms (random, line, curve, web, lattice, tunnel, spiral and kaleidoscope), eight colors (black, violet, blue, green, yellow, orange, red and white) and eight patterns of movement (aimless, vertical, horizontal, oblique, explosive, concentric, rotational and pulsating) for subjects to employ in describing visual imagery. We then selected a group of subjects and trained them with slides illustrating the different categories. For example, in training related to the tunnel form we showed hundreds of different slides of tunnels so that the subjects would have a broad concept of the tunnel form. In this way new instances of the form, which might not have been perceived before, could be appropriately classified on the basis of common features. The subjects were trained to recognize all three categories (form, color and movement) in displays projected for eight milliseconds, with a one-second pause between displays. Our aim in making the displays brief was to simulate the rapid changes of imagery in hallucinations.



WHITE LIGHT seen during the early stages of intoxication with a hallucinogenic drug is portrayed. The visual imagery is reported to explode from the center to the periphery. Pattern appears initially in black and white, but bright colors may develop as the experience progresses.



SPIRAL TUNNEL is another of the early form constants in drug-induced hallucination. The main patterns of movement accompanying the form are reported as pulsation and rotation.



LATTICE-TUNNEL FORM CONSTANT is depicted in a painting made to show a pattern that is often reported during the early stages of intoxication from marihuana and tetrahydrocannabinol (the active principle of marihuana). It is possible that the blue color is related to the initial lowering of the body temperature and to the absorption of blue light by hemoglobin in "floater" cells in the retina. Colors tend to become red with time and with increasing dosage.

All the subjects, including some who were untrained so that we could see if their reports of visual experiences were comparable to those of the trained subjects, then participated in a series of weekly test sessions in each of which they received either a hallucinogen, a stimulant, a depressant or a placebo. No subject knew what he was receiving. Both high and low doses were tested. The subjects were tested in the lightproof chamber I have mentioned. All the subjects were instructed to report what they "saw" with their eyes open, but the trained subjects were restricted to the descriptive code. In the middle of the session the trained subjects were tested with slides of real images to ensure that the drugs and dosages were not impairing the skills they had acquired in training. (We found no indication of impairment.)

The results were intriguing. We found that the trained subjects could keep abreast of the rapid flow of imagery and could readily classify most of the images into the categories of the reporting code. These subjects averaged 20 reports per minute; the untrained subjects reported only about five times per minute. (The duration of a typical session was six hours.)

The imagery associated with placebos, the stimulant *d*-amphetamine and the depressant phenobarbital was described as black-and-white random forms moving about aimlessly. The hallucinogens tetrahydrocannabinol, psilocybin, LSD and mescaline induced dramatic changes. Here the forms became less random and more organized and geometric as the experience progressed. The black-and-white images began to take on blue hues, and movement became more organized and pulsating. At 30 minutes after the administration of the drug the subjects reported a significant increase in lattice and tunnel forms and a slight increase in kaleidoscopic forms. By 90 and 120 minutes most forms were lattice-tunnels. Concomitantly the colors shifted to red, orange and yellow. Movement continued to be pulsating but became more organized, with explosive and rotational patterns.

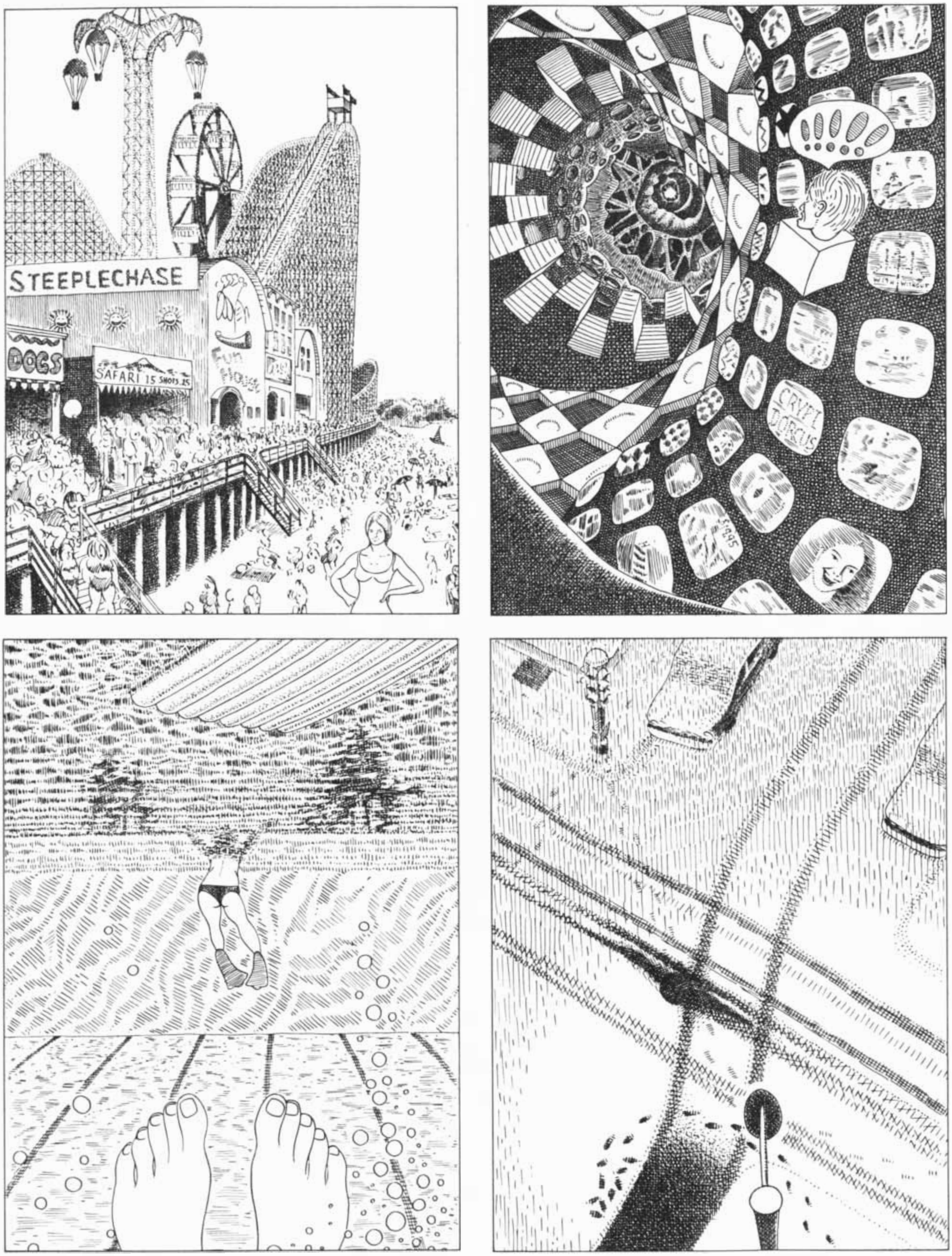
Complex imagery usually did not appear until well after the shift to the lattice-tunnel forms was reported. Thereafter complex forms constituted from 43 to 75 percent of the forms reported by trained subjects who had received hallucinogens. The complex images first appeared in the reports as overlying the lattice-tunnels and situated on the periphery of those images.

Common complex images included childhood memories and scenes associated with strong emotional experiences that the subjects had undergone. These hallucinatory images were more than pictorial replicas; many of them were elaborated and embellished into fantastic scenes. This constructive aspect of imagery can be illustrated by a simple exercise. Recall the last time you went swimming in the ocean. Now ask yourself if this memory includes a picture of yourself running along the beach or moving about in the water. Such a picture is obviously fictitious, since you could not have been looking at yourself, but images in the memory often include fleeting pictures of this kind. Our subjects often reported equally improbable images, such as aerial perspectives and underwater views.

It has not been established where such constructions arise, but contributions are probably made in the encoding, storage and retrieval stages of the memory process. Much of the content of complex imagery can also be influenced by environmental stimuli. Since our subjects were in an isolated chamber, most of their images came from their memory. Occasionally, however, we escorted the subjects (at the peak of the hallucinatory experience) to a botanical garden. There they wore goggles through which they could not see, lay on the grass and reported what they "saw" with their eyes open. At these times the imagery from memory was reduced significantly, and reports of birds, airplanes, trees and so on increased. In other experiments we have shaped and guided drug-induced imagery by giving the subjects suggestive words or music. Nevertheless, even these primed complex images were usually reported as appearing in lattice-tunnel arrangements and moving in explosive or rotational configurations.

During the peak hallucinatory periods the subjects frequently described themselves as having become part of the imagery. At such times they stopped using similes in their reports and asserted that the images were real. This point marked the transition from pseudohallucination to true hallucination. Highly creative and fantastic combinations of imagery were reported, sometimes with as many as 10 changes of image per second. The subjects frequently reported feeling dissociated from their bodies.

