SCIENTIFIC AMERICAN

EXPLORING THE HERBARIUM

$1.50

May 1977

"Friends ask me if we cap natural gas wells instead of sending the gas someplace useful."

"I get questions like that every day," says Gulf Offshore Operations Manager Roy Tillerson.

"People say, 'Is it true there are capped natural gas wells when people need gas so badly?'

"I have to say yes, because there are. But it's not that simple.

One well isn't a field

"When a company thinks there's natural gas down there, it drills an exploratory well. But one gas well doesn't make a gas field. If you find gas with the first well, you have to cap it and drill maybe a dozen more, to see how much gas there is.

"People say, 'If you know there's gas, why not just build a pipeline to where it can be used?'

"Gulf has hundreds of gas wells in the Gulf of Mexico, working at full capacity."

"And I say, a gas field could cover anywhere from one square mile to a thousand. Until you know how big the field is, you don't know how big a pipeline to build. In many cases, federal law doesn't allow pipeline construction until it can be proved there's enough gas in the field to justify it.

"Even if you were allowed to build unlimited pipelines, the cost of the pipelines would send the price of gas out of sight. For instance, Gulf just finished a 60-mile gathering pipeline that cost $25 million. It was built to serve as many as 36 wells. You can see that a separate pipeline for each of those wells would be astronomically expensive.

"This is a capped well. But one well is a long way from bringing in enough gas to make a pipeline practical."

Working full time

"So we cap wells until we find enough natural gas to make it practical to pipe it out.

"Gulf already has thousands of wells producing all the natural gas they can. The challenge we're facing now is to find more natural gas, and ship it to consumers the cheapest way."

Gulf

Gulf people: meeting the challenge.

Gulf Oil Corporation

You have a choice. Chevy Monza Coupe or 2+2 Hatchback. Two different looks, one simple fact: Monzas are for people who know that driving is still America's greatest outdoor sport.

Whether you're a gymkhana expert or a long-distance commuter, Monza available equipment includes everything you need for the kind of performance you want: 5.0-litre, 2-bbl. V8, Chevy F41 sport suspension with rear stabilizer, aluminum wheels, four-speed stick or Turbo Hydra-matic, full center console, sport steering wheel, sport equipment and sport front end—and a lot more.

You've waited long enough for a little road magic. Now you can have it, straight from your Chevy dealer. Monza Coupe or Monza 2+2. Your choice.

The **MONZAS** Chevrolet by Chevrolet.

Established 1845 **SCIENTIFIC AMERICAN** May 1977 Volume 236 Number 5

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.

THE SCOTCH WITH GREATNESS WRITTEN ALL OVER IT.

The hand-lettered label on a bottle of Cutty Sark Scots Whisky makes a statement no other Scotch can make.

Start at the top. Cutty Sark is still blended and bottled only in Scotland. Somehow, the modern Scotch, bulk-shipped and bottled in America, isn't quite the same.

The centre of the label tells you that Cutty Sark consists of 100% Scotch Whiskies from Scotland's best distilleries. This produces that distinctive smoothness which is to be found in this blend alone.

Reading down, you learn that Cutty Sark was first blended at the direction of Berry Bros. & Rudd Ltd., wine merchants. People who, after 280 years of dealing with Royalty, have learned a little something about taste.

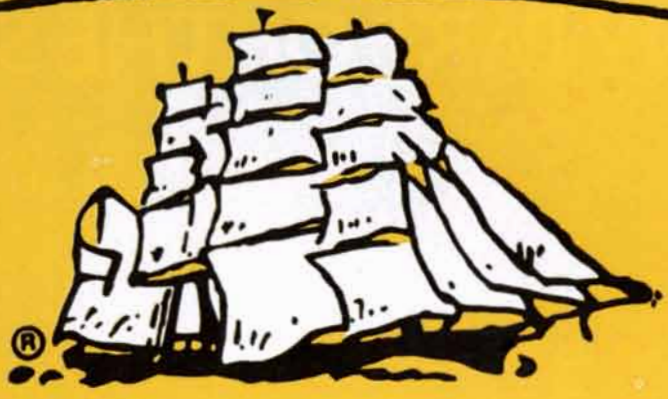
If you have good taste, you must buy the Scotch that tells others you have it.

It's the one in the dark green bottle with the bright yellow label that separates it from all the rest. The one that tells the greatest Scotch story ever told.

"CUTTY SARK." "CUTTY" THE CUTTY SARK LABEL AND THE CLIPPER SHIP DEVICE ARE REGISTERED TRADEMARKS OF BERRY BROS & RUDD LTD. LONDON, ENGLAND. 86 PROOF BLENDED SCOTCH WHISKY DISTILLED AND BOTTLED IN SCOTLAND. IMPORTED BY THE BUCKINGHAM CORPORATION NEW YORK, N.Y

Distilled and Bottled in Scotland
under British Government Supervision

CUTTY SARK



BLENDED SCOTS WHISKY

100% Scotch Whiskies
from Scotland's best Distilleries

86 Proof

4/5 Quart

Blended &

bottled by

BY APPOINTMENT
TO HER MAJESTY THE QUEEN
WINE & SPIRIT MERCHANTS

BERRY BROS & RUDD Ltd

ESTABLISHED IN THE XVII CENTURY

3. St JAMES'S STREET, LONDON, S.W.1

Product of Scotland

WHEN YOU BUILD A SPEAKER TO SOUND GREAT ON EVERY PART OF THE MUSIC, YOU CAN'T CUT CORNERS ON ANY PART OF THE SPEAKER.

A single HPM-100 weighs almost 60 pounds.

The fact it weighs more than a Large Advent speaker, Bose 901 or JBL L100 is not an accident.

Our speaker frames are made of heavy cast aluminum instead of the usual stamped metal, so you hear only the speakers vibrating and never their frames.

Our magnets are oversize to spare your ears needless distortion.

And our cabinet is made out of special compressed wood that's denser and heavier than ordinary wood. So the sound is forced out of the cabinet instead of being absorbed by it.

Of course, not everything that adds to the sound of an HPM-100 also adds to its weight.

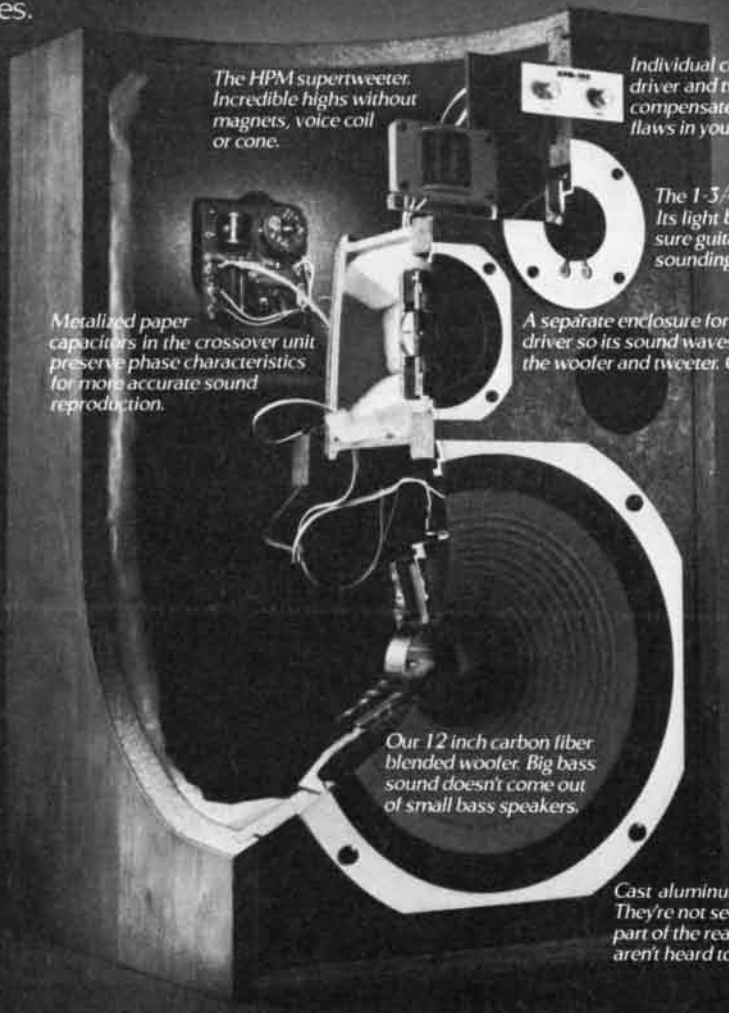
Our supertweeter uses nothing but a piece of High Polymer Molecular film to produce incredibly clear and crisp high frequencies.

Our midrange driver and tweeter have cones that are light enough to give sharp response, but rigid enough not to distort.

And our 12 inch woofer has a long throw voice coil and unique carbon fiber blend cone (instead of the typical cardboard cone) that work to produce the kind of realistic bass you not only hear, but feel.

Naturally, we could go on. About our 12-1/2 feet of damping material. Or about the aluminum screws that keep our speakers from falling out. They're ordinarily used to keep airplanes from falling apart.

But we figure at this point you'd rather hear our speakers in person than hear any more about them from us.



HPM-100™
The all-around great speaker.

PIONEER

THE COVER

The picture on the cover is a reproduction of an illustration in a rare German herbal, *Contrafayt Kreüterbuch* ("book of herb images"), which was printed in Strasbourg (then Strassburg) between 1532 and 1537 and is now in the library of the New York Botanical Garden. The text was written by Otho Brunfels, a physician and botanist, and the illustrations were hand-painted woodcuts based on watercolors by Hans von Weiditz; this one is of the *Leberkraut* ("liver herb"), which we know as the hepatica, *Hepatica triloba*. Herbals were compendiums of plant knowledge with major emphasis on medicinal applications. The hepatica was said to be effective in "drawing off the yellow bile," for example. Until recently botany and medicine were closely related; even now many drugs are derived from (or synthesized to mimic or improve on) plant substances. A promising source of new drugs is therefore the world's herbaria, where hundreds of thousands of dried plant specimens are preserved, many of them carrying ethnobotanical notes describing their uses by indigenous populations (see "Exploring the Herbarium," by Siri von Reis Altschul, page 96).

THE ILLUSTRATIONS

Cover photograph by Fritz Goro

Page	Source
10	Steinhart Aquarium, California Academy of Sciences, San Francisco (*top*); Tom Prentiss (*bottom*)
21–27	Ilil Arbel
29–34	Tom Prentiss
35	Elwyn L. Simons, Yale University
37	Fritz Goro
38–48	Dan Todd
53	Ilil Arbel
63	Andrejs Liepins, Sloan-Kettering Institute for Cancer Research
64–67	Patricia J. Wynne
68	Etienne de Harven, Sloan-Kettering Institute for Cancer Research (*top*); Patricia J. Wynne (*bottom*)
69–70	Patricia J. Wynne
72	William D. Hardy, Jr., Sloan-Kettering Institute for Cancer Research
73–76	Gabor Kiss
78	Lloyd J. Old, Sloan-Kettering Institute for Cancer Research
81	New York Public Library
82–88	Allen Beechel
97–101	Fritz Goro
102	The Bettmann Archive
107–114	Bunji Tagawa
120–126	Gabor Kiss
128	Laszlo Kubinyi
131–135	Ilil Arbel
142	*Carnivorous Plants of the United States and Canada.* © 1976, John F. Blair, Publisher
144	*The Incredible Dr. Matrix.* © 1976, Charles Scribner's Sons

LETTERS

Sirs:

The article on cruise missiles ["Cruise Missiles," by Kosta Tsipis; SCIENTIFIC AMERICAN, February] was illuminating and interesting, not least because it indicated future policies of great value for arms-control and disarmament negotiations.

The pinpointing of the opportunity to exercise "mutual restraint" in regard to the development and deployment of long-range cruise missiles urgently calls for that principle to be incorporated in the continuation of the bipolar SALT talks. I would also strongly agree with the recommendation for an initial period of unilateral restraint, which for obvious reasons should first be the responsibility of the U.S., which is far ahead of any competitor in the field.

In particular, however, I write these lines to remind all readers of the interest of countries other than the two superpowers in any development of cruise missiles, potentially revolutionizing as they are as weapons vehicles. If a land-based or ship-based cruise missile for tactical purposes and ranges were to be developed in line with the designs that Dr. Tsipis describes—a missile of small size, high precision and built for carrying conventional bombs—it ought not

Scientific American, May, 1977; Vol. 236, No. 5. 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

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

"Unquestionably the finest single-volume encyclopedia in the English language" — *Bowker's Encyclopedia Buying Guide 1975-76*

YOURS FOR ONLY $15

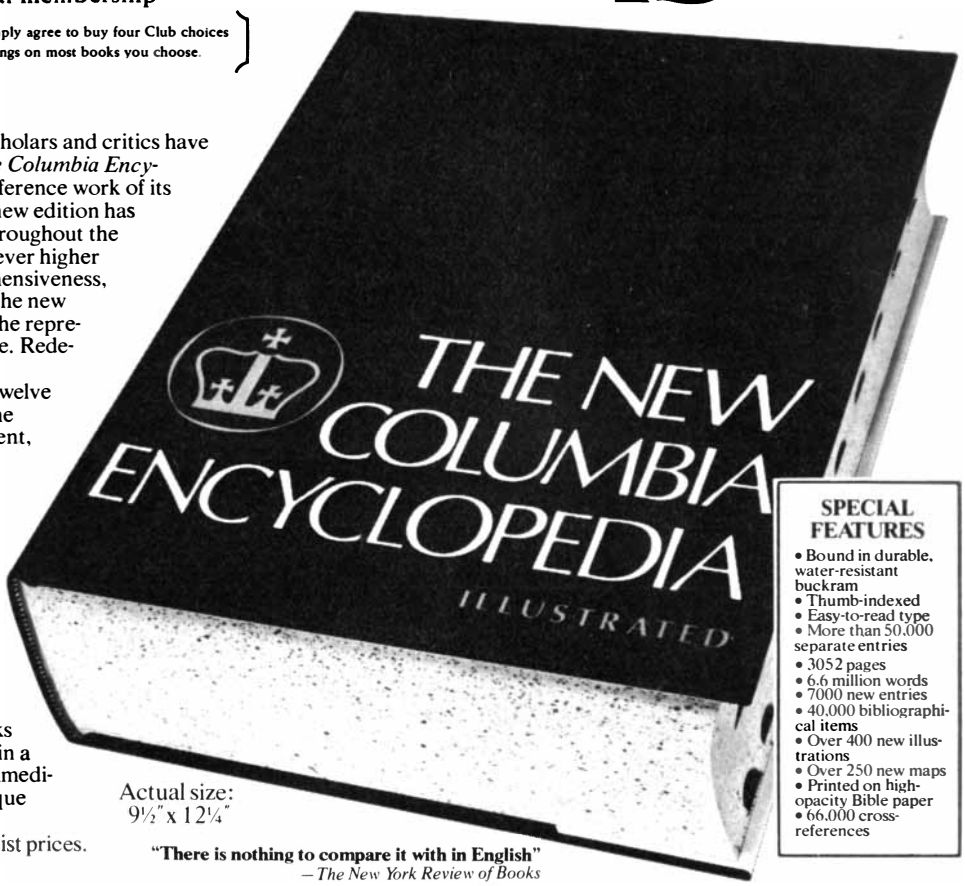
in a short trial membership

PUBLISHER'S LIST PRICE $79.50

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

FOR MORE THAN four decades scholars and critics have been unanimous in calling *The Columbia Encyclopedia* the best single-volume reference work of its kind ever compiled. And as each new edition has appeared it has been acclaimed throughout the English-speaking world as setting ever higher standards of scholarship, comprehensiveness, clarity and skillful compression. The new fourth edition is no exception, as the representative comments below indicate. Redesigned, reorganized, rewritten and freshly illustrated after a lapse of twelve years, it is a worthy successor to the previous edition. And for the student, the businessman, the professional or, indeed, anyone who reads as much as a daily newspaper, it will prove to be one of the most useful books one can own.

The extraordinary opportunity to own this great work *at an 80% saving* is made to demonstrate the continuing benefits of membership in the Book-of-the-Month Club. For not only will you be able to keep abreast of the important new books you are eager to read; if you remain a member after the trial, you will immediately participate in the Club's unique Book-Dividend® plan — and save 70% and more on the publishers' list prices.



Actual size: 9½" x 12¼"

SPECIAL FEATURES

- Bound in durable, water-resistant buckram
- Thumb-indexed
- Easy-to-read type
- More than 50,000 separate entries
- 3052 pages
- 6.6 million words
- 7000 new entries
- 40,000 bibliographical items
- Over 400 new illustrations
- Over 250 new maps
- Printed on high-opacity Bible paper
- 66,000 cross-references

"There is nothing to compare it with in English"
— *The New York Review of Books*

"Surpasses in usefulness even its great ancestor. . . . I judge it to be the reference work of first resort for scholar and layman alike." — *Lionel Trilling*

"The best one-volume repository of useful knowledge in the language. It is *the* indispensable reference work." — *Ashley Montagu*

"Without peer in the English language." — *Scientific American*

"One can keep the whole range of [factual] knowledge on one's lap . . . in an immediately available form." — *Times Literary Supplement* (London)

NEW FOURTH EDITION... COMPLETELY REVISED AFTER A LAPSE OF TWELVE YEARS

A Few of the 7000 New Articles

World Affairs
Valéry Giscard d'Estaing
Chiang Ching
The Transkei
Arab-Israeli Wars

Life and Physical Sciences
Pulsar
Gravitational Collapse
Vasectomy
Group Psychotherapy

Social Sciences
Space Law
Ralph Nader
Betty Friedan
No-Fault Insurance

Humanities
African Languages
Rudolf Nureyev
François Truffaut
Rock Music
Aleksandr Solzhenitsyn

Sports
Bobby Fischer
Billie Jean King
Muhammad Ali
Hank Aaron

U.S. National Affairs
Pentagon Papers
Henry Kissinger
Watergate Affair
American Indian Movement

Religion, Philosophy
Pierre Teilhard de Chardin
Pure Land Buddhism
Native American Church
Hans Küng

PLUS such useful features as: the text of the Constitution of the United States . . . tables of rulers of France, Great Britain, Spain and Russia . . . outstanding U.S. senators, congressmen and cabinet members . . . Nobel Prize winners . . . Third World persons, events, art forms and places . . . and separate articles on more than 2000 U.S. cities.

BOOK-OF-THE-MONTH CLUB, INC., Camp Hill, Pennsylvania, 17012 **7-A170-5**

Please enroll me as a member of the Book-of-the-Month Club and send me *The New Columbia Encyclopedia*, billing me $15. 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.

23

MR. / MRS. / MISS .....
(Please print plainly)

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

City &
State ..... *Zip* .....

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 10 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.

to, and could not, remain a monopoly of the superpowers. Its value would be such that it should be made available to other countries, with either the missiles or their designs openly available in the marketplace.

But most imperative for the sake of promoting both international equity and arms limitation is that any advances in regard to reconnaissance be internationalized and that a multilateral approach instead of just a bipolar one be followed, as I have so strongly argued in my book *The Game of Disarmament.*

ALVA MYRDAL

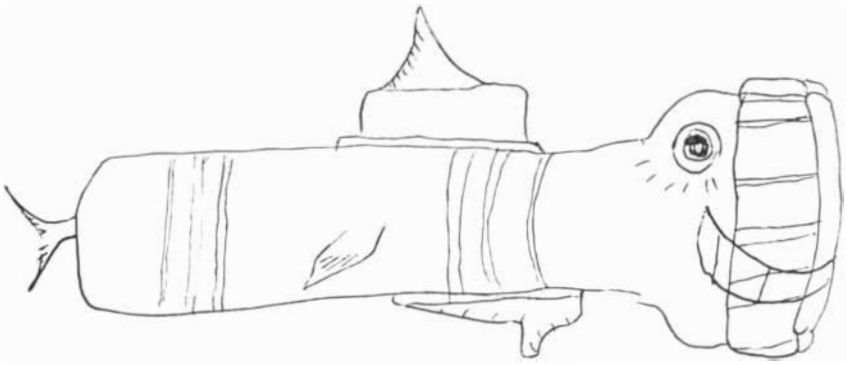
Ministry of Foreign Affairs
Stockholm

Sirs:

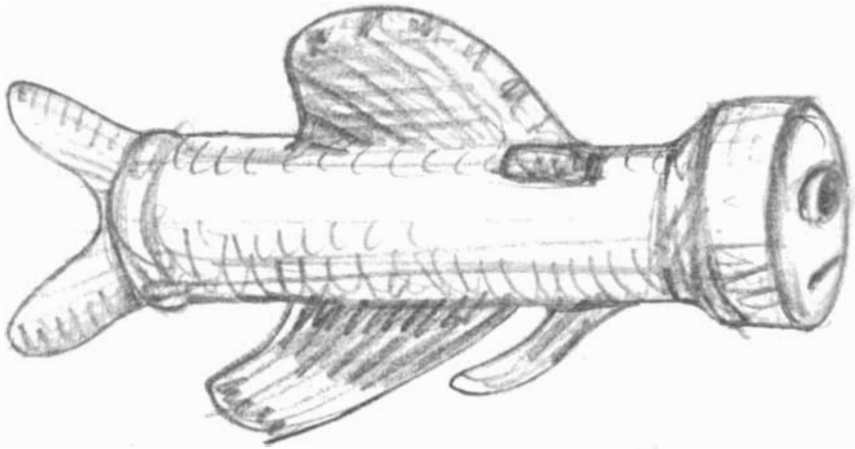
As a postscript to my article on flashlight fishes ["Flashlight Fishes," by John E. McCosker; SCIENTIFIC AMERICAN, March] I should like to describe an extraordinary discovery that lends credence to the concept of convergent evolution. In March, 1975, I returned from the Comoro Islands with the first living specimens of the flashlight fish *Photoblepharon* to be displayed in an aquarium. After spending 30 sleepless hours on airplanes and in airports on three continents I landed in San Francisco with flashlight fishes whose eyes (and light organs) were nearly as bleary as my own. The following day, while unearthing my desk under six weeks of accumulated correspondence, I discovered the enclosed drawing of a flashlight fish whimsically conceived by my exhibits staff [*see upper illustration below*]. Clever, I mused. Now, two years later, I have received a drawing of a remarkably similar organism independently created by Tom Prentiss in the course of doing the drawings for my article [*see lower illustration below*]. This example of battery-powered speciation, evolved in allopatry at opposite ends of the New World, is enough to cause one to reexamine the piscine pillars (fishy foundations?) of one's Darwinian principles.

JOHN E. MCCOSKER

Superintendent
Steinhart Aquarium
California Academy
of Sciences
San Francisco



Battery-powered flashlight fish from the Steinhart Aquarium in San Francisco

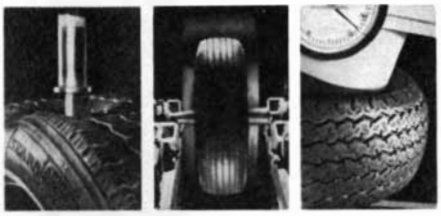


"Photoblufferon," a genus of independent origin that lends credence to convergent evolution

Sears RoadHandler. Better after 40,000 miles than Federal standards say a new tire has to be.

"Proof of Performance" facts about Sears best Steel-Belted Radial, the tire that ran the route of the Pony Express.

With 40,000 miles of relentless day and night driving already behind them, four Sears RoadHandlers set out and tamed the rugged 2,000 mile route of the old Pony Express.



Then we put them to the real test: Department of Transportation Federal Motor Vehicle Safety Standard 109 for *new* tires.

Results? Sears RoadHandlers exceeded every government requirement, *e.g.* for strength, heat resistance and bead unseating (keeping the tire on the rim in hard turns).

Even when subjected to X-ray examination of the complete tire, these RoadHandlers showed no failure of any kind anywhere.

No wonder they're Sears best Steel-Belted Radial.

15% better wet cornering than Sears original Steel-Belted Radial.

A new wet cornering traction record for Sears tires! Under carefully controlled laboratory conditions at Calspan Tire Research Facility, Sears RoadHandler beat (by an impressive 15%) the original Sears Steel-Belted Radial that beat the Baja.

Two more facts: 12.5% more rubber meets the road for quicker, surer stops. And more water is channeled along 5 tread grooves instead of 4 to help prevent *hydroplaning* (skidding on the thin film of water between a rainy road and the tire's surface).

Sears

Tire and Auto Centers and Catalog.

"Straight talk, good values and satisfaction."

A WAGON BUILT MORE PRECIOUS

Any station wagon can take a load of stuff from one place to another.

The Volvo station wagon was designed to take a load off your mind as a parent.

Volvo realizes, for example, that it's impossible to keep both eyes on the road if you have to keep one eye on the back seat.

So to keep your children in place, we provide you with things like child-proof door locks on *all* the rear doors. Including the back one.

And to virtually guarantee that you can focus your attention on the road at all times, we give you defrosters for the front side windows. And the rear window comes with its own wiper, washer, and defroster.

On the road, the first thing you'll notice is how quickly our overhead cam engine can put trouble behind you. Should trouble appear ahead, you'll appreciate the way our rack and

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MAY, 1927: "No stronger endorsement could be given of Vice-President Dawes's determination to make such radical changes in the rules governing debate in the Senate as will prevent 'filibustering' than the recent disgraceful exhibition against the discussion of the Boulder Creek Dam bill. The damming of the Colorado River is the most important public engineering work now before the American people, comparing in its prospective usefulness with the Panama Canal. Until the Boulder Creek Dam is completed, the wide-spreading and fruitful Imperial Valley will be subject to the threat of a devastating inundation such as that of 20 years ago. There is every reason for the construction of this work, and not one can be offered against it. This bill that is awaiting the approval of the Senate not only will rid the Western farmer of an ever present menace but also will gradually place at his disposal hydro-electric energy up to a total of five million horsepower. Five great Western states await the decision of the Senate."

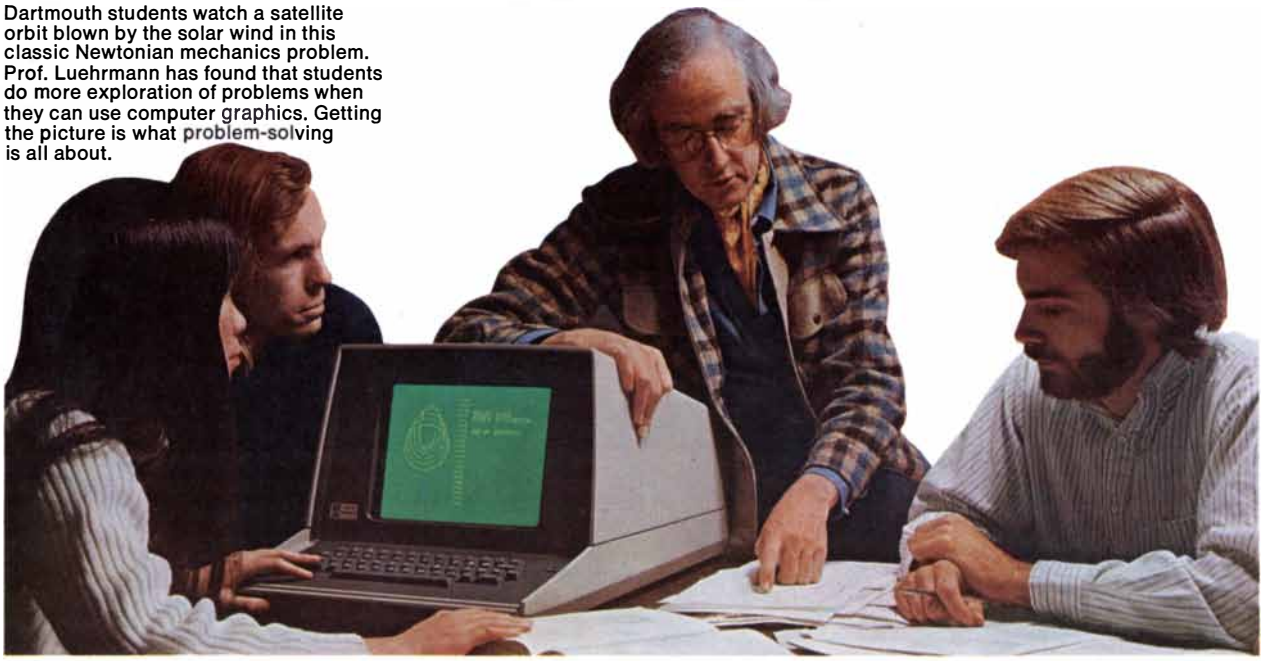
"Professor Michael I. Pupin, physicist and inventor, delivered the presidential address at the annual convention of the American Association for the Advancement of Science. In it he mentioned strange solar activities and suggested that radio static represents 'messages' from the sun, 93,000,000 miles away! As Professor Pupin later explained: 'By electrical messages I meant a solar electromagnetic action that manifests itself as static and fading, things that we do not know the cause of today. These manifestations are messages from the sun; not messages in the literal sense but an action that within 25 years we will be able to decipher by means of radio. The deciphering process will show us the real cause of static and fading.'"