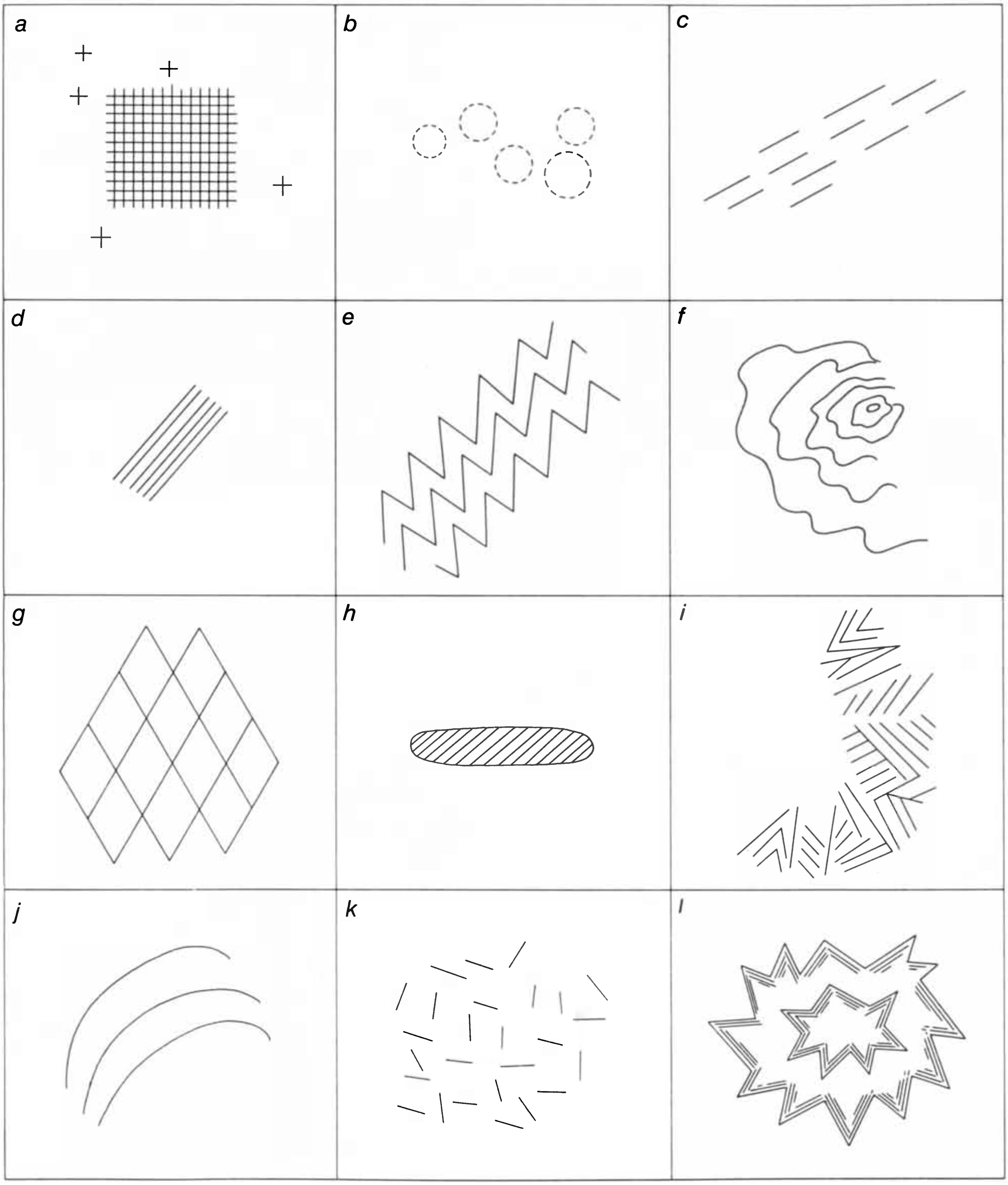
The remarkable constancies of drug-induced hallucinations lead naturally to an inquiry into how universal they



UNUSUAL PERSPECTIVES are reported in hallucinatory images that seem to be drawn from the memory. Characteristic perspectives include a distant scene (with abundant detail) often recognized as an event that was experienced in childhood (*upper left*), a lattice-tunnel pattern with complex memory images at the periphery (*upper right*), a scene viewed as if the subject were under water, looking up toward and through the surface (*lower left*) and an aerial perspective (*lower right*), which may be accompanied by sensations of floating and flying.

may be. Some of them are strikingly similar to the primordial or archetypal forms (such as the mandala, the mystic symbol of the universe employed in Hinduism and Buddhism as an aid to meditation) that the psychoanalyst C. G. Jung described as part of man's collective unconscious. Moreover, as many anthropologists have noted, the hallucinogen-inspired art of many primitive peoples often contains constants of form, color and movement.

We examined this phenomenon by traveling to the Sierra Madre of Mexico to study a group of Huichol Indians who take peyote. They have remained relatively isolated since Aztec times. We interviewed them during their peyote ceremonies, eliciting reports on their visual imagery. The images proved to be virtually identical to the symmetrical, repeat-



VISUAL HALLUCINATIONS seen during controlled intoxication with cocaine were drawn by people who served as subjects in the author's experiments. The patterns were usually seen with the eyes open in a dark room; they appeared as transitory black-and-white spots in the periphery of the visual field. Patterns *c, e, f, g, i, k* and *l* are virtually identical to the patterns seen in the hallucinations accompanying migraine attacks. Indeed, pattern *l* is the "fortification illusion" that is frequently reported by people who are suffering migraine headaches.

The first complete low-cost microcomputer system for lab, business or home.

Radio Shack TRS-80

The TRS-80 is for people who want to use a computer now — without the delay, work and problems of building one board by board. The system is fully wired, tested and U.L. listed — ready for you to plug in and use. Program it to handle scientific calculations, mass data storage, small business or home accounting, innumerable games — and use Radio Shack's expanding line of prepared programs on cassettes. The Z80-based microcomputer comes with 4K read/write memory and Radio Shack Level-I BASIC stored in read-only memory. Memory expandable to 62K bytes. Includes CPU, memory, keyboard, display, power supply, cassette data recorder, 300-page manual, 2-game cassette program. Designed and manufactured in USA by Radio Shack. Only 599.95.

Includes everything you need to start using it immediately!

Clip and Mail Coupon Today!

C008

Mail to: Radio Shack, Dept. TRS-80
205 N.W. 7th St., Ft. Worth, TX 76101

Send me more data on the TRS-80 microcomputer

- **Description of applications, software and peripherals available through Radio Shack • Owners' newsletter**
- **Price list • List of stocking stores and dealers**

NAME ____________________ APT. NO. ______

ADDRESS ______________________________

CITY ______________ STATE ________ ZIP ________

SOLD ONLY WHERE YOU SEE THIS SIGN:

Radio Shack®

A TANDY COMPANY • FORT WORTH, TEXAS 76102
OVER 6000 LOCATIONS IN NINE COUNTRIES

Price may vary at individual stores and dealers

It looks like several different whirlpools, with lots of spirals divided up into checks. It's pretty black. There's purple and green glowing areas in the middle of the spirals, kind of clouds around. There are lines going from top to bottom, kind of a grid, but the lines squiggle around. There's odd shapes, but still lots of right angles in them. Seems really bright.... There's like an explosion, yellow in the middle, like a volcano gushing out lava, yellow, glowing. There's a black square with yellow light coming behind it. There's a regular pattern superimposed on everything, lots of curlicues, with dots in the middle. Lot of little paisley things that fill up the spaces between the patterns of triangles, squares, or crown-shaped things. And there's a little white star that floats around the picture and sometimes goes behind what's on the screen and illuminates from behind.

Now there's a kind of landscape. Very flat, flat country. The picture is very narrow. In the middle part a tree at the left and then flat with green grass and blue sky above. There are orange dots, oranges hanging all over, in the sky, on the tree, on the ground. A bicycle! Oh, my! It's headed down, not horizontal, like someone's holding it up on end.... There's a checkerboard superimposed on everything, like the flags they wave at the races.

I can see the street out there.... Well, it's old--golly--interesting! It's like in the forties, I guess, or maybe the fifties.... And there are people riding their bicycles, and there are, like, boys, in plaid vests and those funny kind of hats.... I was at the side walking on the sidewalk, so it wasn't like I was in the middle of the street and (laughter) you can't laugh very long in the middle of the street in the city, so that image kind of went away (laughter).

PARTIAL TRANSCRIPT of the remarks made by a subject who had taken 20 milligrams of the hallucinogen psilocybin shows an increasing complexity of imagery. The first paragraph is from an initial segment of the experience, which began about 25 minutes after the subject had taken the drug. The simple geometric forms are supplemented several minutes later (*second paragraph*) by complex images. They are later replaced (*third paragraph*) by dreamlike scenes.

ing patterns found in Huichol weaving and art.

In attempting to explain the origin of the simple hallucinatory constants a number of investigators have described them as the product of events within the eye resulting from the visualization of certain structures when light strikes the eye in a particular way. For example, horizontal bands are said to be due to folds in the corneal epithelium that change with motion of the eyelid, and a black lacework seen against a red background is attributed to shadows cast on the rods and cones by blood vessels of the retina.