"The distillation of oil shale, which promises to become very important to the entire nation, is not a new way of getting oil at this time of approaching scarcity but is an old method from which the cobwebs are being brushed after more than half a century of neglect in the United States. In France it is older still; there the production of oil from shale started in 1838. In Scotland oil was obtained from shale before 1850, but the best of the raw material is now gone. Shale pits 3,000 feet deep show how the industry once flourished. In 1850 oil was

TO GET YOUR "FRENCH WINE BASKET" SEND $3.95 TO: B&G-BEAUJOLAIS SAINT-LOUIS, 505 BEACH STREET, SAN FRANCISCO, CA 94133.

"With computers and graphics at Dartmouth we're taking students out of their passive receptor situation and putting them into the active inquiry role of a researcher."

Prof. Arthur Luehrmann
Office of Academic Computing, Dartmouth College, Hanover, N.H.



Dartmouth students watch a satellite orbit blown by the solar wind in this classic Newtonian mechanics problem. Prof. Luehrmann has found that students do more exploration of problems when they can use computer graphics. Getting the picture is what problem-solving is all about.

Arthur Luehrmann was one of the first professors in the country to use time-shared computer graphics extensively in teaching. In his field — physics — a primary stumbling block for beginning students was the lack of an intuitive "feel" that made sense of all the calculus and diagrams. Words and equations simply failed to convey the idea of a system evolving, step-by-step, from an initial condition through a series of neighboring states. Computer algorithms and graphic output did the job.

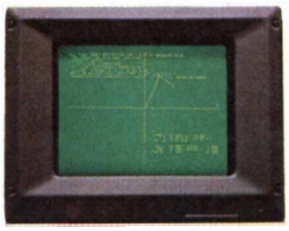
At Dartmouth, in 1968, Prof. Luehrmann began using x-y plotters to convert tables of numbers into pictures. In 1969, Tektronix supplied Dartmouth with an early Model 4002, the first graphic terminal priced under $10,000. Today, Dartmouth has dozens of graphic terminals and the majority bear our name. We are participating in Prof. Luehrmann's development of new graphic syntax in BASIC, the computer language.

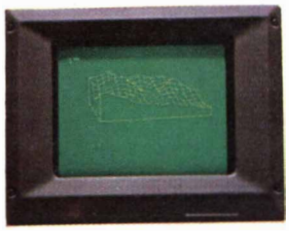
Prof. Luehrmann says that the exciting part of using computer graphics in education is that it causes students to formulate problems in a new language. "Not just programming language," he says, "but conceptual language where the student is thinking in terms of procedures and sub-procedures and loops and branches and all of these complex ideas that are very powerful. Students develop new ways of thinking about problems."

Prof. Luehrmann has been something of an evangelist about computer graphics, writing enhancements of BASIC and showing his colleagues, both at Dartmouth and beyond, how this emerging tool can help both their teaching and research.

Students watch business-cycle and growth-pattern data, plotted in a time series, come to life as moving lines on the projection screen. "For teaching, there's nothing like graphics," Prof. McGee says. Working before the class at a graphic terminal instead of a blackboard, while the terminal's picture is projected via a scan converter, Prof. McGee is here *generating* data according to one of the Box-Jenkins forecasting models. While the numbers are being generated they can also be stored while a second program analyzes the data empirically. Exactly the same boxes appear in the *analyzed* data, but this time the autocorrelation and the partials are derived from the data, not from theory. "I use these programs in front of the class to show how the theoretical properties of a time series model can be obscured by 'data noise' when we do the analysis."

Computer graphics helps Prof. Victor McGee teach statistics to students in Dartmouth's AmosTuck School of Business Administration.





Optimization Prof. Alvin Converse teaches engineering courses involving computer simulation, and regularly uses the lecture hall graphic terminal and TV projector for demonstrating concepts like optimization. "Students are highly motivated by graphic output, as most everyone is," he says. "You can actually watch the progress of the algorithm trying to find the top of the surface. For instance, if the test function is a curved valley, you can watch certain methods get hung up and just not move. And you can watch other methods come in very quickly, or rapidly diverge if they're not very stable. Graphic output has turned out to be an interesting pedagogical tool."

Mapping Prof. Robert Huke, of the Geography Department, finds that using a Tektronix graphic tablet — an electronic grid of thousands of points teamed up with the terminal — saves him and his students days of laborious tracing and drafting. With planimeter mapping, the scientist uses an electronic stylus to trace the area of interest on a map, which has been overlaid on the graphic tablet. The terminal simultaneously shows the shape being traced, and then displays the computer's calculations of the area. Not ready for hard copy? Fine. Store the data, recall it later. For choropleth mapping, graphically expressed values for anything—rainfall, for example—appear exactly where they pertain—measuring stations within a state, for example. The uses go on and on.

From the beginning, students are taught to write programs that make use of the graphic software already in the computer. Program writing is not considered a rote exercise. This is the way Arthur Luehrmann puts it: "When a person is writing a computer program he is cast in the role of a teacher explaining to a perfectly logical infinitely patient, and totally unimaginative student how to carry out a sequence of tasks. It is axiomatic that one never learns a subject so well as when he teaches it to someone else."

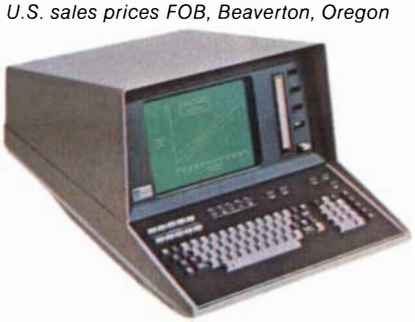
These are just a few of the many ways in which computer graphics is helping to improve the learning process at Dartmouth. It is growing and changing constantly, as more people discover that computer graphics is easier to use than they think.

The graphic tools available to Dartmouth students and faculty stem from a large library of graphic software developed by Arthur Luehrmann and his associates. The package of graphic sub-programs is written in Prof. Luehrmann's enhanced BASIC, so that no user need be locked into using any particular manufacturer's hardware.

The central computer is a large Honeywell 66-40, connected to a network of more than 300 terminals in 25 buildings on the campus. In addition, an educational network reaches dozens of other schools and colleges from New York City to northern Maine, and west as far as Chicago.

Tektronix has been working hard to remove the cost barriers to even wider use of computer graphics. The Tektronix 4006 terminal, shown middle left on the opposite page, is priced at only $2995, a new standard of value. The Tektronix 4051 desktop computer provides graphics power, either standing alone or on-line with a larger computer, from $7,500.

U.S. sales prices FOB, Beaverton, Oregon



These are just two of the 835 Tektronix products and instruments serving the test, measurement and display needs of thousands of customers worldwide. Not only in education but also in science, health, communications and industry.

For a copy of our annual report or product information, write to the Marketing Communications Department, 76-260, Tektronix, Inc., P.O. Box 500, Beaverton, Oregon 97077.
By phone—call 503-644-0161.

Tektronix
COMMITTED TO EXCELLENCE

Tektronix graphic terminals are now located in many parts of the Dartmouth campus. Here in the Mathematics Department students are visualizing functions in three dimension, working on their own. A nearby hard-copy printer captures permanent records, whenever needed, of any picture that shows on the screen.

distilled from shale in Utah and Pennsylvania, and the development of a great shale industry would have resulted but for the discovery of well oil. The 'liquid gold' fever, following the development of oil in drill wells in Pennsylvania, caused the interest in shale oil to dwindle and then die, just as an impending scarcity has brought it to life."

"There are 6,500,000 radio receivers in operation throughout the United States today, compared with 60,000 in 1922, and the audience listening in today is about 26,000,000 as against 75,000 then. The large increase in the audience is attributable to the fact that loudspeakers are generally used instead of headphones, so that the entire family can enjoy the etherial entertainment."

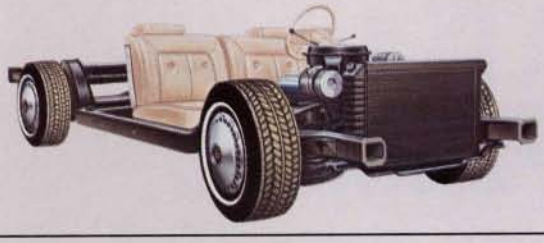
SCIENTIFIC AMERICAN

MAY, 1877: "Dr. C. W. Siemens, the new president, delivered an inaugural address to the London Iron and Steel Institute. Among his remarks were the following: 'Time will probably reveal to us effectual means of carrying power to great distances, but I cannot refrain from alluding to one which is, in my opinion, worthy of consideration, namely the electrical conductor. Suppose water power to be employed to give motion to a dynamo-electrical machine. A powerful electric current is the result. This may be carried to a great distance through a large metallic conductor, and there be made to impart motion to electromagnetic engines, to ignite the carbon points of electric lamps or to effect the separation of metals from their combinations. A copper rod of 3 in. diameter would be capable of transmitting 1,000 horse power a distance of, say, thirty miles, an amount sufficient to supply one quarter of a million candle power, which would suffice to illuminate a moderately sized town. The use of electrical power has sometimes been suggested as a substitute for steam power, but it should be borne in mind that so long as the electrical power depends upon a galvanic battery it must be much more costly than steam power, inasmuch as the combustible consumed in the battery is zinc, a substance necessarily much more expensive than coal. But the question assumes a different aspect if in the production of the electric current a natural force is used, which could not otherwise be made available.'"

"The two great grain-producing lands of the world are southern Russia and the valley of the Mississippi, and between these sections there exists active competition for the supplying of the nine to fourteen million quarters of foreign wheat required by England, and the large additional amounts needed by other European nations. Already in this rivalry our Western farmers are far ahead, and statistics, recently gathered by the Odessa Committee on Trade and Manufactures for the information of the Russian government, show with what remarkable rapidity this advance has been accomplished. In 1867 the proportions of wheat supplied by Russia and the United States to England were respectively 44 per cent and 14 per cent. In 1873 the proportions were respectively 21 per cent and 44 per cent."

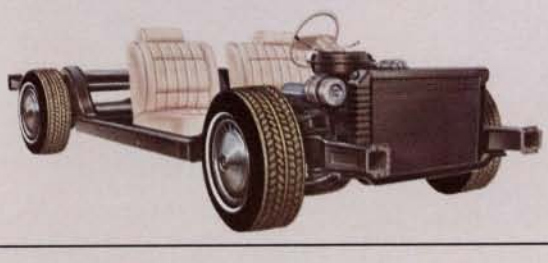
"The distinguished Russian chemist Mendelejeff thinks it very improbable that petroleum is the product of decomposed organic matter. Mendelejeff starts with Laplace's theory of the formation of the earth, applies Dalton's law to the original gaseous condition of the constituents of the earth and calculates the probable arrangement of the metals in the earth from the density of the globe and of the elements. Starting with the assumption that iron is the most abundant of metals and admitting the existence of carbon compounds of this metal, not only will it be easy to explain the formation of petroleum but also one can understand all the peculiarities of its occurrence in those places where the earth's strata have been broken by the elevation of mountain chains. Breaks made in this way allowed the water to permeate to the carbonaceous metals, and at high temperature and heavy pressure it acted upon them, forming oxides of the metals and saturated hydrocarbons. The latter rose as vapors to the highest strata, where they were condensed, saturating the porous sandstones, which are capable of absorbing many oily products."

"Dr. Oliver Holmes, the poet, author, scientist, inventor of the popular stereoscope instrument, recently delivered an address before the Boston Microscopical Society. 'To those of my generation,' he began, 'this modern world which most of you take as a matter of course, it being the only condition of things of which you have experience, is a perpetual source of wonder—a standing miracle. Science and art have in our time so changed the aspect of every-day life that one of a certain age might well believe himself on another planet or another stage of existence. The wand of Prometheus is in our matchboxes; the rock of Horeb gushes forth in our dressing rooms; the carpet of Arabian story is spread in our Pullman car; our words flash from continent to continent; our very accents are transmitted from city to city; the elements of forming worlds are analyzed in our laboratories; and, most wonderful and significant of all, the despotic reign of tradition received its death blow when the angel of anaesthesia lifted from womanhood the worst terrors of the primal malediction.'"



Cadillac Coupe de Ville.

Ford Thunderbird.



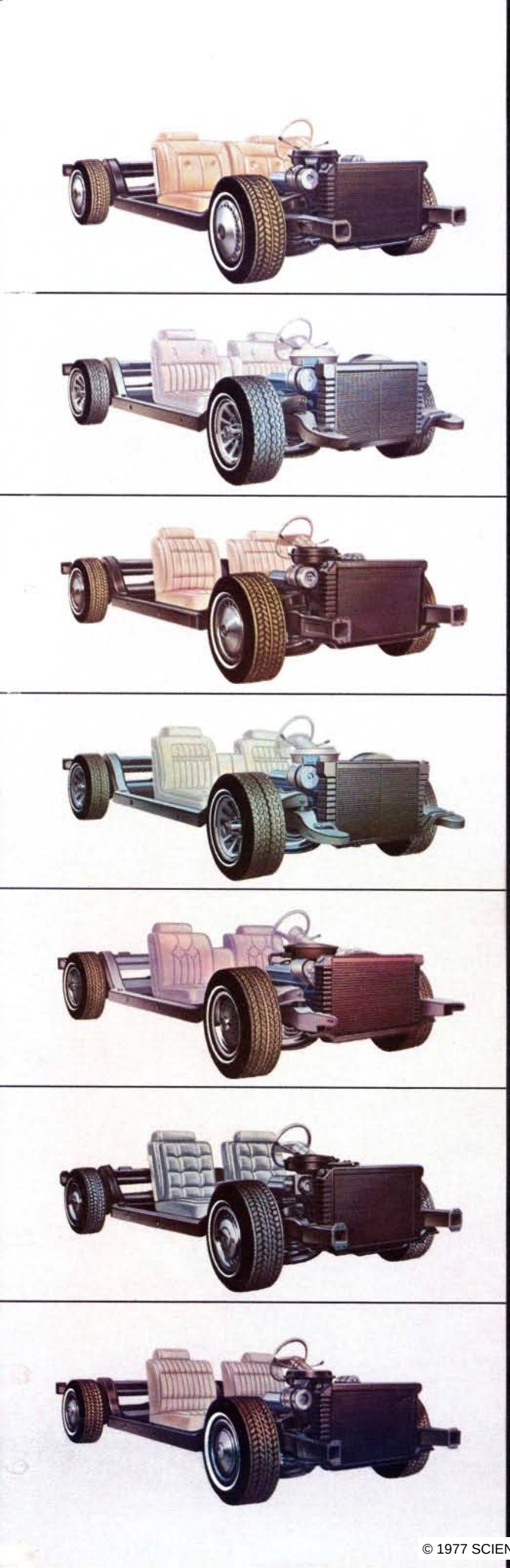
Chevrolet Monte Carlo.

Mercury Cougar XR-7.

Pontiac Grand Prix.

Oldsmobile Cutlass.

Buick Electra 225.



THE FINEST CARS IN THE WORLD USE UNIROYAL

Here are just seven that chose Uniroyal as one of their original equipment radial tires. Can you identify them?

Cadillac Coupe de Ville.

Ford Thunderbird.

Chevrolet Monte Carlo.

Mercury Cougar XR-7.

Pontiac Grand Prix.

Oldsmobile Cutlass.

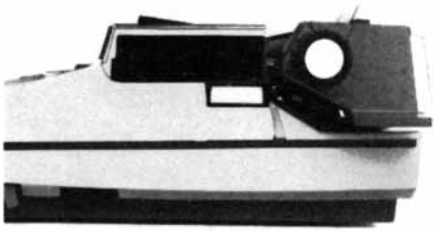
Buick Electra 225.

WHY SMITH-CORONA IS AMERICA'S BEST SELLING PORTABLE.

There are 15 brands of portable typewriters on the market.

But this year, like the last 10 years, more people will buy a Smith-Corona than any other brand.

Here are just five among many important reasons why:



THE TYPEWRITER WITH A HOLE

When we designed our cartridge, we re-designed our typewriters.

In place of spool cups, posts, reversing levers, ribbon guides and messy ribbons, there's a hole.

The hole is for our cartridges.

Next to the hole is a lever. Depress the lever and the cartridge pops out.

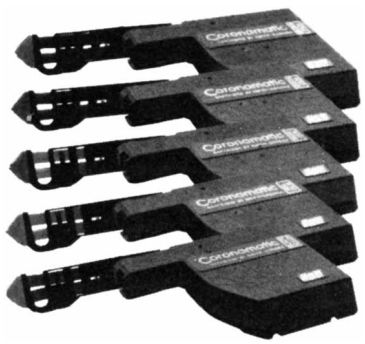
In three seconds, just by inserting a new cartridge, you can change a ribbon without getting your fingers dirty.

mistakr
mistak
mistake

CORRECTS MISTAKES IN 10 SECONDS

The hole also accommodates a correcting cartridge.

So when you make a mistake, you can snap out the typing cartridge, snap in the correcting cartridge, type over the mistake, snap the typing cartridge back in and type the correct character—all in ten seconds or less.



FIVE COLORS

Cartridges come in black, red, blue, green, and brown.

If you're typing along in black, but you want to type a line in red for emphasis, snap out the black cartridge and snap in the red cartridge. In just three seconds you can see red!

NYLON
FILM

EXECUTIVE LOOKING CORRESPONDENCE

The Smith-Corona electric portable with a film ribbon will give you the sharpest typing image of any portable.

When you want to type a letter that looks like an executive's, snap in the black film cartridge and type with real authority.

The black nylon is not quite so authoritative but is more economical. One cartridge lets you type about 325,000 characters.

FOUR YEARS OF COLLEGE AND BEYOND

We could make some remarkable statements about how we test out typewriters. But that's not necessary.

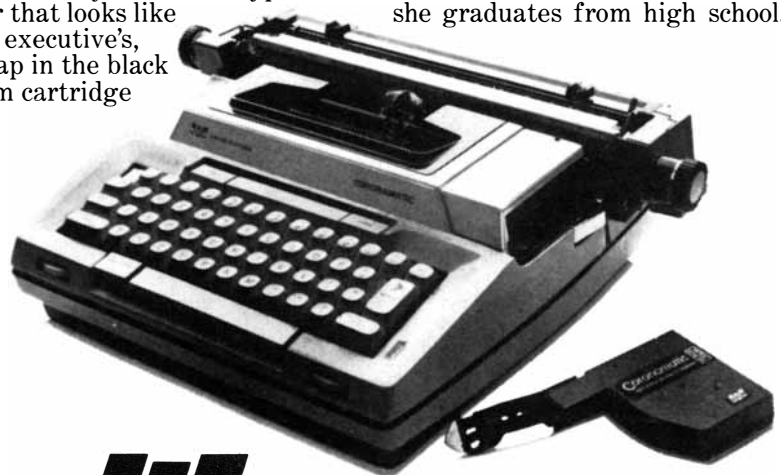
Most typewriter retailers have typewriters on display. Type a few sentences on a Smith-Corona. Snap a cartridge in and out. Smith-Corona looks and feels sturdy because it is sturdy.

So, for your high school graduate, we suggest you follow this simple formula:

Buy a doctor a Smith-Corona when he or she graduates from high school.

Buy a lawyer a Smith-Corona when he or she graduates from high school.

Buy a successful businessperson a Smith-Corona when he or she graduates from high school.



SCM® SMITH-CORONA
SCM CORPORATION

THE AUTHORS

ROBERT P. AMBROGGI ("Underground Reservoirs to Control the Water Cycle") has been senior adviser with the Food and Agriculture Organization of the United Nations since 1961. In addition he serves as a consultant to the United Nations Development Program, the World Bank and the United Nations Environment Program. A French citizen from Corsica, he was graduated from the University of Nancy in 1939 and received his doctorate at the University of Paris (the Sorbonne). Before he joined the Food and Agriculture Organization he was director of the Water Resources Department of Morocco, where in 1942 he began working in the field of ground-water hydrology. Ambroggi's major current interest is the improved management of the scarce resources of water of good quality. He is involved in the Blue Plan for the Mediterranean, a regional undertaking of the 18 Mediterranean countries to assist their governments in the sound development and national management of their natural resources.

ELWYN L. SIMONS ("Ramapithecus") is professor of geology at Yale University and head of the division of vertebrate paleontology at the Yale Peabody Museum. Last year he was on leave doing research at the University of Kassel in Germany, where he held an Alexander von Humboldt Senior Scientist's Prize from the Federal Republic of Germany. During that sabbatical he carried out his new studies of *Ramapithecus* in Europe, Africa and the Near East. Simons has led or has participated in more than 30 fossil-hunting expeditions on three continents, and he has written two books and more than 100 research papers. The most significant achievements of his fieldwork were the discovery of *Aegyptopithecus* and other early apes of the North African Oligocene, and the uncovering of the first specimen of *Gigantopithecus* in northern India.

DAVID ADLER ("Amorphous-Semiconductor Devices") is professor of electrical engineering at the Massachusetts Institute of Technology. After attending the Bronx High School of Science and Rensselaer Polytechnic Institute, he received his Ph.D. from Harvard University in 1964. Before joining the faculty at M.I.T. he spent a year at the Atomic Energy Research Establishment at Harwell in England. He is a member of the executive committee of the Division of Solid State Physics of the American Physical Society and is on the editorial board of *Semiconductors and Insulators*. He has also served as a member of the Basic Research Committee that advises the U.S. Army Research Office in the National Research Council, as chairman of the Solar Photovoltaic Panel of the M.I.T. Solar Energy Workshop and as president of the Harvard Graduate Student Council. In addition to amorphous semiconductors Adler's research interests include solar cells, transitions between insulators and metals, transition-metal compounds and polymeric semiconductors.

LLOYD J. OLD ("Cancer Immunology") is vice-president of the Sloan-Kettering Institute for Cancer Research. Born in California, he is a graduate of the University of California at Berkeley. He received his M.D. at the University of California School of Medicine in 1958 and that year joined the Sloan-Kettering Institute. He was appointed vice-president and associate director of the institute in 1973, and also associate director of research of the Memorial Sloan-Kettering Cancer Center and Memorial Hospital for Cancer and Allied Diseases. In addition he has been professor of biology in the Sloan-Kettering Division of the Graduate School of Medical Sciences at Cornell University. Old is one of the founders of the discipline of tumor immunology. He was among the first to offer evidence for the ability of microbial derivatives to increase immunological responsiveness in a broad manner and thereby heighten resistance against cancer cells.

JOHN A. EDDY ("The Case of the Missing Sunspots") is a solar astronomer at the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory. He was graduated from the U.S. Naval Academy in 1953 and received his Ph.D. from the University of Colorado in 1962. The next year he was a research fellow in radio astronomy at the National Bureau of Standards. In 1963 he joined the research staff of the High Altitude Observatory in Colorado; concurrently he lectured on astronomy and the history of astronomy at the University of Colorado. Between 1967 and 1970 he was a consultant to the College Science Improvement Program. In addition to his work in solar astronomy Eddy has also done research in infrared astronomy, the history of astronomy, and archaeoastronomy (particularly the astronomical alignment of Indian medicine wheels in the U.S. and Canada).

SIRI VON REIS ALTSCHUL ("Exploring the Herbarium") is honorary curator of ethnobotany at the New York Botanical Garden and honorary research associate in ethnopharmacology

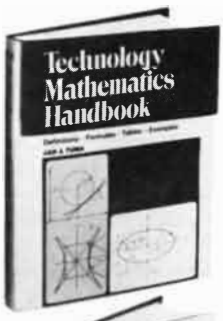
Physics of the earth and planets...finite math ...quantum electronics ...megalithic rings... infrared astronomy...

To keep up with everything that's happening in every scientific area that interests you...

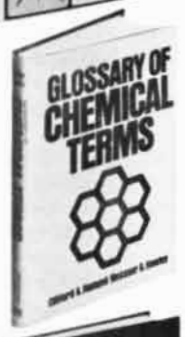
Take any 3 books (values to $67.50) all for only $3.95 as a member of

The Library of Science.

if you will join now for a trial period and agree to accept only 3 more books—at members' discount prices—over the next 12 months.

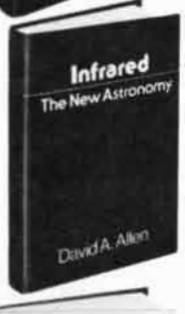


82660. TECHNOLOGY MATHEMATICS HANDBOOK. *Jan J. Tuma.* Instant access to concise summaries of all major definitions, formulas, graphs, table of elementary and intermediate mathematics — from algebra to determinants and vectors and much much more. **$15.95**



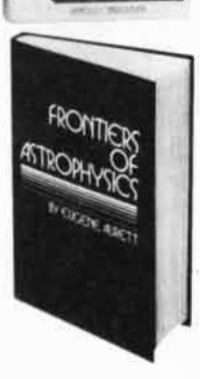
50380. GLOSSARY OF CHEMICAL TERMS. *Clifford A Hampel and Gessner G. Hawley.* Over 2,000 concise definitions: all major chemical groups, important functional terms, basic phenomena and processes, all the chemical elements, and more. **$14.95**

50576. GRAPHS, MODELS AN FINITE MATHEMATICS. *Joseph Malkevitch and Walter Meyer.* A problem-puzzle-solving work that is a delightful introductio to practical finite mathematics. Leads through graph theory, computers, functional models, probability, games and decisions— with examples, tables, illustrations and solutions to selected exercises. **$12.50**



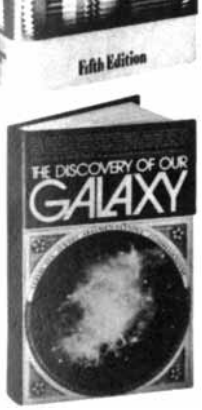
55375. INFRARED: The New Astronomy. *David A. Allen.* The first comprehensive account of the last decade's remarkable breakthroughs in infrared astronomy. **$12.00**

77500. THE SEARCH FOR THE GOLD OF TUTANKHAMEN. *Arnold C. Brackman.* The first complete story of the search, the discovery and the aftermath of one o the great archaeological finds of all time. **$8.95**

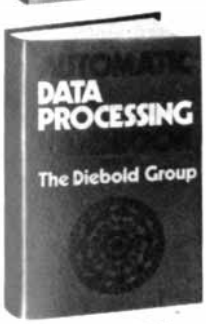


49267. FRONTIERS OF ASTROPHYSICS. *Edited by Eugene Avrett.* A contemporary source on astrophysical topics from recent solar research, to neutron stars, supernovae and intergalactic matter. **$20.00**

00470. VAN NOSTRAND'S SCIENTIFIC ENCYCLOPEDIA. Fifth Edition. Hailed as the most comprehensive one-volume reference work of its kind, this edition has been thoroughly revised and expanded. Illustrated with 2,450 photographs, diagrams, graphs, and over 550 tables. *Counts as 3 of your 3 books.* **$67.50**

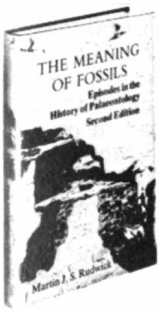


42325. THE DISCOVERY OF OUR GALAXY. *Charles A. Whitney.* Recreates the lives and discoveries of great astronomers — from the sixth century B.C. to black holes, pulsars, supernovae, and Seyfert galaxies. **$10.00**

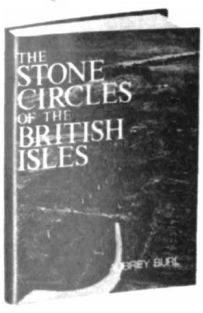


35450. AUTOMATIC DATA PROCESSING HANDBOOK. *Edited by The Diebold Group.* A matchless reference that puts at your fingertips the current issues, concerns and methodology in just about every diverse data processing area, each covered by an expert. *Counts as 2 of your 3 books.* **$29.50**

33540. AN ANCIENT WORLD PRESERVED: Relics and Records of Prehistory in the Andes. *Revised and Updated by Frederic André Engel, Translated by Rachel Kendall Gordon.* Pieces together the complex relationship between men and the Andes and presents us with a vital picture of man at his best, mastering a hostile environment and erecting monuments that testify to his victory. **$12.95**

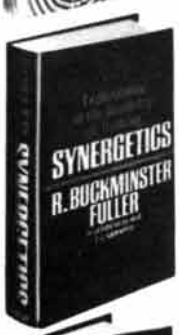


61640. THE MEANING OF FOSSILS. Episodes in the History of Paleontology. Second Edition. *Martin J. S. Rudwick.* Fine, intelligent examination of the cornerstones of paleontology, covering the men, the times and the ideas. **$15.00**



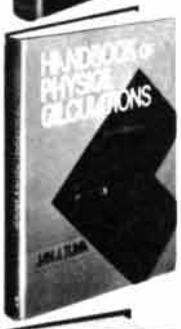
80630. THE STONE CIRCLES OF THE BRITISH ISLES. *Aubrey Burl.* The first systematic, serious investigation of the more than nine hundred megalithic rings that still survive in the British Isles. Contains a catalogue of all stone circles known with accurate data on size, location and possible astronomical significance. **$20.00**

64145. A NEW DICTIONARY OF PHYSICS. *H. J. Gray and Alan Isaacs.* More than 9,000 entries provide a comprehensive, and, where necessary, detailed focus on key concepts, formulas, men and machines that are crucial to understanding modern physics. Updated reference contains hundreds of illustrations, graphs, charts and tables. *Counts as 2 of your 3 books.* **$35.00**



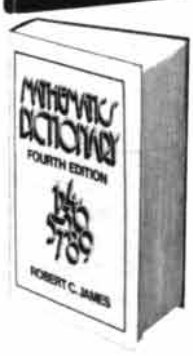
81840. SYNERGETICS. *R. Buckminster Fuller.* Fuller's magnum opus—a mind-stretching work designed to "teach us to understand the universe as a scenario in which we are the actors." *Counts as 2 of your 3 books.* **$25.00**

32300. ADVENTURES OF A MATHEMATICIAN. *S.M. Ulam.* A rare, first-hand account of the birth and childhood of the Atomic age by a man who was instrumental in the development of both the A-and H-bombs. **$14.95**



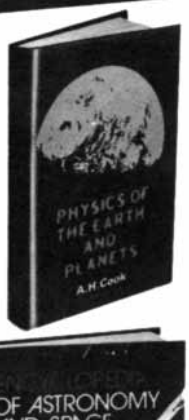
52320. HANDBOOK OF PHYSICAL CALCULATIONS. *Jan J. Tuma.* In one volume, a precise summary of the major definitions, formulas, tables and examples of technology physics. Summary of the major tools of physical calculation, with emphasis on practical applications. **$14.95**

36395. BLACK HOLES, QUASARS AND THE UNIVERSE. *Henry L. Shipman.* Also white dwarfs, neutron stars, supernovae, Seyfert Galaxies—and more. **$12.95**



61530. MATHEMATICS DICTIONARY. Fourth Edition. *Edited by Robert C. James and Edwin F. Beckenbach.* Revised and expanded edition of this leading mathematics dictionary offers clear, accurate definitions of more than 8,000 terms. **$17.95**

34670. ASTRONOMY AN COSMOLOGY. *Sir Fred H* One of the giants of modern astronomy clearly explains to astronomy, and the physical processes and relationships underlying basic astronomica phenomena ranging from the and the sun to the origin of th universe. Over 600 illustratio **$16.50**



68735. PHYSICS OF THE EARTH AND PLANETS. *H. Cook.* Brilliant investigati one of the world's leading au thorities on physical meteoro *Counts as 2 of your 3 books.* **$**

55000. THE ILLUSTRATE ENCYCLOPEDIA OF ASTRONOMY AND SPAC *Ridpath.* Covers the full rang astronomical knowledge and exploration, from ancient tim now. **$16.95**

MEMBERSHIP BENEFITS • In addition to getting three books for $3.95 when you join, you keep saving substantially on the books you buy. • If you continue membership past the trial period, you will be eligible for our Bonus Book Plan, with savings of 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. • 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 Science 2-3AK
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 Selections or Alternates during the first year I am a member, under the club 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. **Send no money. Members are billed when books arrive.**

3 books for $3.95
Indicate by number the 3 books you want.

A few expensive books (noted in book descriptions) count as 2 choices.

Name _____

Address _____

City _____ State _____ Zip _____

(Offer good in Continental U.S. and Canada only. Prices slightly higher in Canada). Book selections purchased for professional purposes may be a tax deductible expense.

at the Botanical Museum of Harvard University. After earning her A.B. in botany at the University of Michigan in 1953, she went to Harvard, where she received her M.A. in biology in 1957 and her Ph.D. in 1961. Between 1962 and 1972 she was a research fellow at the Harvard Botanical Museum.

RICHARD LORE and KEVIN FLANNELLY ("Rat Societies") are respectively director of the graduate program in developmental psychology at Douglass College of Rutgers University and research scientist at the Johnstone Training and Research Center at Bordentown, N.J., a state-supported institution devoted to basic research on mental retardation. Lore received his A.B. degree in psychology at the University of North Carolina in 1959 and his Ph.D. from the University of Tennessee in 1965. Except for a year as associate professor at Virginia Commonwealth University, he has been at Rutgers since 1965. Flannelly received his A.B. in psychology at Jersey City State College and his M.S. in psychobiology at Rutgers in 1975, where he studied under Lore. He has been at Johnstone since 1975.

BRADLEY EFRON and CARL MORRIS ("Stein's Paradox in Statistics") are respectively chairman of the department of statistics at Stanford University and senior statistician at the Rand Corporation in Santa Monica. Both received their undergraduate training at the California Institute of Technology and went on to do graduate work in the department of statistics at Stanford. Efron received his Ph.D. in 1964; Morris received his in 1966, studying under Charles Stein (the discoverer of Stein's paradox). Efron is also professor of preventive medicine at the Stanford School of Medicine, where he works on the analysis of medical data. In addition he was for three years theory-and-methods editor of the *Journal of the American Statistical Association*. He writes: "My current research interests include the use of differential geometry to describe the structure of problems in statistical inference." Morris has been at the Rand Corporation since 1967. There he is working on the statistical design of the Health Insurance Study, which is sponsored by the Department of Health, Education, and Welfare, in order to determine the effects of alternative insurance plans on the demand for medical services and on the status of general health. Before he joined Rand he taught statistics at the University of California at Santa Cruz, Harvard University, Stanford and the Rand Graduate Institute. He is an avid tennis player, and because of his interest in both statistics and sports he has written several articles on the statistics of tennis and other sports.

Dick Tracy never had it so good . . .

and your family never had so much fun!

LCD Chronograph CH1B

LCD Calculator/Watch WF1W

NEW CHANNEL F HAS ALL THE FUN !

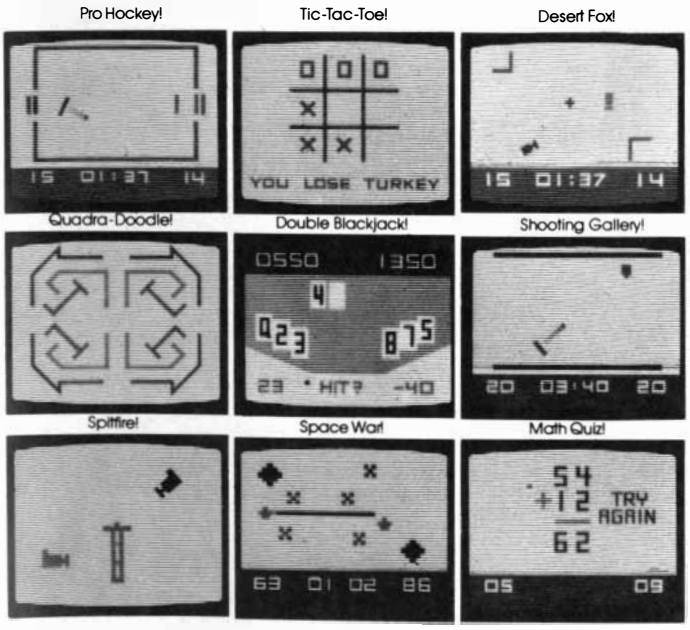
Fairchild's computer-based home entertainment system that plays a limitless variety of games in full color. Attaches easily to any TV size. (plays on channel 3 or 4 through VHF antenna jacks).

GAME CONSOLE: Fairchild's F8 microcomputer-based console comes with built-in 4 dimensional hockey & tennis. Supercharged remote hand-controllers move in 8 directions.

VIDEOCARTS™: Plug-in ROM memory cartridges that never wear out, provide capability for very complex games, educational quizzes, etc. 10 Carts announced with more to come ! Expand your Channel F with the following VideoCarts:

Cart 1: Tic-Tac-Toe, Shooting Gallery, Doodling. 2: Tank 3: Blackjack 4: Spitfire 5: Space War 6 & 7: Math Quizzes 8: NIM 9: Roadracing-(June) 10: Maze (July)

Fairchild Channel F. $169.95
VideoCarts. . . . . each 19.95



Lectro-Media Ltd. introduces Two New Quartz LCDs by National Semiconductor

World's first CHRONOGRAPH with timing to 100th of a second in a liquid crystal digital display (LCD). Displays hrs., mins. and counting secs., day of week plus date on command. Stopwatch times single events or relays, total elapsed time and splits to 6-digit accuracy. Also times events with time- outs. *Other features:* European timing/ Swiss Army-tested black stainless steel case/ Night viewing light/ 1 year battery life/ Scratch-resistant mineral crystal face. $150

SCIENTIFIC CALCULATOR/ WATCH

The first timepiece offering full scientific calculation power with an LCD 6-function watch. Calculator has 18 double-function keys hidden behind an attractive hinged cover. *Features:* CMOS/LSI circuitry for min. power consumption/ FRP nylon scratch-resistant keyboard activated by stylus or pen/ scientific notation with full 8-digit mantissa shows on flip-over display/ Night-viewing light/ Date calculations. $300

LECTRO-MEDIA LTD
P.O. Box 1770 SA
Philadelphia, Pa. 19105
215-925-4477
Showrooms New Market 2nd & Pine Sts.
America's Oldest Market

Yes, send me:

_____Fairchild's Channel F @ $169.95
_____VideoCarts No.___ ___ ___ ___ ___ ___ @ $19.95
_____Nat'l Semi/Novus Chrono CH1B @ $150
_____Nat'l Semi Calc/Watch WF1 (Gold or Stainless)
in _____European or _____Conventional Time @ $300

Charge to: ____MC ____BAC ____AMEX ____DINERS
Acct. No._____Exp.:_____
Signature:_____Bank No.:_____
Enclosed is my_____Cashiers Check/Money Order
Name:_____
Street Address:_____(no p.o. boxes)
City:_____St.:_____Zip:_____
Shipping: $5. -Channel F and Carts. Carts only: $2.50
Watches: $2.50 Pa. Res.: Add 6% tax No CODs

Lectro-Media Ltd. will match any advertised price.
CREDIT CARD ORDERS Only: Call Toll-Free:
800-523-2906
SA5/77

Energy... Canned in America.

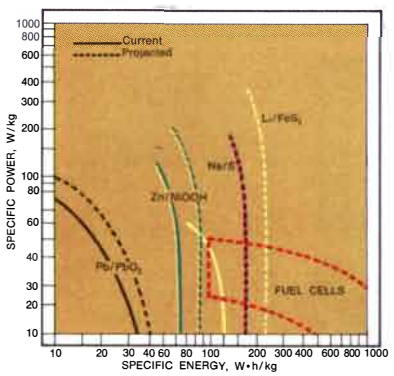
A practical electric car could go a long way toward reducing America's dependence upon foreign oil. To satisfy a large fraction of U.S. urban driving needs, such a car should double or triple the range of present electric vehicles. Urban driving ranges of 30-50 kilometers are now offered by improved lead-acid batteries with specific energies of about 35 watt-hours per kilogram.

The Research Laboratories at General Motors recently has moved more energetic batteries a step closer to reality. Our Electrochemistry Department successfully developed and tested a compact lithium/iron sulfide cell that achieved a maximum specific energy of 115 W·h/kg. This cell is only one-sixth the size and weight of a lead-acid unit with the same capacity.

During a continuous test of more than 14 months, the cell underwent more than 700 deep discharge and recharge cycles. As far as we know, that's some kind of world's record for a high-temperature battery. (It operates at 425-500° Celsius—about the same as a self-cleaning oven.)

Future high-temperature batteries could increase the range of battery-powered vehicles to 300 km. Electric utilities could use such batteries for storing energy generated during low-demand hours to be used during peak periods.

The performance of other electrical energy systems we're studying is shown in the graph. Much closer to commercial application are our ambient-temperature batteries of zinc/nickel oxide. These may almost triple the range of small, battery-powered urban cars.



Energy for America's transportation needs —that's where our advanced battery research is driving.

GM

General Motors Research Laboratories
Warren, Michigan 48090

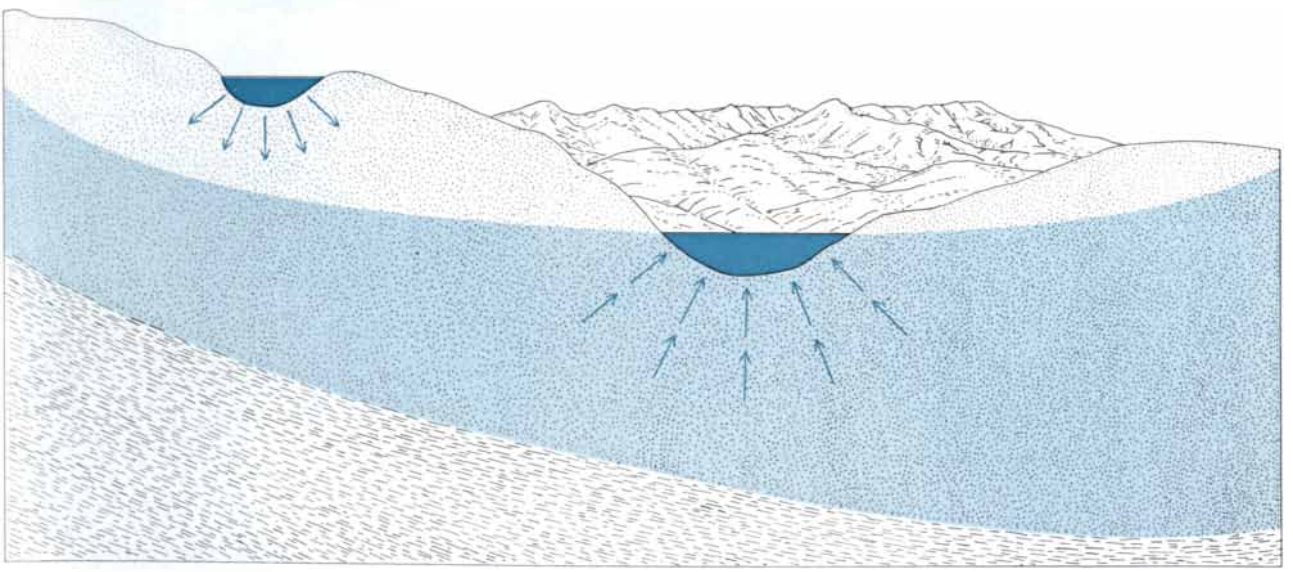
Established 1845 SCIENTIFIC AMERICAN May 1977 Volume 236 Number 5

Underground Reservoirs to Control the Water Cycle

At any one time perhaps two-thirds of the fresh water on the earth is held in underground reservoirs. These reservoirs could be much more intensively drawn on, to be refilled when water is plentiful

by Robert P. Ambroggi

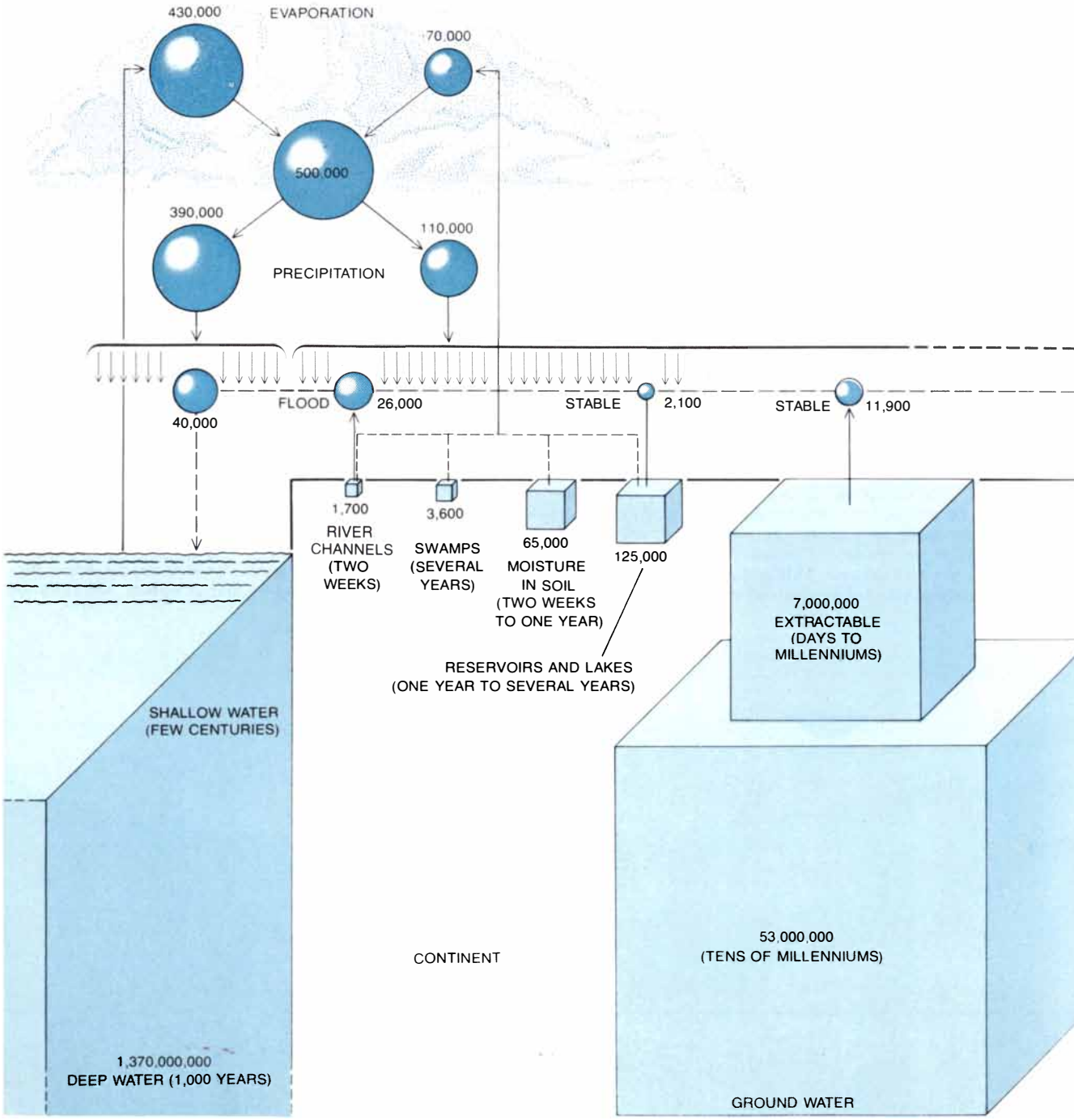
One tends to think of the global water cycle as proceeding in the following way. Water is evaporated from the oceans by solar energy; meteorological processes transport it through the atmosphere, from which it falls to the earth as rain or snow, and eventually it returns to the oceans by way of rivers and streams. This picture overlooks the enormous amounts of water that reside for various periods of time (from weeks to millenniums) in underground reservoirs. Moreover, few of the people involved in planning the distribution of water resources have recognized the opportunity the underground reservoirs provide for dealing with the water shortages that increasingly beset agriculture and other human activities in various parts of the world and for greatly increasing the amount of water that could be put to the service of human endeavor. In this article I contend that mankind should deliberately draw down the underground reservoirs—far more than it already does to some extent—to meet the needs of agriculture, industry and community activities, because sooner or later nature, sometimes abetted by human efforts, will refill the



NATURAL UNDERGROUND RESERVOIR is filled when water percolates into permeable soil or rock formations under a river or stream, as in the example at the left. There it may remain for periods ranging from weeks to millenniums before it finds its way back into the global cycle of evaporation, precipitation and runoff. Ground water also constitutes a significant part of the flow of most streams, as is indicated by the lower river in this illustration. Other natural processes that contribute to underground reservoirs include rain and snow falling on the ground and seepage from lakes and swamps. Reservoirs depleted by intensive use can be refilled artificially by seepage from canals and man-made lakes and by the pumping of water into the ground in periods when the water supply exceeds demand.

reservoirs. The result will be a vastly increased control of the water cycle and an avoidance of the struggles over water that virtually every nation will face within the next few decades if the present wasteful practices in the management of water are continued.

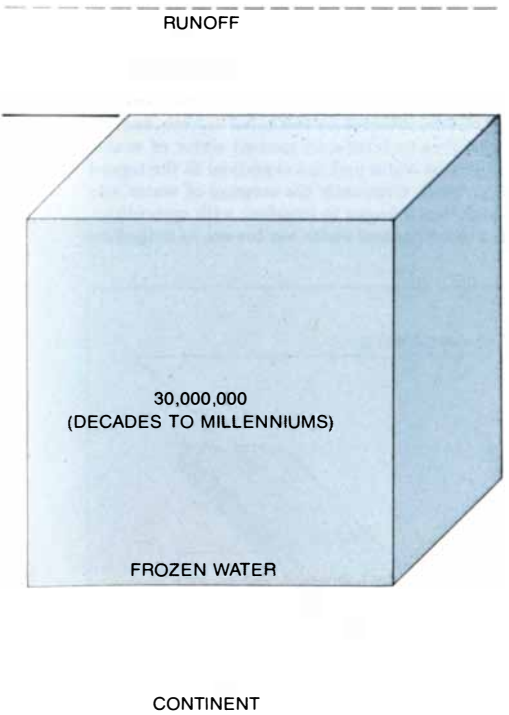
The widespread crop shortages of 1972 provided an indication of the difficulties that lie ahead if water is not managed more efficiently. The shortages were the result of a deficit of some 400 cubic kilometers (400 billion cubic meters) in the continental part of the water cycle. On the global scale of water flow the deficit was small, but it had a heavy impact because most countries use water in such a way that they have little margin for dealing with the consequences of drought. Under present practices the production of food in the world is coming dangerously close to the limit of self-sufficiency. The prevention of food crises will in the future require the



WATER RESOURCES of the world are depicted in a diagram that shows both the annual water cycle, which is represented in the upper part of the diagram, and the water stored in the oceans (*bottom left*) and on the continents (*right*). All numerals represent cubic kilometers of water. The period of time during which water is likely to remain in a particular storage area is also shown. Following the water cycle, one sees that 430,000 cubic kilometers of water is evaporated from the oceans and 70,000 from the continents, whereas precipitation is 390,000 cubic kilometers per year over the oceans and 110,000 over the continents. There is thus a transfer of 40,000 cubic kilometers of

manipulation of water on a far larger scale than nations have practiced up to now. One can envision the need for a great increase in irrigation and for the transfer of water between river basins and even between continents.

People involved in the management of water draw a distinction between renewable water resources and nonrenewable ones. The conventional water cycle I have described carries the renewable resources, which amount to about 500,000 cubic kilometers per year. The nonrenewable resources have a volume of 1.5 billion cubic kilometers. They are stored in three reservoirs: the oceans (97.3 percent of the resources), the continents (2.7 percent) and the atmosphere (.0001 percent).



fresh water per year from oceans to continents, which is balanced in the yearly cycle by runoff. All continental storage areas contribute to runoff, as is shown by numerals next to circles. Runoff from polar ice is omitted.

These distinctions overlook the fact that all the water resources of a continent proceed through a cycle. The only difference between renewable and nonrenewable resources is the speed with which the water is cycled. The resources that are called renewable move fairly rapidly, whereas the nonrenewable resources move quite slowly.

The period of time a particular amount of water spends in its cycle is termed its residence time. Water in the atmospheric reservoir may have a residence time of 10 days. Deep in the oceans the residence time of a body of water may be more than 1,000 years. The residence time of continental water resources ranges from a few weeks to many millenniums, depending on the type of reservoir: a river channel, a lake, a swamp, a glacier, polar ice and ground water. The ground-water reservoirs show the largest diversity of residence time: a few days or weeks in karst (limestone) aquifers, a few weeks or months in gravelly riverbeds, a few months or years in alluvial deposits, a few years or decades in other unconfined aquifers and a few decades, centuries or millenniums in confined aquifers.

Let us look more closely at the movement of water in the conventional water cycle. The annual evaporation from the oceans and the continents is about 500,000 cubic kilometers per year: 430,000 from the oceans and 70,000 from the continents. The other part of the equation is precipitation, which averages 110,000 cubic kilometers per year over the continents and 390,000 cubic kilometers over the oceans. To put it another way, the oceans yield 430,000 cubic kilometers by evaporation and receive 390,000 in precipitation, whereas the continents lose 70,000 by evaporation and receive 110,000 in precipitation. Hence every year 40,000 cubic kilometers of fresh water is transferred from the oceans to the continents. In accordance with the principle of the conservation of water within the water cycle, every year 40,000 cubic kilometers returns to the ocean in the form of surface and underground runoff.

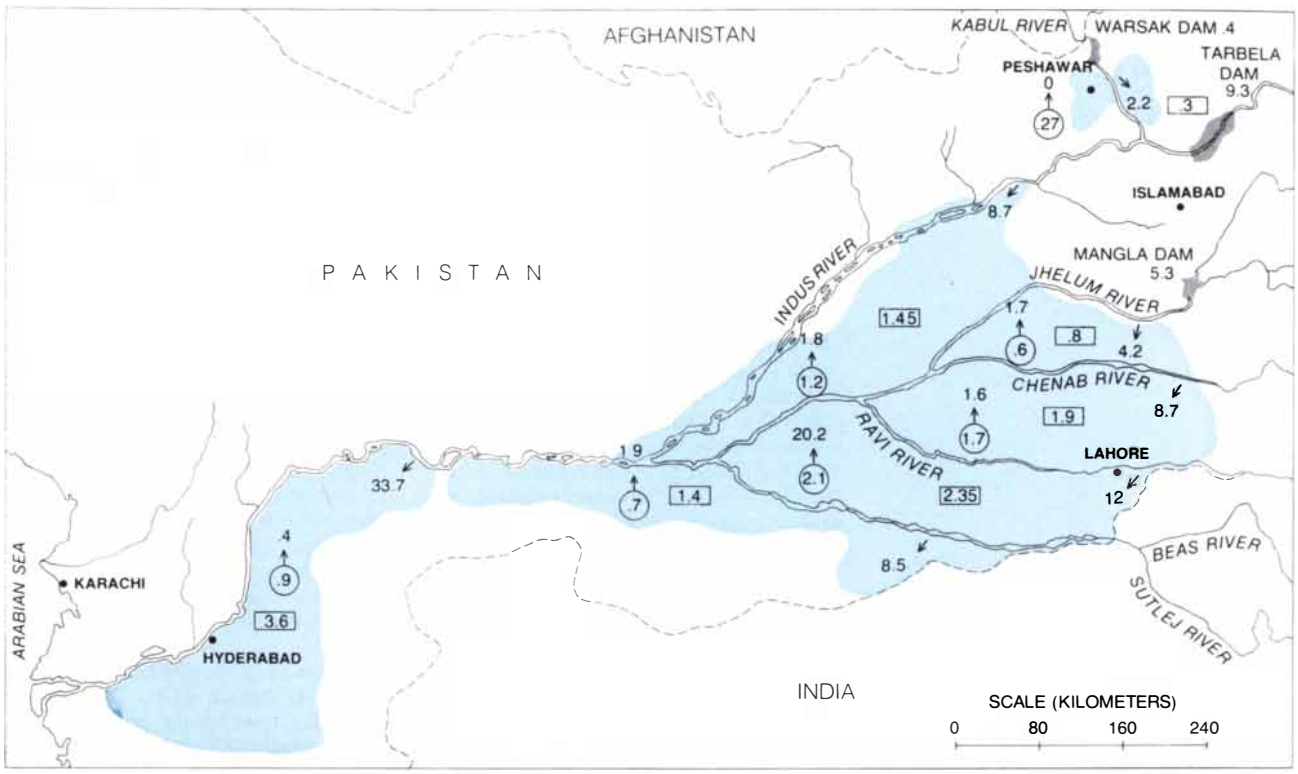
One might therefore suppose 40,000 cubic kilometers of water per year would be available to mankind if water resources were managed with the utmost efficiency. Actually much of that water is inaccessible; it consists of floods, water held in the soil, water running in uninhabited places and so on. The most that mankind might hope to utilize is about 14,000 cubic kilometers, which is the base, or stable, flow in rivers and streams. The base flow is regulated mainly by discharges from underground reservoirs (11,000 cubic kilometers), from man-made reservoirs (1,840) and from lakes (260).

Even the total of 14,000 cubic kilometers is not fully within reach unless transfers of water are made on an enormous scale, since a large part of the total is in uninhabitable areas. As of 1972 the amount of water under human control was approximately 3,000 cubic kilometers per year. Taking into account the fact that about 5,000 cubic kilometers of the stable river runoff is in inhospitable regions, there remains some 6,000 cubic kilometers that under conditions of sound management man could employ for his needs in the future.

The human species normally inhabits only 60 million of the 150 million square kilometers of the continents, being effectively excluded from such places as deserts, polar regions, high mountains and forests. Within the inhabitable zones about 3,200 million hectares of land are potentially usable for agriculture; in 1970 about 1,400 million hectares were under cultivation. (A hectare is 2.47 acres.) It is unreasonable to suppose the amount of additional land that could be brought under cultivation over the next century would be as much as 1,000 million hectares. A better strategy in meeting the rising requirement for food would be to use land more intensively, with irrigation playing a major role. The amount of land under irrigation in 1970 was 180 million hectares. Irrigation consumed some 2,400 cubic kilometers of water per year. Industry consumed 500 and other human activities 200, so that mankind's worldwide consumption of water was 3,100 cubic kilometers.