Since such causes require light, they cannot have figured prominently in our experiments, which were conducted in darkness. Light is not necessary, however, for the production of phosphenes, which are visual sensations arising from the discharge of neurons in structures of the eye. Phosphenes can include spots, disks, concentric arcs or circles and checkerboard patterns.

The constants are also highly similar to the patterns found in hallucinations accompanying migraine attacks. Migraine patterns include lines, grids, concentric circles and the "fortification illusion," which is a horseshoe-shaped area consisting of bright zigzag lines appearing at an expanding outer edge. Migraine hallucinations can also be brightly colored and explosive and can include complex images of people and objects. The most plausible explanation of migraine phenomena is that they reflect the electrical excitation of organized groups of cells in the visual cortex of the brain. Indeed, the work of Wilder Penfield and his colleagues at McGill University has shown that the direct electrical stimulation of the visual cortex or the temporal lobes gives rise to moving colored lights, geometric forms, stars and lines.

Most of the investigators undertaking to explain complex hallucinatory imagery have described the images as the result of an excitation of the central nervous system. As early as 1845 Moreau was maintaining that hallucinations resulted from cerebral excitation that enabled thoughts and memories to become transformed into sensory impressions. Recent electrophysiological research has confirmed that hallucinations are directly related to states of excitation and arousal of the central nervous system, which are coupled with a functional disorganization of the part of the brain that regulates incoming stimuli. Behaviorally the result is an impairment of the discrimination normally based on external stimuli and a preoccupation with internal imagery.

This hallucinatory process has been described by such terms as "memory flashback" and "involuntary reminiscence." Certain psychoanalysts have postulated that it is the result of a general regression to primitive or childlike thinking, coupled with the emergence of repressed information and memories. Students of psychedelic phenomena have postulated that hallucinogens release normally suppressed information and memories.

Perhaps the most integrated explanation has been provided by the perceptual-release theory of hallucinations, which was formulated by the British neurologist Hughlings Jackson in 1931. As recently brought up to date by Louis Jolyon West of UCLA, the hypothesis assumes that normal memories are suppressed by a mechanism that acts as a gate to the flow of information from the outside. An input of new information inhibits the emergence and awareness of previous perceptions and processed information. If the input is decreased or impaired while awareness remains, such perceptions are released and may be dynamically organized and experienced as hallucinations, dreams or fantasies.

West has offered an analogy to illustrate the process. Picture a man in his living room, standing at a closed window opposite his fireplace and looking out at the sunset. He is absorbed by the view of the outside world and does not visualize the interior of the room. As darkness falls outside, however, the images of the objects in the room behind him can be seen reflected dimly in the window. With the deepening of darkness the fire in the fireplace illuminates the room, and the man now sees a vivid reflection of the room, which appears to be outside the window. As the analogy is applied to the perceptual-release hypothesis, the daylight (the sensory input) is reduced while the interior illumination (the general level of arousal of the central nervous system) remains bright, so that images originating within the rooms of the brain may be perceived as though they came from outside the windows of the senses.

Through such research and hypotheses we have begun to understand the nature of hallucinations as stored images in the brain. Like a mirage that shows a magnificent city, the images of hallucinations are actually reflected images of real objects located elsewhere. The city is no less intriguing and no less worthy of study because it is not where we think it is. Further experiments will help localize it.

Our promise is simple:

we'll show you how to photograph everything you love in nature.

See how Ralph Morse used a high-speed strobe to capture this gibbon in swing. (See page 24)

How can you get a dramatic shot like this? See how Bob Walch did it—with a wide-angle lens. (See page 215)

You add new dimensions to your enjoyment of nature when you record memorable scenes and moments on film, to enjoy over and over again. Now, let the LIFE Library of Photography show you how to make your photos as memorable as nature itself.

The introductory volume in the Library, *Photographing Nature,* tells you and shows you how to capture the fleeting, the almost hidden, or the awesome scenes of nature. How to photograph subjects that won't stand still. How to get ultra close-ups. How to shoot underwater. How to capture the grandeur of vast landscapes.

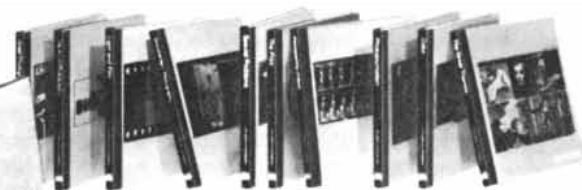
You'll turn again and again to its inspiring collection of nature photographs from the masters, absorbing their special techniques. The birds of Roger Tory Peterson and Eric Hosking. The sea creatures of Douglas Faulkner. The landscapes of Ansel Adams and Eliot Porter. And so many more.

With the guidance of the LIFE Library of Photography, your technical ability will grow and grow...not only in nature photography, but in all areas of photography.

But see for yourself at our expense. We'll send you the introductory volume, *Photographing Nature,* for 10 days free. Details are on the order form.

Learn what Eric Hosking had to do to get this tree-top picture of an eagle family. (See page 22)

Accept PHOTOGRAPHING NATURE for 10 days free as a guest of the LIFE LIBRARY OF PHOTOGRAPHY



Among other volumes in the LIFE LIBRARY OF PHOTOGRAPHY

How did Philip Teuscher photograph this Bronx Zoo tiger without getting the wire cage or bars in the picture? (See page 175)

• Actual book size: 10¼" x 10¼" • Hard cover, silver-stamped black-cloth binding. • Each book contains more than 200 pages with as many as 300 monochrome and color pictures

TIME LIFE BOOKS

YOURS FREE with PHOTOGRAPHING NATURE

This valuable 64-page pocket-sized manual contains hundreds of tips and ideas for taking and making better pictures. Includes 150 photographs and drawings, dozens of charts and tables.

TIME-LIFE BOOKS
Time & Life Building, Chicago, Illinois 60611

Yes. I would like to examine *Photographing Nature.* Please send it to me—together with the illustrated *Photographer's Handbook*—for 10 days free examination and enter my subscription to the LIFE Library of Photography. If I decide to keep *Photographing Nature,* I will pay $7.95 ($8.95 in Canada) plus shipping and handling; I then will receive future volumes in the LIFE Library of Photography series, shipped a volume at a time approximately every other month. Each is $7.95 ($8.95 in Canada) plus shipping and handling and comes on a 10-day free examination basis. There is no minimum number of books that I must buy, and I may cancel my subscription at any time simply by notifying you.

If I do not choose to keep *Photographing Nature,* I will return it within 10 days, my subscription for future volumes will be cancelled, and I will not be under any further obligation. The illustrated *Photographer's Handbook* is mine to keep without obligation.

BMB5Q6

☐ Mr.
☐ Mrs. ....
☐ Miss (please print)

Address .... Apt

City ....

State or Province .... Zip or Postal Code ....

THE AMATEUR SCIENTIST

The salt fountain and other curiosities based on the different density of fluids

by Jearl Walker

The fact that salty water is denser than fresh water provides a basis for a number of experiments, some of them quite surprising. For example, one might suppose that if a layer of salty water was put above a layer of fresh water, the salty water would simply sink. If the layering is done in a certain way, however, the result is not so simple. What one finds is that the system starts to oscillate and may continue to do so for a remarkably long time.

Pour cold tap water into a tall clear glass until the glass is nearly full. Put a hole in the bottom of a paper cup or a plastic cup with the point of a safety pin or something of similar size. Make a salty solution (about equal to half the volume of the cup) from cold tap water and a teaspoon or so of salt, taking care that the salt dissolves fully. Add a bit of food coloring to the solution so that later, when the experiment is under way, it will be easy to distinguish the salty water from the fresh.

Now start lowering the cup into the glass, simultaneously pouring the salty solution into the cup. You want the cup to reach a depth where the fresh water outside it is about halfway up the sides of the cup, and you want the salty solution in the cup to be at about the same level as the fresh water outside. At that point, since the effect is sensitive to jostling, you should fasten the cup. I fastened it with a beaker clamp attached to a ring stand of the kind used in chemistry laboratories, but you might try taping the cup to two flat kitchen knives laid across the top of the glass.

One would suppose that in this situation the dyed salty water would stream through the hole into the fresh water until the level of the salty water was reduced enough so that the pressure at the level of the hole was the same in both bodies of water, whereupon the stream would grow thinner and finally stop. Indeed, you will see these events over a period of several minutes, but do not turn away. In a short time the stream will begin flowing again! Thereafter, with a period of from 15 to 20 seconds, the stream will begin abruptly, wane, disappear and then start up abruptly again.

What is happening is that during the time when the dyed stream is not visible, fresh water is streaming upward into the cup through the hole. The system is oscillating. This effect was discovered in 1970 by Seelye Martin of the University of Washington [see "The Amateur Scientist," June, 1971].