The food crisis of 1972 arose from a deficit of 35 million metric tons of cereals (wheat, rice and coarse grains). A further consequence was that the wheat stocks of the exporting countries dropped from 50 million metric tons in 1971 to 30 million in 1974. Another year or two of deficit in rainfall would create a global food crisis with dramatic international implications. Thus it appears that the food situation in the world has reached a point where the effects of water deficiency are much severer and more widely felt than they were in the past.

Another disturbing problem for food production is the possibility that the world's weather patterns may be changing. An example is offered by the north-

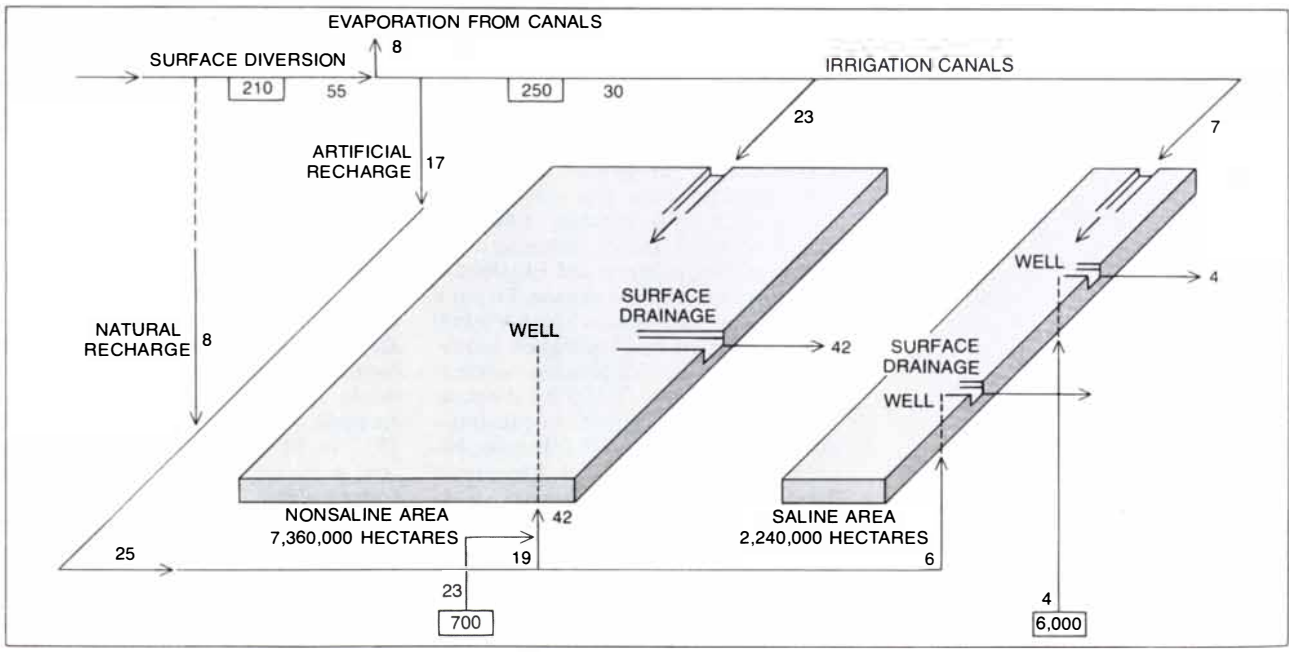


SUPPLY FROM SURFACE WATER (6.3)

SUPPLY FROM GROUND WATER

STORAGE IN RESERVOIR BEHIND DAM

IRRIGATION SYSTEM of the Indus River basin in Pakistan represents an enormous man-made contribution to the reservoir of ground water through a system of some 60,000 kilometers of canals, which are in the regions shown in color. The numerals in rectangles in the canal regions represent in millions of hectares the amount of land affected by the canal system, and the numerals in circles show in millions of hectares the area underlain by ground water of usable quality. All other numerals represent cubic kilometers of water and are explained in the legend at the left. The construction of the system began in 1860. Gradually the seepage of water into the ground from canals raised the water table so high that it began to interfere with agriculture. Now system is managed so as to deliberately pump some ground water out for use in irrigation.



WATER MANAGEMENT in the Indus valley is depicted schematically in an adaptation of a conception of Harold A. Thomas, Jr., and Robert P. Burden of Harvard University. The numerals associated with arrows show the flow of water in cubic kilometers per year. The numerals in rectangles represent (in parts per million) the concentration of total dissolved solids in water and therefore indicate salinity. The entire 9.6 million hectares of irrigated land is divided into a saline area and a nonsaline area because the management of the aquifers differs. The top layer of the ground-water reservoir is pumped in excess of the annual recharge, thereby drawing down the high water table built up during the past century. The effect of these actions is a reduction in problems of waterlogging and salinity that had hampered agriculture. In addition the use of ground water in irrigation has made possible a substantial expansion of area planted to crops.

ern circumpolar vortex, the great cap of high-altitude winds revolving around the pole from west to east. In recent years the lower edge of the northern vortex has been farther south than it was for some years previously. As a result the high-pressure zones have also been farther south, thereby blocking the monsoon rains out of the regions where they are vital to the survival of hundreds of millions of people. The rains therefore have fallen into the oceans or onto land areas that already have more than enough rain.

All these events raise the question of whether man is technologically capable of manipulating the water cycle in such a way as to prevent the calamities that might arise with prolonged shortages of rainfall or changes of climate. Certainly a fair number of technological efforts to deal with such problems have been made or proposed. They include cloud seeding, the long-distance transport of icebergs, the desalination of seawater and the storage of river floodwaters through systems of dams and reservoirs. How effective are these approaches?

The effort to induce precipitation by seeding clouds with silver iodide, frozen carbon dioxide and other substances has not yet been put on a firm scientific basis. It remains a somewhat expensive and haphazard endeavor. Only a few of the many experiments in this field have demonstrated that cloud seeding can increase precipitation in a limited range of favorable conditions. Some experiments have resulted in less precipitation. In sum, weather modification is still in the research stage.

The towing of polar icebergs, mainly from the Antarctic, to arid continental coasts has been proposed. For example, it would be possible to tow a large iceberg (containing six cubic kilometers of fresh water) from Antarctica to the Atacama Desert in Chile in about seven months, with a water loss of 30 percent. The iceberg would be moored at its destination and then melted or quarried for various purposes. Although nothing has actually been done along this line, it would be worthwhile to consider a pilot project for developing countries such as Chile and Peru. The Antarctic runoff amounts to about 2,000 cubic kilometers of icebergs per year, so that the water resource is considerable.

The desalination of seawater has attracted a good deal of attention, particularly because over the past decade the cost has been reduced from about $1.50 per 1,000 gallons to about 50 cents (13 cents per cubic meter). The present trend is to build nuclear plants with the dual purpose of generating electricity and desalinating substantial quantities of seawater. This approach, however, appears feasible only for the domestic water supply of coastal regions in the affluent countries, in view of the capital investment required and the high cost of operating the plants. The present production of fresh water by this means is a mere three cubic kilometers per year.

Up to now the most effective method mankind has employed to manipulate the water cycle has been to build dams and reservoirs that control the flood cycles of rivers. Some 1,840 cubic kilometers of fresh water is stored in this way: 560 in Asia, 490 in North America, 400 in Africa, 200 in Europe, 160 in South America and 30 in Australia. The prospects for a large expansion of the technology appear to be limited, however, particularly in the developing countries, because the cost of regulating one cubic kilometer of water per year by the dam-reservoir method now amounts to about $100 million. Moreover, the residence time of the stored water rarely exceeds a year, so that the system offers inadequate protection against a rainfall deficit that continues for more than a year.

Nevertheless, by building dam-reservoir systems mankind has unwittingly set in motion the long-term control of the water cycle through ground-water reservoirs. For example, the vast irrigation system of the Indus River basin in Pakistan, which was initiated in 1860, gave rise to a dense network of irrigation canals. A large proportion of the surface water diverted from rivers into these canals seeped into the underground reservoirs. The result was that the water table rose steadily for more than a century, creating a huge underground storage. Another example is afforded by the Aswan Dam in Egypt: several cubic kilometers per year is leaking from the reservoir behind the dam into the sandstone aquifer of the Western Desert of Egypt, the largest ground-water reservoir of the Sahara. Ventures of this kind suggest a long-term control of the water cycle through the artificial recharge of underground reservoirs. Is this a sound strategy for the future?

Nature annually cycles about 12,000 cubic kilometers of water through underground reservoirs to form the stable river runoff, or low river flow; this represents 30 percent of the total runoff. Mankind meanwhile is regulating less than 2,000 cubic kilometers by dam-reservoir systems. On the other hand, only 1,200 cubic kilometers of the 3,100 cubic kilometers of water that mankind uses yearly comes from ground-water reservoirs. For more than a century governments have relied on the dam-reservoir approach and have largely ignored the resources available underground. Apparently the reason was that little was known about the condition and behavior of the underground water.

Underground reservoirs have various functions, including the supply, storage, mixing and conveying of water. Until now mankind's main reliance has been on the supply function, that is, on drilling wells that tap the underground supplies. Only in California and more recently in Israel has much attention been given to the storage function. California diverts a good deal of water from the north, where it is abundant, to the south, where it is scarce. Ground water has been extensively depleted by pumping in the south, and much of the water transferred from the north serves to replenish those ground-water reservoirs. The depleted reservoirs therefore provide a mechanism of storage.

Similarly, in Israel the National Water Carrier system transfers 300 million cubic meters of water per year from the north to the south. Some 200 million cubic meters of it is stored in winter in the two main ground-water reservoirs of sandstone and limestone. The objectives are to cover the peak demand for water in the summer and the higher demand in years of low rainfall and also to improve the quality of the water through mixing.

Obviously a ground-water reservoir that is full cannot be employed for additional long-term storage. This was the condition of almost all the world's ground-water reservoirs until the 19th century, when the technology of drilling and pumping became widespread. Nevertheless, even now only a few ground-water reservoirs are depleted enough to provide space for trapping water from the water cycle. Part of the problem is that hydrologists are generally conservative and insist that the exploitation of ground-water reservoirs be kept below the "safe yield," which is defined as the amount of yield that can be expected to be replaced by natural recharge during the year.

Any reservoir managed in this way must be considered a saturated reservoir. It has no room for deliberate recharging for long-term storage, which is the best way to control the water cycle. The first condition for long-term storage therefore is to deliberately increase the exploitation of an aquifer beyond the safe yield. If it is possible, the exploitation should also carry the level of the reservoir below its natural outlets. In short, the policy should be to overexploit the reservoir.

Such a policy implies attention to the problem of recharging the reservoir, which can be done either naturally or artificially. The natural recharge of a ground-water reservoir results from the downward percolation of water from streams and also of water that falls as rain or snow on permeable soils. Experience over the past three decades has shown that a single year of unusually heavy rainfall, which can be expected usually at least once in 15 years, will replenish an aquifer that has been drawn down from 10 to 20 meters during previous years. For instance, the floods of 1969 in Tunisia replenished all the reservoirs that up to then had been described

as overexploited. Similarly, a shallow aquifer in the Souss Valley of Morocco had supported a flourishing production of citrus fruit since 1946 but by 1957 was a source of great concern because the water table had declined at a rate of a meter per year. The rains of 1957 were well above normal; they recharged the aquifer by about 10 meters, almost restoring the initial situation.

Nature is therefore already operating a long-term storage system with some 10 or 15 years of residence time, which represents the best security against droughts or years of low rainfall. I suspect that further experience with ground-water reservoirs will reveal that in some places it may be possible to deplete a reservoir for as long as 40 years and still have it refilled by natural processes. Indeed, we already know that a much longer storage time (centuries or millenniums) was arrived at by nature in the huge ground-water reservoirs that exist under deserts such as the Sahara. Recent studies have shown that the recoverable water from those reservoirs would make possible the irrigation of several hundred thousand hectares of land for centuries, even if it were assumed that the natural recharge of the reservoirs was insignificant. Mankind's perpetual quest for more water means, however, that close attention should be given to the artificial recharge of underground reservoirs.

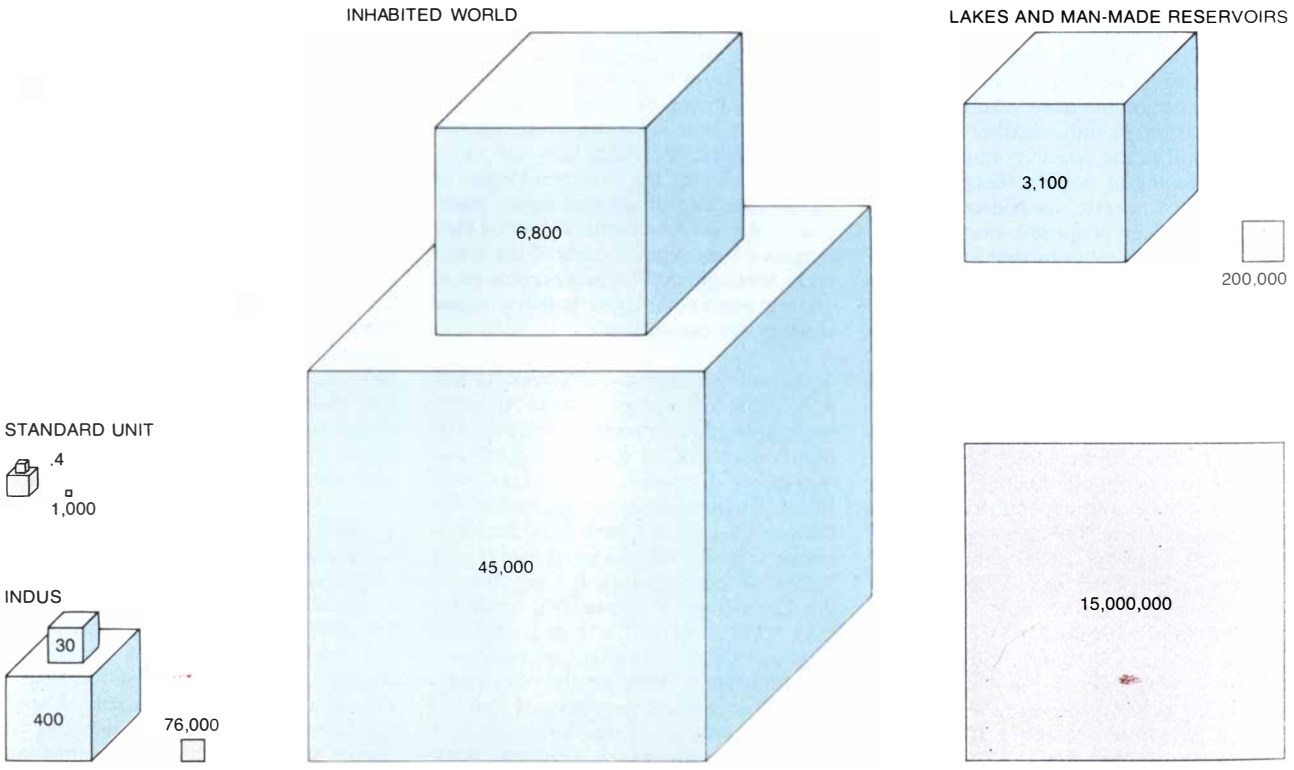
The practice of artificial recharge began in this century. In general it was done on a small scale and with limited objectives: restoring the level of a water table that was declining rapidly or improving the quality of the water. Most of the work was done in California, where some 300 stations recharged a total of about eight cubic kilometers of water between 1900 and 1960. Since 1960 the level of artificial recharge has reached one cubic kilometer per year. A certain amount of artificial recharging has also been done in Europe, mainly to improve water quality.

The later plans of California and Israel that I have mentioned involve artificial recharging on a larger scale and also embrace the concept of long-term storage in ground-water reservoirs. The most striking example of artificial recharge, however, is to be found in the Indus valley, although until about 15 years ago it was an uncontrolled and largely incidental process.

The Indus project is the largest irrigation system in the world, covering more than 10 million hectares. As I have mentioned, it was developed in the Indus River system over the past century. A network of 60,000 kilometers of canals in Pakistan covers the greater part of a vast ground-water reservoir encompassing 16 million hectares and extending 350 meters downward. The reservoir has a bottom layer of relatively saline water (6,000 parts per million, compared with the 35,000 parts per million in typical seawater) and an upper layer with water of higher quality (700 parts per million).

The reservoir is continuously recharged by leakage from the irrigation canals. This artificial recharge was estimated in 1965 at 17 cubic kilometers per year, compared with eight cubic kilometers of natural recharge. As a result of these inputs the water table rose by 30 centimeters per year, bringing about waterlogging in many parts of the Indus valley and causing problems of salinity in the soil.

A group from Harvard University studied this situation as part of a project sponsored by the World Bank. The team proposed making use of the ground water for irrigation and water management. Part of the plan was to mine the ground water for 30 years to lower the water table and alleviate the salinity problem. The water thus pumped must be diluted by canal water in order to keep an acceptable level of salinity near



STORAGE AND RECHARGE OF WATER in the upper 30 meters of ground-water reservoirs are shown for a standard unit, for the Indus region and for the inhabited world. In each diagram the figure by a square represents number of square kilometers of land underlain by ground-water reservoirs; the figure in the upper cube is the amount of annual recharge in cubic kilometers, including the artificial recharge in the Indus, and the figure in the lower cube shows how many additional cubic kilometers of usable and recoverable water are available. For comparison the amount of surface area and the amount of annual recharge in lakes and man-made reservoirs are also shown.

the roots of irrigated plants. In addition the water taken out of the ground-water reservoir must be carried off to the sea through the canals.

The water the Indus system has trapped from the water cycle over the past century and stored in underground reservoirs represents today a volume of about 400 cubic kilometers of usable water stored in the upper 30 meters of the part of the aquifer that is overlain by the irrigated area. The water has been mined by means of a system of tube wells. Most of them were drilled by private enterprise to a depth of about 30 meters; they draw water up at the rate of about 30 liters per second. Tube wells installed by government agencies go to a depth of about 70 meters and discharge at a rate of about 120 liters per second.

The result of these activities is that ground water, heavily supplemented by artificial recharge, is contributing a third of the water supply in the region and has made possible an expansion of four million hectares in the amount of land from which crops can be harvested. Each hectare of irrigated land could also benefit from 40,000 cubic meters of usable water stored underground over the past century. The exploitation of this stored water could allow either an annual extension of the cropped area or a guarantee of full irrigation in times of severe drought or several years of low rainfall. At present prices one cubic kilometer of usable water per year could be produced from the ground-water reservoir for about $20 million, whereas the same amount of water trapped from the water cycle with a dam and reservoir would cost $100 million.

Such a system is the key to the future of agriculture and of water management if food production is to keep pace with the increase in population and the availability of water is to keep pace with the rising demands being put on water resources. The joint operation of surface and underground reservoirs will alleviate both seasonal and long-term deficiencies of water. The success of such a system requires that the underground reservoirs be fully equipped for an exploitation that could continue for any period of time from a few months per year to a few years per decade. To prepare such a system calls for political decisions that up to now have not even been considered in most jurisdictions.

The application of systematic management to aquifers is only in the exploratory stage. Further experience could improve the techniques employed in the management of water resources in general and of ground-water reservoirs in particular.

A case in point is the technology of artificially recharging ground-water reservoirs on a large scale. If that technology were improved, the ground-water reservoirs would serve for both seasonal and prolonged storage. Moreover, the total supply of manageable water would be increased because the system would save water that otherwise would evaporate or escape unused.

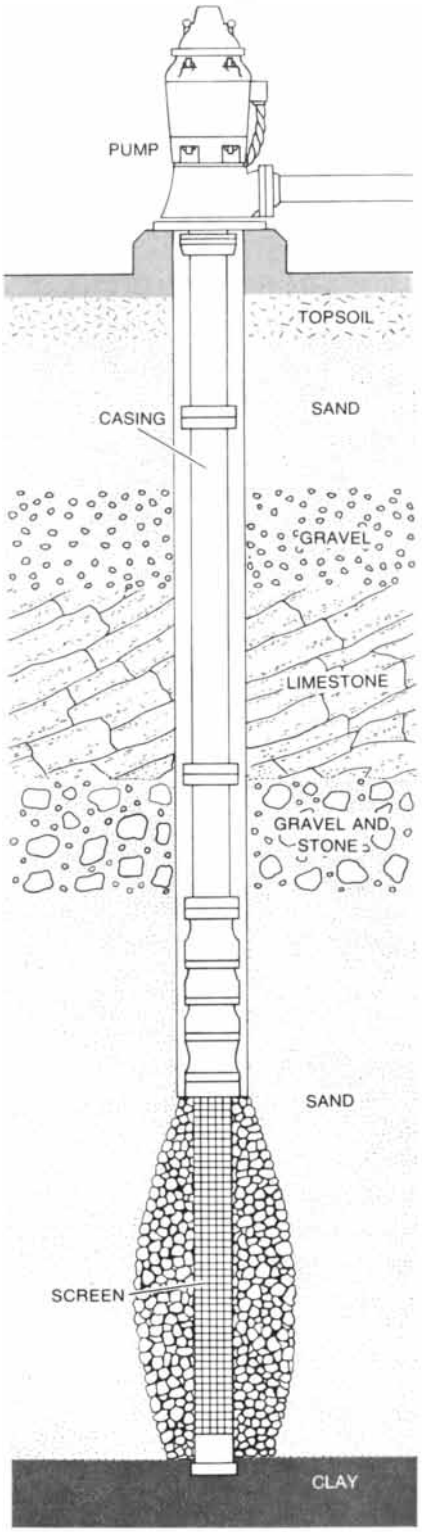
Seepage from irrigation canals is one of the best techniques for artificially recharging ground-water reservoirs on a large scale. Accordingly the lining of canals, which is often done to reduce seepage, is in fact a useless expenditure. It would be better to remodel the canals and enlarge them, provided that the work were closely coordinated with the installation of tube wells to control the rising of the water table.

Another means of artificially recharging ground-water reservoirs in Temperate zones is to spread the high waters from winter thaws over bare irrigable land, that is, to irrigate the land even though no crops are growing. The technique would provide both a temporary storage in the soil that would be useful to the crops when they were planted and a long-term storage in the ground-water reservoir underlying the irrigated area.

A simple and very old technique for recharging ground-water reservoirs is to build lakes and small dams in the upper basins of rivers, so that some of the water that would normally run off will seep into the ground. More expensive techniques are to build infiltration pits (an excavation to catch storm runoff that otherwise would run uselessly out to sea) and recharging wells, by means of which water is pumped into the ground rather than drawn from it. In the U.S. infiltration pits and recharging wells have been installed extensively on Long Island, in Arkansas, on the northern Great Plains and in certain Western states.

On a global scale the problem facing mankind is not a lack of fresh water but a lack of efficient regimes for using the water that is available. The need to produce more food will continue, and it can be met only by an assured supply of water in agricultural areas. The water deficit of 400 cubic kilometers that caused the food crisis of 1972 is insignificant when one considers that about 45,000 cubic kilometers of water is still in storage in the top 30 meters of ground-water reservoirs.

Mankind can solve the water problem if governments and decision makers are prepared to embark on the proper management of their water resources and particularly of their reserves of ground water. Moreover, it is becoming technologically feasible to deal with such adverse developments as water deficiencies resulting from changes in weather patterns by initiating the large-scale and long-term control of the water cycle through ground-water reservoirs.



CASED WELL, which in the Indus region is called a tube well, is portrayed. Such wells are the means whereby ground water is drawn down in the Indus valley as a source of irrigation and also of management of the ground-water resources. The well is made by drilling to a depth of from 30 to 70 meters and putting into the hole a casing ending in a screen that is some distance into the water table. Screen keeps out the large particles of solid matter.

Ramapithecus

This extinct primate is the earliest hominid, or distinctively manlike, member of man's family tree. The finding of many new specimens of it has clarified its place in human evolution

by Elwyn L. Simons

Fifteen years ago the only known evidence bearing on when and how the distinctively manlike branch of man's family tree first arose was a single fragment of upper jaw that had been found in northern India in 1932. Today numerous other fossils of this earliest-known hominid and genera closely related to it are known from East Africa, Greece, Turkey, Hungary, India and Pakistan. In the early 1960's the count of these hominid fossils, most of which were turning up not in the field but in museum collections of primate fossils around the world, increased from the single India specimen to nearly a score. Since 1972 that number has at least doubled. For primate paleontology it has been a remarkable progression from rags to riches. For the student of human evolution the new finds make it possible to more clearly discern an evolutionary pathway traversing the past 14 million years. That pathway can now be traced with little fear of contradiction from generalized hominoids (the larger branch of man's family tree that includes the apes) to the hominids and from the hominids to the genus *Homo.*

The pathway begins in Miocene times with an Old World population of apes whose existence became known more than a century ago. In 1856, Édouard Lartet, a French lawyer and paleontologist, reported on a primate jaw found in a clay of Miocene age at Saint Gaudens in the French Pyrenees. Lartet named the fossil species *Dryopithecus fontani.* The generic name, a combination of the Greek for "oak" and "ape," reflected Lartet's belief that the primate had lived in the forest. The animal and plant remains found in association with other *Dryopithecus* fossils since Lartet's day strengthen his conjecture. These cosmopolitan apes evidently preferred wooded tropical and subtropical environments, where they lived by browsing on leaves and fruit.

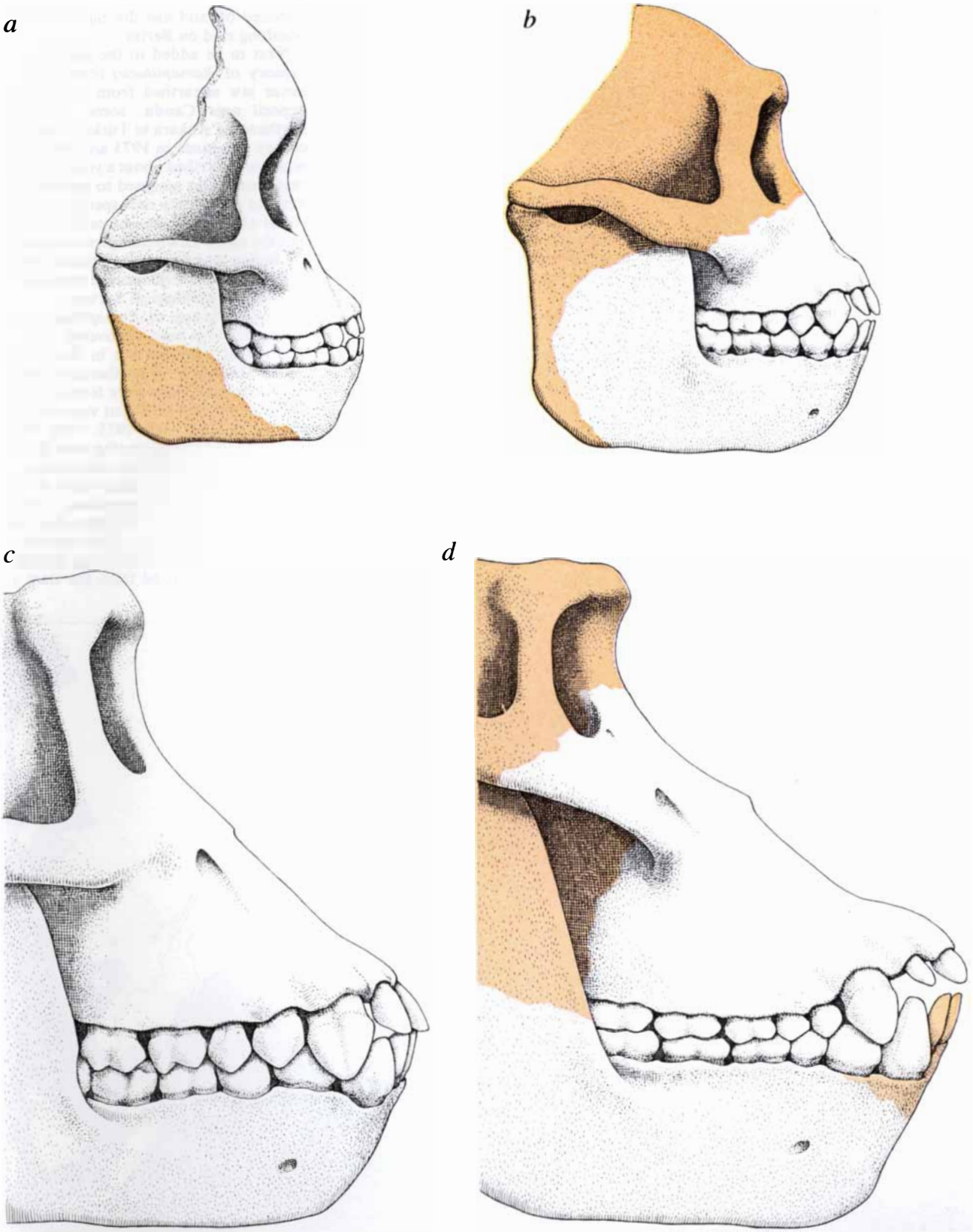
The first *Dryopithecus* fossils from France consisted of three partial lower jaws, one of which had retained all but one of its teeth. No upper jaws were found, and most other fossils of *Dryopithecus* found in Europe consisted only of isolated teeth. As a result almost nothing was known about the oak ape's skull, face or other body parts until the late 1940's. At that time L. S. B. Leakey and his co-workers in East Africa began to find more complete primate specimens in fossil-rich Miocene deposits on islands in Lake Victoria and at sites inland from the northeastern shore of the lake. These ape remains included parts of jaws that ranged in size from those of living gibbons (the smallest of today's apes) to those of living gorillas. In 1948 Leakey's wife Mary found a beautifully preserved fossil skull on Rusinga Island in Lake Victoria, and three years later the Leakeys' colleague Tom Whitworth discovered parts of a second skull on the island, associated with a forelimb, a hand and some other limb bones, including part of a foot. These fossil African apes were assigned to the genus *Proconsul;* the two skulls and the limb bones were assigned in particular to the species *Proconsul africanus.* Studies in recent years lead to the conclusion that *Proconsul* is not a unique genus but an African member of the cosmopolitan genus *Dryopithecus.*

Two more facts about the dryopithecines should be noted before we follow the branching of man's family tree any further. One is that dryopithecine fossils are found both in Miocene deposits and in late deposits of the preceding epoch, the Oligocene. This means that dryopithecine apes flourished over a period of some 20 million years. The other fact is that such fossils have been found not only in France and East Africa but also at sites in other regions extending across a vast area of the Old World: the western desert of Egypt, the region of Barcelona, the valley of the Rhine, the region of Vienna, the mountains of northeastern Hungary, Macedonian Greece, Asia Minor, the Potwar plateau of Pakistan, the Siwalik Hills of India, the coalfields of Yunnan in western China and several localities in south-central China.

In the course of their wide and long-lasting radiation these apes seem to have encountered increasingly cooler environments where tropical and subtropical forests gave way to temperate environments with open woodlands and woodland savanna. Fossils found in Europe and Asia since 1970 suggest that between 10 and 15 million years ago *Dryopithecus* gave rise to at least three other genera. Two of them, *Sivapithecus* and *Gigantopithecus,* were primates with a face as large as that of a modern chimpanzee or gorilla. The third genus, *Ramapithecus,* had a small face.

The exact relations among the three advanced Miocene primates are likely to be clarified only when skulls complete with brain cases and other remains such as limb bones are found. *Sivapithecus* has often been classified as a kind of dryopithecine and therefore an ape rather than a hominid. *Ramapithecus* has most often been identified as a member of man's own hominid line. *Sivapithecus* and *Gigantopithecus* do, however, show some hominid characteristics. For example, all three genera have cheek teeth with thick enamel. Such resemblances are presumably due to the fact that the three genera are related, but they may also reflect similar responses to the same environmental changes. Of the three genera, *Ramapithecus* clearly shows the greatest similarity to later hominids. Nevertheless, the discovery of more fossils of *Sivapithecus* and *Gigantopithecus,* not to mention more of *Ramapithecus,* will undoubtedly alter concepts of how manlike animals branched from apelike ones. In any event, let us now turn to some later branches of man's family tree, beginning with *Ramapithecus.*

Haritalyangar is the name given to a cluster of villages in India some 100 miles north of New Delhi in the Siwalik Hills, an area where exposed Miocene fossil beds extend from northwestern India into adjacent Pakistan. There in



PROFILES OF FOUR PRIMATES contrast the size and structure of their faces. First is the Taung juvenile (*a*), the specimen of *Australopithecus* first described by Raymond Dart in 1925. Second is a reconstruction of *Ramapithecus* (*b*), a hominid hardly larger than the *Australopithecus* juvenile. The profile is based on recently discovered specimens (*color indicates conjectural areas*). Third is the face of a modern female gorilla (*c*) and fourth is a reconstruction of a large contemporary of *Ramapithecus,* the Miocene ape *Sivapithecus* (*d*). Changes in the anatomy of the upper and lower jaws and the teeth of *Ramapithecus,* which evidently relate to diet, have made the hominid less snouty than either fossil or living apes and more like *Australopithecus* juveniles and adults (*see illustration on pages 32 and 33*).

1932 G. Edward Lewis, a young Yale University graduate student who had gone into the area alone with a packhorse, discovered fossils of what he called "manlike apes." He assigned one of the fossils, an upper jaw, to a new genus and species he named *Ramapithecus brevirostris*. The generic name simply means "Rama's ape," Rama being the mythical prince who is the hero of an Indian epic poem. The species name that Lewis chose was more meaningful; it is the Latin for "short-snouted," a feature uncharacteristic of apes. Lewis was impressed by many manlike aspects of the jaw and its array of teeth. Writing his doctoral dissertation in 1937, he placed the new genus in the family Hominidae, the division of the order Primates whose sole living representative is *Homo sapiens*.

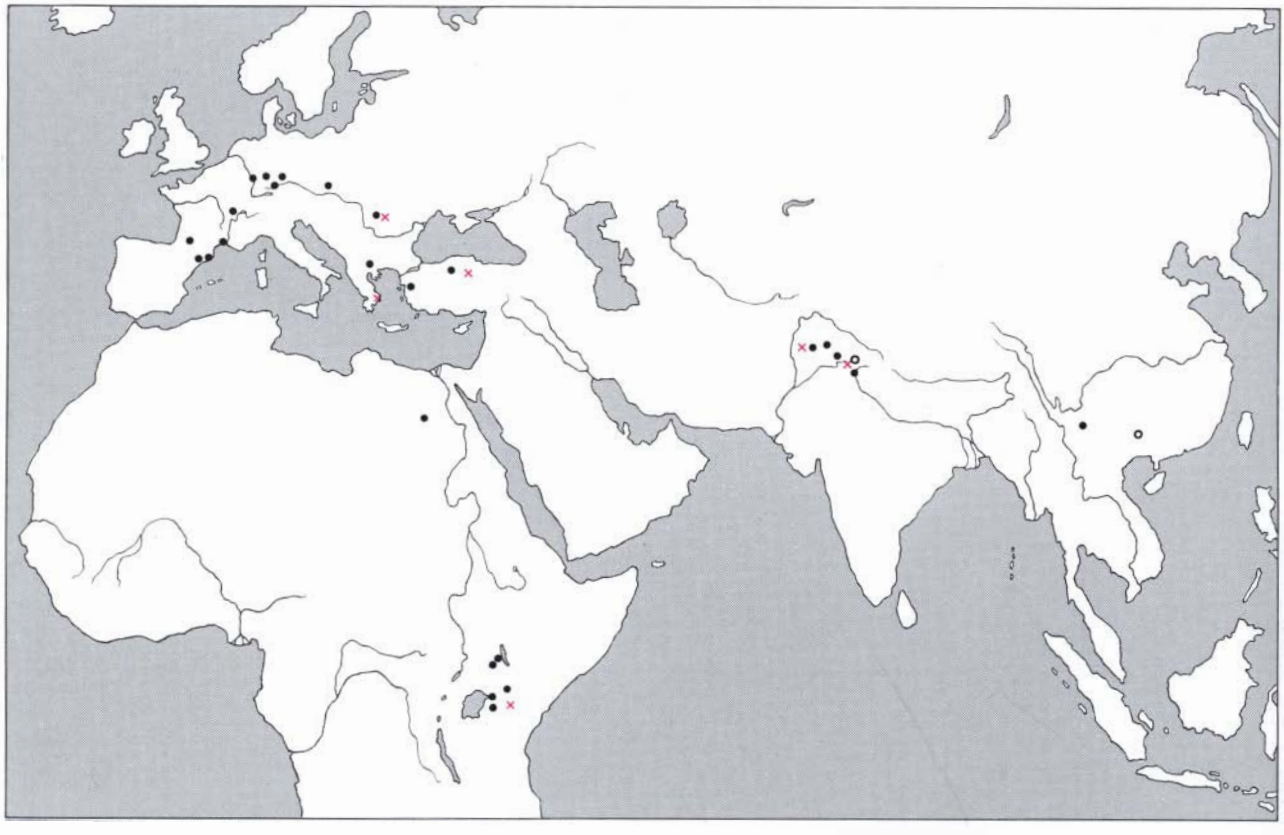
The next *Ramapithecus* fossil find was not made until 1961. Discovered by L. S. B. Leakey near Fort Ternan in southwestern Kenya, it included parts of both sides of an upper jaw. Leakey, with customary panache, treated his discovery as though it were a totally new genus and species, giving it the name *Kenyapithecus wickeri* after his friend Fred Wicker, on whose farm the fossil was found. With hindsight it seems likely that if a lower jaw had been found in direct association with either the Haritalyangar or Fort Ternan upper jaws, or if Leakey had been able to compare his discovery directly with the specimen from Haritalyangar, housed at Yale, there would later have been fewer generic names to deal with.

This, however, is by no means certain. The next *Ramapithecus* specimen to surface after that, which had been excavated in Greece during World War II but was not formally described until 1972, was assigned to another new genus and species: *Graecopithecus freyburgi*. The species name honored Bruno von Freyburg, a German geologist who had pried the fossil out of a Miocene stratum while he was stationed in Athens during the German occupation of Greece. Von Freyburg's find was the complete tooth-bearing part of a lower jaw, and at the time of its discovery it contained all the teeth. It was considered relatively unimportant because another scholar wrongly identified the fossil as being the jaw of *Mesopithecus*, a common Miocene monkey. Before the specimen had even been photographed most of the teeth were knocked off and lost during a wartime bombing raid on Berlin.

Next to be added to the growing inventory of *Ramapithecus* fossils was a lower jaw unearthed from a Miocene deposit near Çandir, some 40 miles northeast of Ankara in Turkey. The discovery was made in 1973 and the specimen was described about a year later. At that time it was assigned to another genus and an entirely new species: *Sivapithecus alpani*. *Sivapithecus*, or Siva's ape, was the name Lewis had given to one of his Indian specimens; the species name of the Çandir jaw honors the director of the Turkish Geological Survey.

A major group of *Ramapithecus*-like fossils has also been uncovered in coal deposits of Miocene age in the Rudabanya Mountains of northeastern Hungary. The specimens were found over a period of several years but were not described in detail until 1975. They have been assigned to still another new genus and species: *Rudapithecus hungaricus*.

Of all the recently discovered or recently described *Ramapithecus* fossils only those found in Pakistan over the past 18 months by my Yale colleague David R. Pilbeam and his co-workers have been recognized from the start as



ADVANCED PRIMATES of Oligocene, Miocene and even later times included an abundant and cosmopolitan group of large apes: the dryopithecines (*black dots*). Most may be assigned to two genera: *Dryopithecus*, found mainly in Africa and Europe, and *Sivapithecus*, found mainly from Asia Minor eastward. These great apes flourished for more than 25 million years. The stock gave rise both to the great ape *Gigantopithecus* (*black circles*) and to hominids such as *Ramapithecus*, a genus now known from five areas in Eurasia and one in East Africa (*colored crosses*). Like *Ramapithecus*, *Gigantopithecus* exhibits anatomical changes in the jaw and teeth related to diet.

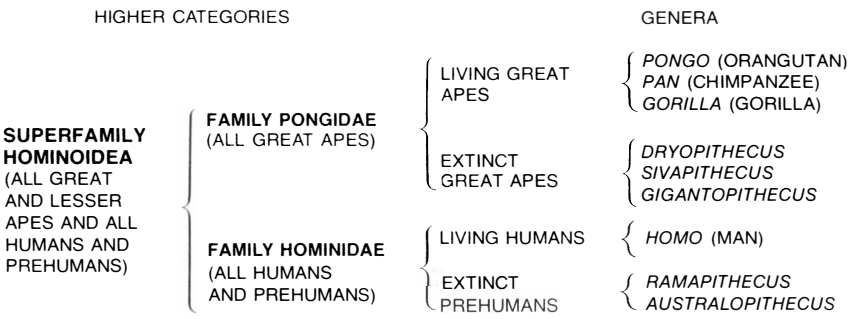
belonging to *Ramapithecus*. Even a fragment of a lower jaw from Fort Ternan, reclassified in 1971 as *Ramapithecus* by Peter Andrews of the British Museum (Natural History), was originally believed to represent *Dryopithecus*.

Not all this profligate naming was due to overemphasis on what taxonomists call "splitting" (in contrast to "lumping"). Ibrahim Tekkaya, the Turkish paleontologist who assigned the Çandir jaw to *Sivapithecus,* had little material for comparison. The principal information at his disposal was a discussion of a *Dryopithecus* jaw that had been mistakenly assigned to *Ramapithecus* in 1938. This mistaken assignment had been persistently repeated in the scientific literature, and Tekkaya thought the lack of similarity between that jaw and his find disqualified the find as a *Ramapithecus* specimen. He and Andrews have since published a joint report correctly reclassifying the find as *Ramapithecus.*

By the same token Miklos Kretzoi, the Hungarian paleontologist who named his new finds *Rudapithecus,* might not have done so if the first lower jaw retrieved in Hungary in 1969 had not been heavily eroded and incomplete. With the later Rudabanya fossils in hand (these include both upper and lower jaws) the resemblance to *Ramapithecus* is unmistakable. Kretzoi now recognizes his fossils as being "ramapithecine" even if they are not generically *Ramapithecus.*

Finally, the eminent paleoanthropologist G. H. R. von Koenigswald, who named the Athens lower jaw *Graecopithecus,* separated it from *Ramapithecus* because relatively complete lower jaws of *Ramapithecus* were then unknown. The first good specimen to be found, the Çandir mandible, was not described until 1974. Therefore even as recently as 1972 von Koenigswald could not have perceived the close similarities that are evident when the Athens specimen is compared with the ones from Fort Ternan and Çandir.

There are still other "rehabilitated" remains of *Ramapithecus.* For example, Pilbeam and I, writing in 1965, followed the received opinion that hominids do not have the "simian shelf" that is characteristic of apes and monkeys. The simian shelf is a torus, or horizontal ridge of bone, that projects inward from the inside of the lower jaw. Its presence in a half lower jaw from near Domeli in Pakistan, we thought at the time, excluded the fossil from consideration as being a specimen of *Ramapithecus.* We later realized that shelflike ridges are present inside the lower jaw of both *Ramapithecus* and the later African hominid *Australopithecus.* In 1969 Pilbeam rescued the Domeli mandible from its undeserved obscurity, enumerated its hominid features and added it to the inventory of *Ramapithecus* fossil remains.



SUPERFAMILY HOMINOIDEA, a major subdivision of the order Primates, consists of three families; two of them, the pongids and the hominids, are shown in the diagram. The family Pongidae includes, in addition to three genera of living great apes, the three fossil genera shown. The family Hominidae includes the living genus *Homo* and two extinct prehuman genera.

At this writing some two dozen of these assorted *Ramapithecus* jaws and teeth have been formally described. What do they tell us about man's most remote hominid ancestor? The evidence consists of a number of anatomical fine points, but they lead to an inevitable conclusion: *Ramapithecus* had adapted to a way of life quite different from that followed by most of its forest-dwelling relatives of the *Dryopithecus* group. Where its jaws and teeth reflect that adaptation they resemble those of the African hominid *Australopithecus.* Between the two, however, is a large gap in time. *Ramapithecus* species are not known to have flourished in Eurasia more recently than eight million years ago. The earliest fossils of *Australopithecus* and *Homo* appear to be those found since 1974 in Ethiopia by Donald C. Johanson of Case Western Reserve University and in Tanzania by Mary Leakey; of these the most ancient are all less than four million years old.

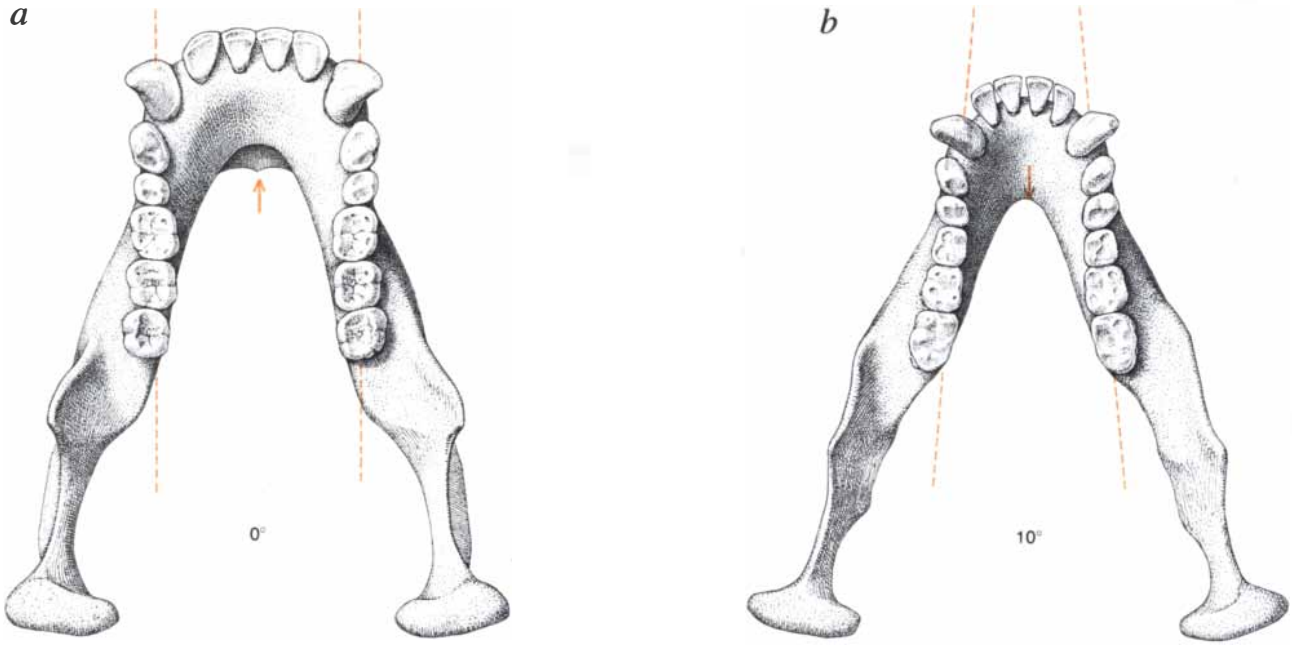
To turn to the anatomical details, in 1967, while I was reviewing the jaw morphology of the best-known fossil apes, I found to my surprise that apes in the Miocene epoch had a dental arcade (the arch formed by the row of teeth seen from above) quite unlike that of typical living apes. Among the living apes the arcade is *U*-shaped because the two sides of the jaw are parallel. In typical Miocene apes the arcade is *V*-shaped because the two sides of the jaw diverge to the rear.

Two years later E. Genet-Varcin, writing in Paris, pointed out that the dental arcade of most *Australopithecus* jaws is also *V*-shaped. (In both groups the point of the *V* is squared off in front.) Thus the same kind of arcade is shared by two groups separated by some millions of years. The dental arcade of *Homo sapiens* is neither *U*-shaped nor *V*-shaped but more semicircular in outline. Therefore both the semicircular arcade of modern man and the *U*-shaped one of modern apes presumably arose from a *V*-shaped original.

Where do the jaws of *Ramapithecus* fit into this range of variations? It turns out that they are intermediate. The rows of teeth diverge about 20 degrees from the parallel. This angle is greater than that in Miocene apes (about 10 degrees) and less than that in *Australopithecus* (about 30 degrees).

Both *Ramapithecus* and *Australopithecus* have a lower jaw that is typically about as thick, from the cheek side to the tongue side, in the molar region as it is deep from top to bottom. This robustness is in contrast to the lower jaw of living apes, which in the molar region is much thinner with respect to its depth. The molar teeth of *Ramapithecus* and *Australopithecus* convey the same impression of robustness. They are large and flat, and viewed from above most of them have rounded outlines. Unlike the molars of chimpanzees and gorillas, and of *Dryopithecus* in the narrow sense (excluding *Sivapithecus*), they have a thick enamel and hence are more resistant to wear. The canine teeth and incisors are smaller than those of living apes. Furthermore, the lower jaw of *Ramapithecus* and *Australopithecus* is buttressed against the stresses of heavy chewing by not one shelf on the inside but two. The shelves, known respectively as the superior and the inferior torus, were still present in the more robust species of *Australopithecus* as recently as a million years ago.

Another unusual feature shared by these two hominids is a marked difference in the amount of wear shown by the first and second molars. In *Dryopithecus* and modern apes when the enamel on the cusps of the first molar is worn away to expose the underlying dentine, the cusps of the second molar usually show nearly as much erosion. In *Ramapithecus* and *Australopithecus* the difference in wear between the first and second molars is evidence that the eruption of these permanent teeth was separated by a considerable length of time. This in turn



FOUR LOWER JAWS show variations in the amount of rearward divergence of the tooth arcades in three fossil primates. At far left (*a*), for comparison, is the mandible of a modern chimpanzee; its typically *U*-shaped dental arcade has parallel tooth rows; thus the degree of divergence is zero. Next (*b*) is a reconstructed *Dryopithecus* mandible; the tooth rows show an angle of divergence (*color*) averaging some 10 degrees. Next (*c*) is a composite reconstruction of a *Ramapithecus* mandible. Its tooth rows, when preserved, show an angle of

may mean that both hominids matured slower than Miocene or modern apes.

This group of related features—thickened and buttressed jaws, flattened molars and reduced canines and incisors—can be compared with the dental system of certain other mammals: hoofed herbivores, herbivorous rodents and even the elephant. These mammals chew mostly by means of a rocking or grinding action, so that the movement of the teeth is partly sideways. Where chewing by chopping and biting mainly exerts vertical forces, grinding includes some nearly horizontal forces. A further adaptation in these mammals is a vertical lengthening of the ascending ramus, or upward extension, at the rear of the lower jaw, which is set at a right angle. This contrasts sharply with the short and backward-sloping ascending ramus of modern man and the chimpanzee. Man and the chimpanzee also lack thick tooth enamel and robust jaws.

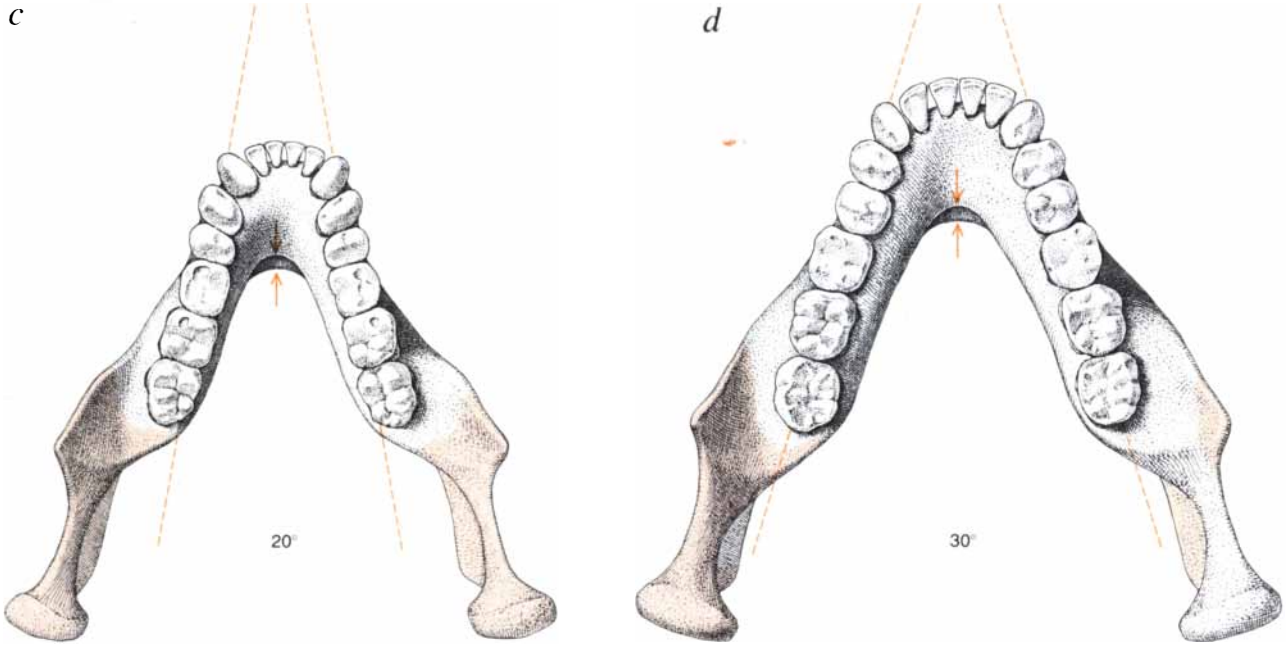
It is therefore not surprising to find that the ascending ramus of the lower jaw in both *Ramapithecus* and *Australopithecus* is typically at more of a right angle to the horizontal than that in modern man and the living apes of Africa. In *Ramapithecus* and *Australopithecus* the ascending ramus also seems to have been proportionately higher. The front of the ramus, the coronoid process, is shifted slightly forward with respect to the cheek teeth. This angularity and shifting is correlated with changes in the upper jaw and lower face. The bone of the upper jaw is deepened and thickened around the roots of the teeth, and the snout is flat and short, with the nasal openings located close to the upper incisors. The cheekbones are set farther forward, and the upper incisors and canine teeth tend to be directed downward rather than forward. Such striking similarities between two genera so widely separated in time must have some meaning.

The first intimations of a shift in a major adaptation of an organism, reflected in a change in the function of an existing anatomical system, provide some of the clearest signposts in the fossil record. The anatomical change often signals a dramatic evolutionary event to come. In primate evolution one of the latest of these adaptive shifts and perhaps the most significant is the one that gave rise to the quantum advance that sets man apart from the other primates. I believe this key shift is the one intimated by the functional change in the dental system of *Ramapithecus.*

When and why did *Ramapithecus* diverge from its cosmopolitan relatives, the dryopithecines, developing a jaw anatomy foreshadowing that of the two later hominid genera, *Australopithecus* and *Homo?* We can begin by looking for places where *Dryopithecus* encountered altered environmental circumstances that would have increased the probability of such a divergence.

Among the dryopithecines the group with the closest ties to *Ramapithecus* seems to be *Sivapithecus,* which is mainly Asian in distribution; the African dryopithecines were considerably earlier. Many of the more recent fossil apes from Eurasia are rather large, with massive, deep jaws and thickened tooth enamel; such specimens can be placed in the genus *Sivapithecus.* Two of these Eurasian fossil apes are now well known. One of them, *Sivapithecus indicus,* is represented by remains found mainly in northern India, Kashmir, Pakistan and Turkey. The other, which probably belongs to a new and still undescribed genus, is represented by a find from northern Greece that was named *"Dryopithecus" macedoniensis* in 1974. Its discovery was followed by the unearthing at the same location of several more partial jaws of the same species; all resemble *Sivapithecus* more than they do *Dryopithecus.* Some of the individuals among these far-flung populations of *Sivapithecus* and its close relatives have a face comparable in size to that of a gorilla; very few have a face as small as that of a chimpanzee.

In late Miocene times, 10 to 12 million years ago, much of Eurasia was covered with forest. The cover was not, however, tropical forest. Thus it did not provide the kind of year-long fruit production and continuous vegetative renewal that is typical of the relatively seasonless forests where apes now reside. Under those circumstances it seems likely that the larger of the Mio-



divergence averaging 20 degrees. Last (*d*) is a reconstructed *Australopithecus* mandible. Its typical angle of tooth-row divergence is 30 degrees. The tooth rows of later hominids show even greater angles of divergence. Colored arrows show differences in the two jaw-ridge buttresses known as the superior and the inferior torus. Modern apes possess a large, shelflike inferior torus; in *Dryopithecus* the superior torus was dominant. Both of the ridges are developed in *Ramapithecus* and *Australopithecus* (*see illustration on next page*).

cene apes would often have found that the food available to them in the trees was inadequate to their needs. A tendency to pursue an alternative food-gathering strategy—foraging on the ground and along the edge of the forest for small, tough foods such as nuts or roots—could have provided natural-selection pressures favoring the survival of individuals with more robust jaws and thickened tooth enamel.