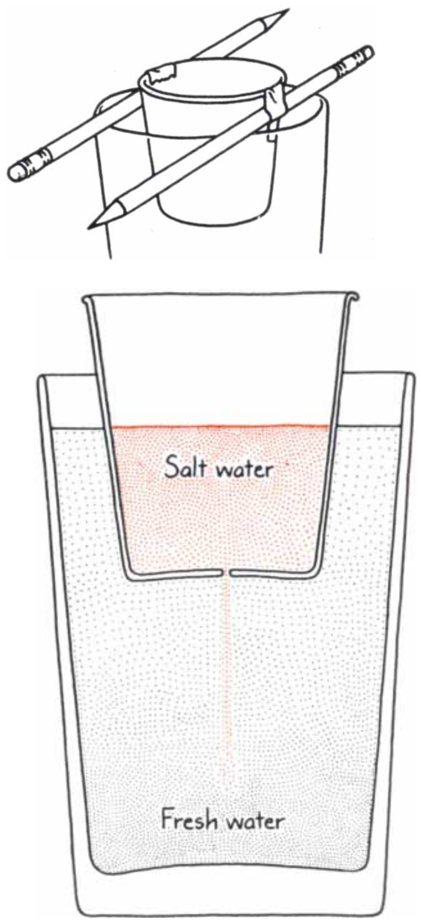
The period of oscillation depends mainly on the size of the hole and to a lesser extent on the concentration of salt in the cup. My oscillator continued overnight with approximately the same period in spite of the continuous reduction in the concentration of salt due to mixing. Martin's oscillator, which was in a tin can instead of a paper cup, continued for four days. He also ran a test with a hypodermic syringe replacing the cup; with this system he obtained a period of about four seconds and a lifetime of 20 cycles.

It is easy to understand the first part of the flow. The salty water was initially as high in the cup as the fresh water outside the cup. Since the salty water was denser than the fresh water, the pressure at the hole was higher from the salty water than from the fresh water. The salty water therefore poured through the hole until the pressures equalized. The puzzle is why the salty water did not merely stabilize then as the downward flow slowed to a stop.

In spite of the equalization of pressures at the hole a layer of denser fluid overlying a layer of lighter fluid is unstable and subject to any small, random disturbances to the system. Any such disturbance creates a small wave on the interface between the two types of fluid. The wave grows in amplitude exponentially with time, at least initially, because of the density difference. As a result some of the lighter fluid protrudes upward across the old interface and some of the denser fluid protrudes downward. This instability to small disturbances and the resulting intrusion of each fluid on the other's domain are responsible for the oscillation in the salt oscillator.

The fresh water protruding upward accelerates through the hole because it finds itself lighter than the salty water at the same level on the other side of the hole. Apparently the fresh-water stream pinches off the salt-water stream soon thereafter, and there is then a single upgoing stream of fresh water moving through the hole. This addition of water to the cup gradually increases the height of the fluid in the cup and thus the water pressure at the level of the hole. The loss of water from the glass, however, barely lowers the water level there because the glass is wider. Eventually the pressure from the salt-water side becomes great enough at the level of the hole to overwhelm and halt the upgoing stream of fresh water. We are then essentially back at the start of the cycle. There is too much water in the cup, and a stream of salt water emerges downward for a while. Gradually the flow of the stream decreases as the pressures once again become equal at the level of the hole. Then some random disturbance sets up a wave on the interface that once again sends up a stream of fresh water. With a rate depending on the diameter of the hole and of the cup and on the viscosity of the water, the streams alternate up and down. You have a salt oscillator.

The instability at the interface be-



The basic salt oscillator

Setting the Standard in Biochemistry

Biochemistry

Lubert Stryer, Stanford University

"This must be one of the best biochemistry books on the market. It is certainly the easiest to read and study. The text is lucid and digestible. The figures complement the text in a way which is not normally found in books at this level . . . For me, difficult concepts in biochemistry (those which I've always avoided) were made so simple it was a pleasure revisiting tricky grounds . . . " *—Bios*

"One can open this book almost at any page and discover a brilliant example of artistic expression used to illuminate scientific concepts. But in addition, the author has an attractive style of writing; and most important, he has something new to say." *—Biochemical Education*

"Professor Lubert Stryer has succeeded in producing an introductory biochemistry textbook that will probably be the standard text for life science, biochemistry, and medical students for many years to come."
—Yale Journal of Biology and Medicine

1975, 877 pages, 626 illustrations, 55 tables,
ISBN 0-7167-0174-X, $22.00

The Stryer slide series is a teaching aid that will enhance the student's understanding of biochemistry.

Molecular Architecture

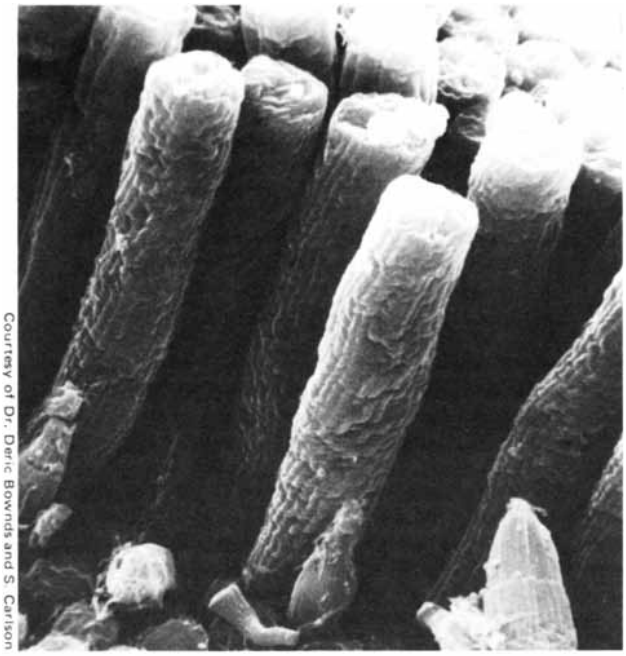
This collection of 50 full-color 35mm slides, selected from the illustrations in Lubert Stryer's *Biochemistry,* is organized to illustrate the theme of molecular architecture. Price $50.00

Energy Generation and Storage

A new addition, these 90 full-color 35mm slides illustrate processes related to the generation and storage of metabolic energy, including glycolysis, the citric acid cycle, photosynthesis, glycogen metabolism, fatty acid metabolism, the pentose phosphate pathway, amino acid degradation, and the urea cycle. Price $90.00.

Special offer to adopters

Each department that adopts 50 or more copies of Stryer's *Biochemistry* is entitled to receive both sets of slides at no cost.



Courtesy of Dr. Deric Bownds and S. Carlson

Physical Biochemistry

Applications to Biochemistry and Molecular Biology

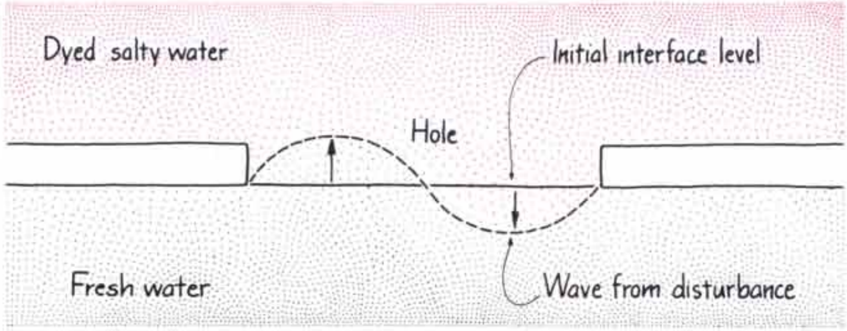
David Freifelder, Brandeis University

With clear descriptions, a wealth of practical examples, and a de-emphasis of mathematical derivations, David Freifelder describes the wide range of techniques employed in the modern study of the physical properties of biological macromolecules, including separation and identification, hydrodynamic methods, and spectroscopy. Readers will appreciate the review of filtration techniques, pH measurement, and other basic laboratory procedures. Observational methods such as light and electron microscopy and autoradiography are also included.

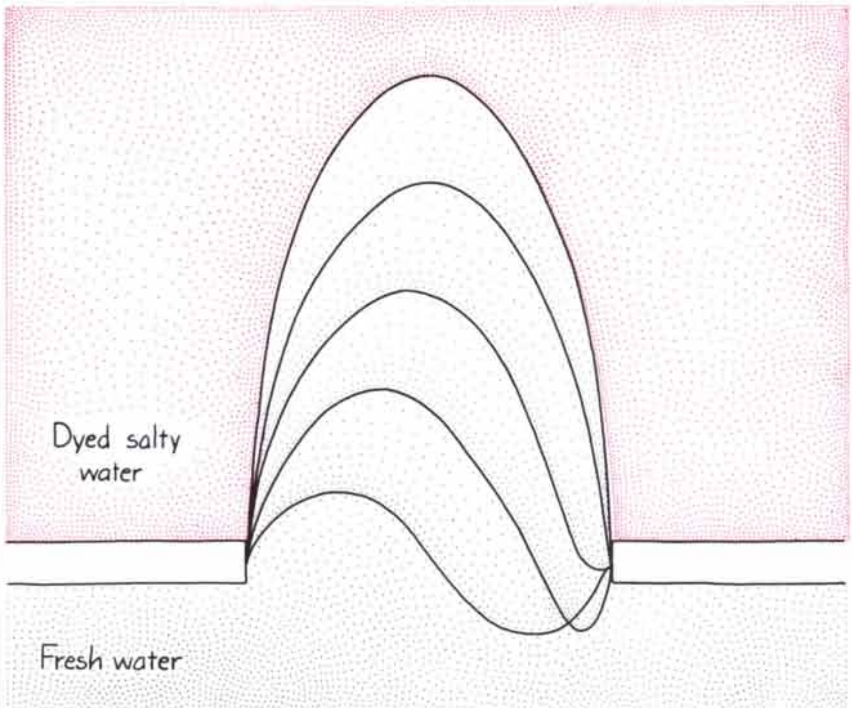
Readers will come away from this book able to read, understand, and judge the work reported in contemporary scientific literature. They will also gain a better understanding of general principles, since many of the techniques discussed were originally used to establish major principles of modern biochemistry.