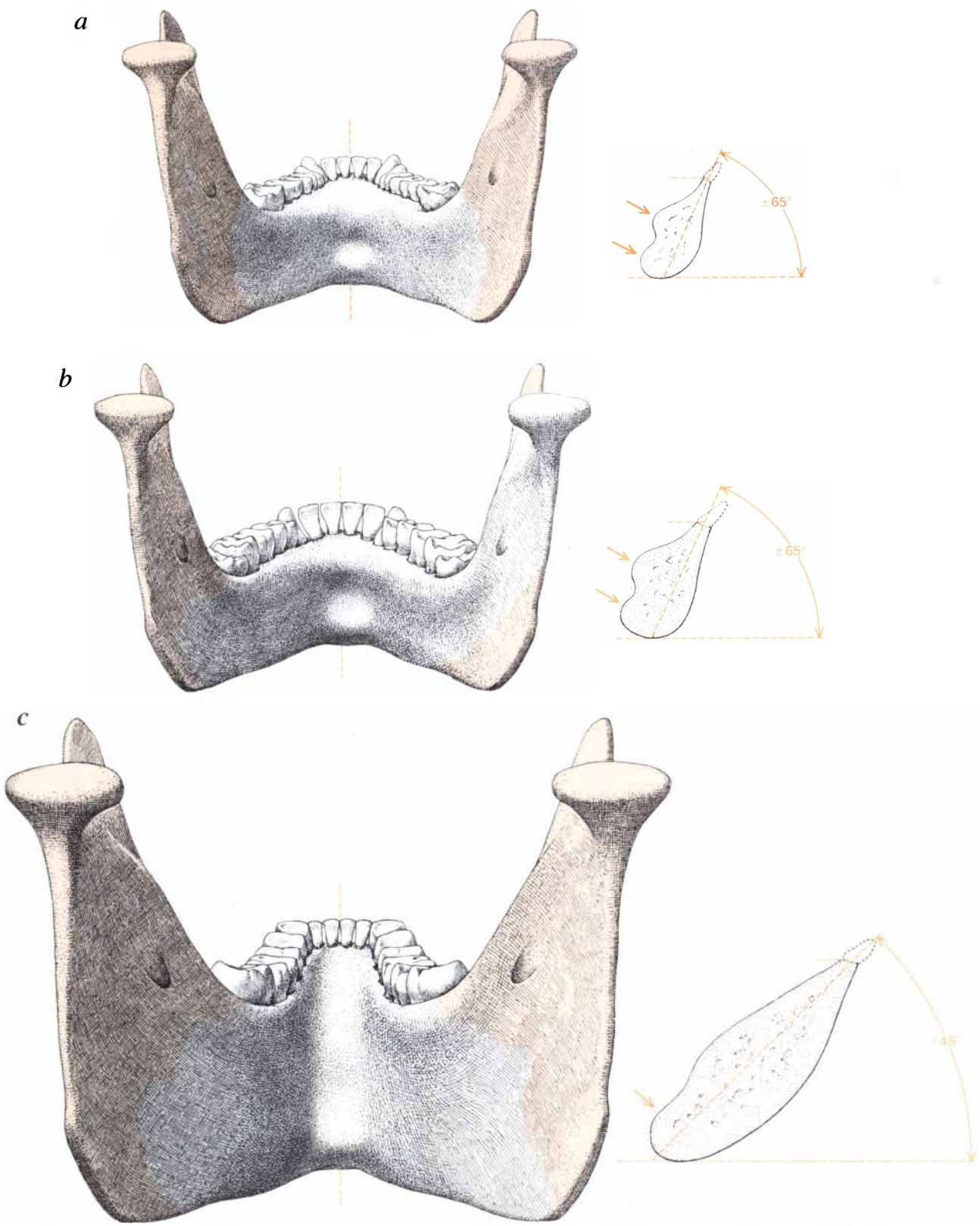
The animal remains associated with *Ramapithecus* at certain sites in Eurasia suggest an even more distinct forest-fringe and wooded-savanna adaptation. For example, the animal remains at Çandir include species that were at home in both forest and savanna environments. Although *Ramapithecus* was present in those habitats, no Miocene apes were. The animals in the Athens Miocene formation are species indicative of a savanna-and-steppe environment; there too *Ramapithecus* was present but apes were not. In Hungary, in Pakistan and in India the fossil plants and animals associated with *Ramapithecus* suggest subtropical to warm-temperate environments, and there *Ramapithecus* and apes lived together. Such habitats could have supported mixed populations of higher primates, including not only *Ramapithecus* but also the larger *Sivapithecus* and *Gigantopithecus,* ranging in size from the size of a chimpanzee to that of a gorilla. These Miocene primates would have been in the midst of the dietary shift that is recorded in their altered dental system. Even as they were still living in forests and on the margins of forests they had given up feeding in the trees in favor of feeding on the ground. Indeed, *Ramapithecus* was already at home away from the treetops in the open woodland and savanna.

It seems unnecessary to postulate two separate adaptations to the same circumstances, one affecting *Ramapithecus* and the other the larger dryopithecines of Eurasia, in order to account for these anatomical developments. Rather, one can speculate that somewhere in southern Eurasia about 15 million years ago a stock of ground-adapted apes that already showed a tendency to forage on small and tough foodstuffs initiated a divergence that led to the appearance of two groups, one manlike and the other still apelike, with somewhat similar dental systems. Although both groups are now extinct, *Ramapithecus,* the hominid, may well have given rise to the later hominid *Australopithecus.*

As we have seen, a gap of at least four million years, a length of time greater than the documented span of the genus *Homo,* separates the youngest-known fossil remains of *Ramapithecus* from the oldest-known representatives of *Australopithecus.* Although fossils of *Australopithecus* have so far been found only in Africa, could *Australopithecus* have been as cosmopolitan as *Ramapithecus?* This is certainly possible, but the possibility will remain a conjecture until some fossil evidence is unearthed in support of it. A comparable situation may nonetheless be worth mentioning. The first fossils of *Homo erectus,* the immediate precursor of modern man, were recognized by Eugène Dubois in Java in the 1890's. Thereafter it was not until the early 1930's, when fossils of the same general type were unearthed in China, that the existence of *Homo erectus* outside Java was demonstrated. Many more years were to pass before fossil forms of *Homo erectus* were recognized in southern, eastern and northern Africa and possibly in Europe as well. Today the genus is seen as having a true cosmopolitan status. Thus it would not be surprising if fossils of *Australopithecus* were found on other continents. What would be surprising, biologically speaking, would be if only that part of the cosmopolitan *Ramapithecus* population living in Africa gave rise to *Australopithecus.*

So far this tracing of pathways has led from generalized hominoids to the hominids by way of an unknown offshoot of *Dryopithecus* that gave rise both to the large ground apes of southern Eurasia and to the hominid *Ramapithecus.* This hominid in turn may well have given rise to the hominid genus *Australopithecus.* What pathway leads to the genus *Homo?*

Here again we find resemblances between *Ramapithecus* and the most primitive recognized species of *Homo,* although the resemblances are not always the same. The name *Homo habilis,*



THREE FOSSIL MANDIBLES are seen from the rear to show the location of the bony jaw buttresses indicated in the cross sections (*right*). The sections show the plane where the left and right halves of the mandibles meet. The buttresses evidently strengthen the mandible against the lateral stresses produced by side-to-side chewing. In *Ramapithecus* (*a*) and *Australopithecus* (*b*) the buttressing is much the same. *Gigantopithecus* (*c*) more resembles the Miocene ape *Sivapithecus* and the modern gorilla: the lower buttress is more developed than the upper. In the two hominids the long axis of the cross section is also at a relatively high angle with respect to the horizontal; in the specimens shown here it is about 65 degrees, but it is often higher. This contrasts sharply with the *Gigantopithecus* cross section; the entire front end of that mandible is lengthened and tilted forward, making a lower angle with the horizontal (45 degrees in this example).

coined in 1964 by Leakey to categorize certain fossils from his famous site at Olduvai Gorge in Tanzania, is usually applied to a group of African fossils that many students of human evolution believe represent the forerunners of *Homo erectus*. They may also include the prehuman fossils recently found by Mary Leakey in Tanzania and some of the fossils uncovered by Johanson in Ethiopia; both groups of specimens are believed to be between three and four million years old. The specimens from Tanzania and Ethiopia resemble somewhat younger fossils of primitive members of the genus *Homo* uncovered in northeastern Kenya by the Leakeys' son Richard E. Leakey, including the well-known cranium and face designated E.R. (for East Rudolf) 1470.

Homo habilis is a name that needs to be overhauled. Few workers agree on what specimens belong in the category. Nevertheless, it is clear that many of these early members of the genus *Homo* in Africa had canine and incisor teeth somewhat larger than those typical of the specimens of *Australopithecus* that are contemporary with them. There are other apparently correlated differences between the early members of the genus *Homo* and *Australopithecus:* the brain of the *Homo* specimens is somewhat larger, and so is their body. The emergence of *Homo* from *Australopithecus,* however, is a complex subject that will not be taken up here. What remains to be dealt with are the resemblances between *Ramapithecus* and primitive forms of *Homo,* such as *Homo habilis* and *Homo erectus,* that are not seen in typical specimens of *Australopithecus.*

FOSSIL SITE IN INDIA, near the area where G. Edward Lewis first discovered *Ramapithecus* in 1932, was visited by a group of paleontologists from Yale University in 1969. The site, a part of the fossil-rich Miocene formation in the Siwalik Hills, contained *Ramapithecus* teeth.

Here again the evidence mainly has to do with jaws and teeth. That evidence was first pointed out by von Koenigswald in relation to a *Homo erectus* specimen found by his collectors at the Sangiran site in Java in 1939. The specimen is usually designated either Sangiran II or Pithecanthropus IV. As in the earliest African species of *Homo* from Ethiopia, Kenya and Tanzania (and in L. S. B. Leakey's first specimen of *Homo habilis*), the upper canine teeth of Sangiran II are comparatively large. Moreover, whereas many *Australopithecus* premolars are "molarized," that is, enlarged almost to the size of molars, the premolars of Sangiran II are not. The upper canine tooth of Sangiran II overlapped the teeth of the lower jaw to such an extent that it caused facets of wear to appear on the opposing canine and first premolar in the lower jaw. There is also a diastema, or small gap, in front of the upper canine, separating it from the next tooth forward, the lateral incisor.

Such gaps between the teeth have traditionally been viewed as an apelike characteristic. For many years the Sangiran II diastema was often explained away as an anomaly. One of the best of the new finds from Ethiopia, a complete palate with upper teeth, shows that these dental relations are not anomalous: the fit of the upper teeth closely resembles the fit of the same teeth in *Ramapithecus.* In contrast, the canines of a typical *Australopithecus* specimen are so small that they do not overlap and cause wear, and there is almost no diastema in the upper jaw. It should also be pointed out that where the upper-jaw diastema is found in *Homo erectus* and *Homo habilis* (and in *Ramapithecus*) the teeth of the lower jaw are closely packed together. This close packing is typical of *Australopithecus* but, as we have seen, not of apes.

A further observation can be made. The earliest fossils of *Homo* from Africa such as E.R. 1470, if indeed they should be designated *Homo,* show a facial and cranial modeling that is quite reminiscent of *Australopithecus.* If it were not for such emergent human characteristics as an enlarged brain and teeth resembling those of *Homo erectus,* all these fossils might well be classified as another species of *Australopithecus* that could be called *Australopithecus habilis.*

The simplest way to resolve these complex relations is to postulate that the hominid stock ancestral both to primitive *Homo* and to *Australopithecus* resembled *Ramapithecus* more closely than later representatives of *Australopithecus* did. The resemblance centers on the possession of large canine teeth. The only evolutionary room available in the fossil record for such a postulated ancestral form is the period between the last appearance of *Ramapithecus* and the first appearance of *Homo* and *Australopithecus.* This is the period between four and eight million years ago, or exactly where there is now a large gap in the fossil record.

There is an alternative to this simple resolution of the matter, and it is one that L. S. B. Leakey favored. It postulates that *Homo* and *Australopithecus* branched independently and directly from *Ramapithecus.* There is a third and less likely possibility that does not fit with the interpretation of *Ramapithecus* presented here. It is that not only *Ramapithecus* but also other hominids emerged independently from the parental stock of Miocene apes. If future discoveries should prove this to be the case, and significant gaps in present knowledge prevent its being entirely ruled out, then current conceptions of the origin of the hominids will have to be drastically revised. Even if that should happen, however, *Ramapithecus* would almost certainly retain its position as a very early hominid.

Amorphous-Semiconductor Devices

Glassy materials that act as electronic switches rival crystalline materials for certain tasks. Their electrical behavior is now better understood, opening the way to a wide range of new applications

by David Adler

It is now more than 15 years since the first demonstration of switching and memory devices based on the amorphous semiconductors known as chalcogenide glasses: glasses containing a large percentage of one or more of the "chalcogen" elements (sulfur, selenium and tellurium). The idea of using amorphous materials rather than carefully prepared crystalline materials as semiconductors in electronic devices was appealing, but the development of amorphous electronic devices progressed slowly for lack of adequate theoretical understanding. Over the past few years, however, much of the mystery surrounding the behavior of electrons in amorphous solids has been dispelled. Perhaps the most important concept that has emerged is that the electronic properties of amorphous semiconductors are controlled by well-defined "active centers," or regions occupied by atoms in unusual energy states, just as the electronic properties of crystalline semiconductors are. In the same period there have been major advances in understanding the switching mechanism in chalcogenide glasses and also in the development of commercial devices based on that mechanism. The rapid progress made in a fairly short time promises to rekindle interest in a field sometimes described as the last frontier of solid-state physics.

The major distinguishing characteristic of a crystal is that its constituent atoms or molecules are aligned in ordered three-dimensional arrays that exhibit a long-range periodicity. Because of the mathematical simplifications that result from such periodicity physical phenomena in crystalline solids were well understood shortly after the development of quantum mechanics 50 years ago. This detailed understanding eventually led to the development of solid-state devices such as transistors, microwave oscillators and lasers. In spite of these great successes the careful techniques of preparation and crystal growth necessary to fabricate devices based on crystalline materials have important economic drawbacks. A good example is to be found in the fabrication of solar cells, devices that directly convert solar radiation into electric power. The reliability of solar-cell technology has been proved in scores of space vehicles. The most important material needed for conventional solar cells is silicon, the second most abundant element in the earth's crust. Nevertheless, the large-scale terrestrial use of such solar cells seems unlikely to arrive for some time, in spite of the urgent need for new energy sources. The major impediment is the expense of preparing silicon in high-quality crystalline form. With present technology the cost of generating electricity with solar cells is about $1.50 per kilowatt-hour, nearly 100 times the cost of electricity from fossil fuels.

From the overall perspective of industrial technology highly periodic crystalline solids occupy a tiny niche, although it is an important one. Nonperiodic, or amorphous, solids form a class of materials with a greater diversity of physical properties. Moreover, their preparation usually does not require carefully controlled growth techniques, a vital economic advantage in many applications. Amorphous solids are often easily prepared in the form of glasses by rapidly cooling a material from its liquid phase. An even wider range of compositions can be formed as thin films by allowing their vapor to deposit onto a cold substrate.

The mathematical complexity inherent in dealing with a nonperiodic system inhibited a theoretical understanding of the microscopic properties of amorphous semiconductors, seriously retarding their potential application as electronic devices. A major exception was in the field of electrophotography, where the photoconductive properties of certain selenium-based glasses were exploited by the Xerox Corporation to build a multibillion-dollar industry. Even that remarkable success did not get under way until more than 20 years after Chester F. Carlson described the original concept in 1938. Even today the process is not completely understood. The lack of understanding proved an even more formidable handicap to Stanford R. Ovshinsky, who began his pioneering work with amorphous semiconductors in 1958. His 1968 publication, in which he described his earlier discovery of two types of reversible switching phenomena, stimulated much interest, but the interest waned when adequate explanations of the phenomena proved elusive. I shall describe the new perceptions that have persuaded many workers in the field that the long period of mystification has come to an end.

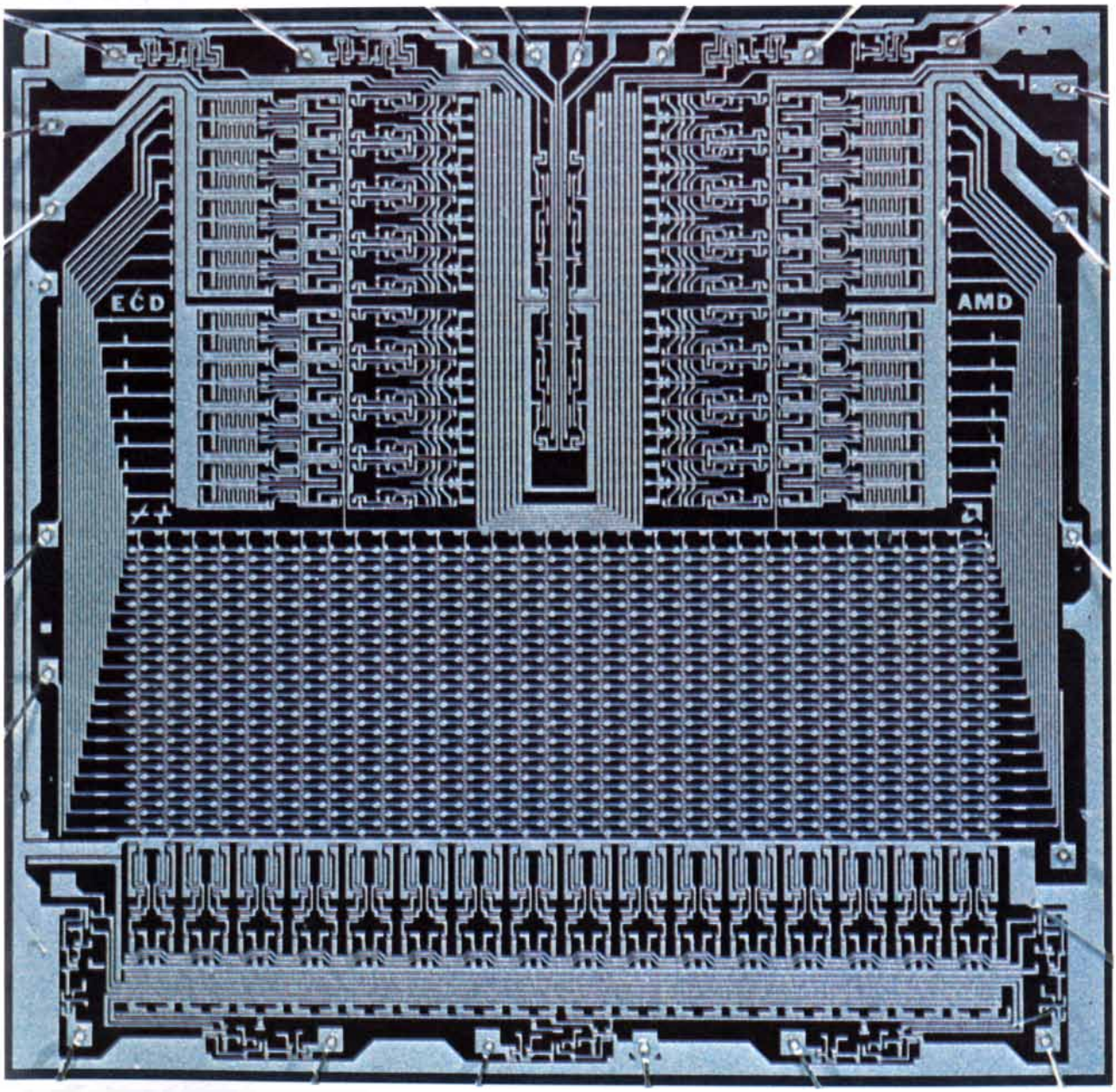
Amorphous solids, like crystalline solids, can be insulators, semiconductors or metals, and in some cases, at very low temperatures, they can even be superconductors. The quantum theory of solids, as it originally developed (and still is presented in textbooks), is based completely on the presence of long-range order. The existence of semiconductors is intimately connected with this theory. It was long believed, therefore, that amorphous solids could not be semiconductors. In 1955, however, B. T. Kolomiets and his co-workers in the U.S.S.R. discovered that chalcogenide glasses can act as semiconductors. In an effort to explain that unexpected behavior the Russian workers A. F. Ioffe and A. R. Regel suggested in 1960 that the electronic properties of chalcogenide glasses originate from short-range order rather than from long-range order, but no detailed understanding followed, even though the problem was tackled by many.

Now, many years later, it is finally clear that the key to understanding the properties of solids is not their periodic structure but the chemical nature of their constituent atoms. The periodicity of crystalline solids greatly simplifies

quantitative calculations, but it is not essential for explaining the major features of their electrical, optical and magnetic behavior. A more productive approach is to try to understand the electronic structure of atoms in different columns of the periodic table of the elements before attempting to analyze the much more difficult problem of solids containing 10^{24} atoms or so.

As an initial step in this direction we can confine ourselves to the elements that are neither rare earths nor transition metals, that is, the elements that appear in columns I through VIII of the ordinary periodic table [*see illustration on page 40*]. A major result of quantum theory is that only a discrete number of states is possible for the electrons in any atom or molecule, and that these states can be grouped into shells, in which the energy of the electrons lies within a small range of values. The diversity of the elements in our universe follows from the exclusion principle, first enunciated by Wolfgang Pauli, which prevents any two electrons in the same atom from occupying the same state. As a consequence the properties of each atom are determined primarily by the number of electrons it possesses. The number of available states in each shell becomes crucial. For the atoms being discussed here the outermost shell (that is, the highest-energy shell) always has exactly eight possible states. The eight states of the outermost shell can be further subdivided into two groups: two lower-energy, or *s,* states and six higher-energy, or *p,* states. Actually the energy separation between the *s* and the *p* states is fairly small. The properties of every atom are primarily determined by the number of electrons in the outermost shell, which accounts for the eight columns of the periodic table.

Each of the eight possible states has a characteristic nature that differs from all the others. It can be categorized by its orbital angular momentum (analogous to the path a planet takes in orbiting the



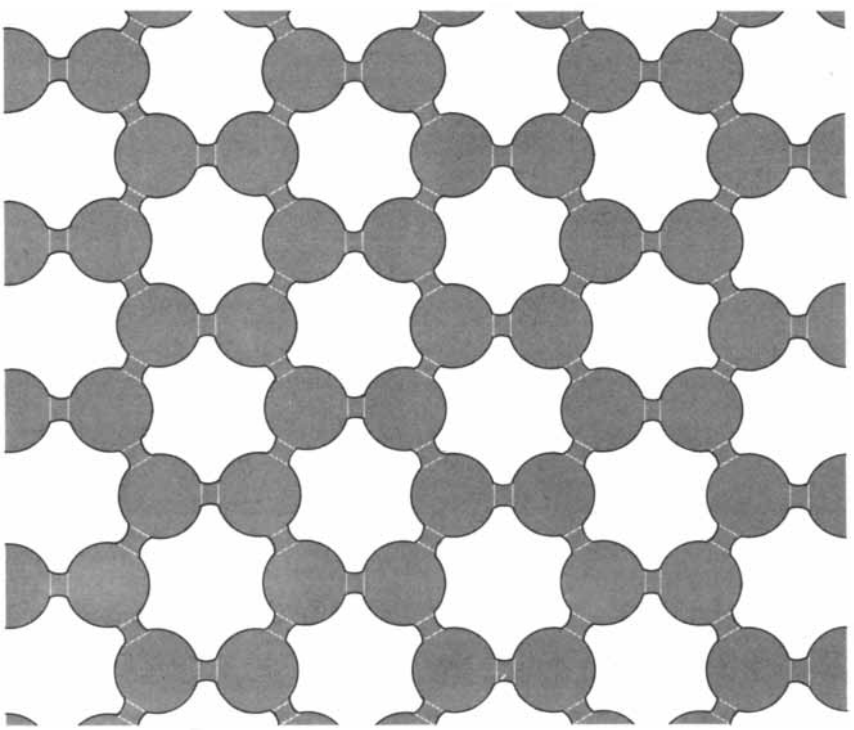
"OVONIC" MEMORY with a storage capacity of 1,024 bits measures 5.6 millimeters on an edge. The amorphous-semiconductor memory elements form the regular pattern at the bottom of the chip. Drivers and decoders employing conventional silicon-crystal technology occupy other half. This device, manufactured by Burroughs Corporation, serves as an alterable memory in computer applications.

sun) and its internal angular momentum, or spin (analogous to the direction of rotation of the planet around its axis). The *s* states have only one possible value of orbital angular momentum; the *p* states have three. All electrons, however, have only two possible values of spin, usually designated spin "up" and spin "down." The magnetic properties of solids are determined by the electron spins, and these in turn follow from the chemical nature of the material.

Combinations of atoms will form stable molecules, provided the energy of the molecular combination is lower than that of the uncombined atoms. The energy can be lowered in two ways. In molecules held together by the kind of bonds known as ionic bonds the reduction in energy arises from the transfer of one electron or more from one atom to another and the electrostatic attraction between the oppositely charged atoms due to the transfer of charge. In molecules held together by covalent bonds the reduction in energy has a more subtle origin: the outer electrons of two adjacent atoms become spread out over both of them. Whenever any two atoms are close enough for their outer electrons to interact, the energies of their states can be shifted by the interaction. Typically some of the states are reduced in energy and an equal number are increased; the former are "bonding" states and the latter are "antibonding." When the bonding states are filled (always by pairs of electrons of opposite spin) and the antibonding states are empty, the molecule is maximally bound. Any states that are essentially unshifted in energy are called nonbonding states. Their occupancy does not significantly affect the stability of the molecule.

The *s* states are nondirectional; the *p* states have a definite directionality. (Ordinarily the three *p* states can be thought of as being mutually perpendicular.) Because of this directionality the bonds between electrons in *p* states are generally stronger than those between electrons in *s* states. A more important result is that a pair of electrons (one with spin up and one with spin down) cannot reduce its energy by covalent bonding with another pair. This would seem to imply, for example, that the elements in Column II of the periodic table cannot form covalent bonds. Actually, because the separation in energy between the *s* and the *p* states is usually of the same order of magnitude as the bonding energy, it is often possible to "promote" one of the paired *s* electrons to a *p* state and then to form two covalent bonds instead of none. This can be done advantageously up through the Column IV elements, in which the configuration sp^3 gives the maximum possible number of *s*-*p* bonds, namely four.

Beginning with the Column V (pnictide) elements, it is no longer energetically favorable to promote one of the *s* electrons, because there are already three unpaired *p* electrons without the promotion and therefore no additional bonds can form. The *s* electrons remain paired (they are often called a lone pair), and only three *p* bonds are possible. The Column VI (chalcogen) elements ordinarily have a *p* lone pair in addition to the *s* lone pair, so that the lowest-energy configuration involves only two *p* bonds. Similarly, the Column VII (halogen) atoms have two *p* pairs, leaving only one *p* electron available for bonding. Finally, the Column VIII elements are essentially inert and form no covalent bonds under ordinary conditions. In addition to all of this the directionality of the *p* states leads to fixed equilibrium bond angles for the elements in columns II through VI.



CRYSTALLINE SOLID is shown schematically in two dimensions. In this hypothetical material each atom forms three identical bonds with its nearest neighbors; thus the three bonds from each atom are spaced 120 degrees apart. Resulting periodic structure, consisting of regular hexagons, is said to show long-range order because environment of every atom is the same.

A solid can be regarded as a gigantic molecule, typically consisting of something like 10^{24} atoms. As in molecules, solids can exist only if the energy of the entire system is lowered by their formation. The origin of this reduction in energy is completely analogous to that in simple molecules. The system is computationally much more difficult to describe, however, since there are on the order of 10^{24} states to deal with instead of only a few. In any event it is clear that the discrete states of the individual atoms are shifted by the interatomic interactions of electrons. Because of the enormous number of states in a small energy range it is possible to treat them as if they were continuously distributed in energy over a given range. Instead of labeling every state, as in an atom or a small molecule, the number of states per unit volume per unit energy (a quantity termed the density of states) is ordinarily plotted as a function of energy. For any solid, since the states all spread from discrete atomic levels, there are always regions of energies in which states exist, called bands, separated by regions in which no states are possible, called gaps. Every solid contains a fixed number, *N*, of electrons, which have to occupy the allowed states. Because of the exclusion principle the lowest energy of any solid is attained when the electrons fill the lowest *N* states, leaving all higher states empty. Conventionally a line called the Fermi level (after Enrico Fermi) is drawn on the diagram halfway between the highest filled state and the lowest empty state [*see illustration on page 42*].

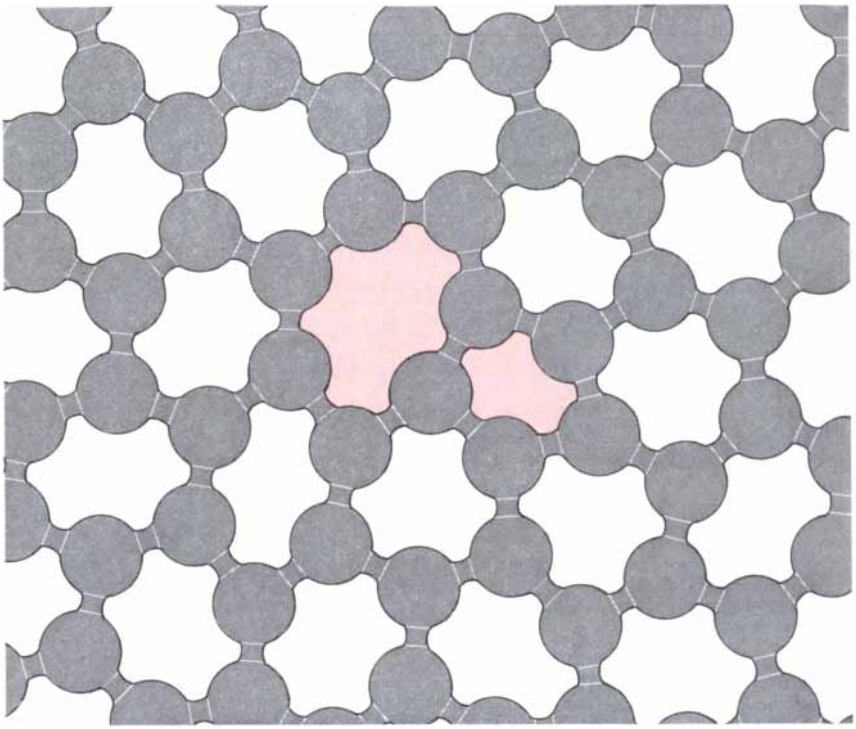
It can easily be shown that in a crystalline solid the long-range periodicity has two important consequences: first, there are definite energies, called the band edges, at which the density of electronic states sharply decreases to zero; second, all the electronic states in the bands extend throughout the material. The latter property is extremely important, since it leads directly to the high electrical conductivities observed in metals. The prevailing theory of solids is completely founded on the assumption of perfect or near-perfect long-range periodicity. The

mere existence of amorphous solids has long been an embarrassment for this approach, but the discovery that they can exhibit the same range of electrical properties as crystalline solids has finally led to a critical reexamination of the importance of periodicity. In this respect the study of amorphous materials continues to have a major impact on our understanding of crystalline materials.

Electrical conduction requires the presence of charged particles that can move easily in the presence of an applied electric field. The conductivity itself is proportional to the product of the number per unit volume, or concentration, of such carriers and the average velocity they attain per unit of applied field, a quantity called the carrier mobility. In order to obtain metallic conduction both a high mobility and a large carrier concentration are necessary. When an energy band is partially filled, the electrons in it can contribute to conduction by redistributing themselves appropriately among the empty states in the band; as a result crystalline solids with partially filled bands are metallic. On the other hand, electrons in a filled band cannot redistribute in any significant way because of the exclusion principle, and there is no conduction. Thus solids where the Fermi level lies in the middle of a band are metals, but those where it lies in a gap are insulators.

Semiconductors are simply insulators where the gap is relatively small: less than about three electron volts. In such cases thermal energy is sufficient to promote some of the electrons in the highest filled (or valence) band to the lowest empty (or conduction) band, where they can move freely in an applied field. When a few electrons are boosted into the conduction band, they leave behind an equal number of empty states near the top of the valence band. These empty states, or "holes," act as if they were positively charged carriers. Both electrons and holes can contribute to conduction in a semiconductor, and both can be excited at high temperatures from the thermal energy within the material.

The most important electronic applications of crystalline semiconductors depend not on the thermally excited carriers but on carriers that exist because of the presence of impurity atoms in the solid. A typical case is silicon to which has been added a small amount of phosphorus. Phosphorus has just one more electron than silicon. This is very important chemically, since the lowest energy state of a silicon atom in a covalent material is one where it is surrounded by a tetrahedron of its four nearest neighbors, each covalently bonded to one of the four outer electrons on the central atom. This leads to an electronic structure where the eight states corresponding to the outermost electronic shell split into bonding and antibonding levels [*see "a" in illustration on page 43*]. Because phosphorus has a fifth outer electron its energetically preferred bonding pattern is trigonal (three nearest neighbors) rather than tetrahedral. Nevertheless, when a few phosphorus atoms are placed in crystalline silicon, the geometric constraints of the crystal force them to enter the material at silicon positions, so that the phosphorus atoms are obliged to bond tetrahedrally. The conduction band of silicon arises from the antibonding states that are normally empty. The extra electron in the phosphorus atom ordinarily stays nearby, but it has an energy close to that of an electron at the bottom of the silicon conduction band [*see "c" in illustration on page 43*]. The resulting state thus has two important characteristics: first, it is localized in space; second, it can be ionized by absorbing only a small amount of energy, whereupon a "free" carrier in the conduction band is created. The development of crystalline semiconductors with a wide variety of electrical properties follows from the fact that it is possible to incorporate in them different types and concentrations of impurity atoms, the procedure called doping.



AMORPHOUS SOLID consisting of the same hypothetical material as that represented in the preceding illustration would consist predominantly of irregular hexagons. The trivalent-bonding requirements of each atom are locally satisfied, and the bonds have the same length they have in the crystal. The angles can vary, however, by as much as 20 degrees, plus or minus, from an average value of 120 degrees. Moreover, sprinkled among the hexagons are a few rings with five and seven members and some with eight members. Although amorphous solid exhibits short-range order (since distance between atoms is known), long-range order is lacking.

The first semiquantitative step in generalizing the theory of crystalline semiconductors to amorphous materials was taken by Sir Nevill Mott of the University of Cambridge. He noted that the sharp band edges of crystalline solids are purely the result of long-range periodicity and hence should disappear in amorphous materials, being replaced by bands with "tails" [*see illustration on page 45*]. One would ordinarily expect the tailing to lead to metallic conduction, but Mott postulated that states in the band tails are not extended throughout the solid as they are in pure crystalline materials; instead they are localized in space, exactly like the state occupied by the fifth phosphorus electron in phosphorus-doped silicon. Mott further suggested that there is a particular density of electronic states above which the states, even in an amorphous solid, do become extended. This leads to the existence of critical energies in each band where there is a sharp jump in mobility, from negligible values to finite ones. These critical energies, or "mobility edges," play the same role in amorphous solids that band edges play in crystalline solids. The energy difference between the mobility edges of the valence band and those of the conduction band is the "mobility gap." By postulating the existence of a mobility gap to take the place of the band gap in crystalline materials the Mott model clearly explains how it is possible for amorphous materials to function as semiconductors.

Mott's model was extended by Morrel

COLUMN	I	II	III	IV	V	VI	VII	VIII
FAMILY NAME	ALKALI METALS	ALKALINE EARTHS			PNICTIDES	CHALCOGENS	HALOGENS	NOBLE GASES
ELEMENTS	LITHIUM SODIUM POTASSIUM RUBIDIUM CESIUM	BERYLLIUM MAGNESIUM CALCIUM STRONTIUM BARIUM	BORON ALUMINUM GALLIUM INDIUM THALLIUM	CARBON SILICON GERMANIUM TIN LEAD	NITROGEN PHOSPHORUS ARSENIC ANTIMONY BISMUTH	OXYGEN SULFUR SELENIUM TELLURIUM POLONIUM	FLUORINE CHLORINE BROMINE IODINE ASTATINE	NEON ARGON KRYPTON XENON RADON
NUMBER OF ELECTRONS IN OUTERMOST SHELL	1	2	3	4	5	6	7	8
LOWEST-ENERGY ELECTRONIC CONFIGURATION	s ↑	s^2 ↑↓	$s^2\ p$ ↑↓ ↑	$s^2\ p^2$ ↑↓ ↑↑	$s^2\ p^2$ ↑↓ ↑↑↑	$s^2\ p^4$ ↑↓ ↑↑↑↓	$s^2\ p^5$ ↑↓ ↑↑↑↓↓	$s^2\ p^6$ ↑↓ ↑↑↑↓↓↓
ELECTRONIC CONFIGURATION FOR MAXIMUM BONDING	s ↑	sp ↑↑	sp^2 ↑↑↑	sp^3 ↑↑↑↑	p^3 ↑↑↑	p^4 ↑↑↑↓	p^5 ↑↑↑↓↓	p^6 ↑↑↑↓↓↓
MAXIMUM NUMBER OF BONDS PER ATOM	1	2	3	4	3	2	1	0
OPTIMUM BOND ANGLE	—	180°	120°	109.5°	90°-100°	100°-105°	—	—
GEOMETRIC CONFIGURATION	DIATOMIC MOLECULE	CHAIN	SHEET	RIGID STRUCTURE	PUCKERED SHEET (OR RIBBON)	PUCKERED CHAIN (OR RING)	DIATOMIC MOLECULE	—
DIMENSIONALITY OF BONDED SOLID	ZERO	ONE DIMENSION	TWO DIMENSIONS	THREE DIMENSIONS	TWO DIMENSIONS	ONE DIMENSION	ZERO	ZERO
COHESIVE ENERGY OF COVALENT SOLID	VERY SMALL	SMALL	MODERATE	VERY LARGE	LARGE	MODERATE	SMALL	NONE

PERIODIC TABLE OF THE ELEMENTS, given in shortened form, shows how electrons added to the outermost shells of the atoms in successive columns of the table are divided among the eight possible states: two *s* states of lower energy and six *p* states of higher energy. (The table omits all the "transition" metals, such as iron, copper, silver and gold, and all the rare-earth elements.) The *s* states have only one possible value of orbital angular momentum; the *p* states have three. Hence the *s* states are filled when they contain one electron with spin "up" and one with spin "down," as is indicated by the arrows. The three *p* states can accommodate a total of three electrons with spin up and another three with spin down. The electron-spin assignments for the lowest-energy configuration appear in the second row. The configurations for the maximum covalent bonding between atoms are shown in the third row. For elements in columns II, III and IV, one of the *s*-state electrons is promoted to a *p* state, forming hybrid states (*color*), each of which can form a bond. Thus carbon, in Column IV, can form four bonds, corresponding to the four unpaired states sp^3. Elements in Column V have a maximum of three bonds, all *p* states. In Column VI the pairing of two electrons in the *p* state leaves only two *p* states for bonding. In Column VII only one *p*-state electron is unpaired and in Column VIII none is unpaired. Only the elements in Column IV form rigid three-dimensional structures. In general when two or more elements form amorphous combinations, each constituent locally retains its own characteristic bonding property.

H. Cohen and Hellmut Fritzsche of the University of Chicago, working in collaboration with Ovshinsky. Their model was based on four principles: First, amorphous materials have band tails whose exact nature depends on the extent of the deviations from perfect periodicity; second, there are sharp mobility edges, separating localized and extended states in each band; third, the localized band tails overlap in the mobility gap, resulting in a finite density of states at the Fermi level; fourth, since amorphous solids do not have the rigid constraints of a crystalline structure, each atom ordinarily can be expected to fulfill its proper valence requirements locally, thus eliminating any sharp structure in the density of localized states in the gap.

In the next few years the main advances were primarily experimental rather than theoretical. The picture that emerged was that there are three classes of covalently bonded amorphous semiconductors with quite different properties. At one end of the spectrum are the primarily tetrahedrally bonded amorphous solids, such as silicon and germanium, whose properties are not radically different from those of the corresponding crystals. They can be doped with chemical impurities, they ordinarily possess some unpaired electron spins that yield magnetic effects and they can be made into solid-state diodes.

In striking contrast to these materials are the chalcogenide glasses, which are largely insensitive to impurities, do not have any unpaired spins except under the most extreme nonequilibrium conditions and ordinarily show symmetrical current-voltage characteristics. In addition they can be made to exhibit reversible switching and memory effects in high electric fields, and also structural changes under the influence of light. The pnictide glasses (glasses whose primary component is an element in Column V of the periodic table) have properties intermediate between the other two groups. Until recently there was no satisfactory theoretical understanding of the differences among these three classes.

An important first step in obtaining an explanation of the diverse behavior of the various types of amorphous semiconductor was taken by Marc A. Kastner, one of my colleagues at the Massachusetts Institute of Technology. He pointed out that the electronic structure of chalcogen atoms is fundamentally different from the structure of all the other atoms ordinarily present in amorphous semiconductors: they have two outer electrons that do not participate in chemical bonding. Ovshinsky showed how these lone-pair electrons could account for the reversible switching properties of amorphous semiconductors: they make free carriers available for excitation without requiring that the chemical bonds holding the material together be broken. Other unique properties of chalcogenide glasses, however, remained unclear.

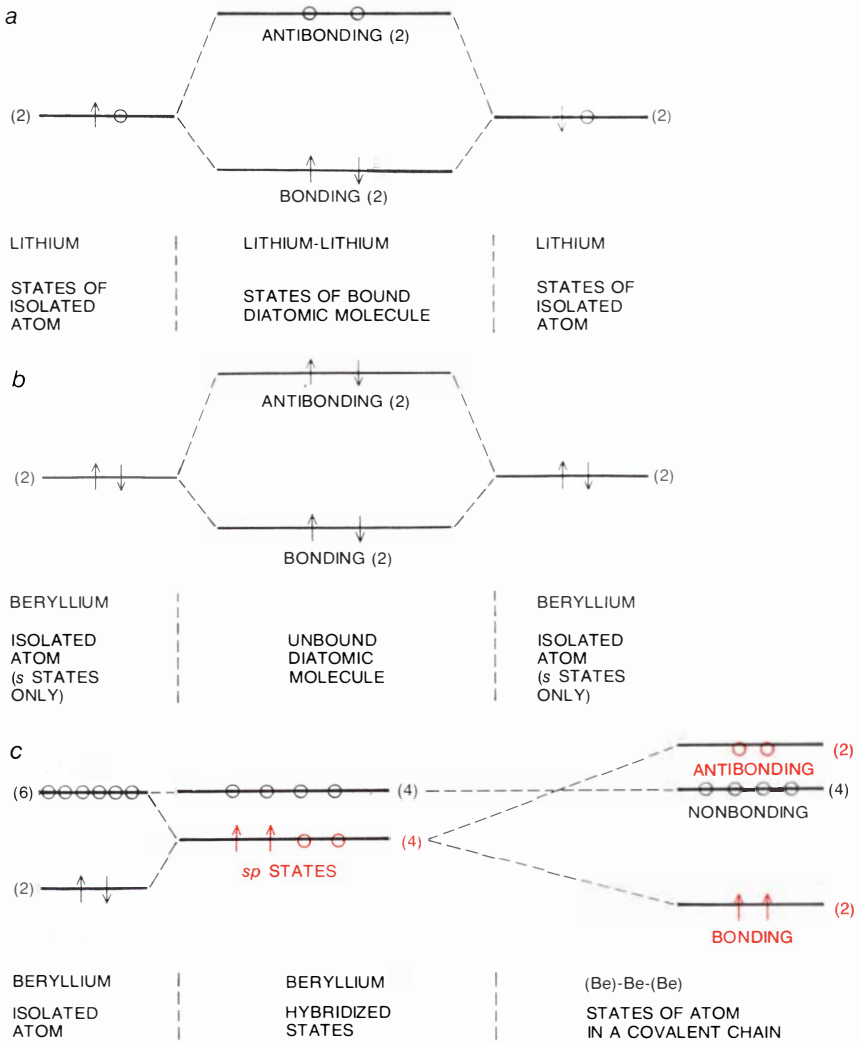
Perhaps the most unusual characteristic of chalcogenide glasses is their lack of a measurable density of unpaired spins. Even if every atom in the solid locally satisfies its bonding requirements, one would expect that at room temperature the available thermal energy would be sufficient to break up some of the highest-energy pairs, as happens in almost all other materials. In 1975 Philip W. Anderson of Bell Laboratories noted that if two electrons with opposite spins close to the same atom of a chalcogenide glass actually attract each other, the lowest-energy state would have no unpaired-spin density. Since all electrons are negatively charged, they should mutually repel one another, but under certain conditions in a metal in-

teractions between electrons and the vibrations of atoms in the metal can lead to an attraction among the electrons and give rise to the phenomenon of superconductivity. An analogous effect in nonmetals, however, had never been previously suggested.

Mott and his collaborators proposed that the important localized states in chalcogenide glasses arose not because of the disorder of the material, as Anderson had postulated, but because of well-defined defects much like those in a crystalline semiconductor. They suggested a dangling bond on a chalcogen atom as the appropriate defect, and they showed that the assumption of an effective attraction between localized electrons could also explain the wide range of behavior chalcogenide glasses exhibit under the influence of light. Ellen Yoffa and I were able to show why such a model further accounts for the relative insensitivity of conduction in such glasses to chemical impurities. Nevertheless, there remained two important problems: How is it possible for localized electrons in an insulator to exhibit an effective attraction? And why does this happen in chalcogenide glasses but not in the tetrahedral amorphous semiconductors?

Some years ago Ovshinsky suggested that interactions between the lone-pair electrons on different atoms and interactions with their local environment can create localized states in the energy gap of chalcogenide glasses. With this concept Kastner, Fritzsche and I were recently able to identify the nature of the appropriate active centers in chalcogenide glasses and also to show that the presence of atoms bonding with fewer than four nearest neighbors can automatically lead to an effective attraction between electrons localized on such centers. The first step toward understanding the properties of chalcogenide glasses is to note that the lowest-energy electrically neutral defect for a chalcogen atom in a covalent solid is not a dangling bond but an extra bond [*see illustration on page 46*]. The presence of either type of defect, however, would result in unpaired spins in the glass. If an electron were removed from one defect and paired with the electron on another, the resulting electrostatic repulsion would ordinarily increase the energy considerably. Removing an electron from a trigonally bonded chalcogen atom leads to a low-energy state, because the atom then becomes structurally similar to a pnictide atom in its proper configuration. On the other hand, if the removed electron is placed on another trigonally bonded chalcogen, a highly energetic state results, since there are then two antibonding electrons in the same vicinity.

There is, however, another possibility. The additional electron can move to any of the three nearest neighboring atoms, and the bond between that atom and the trigonally bonded neighbor can simply break [*see illustration on page 47*]. That leaves a chalcogen atom with an extra electron bonding to only one nearest neighbor. This, however, is simply an atom electronically similar to a halogen (Column VII) in its proper bonding configuration, and hence it represents a relatively low-energy state. In this case the transfer of an electron from one trigonally bonded center to another, thereby breaking a bond on the second center, lowers the energy of the material. This is the origin of the effective attraction between localized electrons suggested by Anderson. Moreover, it is easy to see that each of the resulting alternating positively and negatively charged centers has only paired spins, thus explaining the lack of any observable unpaired-spin density in these glasses. Finally, the model makes clear why the tetrahedral amorphous materials, such as amorphous silicon and germanium, do not exhibit the same behav-



ELECTRONIC STATES THAT FAVOR BONDING (*a*, *c*) **are contrasted with an electronic configuration where bonding does not occur** (*b*). **Since the lithium atom has only one electron in its outer shell, the electrons of two atoms are just enough to fill the bonding state, producing a diatomic lithium molecule** (*a*). **The energy of the molecule is much lower than that of the isolated atoms. (In these diagrams horizontal lines correspond to energies of various states.) The electron spins in the lithium molecule are paired, so that no magnetic effects are observed. The mechanism that bonds the lithium molecule is not present in the beryllium atom** (*b*), **which has two electrons in its outer shell. When two beryllium atoms are close enough to interact, two *s* states are reduced in energy and two are increased, forming "antibonding" states. (Only the *s* states are shown.) The antibonding states are always increased in energy by more than the bonding states are decreased, so that the electronic configuration does not lead to the overall decrease in energy needed to form a diatomic molecule. Beryllium atoms, however, can form chains by the mechanism of *sp* hybridization** (*c*). **Only a small amount of energy is needed to promote one of the *s* electrons to a *p* state. Since the two outer electrons no longer need to be paired, both can form covalent bonds with nearest-neighbor atoms. If the two bonding energies are larger than single promotion energy, a strongly bound covalent chain will form. Beryllium can polymerize, but it normally forms a metallic solid with lower energy. Certain chalcogen elements (sulfur, selenium and tellurium), however, can polymerize by this mechanism.**

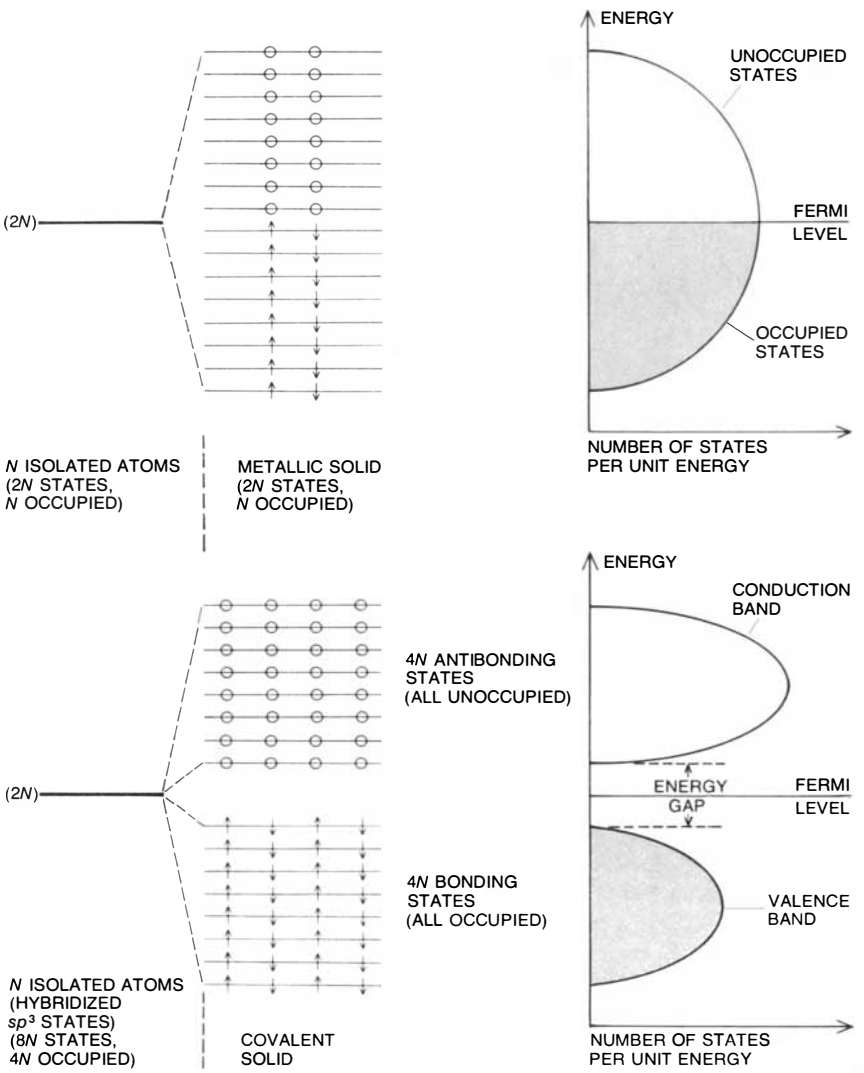
ior, since five bonds are impossible with only *s* and *p* electrons.

At almost the same time that our understanding of chalcogenide glasses expanded there was a major experimental advance in the area of tetrahedral amorphous semiconductors. Walter E. Spear and Peter Le Comber of the University of Dundee succeeded in doping a particular form of amorphous silicon: the form resulting from the decomposition of silane gas (SiH_4). They proceeded to fabricate rectifying diodes from this material, an advance suggesting that ordinary electronic devices based entirely on amorphous materials could be developed. That was strikingly realized soon thereafter, when David E. Carlson and Christopher R. Wronski of the RCA Laboratories fabricated solar cells of up to 6 percent efficiency from doped amorphous silicon. It now appears that the economic breakthrough necessary to achieve the large-scale generation of electric power with solar cells may well come about through the inherently low-cost amorphous technology rather than through a reduction in the cost of making uniformly crystalline devices.

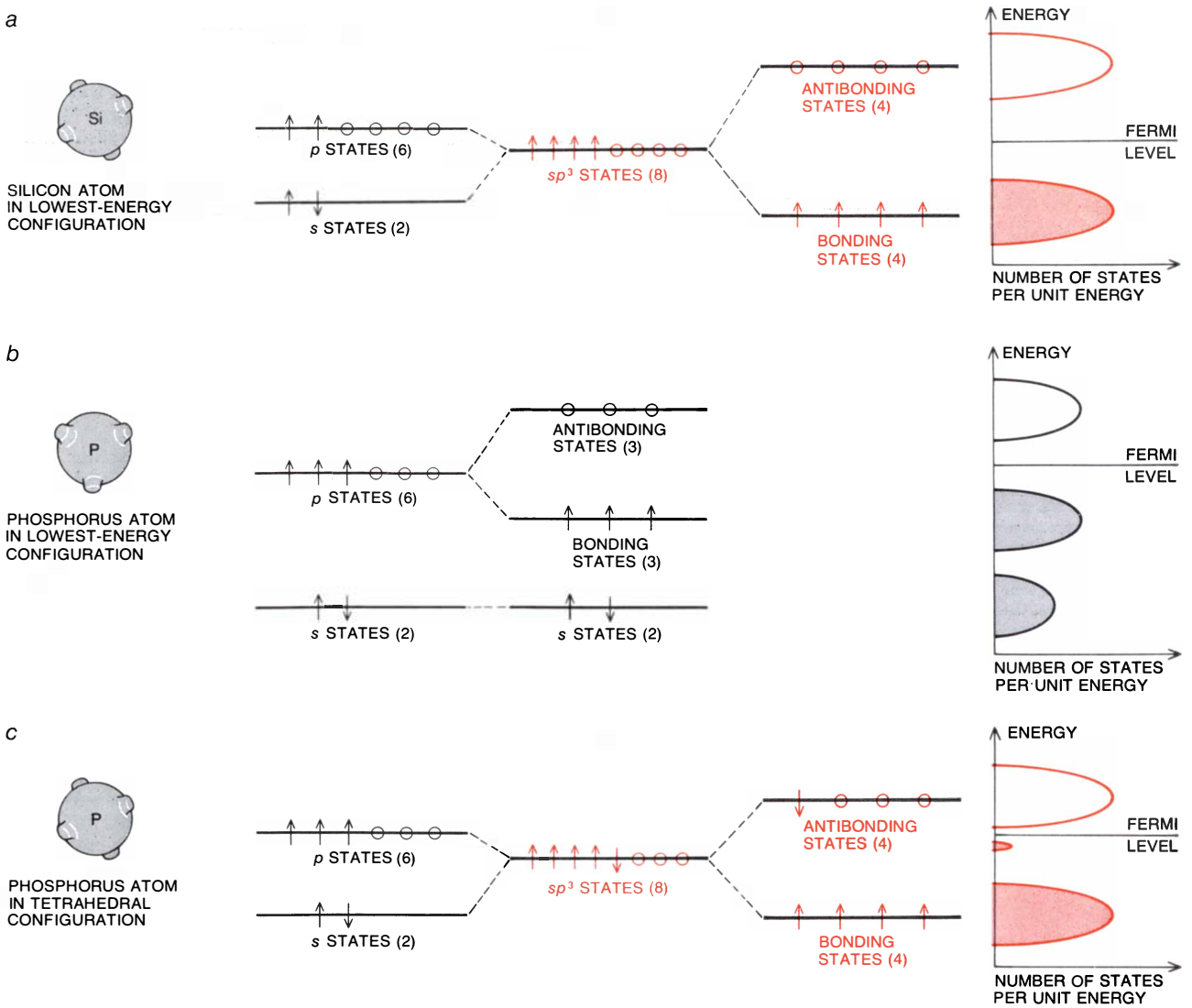
The experimental results on tetrahedral amorphous semiconductors can all be understood in terms of the active-center model I have described, except that in these materials the interaction of localized electrons must be repulsive rather than attractive. As amorphous-semiconductor films are ordinarily prepared, they have rather high densities of dangling bonds, the lowest-energy defect possible. These bonds provide the active centers that not only control the electronic properties but also are responsible for the large unpaired-spin density and make it difficult to dope the materials. Most of the dangling bonds can be removed by annealing the films below their crystallization temperature, but a density sufficiently low to allow the electrical conductivity to be strongly modified by doping with impurities has not yet been attained by any heat treatment. The fact that doping can be achieved in films deposited from silane suggests, however, that residual hydrogen ties up the dangling bonds in an efficient manner and thereby allows for the control of electrical conduction by suitably chosen impurities. This view is supported by the recent work of William Paul and his associates at Harvard University, who were able to dope tetrahedral amorphous films by exposing them directly to hydrogen gas during the deposition process.

As I have mentioned, the conductivity of chalcogenide glasses is rather insensitive to impurities. Ovshinsky and his coworkers, however, have recently succeeded in increasing the conductivity of such materials many orders of magnitude by chemical modification. This development shows that it is possible to saturate either the positively or the negatively charged sites in chalcogenide glasses. The successful modulation of conductivity in these stable glasses promises to open up a new area of commercial applications, from low-cost solar cells to all-amorphous solid-state devices.

The belated realization that the electronic properties of amorphous semiconductors are controlled by well-defined active centers, just as they are in crystalline semiconductors, should eventually lead to the fabrication of amorphous analogues of many of the electronic devices now available only with crystalline technology. Moreover, in the case of chalcogenide glasses the absence of a rigid structure introduces an additional flexibility that greatly broadens their range of possible behavior. Ovshinsky pointed out many years ago that the continuously variable composition of the chalcogenide alloys allows them to be tailored to optimize the figure of merit for a particular application. For example, both the energy gap and the dielectric constant of chalcogenide glasses have been varied as desired. It is also possible to independently change the density and nature of the alternating positively and negatively charged centers by varying the composi-



CONDUCTIVITY OF SOLID depends on the configuration of energy bands produced when the interaction among electrons on neighboring atoms spreads the energy states of the atoms into a series of closely spaced levels. A metallic solid such as lithium (*top*) forms a single continuous energy band in which only the states in the lower half are filled. The Fermi level separates the occupied states from the unoccupied ones. When the Fermi level falls inside a band, the solid is a metal and a conductor. In a covalently bonded solid such as silicon (*bottom*) the hybridized sp^3 states of the isolated atoms are split into bonding and antibonding states, which in turn are spread into bands. The bonding band is completely filled and the antibonding band is completely empty. The Fermi level lies in the energy gap, a region where no electronic states exist. In such cases the solid is an insulator. If the energy gap is small enough so that thermal energy can raise some electrons into the conduction band, the electrical conductivity of the solid will increase with temperature. Solid is then called a semiconductor. In both diagrams shaded states are filled. Unshaded states are empty when material is in lowest-energy configuration.