1976, 570 pages, 304 illustrations, 28 tables,
cloth: ISBN 0-7167-0560-5, $16.95
paper: ISBN 0-7167-0559-1, $10.95

W. H. Freeman and Company
660 Market Street, San Francisco, CA 94104
58 Kings Road, Reading, England RG1 3AA



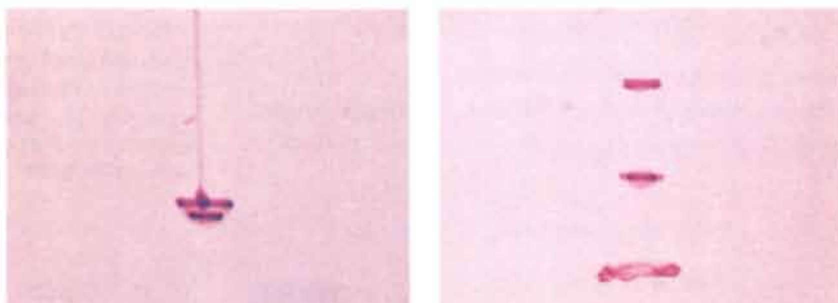
An oscillation begins



A rising stream of fresh water pinches off the descending stream of salt water

tween a denser fluid overlying a lighter fluid, when the interface is otherwise in hydrostatic equilibrium, is called a Rayleigh instability (or sometimes a Rayleigh-Taylor instability). The salt oscillator is an example of a system that oscillates after exciting itself, in this case through the Rayleigh instability and the consequent rapid growth of a disturbance on the interface between the two fluids.

Apparently Martin's paper is the only one that has been published on salt oscillators, and you might be interested in doing more work on them. For example, how does the period of oscillation de-



Shapes of the falling stream

pend on the diameter of the hole? Martin's use of a hypodermic syringe and needle is advantageous because needles of different radii could be easily interchanged, the radius of each was well known and the needle was certainly more circular in cross section than my pinholes. Nevertheless, you can establish the general dependence of the period on the diameter of the hole with a range of pinholes.

The dependence of the period of oscillation on the concentration of salt can be checked by adding salt to an initially weak concentration of salt in the cup. Take care, however, that the added salt does not merely clog the hole.

You might also try other fluids. They do not have to mix. All that is required is that they differ in density and that the denser fluid be in the cup. I tried water and Karo light corn syrup (a clear syrup used in cooking or on pancakes), dyeing the syrup a light red before I poured it into the cup. With a hole of suitable size the oscillations of the viscous red fluid and the clear fresh water were almost majestic. The upward plume was much easier to see in the corn-syrup oscillator than in the salt-water oscillator. I lightly dyed the fresh water blue and could then easily see the upward plume of fresh water periodically break through the hole, push its way up through the syrup and then lie on top of it.

When the downward stream begins during the cycle of a salt-water oscillator, the dyed salt water appears to burst downward, taking the form of an inverted umbrella. During the part of the cycle when the flow is steadier, and in steady flow without oscillations, the stream may fall several centimeters and then separate into ring vortexes. The rings behaved much as my grandfather's rings of cigar smoke did on those quiet days in Aledo, Texas, when his ring-blowing was our only amusement. The rings would race after each other, wrap themselves about each other and then race away. This play is easier to see in the rings of dyed water if you use a round glass container, such as a large tea glass, because the curved surface magnifies the image.

Another example of Rayleigh instability can be observed in a common parlor trick. Partly fill a glass with water, place a piece of paper over the mouth and, while holding the paper in place, invert the glass of water. Now remove your hand from the paper. The water stays in the glass thanks to two effects. First, surface tension between the water and the paper and between the water and the rim of the glass helps to hold the paper in place. Second, in larger glasses the water column falls somewhat and thus reduces the pressure in the air still trapped in the glass. The difference in air pressure between the top and bottom of the column of water also helps to hold the water in place.

Suppose the paper were to disappear suddenly without disturbing the water surface. What would happen? Common sense tells you that the water would fall out. If the water surface could remain perfectly undisturbed, however, there is no reason why the water should fall. The pressure difference between the top and the bottom of the column should still maintain the water in the glass.

The water does fall, though, because it cannot remain undisturbed. Soon after the paper is removed a small wave will develop on the bottom surface because of some small random disturbance to the water. The wave will grow in amplitude, at first exponentially with time, exactly as we have seen at the interface in the salt oscillator, because once again we have a Rayleigh instability. A bubble rises up on one side of the column of water as part of the water begins to descend on the other side, and then we rather quickly have a wet floor.

Salt fingers involve a layering of saltier water over fresher water, similar to the layers in the salt oscillator. Although the fingers do not have the fascination of an oscillation, they are important in the microstructure of the ocean. Consider a layer of hot salty water over a layer of cooler fresh water. The salt makes the top layer denser than the bottom layer, but the difference in temperature more than compensates to give a net situation where the top layer is less dense than the bottom one.

One would suppose that if a small disturbance sends part of the top layer downward and part of the bottom layer upward, the difference in overall densities should restore the original boundary. For example, the hot salty water intruding downward would find itself less dense than its surroundings in the cool fresh water and would be buoyed back upward.

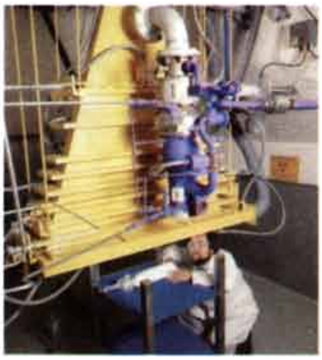
During the intrusion from the initial disturbance on the boundary, however, heat is exchanged relatively quickly between the hot water and the cool water. The hot water releases heat as it moves downward. The cool water absorbs heat as it moves upward. As a result the intruding salt water suddenly finds itself more dense than its surroundings rather than less, and instead of being buoyed back it is accelerated downward. Similarly, the intruding fresh water quickly becomes warm, finds itself less dense than the surrounding hot salty water and is accelerated upward. The initial disturbance to the interface between the two layers is therefore enhanced to produce fingers of salty water stretching downward and fresh water stretching upward. The fingers eventually stretch themselves to the point where they turn over to create a layer of mixed water with a salinity and temperature intermediate to those in the top and bottom layers.

The salt fingers are not themselves directly observable in the ocean, but the

Making plastic-filler marriages last longer.

Incompatibility and stress have been known to cause rifts in many an otherwise perfect union.

Take polypropylene, for instance. This thermoplastic is often wedded with inorganic fillers such as talc or kaolin to increase heat resistance and stiffness.



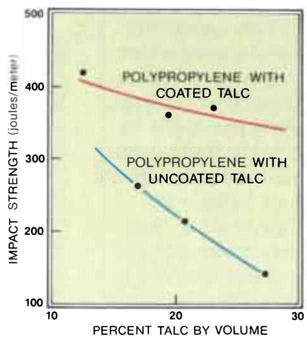
But the resulting union responds poorly to the stress of impact. Shock waves tend to start cracks around the filler particles—mostly because the plastic doesn't cling to the filler. It just surrounds it.

Our polymer scientists here at the General Motors Research Laboratories set out to strengthen the plastic-filler composite by introducing a compatible, shock absorbing interface. The trick was to get the interface material to stick to the filler.

How did they do it? First they adsorbed a monomer, 2-ethylhexyl acrylate, on the filler particles. Then they irradiated the coated filler, under vacuum, with high energy electrons (photo above).

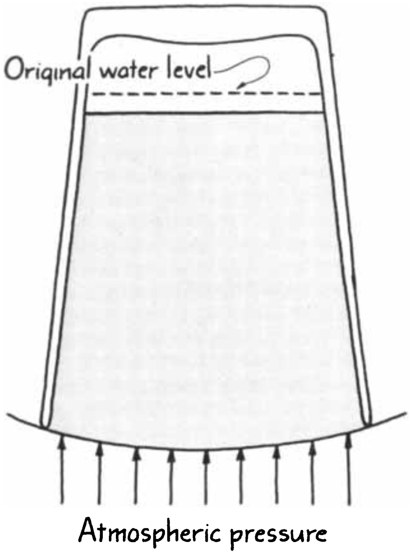
The radiation did two things: It polymerized the monomer coating. And it chemically linked this coating to the filler, as revealed by thermogravimetric analysis and carbon-14 tracer techniques.

As a result, some plastics are now able to bond tightly to fillers through an elastic interface. In the case of polypropylene and talc, that means about twice the impact strength as before (see graph).

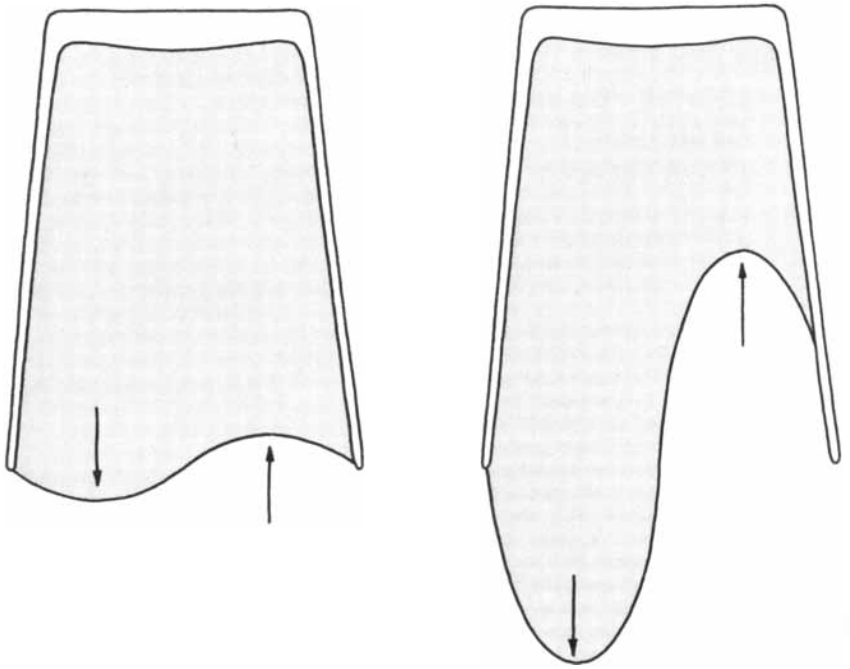


We call this experimental technique "radiation grafting on fillers." It's only one of the approaches we're exploring to help make plastic products live happily ever after.

General Motors Research Laboratories
Warren, Michigan 48090



The inverted-glass trick



The interplay of air and water causing the water to fall

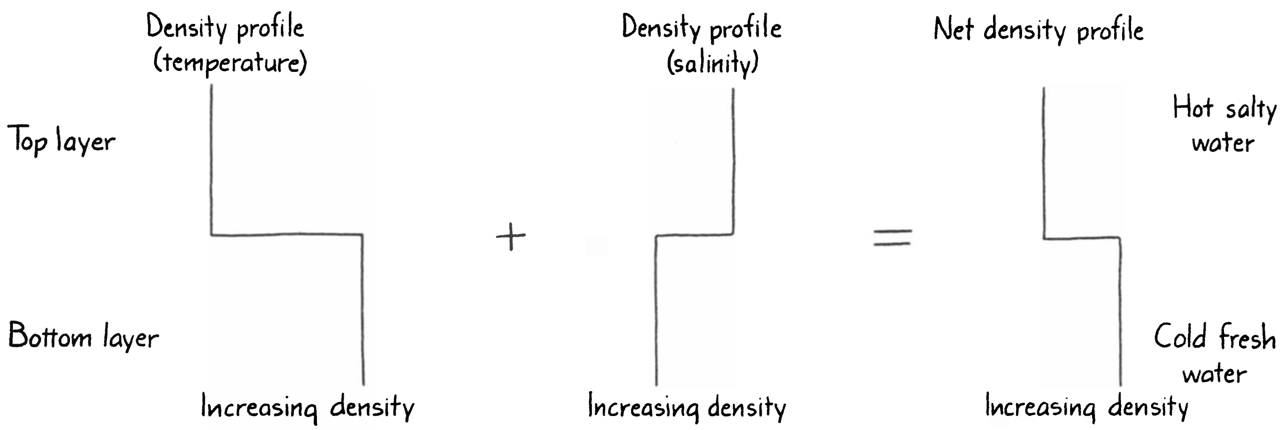
resulting temperature and salinity profiles have been detected [see "The Microstructure of the Ocean," by Michael C. Gregg; SCIENTIFIC AMERICAN, February, 1973]. For example, the warm salty water of the Mediterranean sets up the conditions for salt fingering as it flows through the Straits of Gibraltar and over the fresher, cooler water of the Atlantic. Probes measuring the salinity and temperature as functions of the depth show steplike profiles where the salinity and temperature decrease with depth. Such a step has one layer of higher salinity and temperature than the layer below, with a layer between the two having intermediate values.

You can produce salt fingers quite easily in your kitchen by carefully pouring hot salty water (dyed with food coloring) over cooler fresh water. In pouring you want to minimize the initial turbulent mixing of the two types of water. If the container is a standard drinking glass, I tilt it so that the poured water falls only a centimeter or less. With a bigger container I can lower the container of salty water to nearly the level of the fresh water to accomplish the same thing. You can instead float a thin piece of wood on the surface of the fresh water and then pour onto the wood.

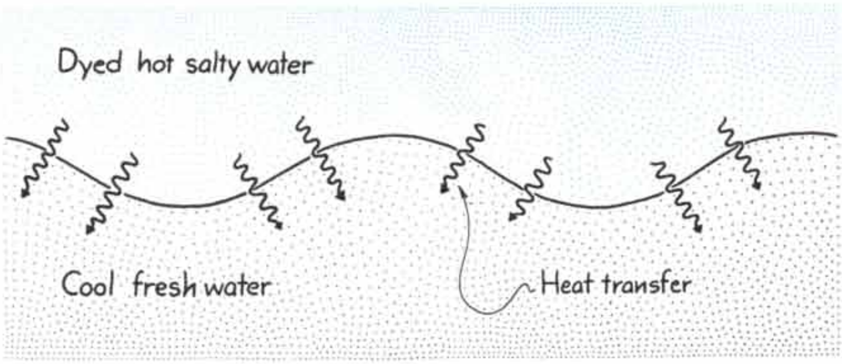
At first the interface between the dyed hot salty water and the clear fresh water merely indicates some mixing. After a few minutes the salt fingers, from one to five centimeters long and about a millimeter wide, develop at the interface, lasting from several minutes to several hours. I find that the fingers are easier to see if I shine a flashlight through the water somewhat toward my eyes.

The key to the salt fingering is the difference in the rate of diffusion of heat and salt. The heat that made the top layer lighter in spite of its salt concentration moved to the cooler fresh water about 100 times faster than the salt. The rather quick transfer of heat is what made the downward-intruding salt water suddenly find itself heavier than its surroundings and what made the upward-intruding fresh water suddenly find itself lighter than its surroundings.

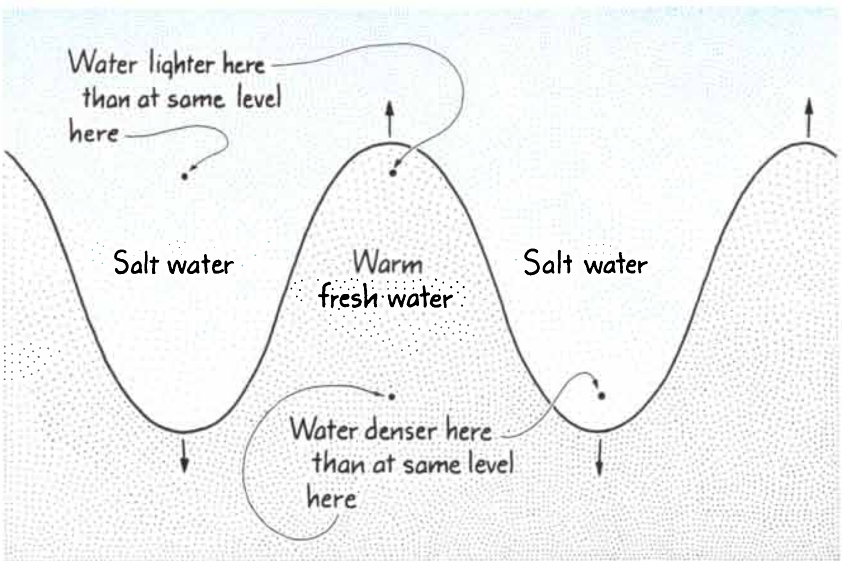
You can accomplish the same sort of convective fingering with solutions of sugar and salt. Make the sugar solution less dense than the salt solution by putting somewhat less sugar than salt in a given amount of water. Dye the sugar solution so that you can follow the motion, and gently put the sugar solution over the salt solution in a container. (Since both solutions are at the temperature of cold tap water, this salt fingering is easier to do in the kitchen



Density profiles when hot salty water overlies cold fresh water



Beginning of a random disturbance at the interface of the layers



Enhancement of the wave as heat diffuses faster than salt

than the preceding experiment, where rapid cooling of the hot water can be a problem.)

Once again, as with the arrangement of hot salty water and cold fresh water, the layers of sugar solution and salt solution should be stable because the top layer is the less dense one. As before, however, any initial disturbance is enhanced because of a difference in the rates of diffusion: Salt diffuses faster than sugar. Thus when the initial disturbance sends a small amount of sugar solution downward, the salt from the surroundings moves into that intruding bulge faster than the sugar can move into the surroundings. The bulge, with salt added, suddenly finds itself denser than its surroundings and is accelerated downward to produce a finger. Similarly, a small bulge of salt water intruding upward loses its salt faster than it gains sugar, finds itself lighter than its surroundings and is accelerated upward to produce a finger.

The interaction of salt and heat on the density of seawater and the difference in their rate of diffusion also lead to a curious "perpetual" fountain you can simulate in the kitchen. In a tropical ocean the water near the surface can be relatively warm and salty whereas the water near the bottom is relatively cold and fresh. Imagine a pipe lowered vertically almost to the bottom and a pump initially used to bring bottom water to the surface. Theoretically the pump could be removed and the flow would continue by itself. Since a single pipe would hardly alter the characteristics of the ocean, this flow would continue "forever."

To understand how the flow continues to pump itself consider a small parcel of water beginning its journey up the pipe. As that bit of cold water rose, it would gradually gain heat from the warmer water outside the pipe at the same depth. The parcel of water would then be lighter than the saltier water at the same depth and temperature just outside the pipe. As a result the parcel of water would accelerate upward. Assuming the parcel is always warmed in this manner throughout the trip upward, it would always find itself lighter than the outside water and would continue to accelerate. In principle the water would spurt as much as a couple of meters above the

Get updated . . . keep updated with BYTE

the leading magazine in the personal computer field

The personal computer age is here.

Join Byte's 110,000 subscribers and catch up on the latest developments in the fast-growing field of microprocessors. Read BYTE, The Small Systems Journal that tells you everything you want to know about personal computers, including how to construct and program your own computer (over 30,000 BYTE readers have already built, or bought, their own systems and half of these have 8K bytes or more).

You'll find our tutorials on hardware and software invaluable reading, also our reports on home applications and evaluative reviews based on experiences with home computer products.

Home computers . . . practical, affordable.

Large scale integration has slashed prices of central processors and other computer components. This has encouraged the development of new, low-cost peripherals resulting in more hardware and software — more applications than you could imagine, more opportunities for you. BYTE brings it all to you. Every issue is packed with stimulating and timely articles by professionals, computer scientists and serious amateurs.

BYTE editorials explore the fun of using and applying computers toward personally interesting problems such as electronic music, video games and control of systems for alarms to private information systems.

Subscribe now to BYTE . . . The Small Systems Journal

Read your first copy of BYTE, if it's everything you expected, honor our invoice. If it isn't, just write "CANCEL" across the invoice and mail it back. You won't be billed and the first issue is yours.

Allow 6 to 8 weeks for Processing.

© Byte Publications, Inc. 1977

BYTE Subscription Dept. 30D • P.O. Box • Arlington, Mass. 02174

BYTE

PLEASE ENTER MY SUBSCRIPTION FOR:

☐ One year $12 (12 issues) ☐ Two years $22 ☐ Three years $32

☐ Check enclosed (entitles you to bonus of one extra issue)

☐ Bill me ☐ Bill BankAmericard/Visa ☐ Bill Master Charge

Card Number: ☐☐☐☐☐☐☐☐☐☐☐☐☐☐ Expiration Date: ____________

Signature: ____________ Name *(please print)* ____________

Address ____________

City ____________ State/Country: ____________ Code ____________

FOREIGN RATES FOR ONE YEAR: *(Please remit in U.S. Funds)*

☐ Canada or Mexico $17.50 (Air *delivered)* ☐ Europe $25 (Air delivered)

☐ All other countries except above: $25 *(Surface delivery)*

Air delivery available on request

surface of the ocean, forming a perpetual fountain.

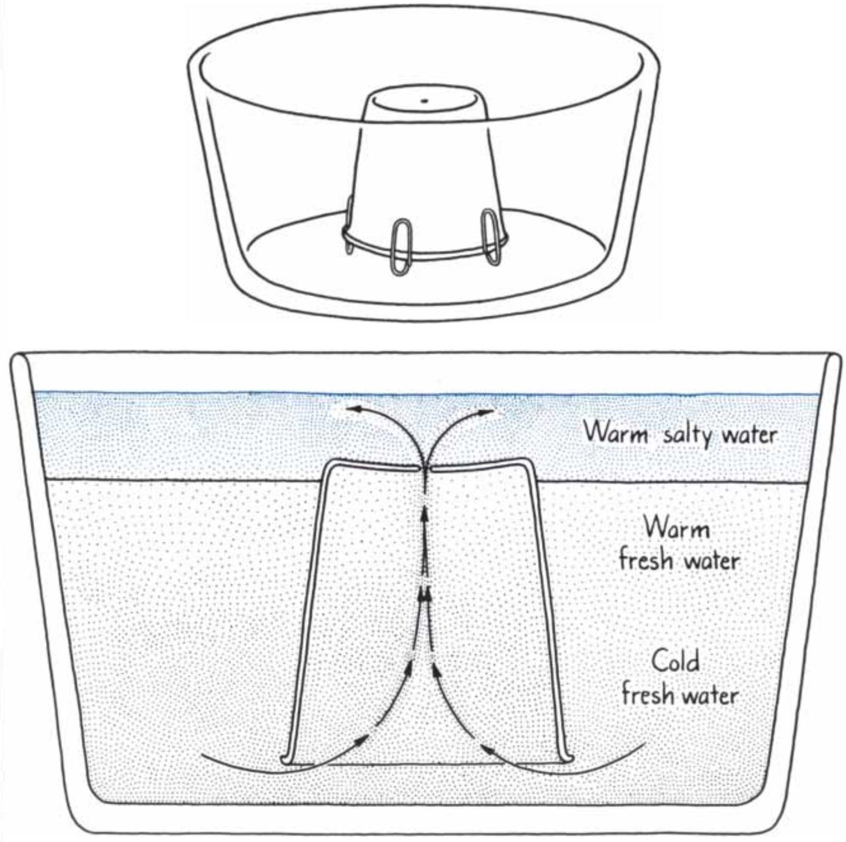
Henry M. Stommel of the Massachusetts Institute of Technology tried some years ago to set up such a fountain in the deep ocean near Martinique. With about 1,000 meters of flexible hose 5/8 inch in inside diameter he and his colleagues obtained a fountain some 24 inches high, but they are doubtful that it resulted from differences in density. The top of the hose was attached to a float, which rose and fell with the waves, stretching the upper part of the hose and creating a sort of pumping action. Stommel thinks the pumping action may have created the fountain.

To simulate the perpetual fountain in your kitchen partly fill a wide container with tap water. Punch a hole in the bottom of a paper or plastic cup and then put the cup in the water upside down. Now pour warm water carefully onto the cold water outside the cup. Try to avoid any excessive mixing of the two layers. Continue pouring until the addition of the warm water has forced the cold water up into the inverted cup to the level of the hole.

Now pour a thin layer of hot salty water over the top of the warm water and thus also over the top of the cup, again being careful to avoid excessive mixing. Place one or two drops of dye just over the hole in the inverted cup. The dye indicates that a slow fountain of water spreads from the hole through the top water, a simulation of the oceanic salt fountain. The water rising through the cup is warmed to about the same temperature as the water at the same depth outside the cup, but it lacks the salt that the outside water has and is therefore lighter than the outside water. Hence the water in the cup is forced upward through the hole. This kitchen fountain will continue until the heat and salt are more evenly distributed.

As a final note on effects due to the difference in densities of salty and fresh water, let me describe the push a ship receives when the gates of the last lock in the Panama Canal open. As a ship approaches the Pacific end of the canal it is progressively lowered through locks to sea level. The water in these locks is mostly fresh water fed by rain-filled lakes in Panama. In each lock the ship must wait until the water level is lowered to that of the next lock. Then the water pressure on the two sides of the gate separating the two locks is equal, and the gate is easily pulled open.

In the last lock the water on the other side is the salty ocean water. Since saltier water is denser than fresher water, equilibrium pressure on the gate is reached when the fresh water is still a bit higher than the ocean level. As the gate swings open, this extra height of fresh water flows seaward, carrying the ship along with it on a brief but free ride.



Arrangement for generating a salt fountain

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

MATHEMATICAL GAMES

THE LOGIC OF SCIENTIFIC DISCOVERY. Karl R. Popper. Hutchinson of London, 1959.

THE JUSTIFICATION OF INDUCTION. Edited by Richard Swinburne. Oxford University Press, 1974.

PEER REVIEW AND THE SUPPORT OF SCIENCE

SOCIAL STRATIFICATION IN SCIENCE. Jonathan R. Cole and Stephen Cole. The University of Chicago Press, 1973.

NATIONAL SCIENCE FOUNDATION PEER REVIEW: VOL. I. COMMITTEE ON SCIENCE AND TECHNOLOGY OF THE U.S. HOUSE OF REPRESENTATIVES. U.S. Government Printing Office, January, 1976.

SCIENTIFIC ELITE. Harriet Zuckerman. The Free Press, 1977.

X-RAY STARS IN GLOBULAR CLUSTERS

GLOBULAR-CLUSTER STARS. Icko Iben, Jr. in *Scientific American,* Vol. 223, No. 1, pages 26–39; July, 1970.

X-RAY-EMITTING DOUBLE STARS. Herbert Gursky and Edward P. J. van den Heuvel in *Scientific American,* Vol. 232, No. 3, pages 24–35; March, 1975.

X-RAY BINARIES IN GLOBULAR CLUSTERS. George W. Clark in *Astrophysical Journal (Letters),* Vol. 199, pages L143–L145; August 1, 1975.

FUNDAMENTAL PARTICLES WITH CHARM

DISCOVERY OF A NARROW RESONANCE IN e^+e^- ANNIHILATION. G. S. Abrams, J.-E. Augustin, A. M. Boyarski, M. Breidenback, D. Briggs, F. Bulos, W. Chinowsky, J. T. Dakin, G. J. Feldman, G. E. Fischer, C. E. Friedberg, D. Fryberger, G. Goldhaber, G. Hanson, R. J. Hollebeek, B. Jean-Marie, J. A. Kadyk, R. R. Larsen, B. Lulu, V. Lüth, H. L. Lynch, D. Lyon, C. C. Morehouse, J. M. Paterson, M. L. Perl, F. Pierre, P. Rapidis, B. Richter, R. F. Schwitters, W. H. Tanenbaum, G. H. Trilling, F. Vannucci, J. S. Whitaker, J. Wiss and J. E. Zipse in *Physical Review Letters,* Vol. 33, No. 23, pages 1406–1408; December 2, 1974.

SEARCH FOR CHARM. Mary K. Gaillard, Benjamin W. Lee and Jonathan L. Rosner in *Reviews of Modern Physics,* Vol. 47, No. 2, pages 277–298; April, 1975.

QUARKS WITH COLOR AND FLAVOR. Sheldon Lee Glashow in *Scientific American,* Vol. 233, No. 4, pages 38–50; October, 1975.

OBSERVATION IN e^+e^- ANNIHILATION OF A NARROW STATE AT 1865 MeV/c^2 DECAYING TO $K\pi$ AND $K\pi\pi\pi$. G. S. Abrams, M. S. Alam, A. M. Boyarski, M. Breidenback, W. C. Carithers, W. Chinowsky, S. C. Cooper, R. G. DeVoe, J. M. Dorfan, G. J. Feldman, C. E. Friedberg, D. Fryberger, G. Goldhaber, G. Hanson, J. Jaros, A. D. Johnson, J. A. Kadyk, R. R. Larsen, D. Lüke, V. Lüth, H. L. Lynch, R. J. Madaras, C. C. Morehouse, H. K. Nguyen, J. M. Paterson, M. L. Perl, I. Peruzzi, M. Piccolo, F. Pierre, T. P. Pun, P. Rapidis, B. Richter, B. Sadoulet, R. H. Schindler, R. F. Schwitters, J. Siegrist, W. H. Tanenbaum, G. H. Trilling, F. Vannucci, J. S. Whitaker and J. Wiss in *Physical Review Letters,* Vol. 37, No. 5, pages 255–259; August 2, 1976.

SIDE-LOOKING AIRBORNE RADAR

ON THE APPLICATION OF COHERENT OPTICAL PROCESSING TECHNIQUES TO SYNTHETIC-APERTURE RADAR. L. J. Cutrona, Emmett N. Leith, Leonard J. Porcello and W. E. Vivian in *Proceedings of the IEEE,* Vol. 54, No. 8, pages 1026–1032; August, 1966.

AN INTRODUCTION TO SYNTHETIC-APERTURE RADAR. William M. Brown and Leonard J. Porcello in *IEEE Spectrum,* Vol. 6, No. 9, pages 52–62; September, 1969.

SYNTHETIC INTERFEROMETER RADAR FOR TOPOGRAPHIC MAPPING. L. C. Graham in *Proceedings of the IEEE,* Vol. 62, No. 6, pages 763–768; June, 1974.

SYNTHETIC APERTURE RADAR. Edited by John Kovaly. Artech House, 1976.

REMOTE SENSING. Allen L. Hammond in *Science,* Vol. 196, No. 4289, pages 511–516; April 29, 1977.

THE STRUCTURE AND FUNCTION OF HISTOCOMPATIBILITY ANTIGENS

MARKERS OF BIOLOGICAL INDIVIDUALITY. Ralph A. Reisfeld and Barry D. Kahan in *Scientific American,* Vol. 226, No. 6, pages 28–37; June, 1972.

BIOLOGY OF THE MOUSE HISTOCOMPATIBILITY-2 COMPLEX. Jan Klein. Springer-Verlag, 1975.

STRUCTURE AND SIGNIFICANCE OF β_2-MICROGLOBULIN. Bruce A. Cunningham in *Federation Proceedings,* Vol. 35, No. 5, pages 1171–1176; April, 1976.

SURFACE MODULATION IN CELL RECOGNITION AND CELL GROWTH. Gerald M. Edelman in *Science,* Vol. 192, No. 4236, pages 218–226; April 16, 1976.

THE SOLUTION OF THE FOUR-COLOR-MAP PROBLEM

THE FOUR-COLOR PROBLEM. Oystein Ore. Academic Press, Inc., 1967.

GRAPH THEORY: 1736–1936. Norman L. Biggs, E. Keith Lloyd and Robin J. Wilson. Clarendon Press, Oxford, 1976.

EVERY PLANAR MAP IS FOUR COLORABLE: PART 1, DISCHARGING. Kenneth Appel and Wolfgang Haken in *Illinois Journal of Mathematics,* Vol. 21, No. 3; September, 1977.

EVERY PLANAR MAP IS FOUR COLORABLE: PART 2, REDUCIBILITY. Kenneth Appel, Wolfgang Haken and John Koch in *Illinois Journal of Mathematics,* Vol. 21, No. 3; September, 1977.

HOW THE IRON AGE BEGAN

THE DARK AGE OF GREECE. A. M. Snodgrass. Edinburgh University Press, 1971.

A STEEL TOOL OF THE FOURTH CENTURY B.C. FROM AL MINA IN SYRIA. Tamara S. Wheeler, James D. Muhly and Robert Maddin in *Levant,* Vol. VIII, pages 107–112; 1976.

AN IRON ADZE OF THE FIFTH–FOURTH CENTURIES B.C. FROM AL MINA. James D. Muhly, Tamara S. Wheeler and Robert Maddin in *Levant,* Vol. IX, pages 156–161; 1977.

HALLUCINATIONS

MESCAL AND MECHANISMS OF HALLUCINATIONS. Heinrich Klüver. The University of Chicago Press, 1966.

HASHISH AND MENTAL ILLNESS. J. J. Moreau. Raven Press, 1973.

HALLUCINATIONS: BEHAVIOR, EXPERIENCE AND THEORY. Edited by Ronald K. Siegel and Louis Jolyon West. John Wiley & Sons, Inc., 1975.

THE AMATEUR SCIENTIST

AN OCEANOGRAPHICAL CURIOSITY: THE PERPETUAL SALT FOUNTAIN. Henry Stommel, Arnold B. Arons and Duncan Blanchard in *Deep Sea Research,* Vol. 3, No. 2, pages 152–153; February, 1956.

THE "SALT-FOUNTAIN" AND THERMOHALINE CONVECTION. Melvin E. Stern in *Tellus: A Quarterly Journal of Geophysics,* Vol. 12, No. 2, pages 172–175; May, 1960.

OBSERVATIONS OF THE CELL STRUCTURE OF SALT FINGERS. T. G. L. Shirtcliffe and J. S. Turner in *Journal of Fluid Mechanics,* Vol. 41, Part 4, pages 707–719; May 15, 1970.

The roomy 1978 Ford LTD.
It could be just the car your family needs.

FORD
When America needs a better idea, Ford puts it on wheels.

Anyone who needs the roominess and comfort of a family-size car should consider the 1978 Ford LTD. Here are three good reasons why:

6-passenger comfort

You've come to expect luxury from the 6-passenger Ford LTD. And the 1978 Ford LTD is no exception. You'll find luxury details everywhere.

Then try the LTD ride with its wide stance—and luxurious stretch-out space for six! There's more hip-and-shoulder room than in Chevrolet Impala and Chevrolet Caprice Classic.

Roomier trunk

What's more, Ford LTD once again has a roomier trunk than a Chevrolet Impala or Caprice. It's a spacious deep-well trunk that accommodates eight pieces of luggage (including four men's 2-suiters) and two sets of golf clubs!

Greater rated towing capacity

For those who do heavy towing, here's good news: Ford LTD with the optional towing package has a rated towing capacity of 7,000 pounds—1,000 pounds more than Chevrolet.

So this year more than ever, consider a Ford LTD. You can put more into it, and get more out of it.

More room here

More room here