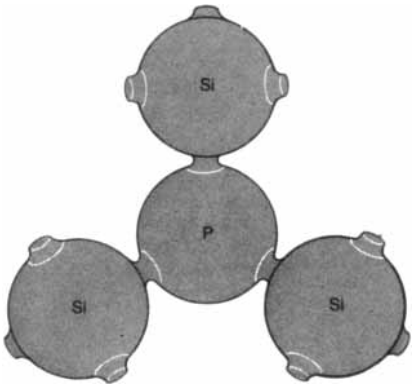


ELECTRONIC STRUCTURE of semiconductors is influenced by their bonding patterns. In the case of silicon (*a*) the configuration of lowest energy is attained when each atom is surrounded by a regular tetrahedron of other atoms. The *s* and *p* states hybridize and then split into bonding and antibonding states. In phosphorus (*b*) the configuration of lowest energy requires each atom to be bound to three nearest neighbors. Since the *s* and *p* states do not hybridize, each forms a filled energy band in the solid. For phosphorus to bond tetrahedrally (*c*) the *s* and *p* states must hybridize to provide four bonding states. Since one electron is forced into an antibonding state, the overall configuration is more energetic than it is for trigonal bonding. Hence trigonal bonding is the preferred configuration. When small amounts of phosphorus are diffused into crystalline silicon, however, the geometric constraints of the periodic crystal force the phosphorus atoms into tetrahedral configurations. The electron in the antibonding state can lower its energy slightly by moving in an orbit around neighboring silicon atoms, so that its energy lies just below the bottom of the conduction band. Antibonding electrons are easily excited into conduction band by thermal energies, so that small concentrations of phosphorus lead to sharp increases in electrical conductivity of silicon.

tion of the glasses, by different preparation procedures or by subsequent heat treatment. This promises to greatly expand the range of devices that can be fabricated from these easily prepared materials.

The most important commercial application of chalcogenide glasses so far has undoubtedly been in electrophotography (or xerography), a process in which the photoconductive properties of a selenium glass are exploited. The upper surface of a film of the material is ordinarily charged positively to about 1,000 volts by a wire that sprays positive ions onto it. An "image" negative charge is built up in a metallic substrate, on which the glass is deposited. Light reflected from the document to be copied impinges on the film. Where there was print on the original the light is absorbed rather than reflected; where there was no print the reflected light energy gives rise to electron-hole pairs near the upper surface of the glass. The large electric field across the glass separates the pairs. The electrons move up and neutralize the positive ions at the upper surface; the holes propagate through the glass and neutralize the negative charge in the metallic substrate. Thus the upper surface of the glass is electrically neutral where there was no printing on the original but is still positively charged where there was print. Small, negatively charged black particles of "toner" are attracted to positively charged areas of the glass. The toner is then transferred onto a sheet of positively charged paper. The toner is permanently fixed to the paper by heat, completing the copying process.

The past year has seen significant advances in our understanding of electrophotographic processes. One persistent problem has been the absence of a well-defined time for the holes excited by the light to reach the metallic substrate. Recent work by Harvey Scher and his collaborators at the Joseph C. Wilson Technology Center has shown that this can result from any process where there is a distribution of characteristic times



CHEMICAL MODIFICATION, or doping, to modulate conductivity is more difficult in an amorphous solid than it is in a crystalline solid because it is possible for all the covalent-bonding requirements of every atom to be satisfied in the absence of geometric constraints provided by a periodic crystal. Here a phosphorus atom is incorporated in its normal, or lowest-energy, trigonal state in a silicon matrix. As can be seen in the illustration on the preceding page, the energy gap in trigonally bound phosphorus is large, and the Fermi level remains near the center of the gap. Therefore the addition of phosphorus to amorphous silicon does not alter conductivity. In order to dope amorphous silicon effectively phosphorus must be forced to bond tetrahedrally in spite of the concomitant increase in energy. This has recently been achieved with the help of a substantial concentration of hydrogen.

for the holes to reach the substrate. Marvin Silver of the University of North Carolina has suggested that the process could be the continual trapping of holes by negatively charged centers in the glass. More recently the valence-alternation model for chalcogenide glasses has clarified several aspects of electrophotography. In this model the presence of large densities of positively and negatively charged traps makes it easy to understand how the holes can be continually trapped. Furthermore, the charged centers provide a simple mechanism for the attachment of both the positive ions and the toner particles to the surface of the glass.

Applications of the photoconductive properties of chalcogenide glasses continue to grow. The Xerox Corporation has developed a color copier based on a three-layer structure of different chalcogenide glasses that responds to light over the entire visible spectrum. This device is used with combinations of colored filters and toner particles to make rapid copies of colored originals. Very recently Hitachi and the Japan Broadcasting Corporation have developed a miniature high-sensitivity image-pickup tube (a device for converting an image into electrical signals) whose active element is a graded-composition alloy of amorphous selenium, arsenic and tellurium. This tube, called the Saticon, is less than an inch long, and it is already widely used in small color-television cameras.

Up to this point I have not discussed the unusual switching properties of chalcogenide glasses, which were discovered and developed by Ovshinsky in the period from 1958 to 1968. Publication of the switching results generated a great deal of interest in the entire field of amorphous semiconductors but also led to considerable controversy. There were immediate assertions that the devices were irreproducible, irreversible, unreliable, short-lived and unstable. On numerous occasions "proof" of such shortcomings was presented in the form of data from experiments that had nothing to do with switching. It is always possible to "switch" any nonmetal at least once, in the sense that application of a sufficiently strong electric field will eventually break down the insulating character of the material, leading to a sharp increase in conductance. This type of breakdown is ordinarily irreversible, and true switching must suppress such effects rather than depend on them. As is now evident, the electronic structure of chalcogenide glasses tends to suppress breakdown while fostering true reversible switching, which is the basis for their superiority in electronic devices. In addition the switches have an extremely large conductance ratio between the off-on states (approximately a million), a rapid switching time (less than a billionth of a second) and a long lifetime (some devices have been switched for more than 10^{14} cycles without failure).

It is somewhat difficult to understand the irrationality that followed the publication of Ovshinsky's original paper, particularly since the subsequent condemnation of the technology by many scientists and engineers damned the entire field of amorphous semiconductors as a potential area for commercial electronic devices. This is particularly remarkable in view of the success of xerography, a process that subjects an amorphous-semiconductor material to extreme variations in electric field, light, temperature and humidity and requires the material to cover a large area and yet permanently resist cracking, decomposition and crystallization.

Two legitimate scientific controversies arose shortly after the publication of Ovshinsky's work. The main one was concerned with the nature of the switching process, in particular with whether switching was primarily an electronic phenomenon or a thermal one. If it were electronic, the conducting state would represent a nonequilibrium increase in the concentration of free carriers (and perhaps also in the mobility of the carriers). Although electronic nonequilibrium states were well known in crystalline solids, the maintenance of such a state with only a small fraction of the switching voltage across the device had never been attained with a homogeneous material. On the other hand, a possible alternative explanation of the conducting state was that it was simply due to a kind of hot central filament, with the drop in resistance resulting entirely from the sharp increase of conductivity with increasing temperature that characterizes all semiconductors. The temperature rise in this mechanism would simply be a by-product of the resistance heating that occurs whenever current is passed through any material. The breakdown phenomenon that results, called "thermal runaway," is well known in many crystalline semiconductors, and it can be initiated in many materials if the sample of the material is thick enough and conducts heat poorly. Although there is no absolute reason why such a thermal mechanism would result in inferior devices, it is true that atoms migrate faster at high temperatures than at low ones; hence devices that operate at high temperatures are less likely to have long operating lifetimes than those that operate at ambient temperature.

As far back as 1971 H. K. Henisch and his collaborators at Pennsylvania State University had experimental evidence in favor of an electronic mechanism for switching in amorphous semiconductors, but it was not until recently that the controversy was definitively settled. First Melvin P. Shaw of Wayne State University, working with Scott Holmberg and W. Derek Buckley of Energy Conversion Devices, showed that the conducting state in an amorphous-semiconductor switch could be induced in less than a billionth of a second, which is too short a time for any significant heating to occur. This result indicates that an electronic switching mechanism exists, but it does not rule out the possibility that there is a hot central filament in the switch. Next Gary C. Vezzoli and his co-workers at the Picatinny Arsenal observed luminescence emanating from the conducting filament in the switching material and found that it had the character of luminescence from an electronically excited cool material rather than of luminescence from a hot one. Then Kurt E. Petersen and I measured the size of the conducting filament by four different methods and found it is such that the temperature of the filament cannot increase any more than 60 degrees Celsius when the switch is operating under normal conditions. Thus it appears that threshold switching in chalcogenide glasses is fundamentally electronic, as Ovshinsky originally maintained.

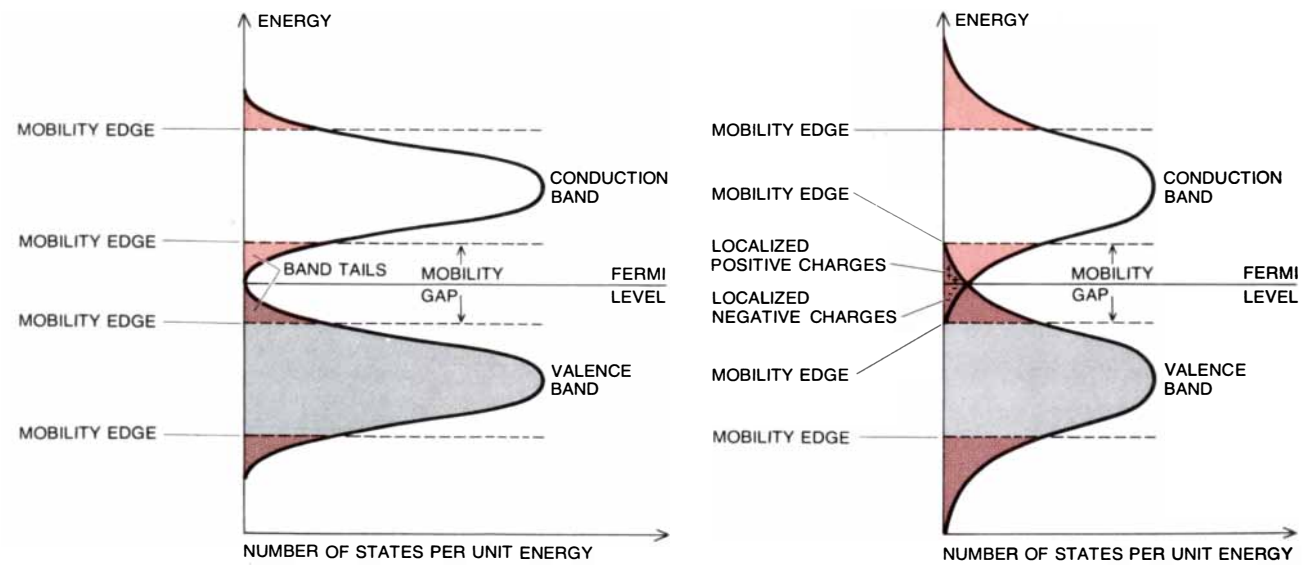
A second scientific controversy involved switching phenomena known as formation, or "first fire," effects. In some devices the first switching event requires a significantly larger voltage than all the succeeding events. Although it had been known that not all devices showed any

formation effects, and indeed that the best switches had a completely stable threshold voltage, it was nevertheless proposed by several groups that formation effects are an essential part of the amorphous-semiconductor switching process. Petersen and I have recently shown, however, that formation effects arise in the best switches only when the cross section of the device is too small or the applied current is too large.

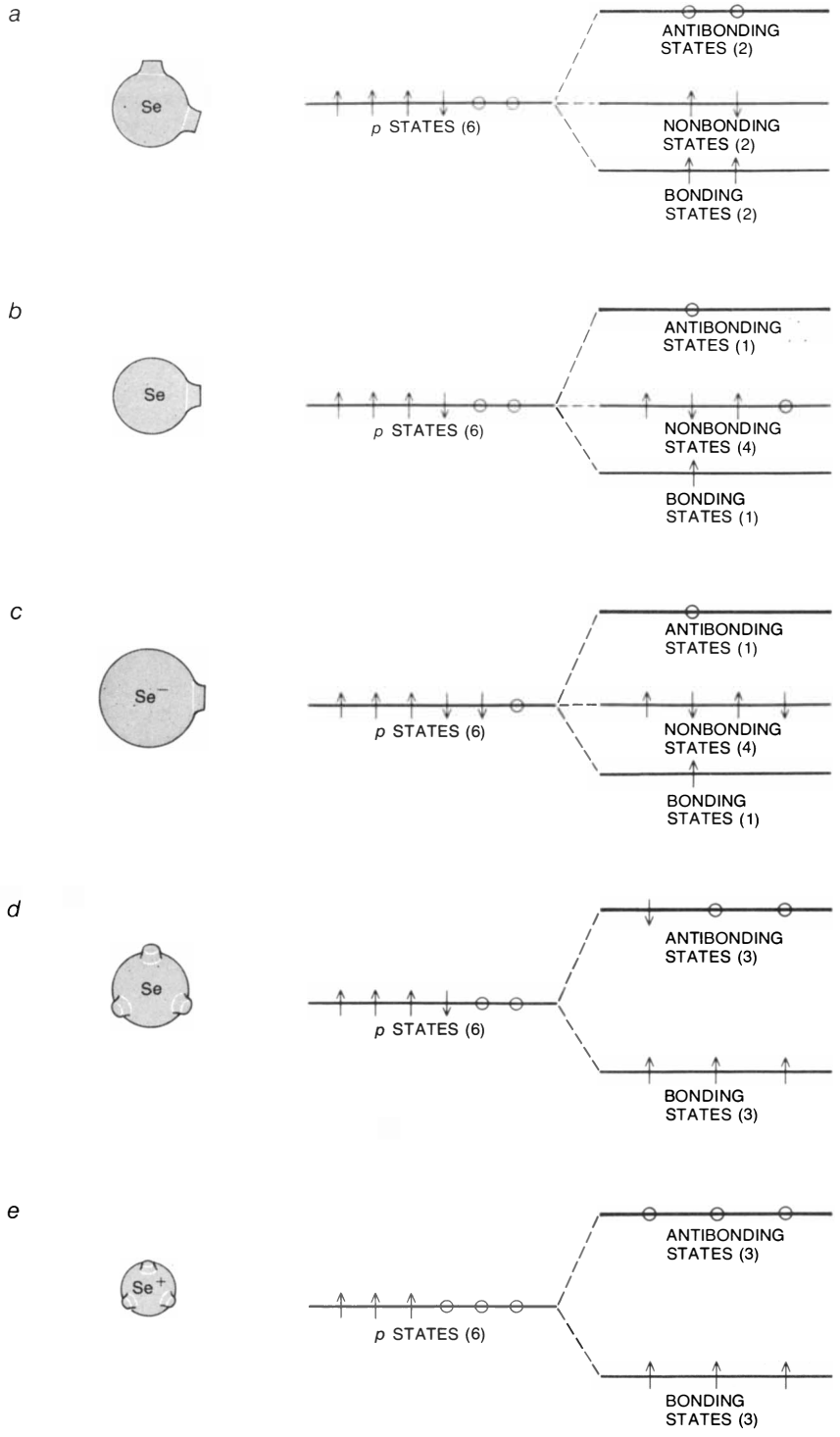
The recent switching experiments taken together with the new advances in our understanding of the electronic structure of amorphous semiconductors have clarified the unique combination of properties in the chalcogenide glasses that lead to particularly desirable devices. First, the glasses should contain fairly large percentages (from 30 to 70 percent) of atoms that bond tetrahedrally or trigonally, thus cross-linking the chains of chalcogen atoms. This provides a structural stability that retards crystallization of the glass. Second, the particular compositions should be such that the electrical conductivity of the devices at room temperature is extremely low. If the conductivity is low, there is a minimum of resistance heating, which can trigger a runaway thermal effect. Finally, it is important to have a high enough concentration of chalcogen atoms to provide a large enough number of positively and negatively charged traps, which are needed to retard an irreversible electrical breakdown of the material.

Once the experimental results became clear, an understanding of the detailed mechanism of amorphous-semiconductor switching began to emerge. Shaw and his collaborators developed a circuit model that successfully explains the transient behavior of current with respect to voltage. Petersen and I proposed a quantitative electronic model for the recovery of the low-conductivity state after the voltage across the device is removed. This model made several predictions that were found to be in good agreement with subsequent measurements. In addition a detailed consideration of the special nature of the electronic structure of chalcogenide glasses has led to the development of a tentative model for the mechanism of switching itself. In this approach the conducting state is achieved only after the positively and negatively charged traps that ordinarily exist in the glass are filled by carriers excited by the applied electric field. When all the traps are filled, the lifetime of an injected carrier sharply increases from a value much lower than the time the carrier needs to cross the entire thickness of the film to a value higher than the time the carrier needs to cross the film, leading to the drop in voltage and the associated increase in current that characterizes the conducting state. This model is in agreement with recent experiments indicating that the carrier mobility in the highly conducting state is the same as it is between trapping events in the low-conductivity state. The response times for both the filling and the emptying of the traps are in agreement with the predictions of the valence-alternation model of chalcogenide glasses. It now appears that switching is as well understood as most nonequilibrium effects in crystalline semiconductors.

In his original paper Ovshinsky also announced the discovery of a second type of switching phenomenon, in which the conducting state is retained even after complete removal of the applied voltage. The nonconducting state can be reestablished only by the application of a strong pulse of current. This behavior, which is now known as memory switching, is found in chalcogenide glasses where the structure is stabilized by only a small amount of cross-linking, thereby making crystallization considerably easier. Ovshinsky suggested that memory switching is therefore a consequence of a transition between the amorphous state and the crystalline one, a phenomenon that in these glasses is intrinsically reversible. Indeed, soon afterward detailed structural studies conducted by Arthur I. Bienenstock of Stanford University, Simon C. Moss of Energy Conversion Devices and their co-workers showed that in the class of materials ordinarily used for memory switches (based on tellurium-germanium alloys) the high nonequilibrium conditions of threshold switching induced the formation of small crystalline regions of semimetallic tellurium, which are then responsible for the memory-on state. Nuclear-magnetic-resonance studies have clearly demonstrated the intrinsic re-



AMORPHOUS SEMICONDUCTORS that are not strongly disordered (*left*) have valence and conduction bands similar to those in the corresponding crystalline semiconductor. The distinguishing feature of the bands in amorphous solids is the replacement of the sharp band edges present in crystals by what are called "band tails," or localized states, that extend into the energy gap. The localized states are separated from the extended states in the main part of the bands by "mobility edges." The region that lies between the mobility edges of the valence and conduction bands is the "mobility gap." It plays the same role in amorphous semiconductors that the energy gap plays in crystalline semiconductors. Chemical impurities or defects in the configuration of local bonds can lead to sharp structural changes (*not shown*) in the mobility gap. If the disorder is large, as is expected in multicomponent glasses (*right*), the band tails of the valence and conduction bands can overlap in the mobility gap. This leads to a redistribution of electric charge as electrons move from one localized state to another in order to lower their energy. The result is a high density of positively and negatively charged traps, which decrease the mobility of the carriers and make the material less sensitive to efforts to modulate its conductivity by chemical means, that is, by doping.



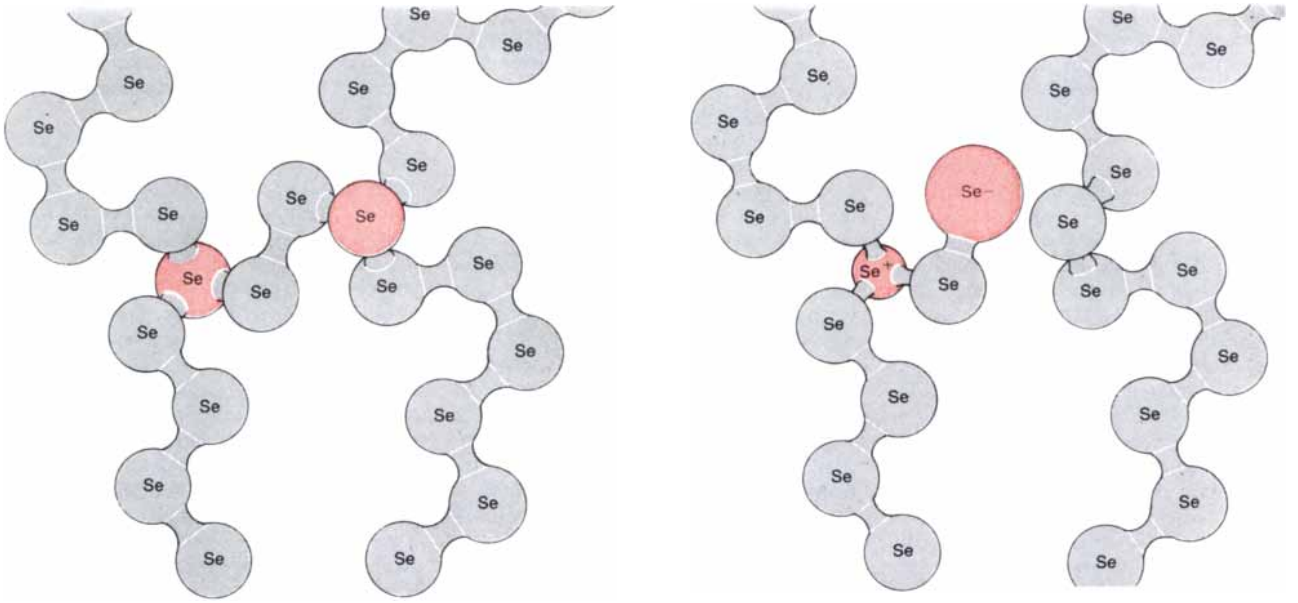
SELENIUM ATOM, one of the chalcogen elements, can assume a variety of covalent-bonding configurations in chalcogenide glasses. When selenium bonds with two of its nearest neighbors, the atom is in the configuration of lowest energy (*a*). The optimum bond angle is between 100 and 105 degrees. When selenium forms only one bond with a neighbor (*b*), the other bond is left "dangling" in an unpaired state. (Bonding electrons always pair their spins with the spin of the neighbor sharing the bond.) If the singly coordinated selenium were to pick up an extra electron (*c*), it would resemble a halogen (Column VII) atom in its lowest-energy configuration. In this charge state, with no unpaired spins, selenium acts to end a chain. When neutral selenium is trigonally coordinated (*d*), the extra electron is forced to enter an antibonding state. This configuration therefore has higher energy than a selenium atom with one less bond, but it has lower energy than the dangling-bond configuration shown in *b*. If, however, an electron is removed from trigonally coordinated selenium (*e*), an extremely low energy state results. Trigonally coordinated selenium forms cross-links between chains in glasses. A combination of selenium atoms in configurations *c* and *e* has an energy almost as low as two atoms of configuration *a*. As a result *c*-*e* combinations, called valence-alternation pairs, occur often in glasses.

versibility of the memory transition. In fact, the conducting state is enough like that of a metal to be superconducting at low temperatures. A quantitative model of the switching process has been formulated, and it is in excellent agreement with the experimental data.

Perhaps as a result of the fact that memory switching was understood fairly early, devices based on this phenomenon were developed quickly. Late in 1970 a 256-bit nonvolatile, electronically alterable memory ("Ovonic" memory) was placed on the market by Energy Conversion Devices. Sometimes called a read-mostly memory (RMM), the device is now being manufactured under the sponsorship of the Burroughs Corporation. The Ovonic memory has turned out to be important in bipolar-memory technology, filling a large gap between the volatile, alterable random-access memory (RAM) and the nonvolatile, nonalterable read-only memory (ROM). In addition it is directly compatible with the logic currently used in computer memory systems, is resistant to ionizing radiation and other environmental stresses, has an access time of 50 billionths of a second and has a write/erase lifetime of more than 100 million cycles per bit. In the past five years the Ovonic memory has been steadily improved, particularly with regard to lifetime, reset current, packing density and maximum operating range. The "learning curves" have been comparable or superior to those attained in other areas of the semiconductor industry.

Another area for the eventual application of memory switching is ultra-large memories, including those for archival storage. Peter Klose, Julius Feinleib and Ovshinsky demonstrated that the two memory states are optically quite distinct, that is, when the material is in one state it is much more transparent to light than it is when it is in the other state. This means that the state of the material can be read by the application of light. In addition it was found that both writing and erasing, as well as reading, can be accomplished by the same optical means. Since visible light can be focused down to a spot with a diameter of about a micrometer, it becomes possible with these memory-switching materials to fabricate mass memories that have a capacity of a trillion bits per square meter.

Initially the fast write/erase time (approximately a millionth of a second) and the symmetry of the write/erase cycle were not understood, but a model that explained the results was soon developed. Optical memory switching is also an example of a reversible amorphous-crystalline transition; the fast write time is due to the enhancement of crystallization by light, and the symmetry of the write/erase process follows from the fact that the glassy areas are more transparent than the crystalline



VALENCE-ALTERNATION PAIRS can form in a neutral chalcogenide-glass matrix without any major displacement of its atoms, leading to a sharp reduction in total energy. At the left two selenium atoms, each of which is triply bonded, serve to cross-link two molecular chains of doubly bonded atoms. A valence-alternation pair can be produced (*right*) by a spontaneous breaking of the cross-linkage, combined with the simultaneous transfer of an electron from one of the triply bonded selenium atoms to an atom near the one where the cross-link was broken. Since such electronic transfers reduce total energy of the solid, nearly all the trigonally bonded selenium atoms become members of a valence-alternation pair. Important physical consequences follow, including virtual disappearance of electrons with unpaired spins and appearance of large but equal concentrations of positively and negatively charged traps in chalcogenide glasses.

ones. After the initial work an alternative process was developed that did not depend on crystallization. This process gave increased resolution, a shorter write time and a large reduction in the energy needed for writing. The alternative process has been explained in terms of the reversible formation of vapor bubbles in the glass by light.

A separate possibility for achieving mass-memory systems is through the use of electron-beam writing and reading. In the reading mode the amorphous and crystalline states of memory-switching glasses differ sharply in their emission of secondary electrons when an electron beam impinges on them. With anticipated improvements in electron optics it should be possible to achieve a dot size of 10 nanometers, or a hundredth the minimum dot size attainable with visible-light optics.

The fact that chalcogenide glasses can be put in a permanent conducting state by light also suggests an important application in the field of printing. As users of electrostatic copiers are aware, multiple copies require multiple applications of light. The memory-switching glasses, however, can provide a permanent plate from which an unlimited number of electrophotographic prints can be made without the necessity of additional exposures. The technique seems promising for rapid large-scale printing.

In fact, one of the most exciting areas of development in the amorphous-semiconductor field at present is the area of imaging in general. As with memory switching, the basis for many of the imaging schemes is the amorphous-crystalline transition. Many imaging applications call for a gradation of tones, so that the important characteristics become the size of the grains of the material and the fraction of the material that is crystallized. Here again the original concept was Ovshinsky's. Either highly contrasting or continuous tones can be produced, depending on the particular application. An additional degree of flexibility is provided by the fact that the image can be obtained either directly after exposure or as a latent image that can be developed later by heat or radiation. In neither case is a fixing step necessary, since both the amorphous and the crystallized states are completely stable at ordinary temperatures. Furthermore, the mechanism allows add-on writing or overprinting without the necessity for any erasing. An additional feature is that various changes in physical properties can be detected and utilized for imaging purposes. For example, the two states differ in their wettability: the crystalline material accepts water and rejects oily ink, whereas the amorphous material rejects water and accepts ink. By this means one can obtain a photographic print from the exposed film.

The amorphous-crystalline concept has been extended to the use of organotellurides as the active element. Here exposure to light leads to the formation of a particular chalcogenide glass, which forms the latent image. When the film is heated above the temperature at which the glass melts, small crystalline regions of tellurium precipitate out of the glass, thus forming the image. An entirely new concept of instant photography seems feasible.

Two other light-induced processes characteristic of certain chalcogenide glasses are photodarkening and photodoping. In photodarkening either reversible or irreversible structural changes induced by the absorption of light lead to a significant decrease in the optical transparency of the material. This suggests many important applications from the modulation of light to holographic information storage. In photodoping the diffusion of certain metals into chalcogenide glasses is enhanced by light. Chalcogenide glasses ordinarily dissolve rapidly in alkaline solutions, but when there is enough silver in them, they become virtually insoluble. This effect, explored by Akira Yoshikawa and his co-workers at the Nippon Telegraph and Telephone Public Corporation, can be exploited to great advantage throughout the semiconductor industry, where photolithography is vital to the fabrication of integrated circuits.

There are other structural effects in film that can be utilized for imaging, and Energy Conversion Devices has developed a wide variety of materials designed to respond to energy in various ways, ranging from subtle changes such as the alteration by light of the length of chains of chalcogenide atoms to alterations in flow properties. Some of these structural effects are not limited to amorphous films but can also be accom-

plished in analogous crystalline films. All these imaging systems offer great flexibility in response, a dry development process, high resolution and excellent stability and durability.

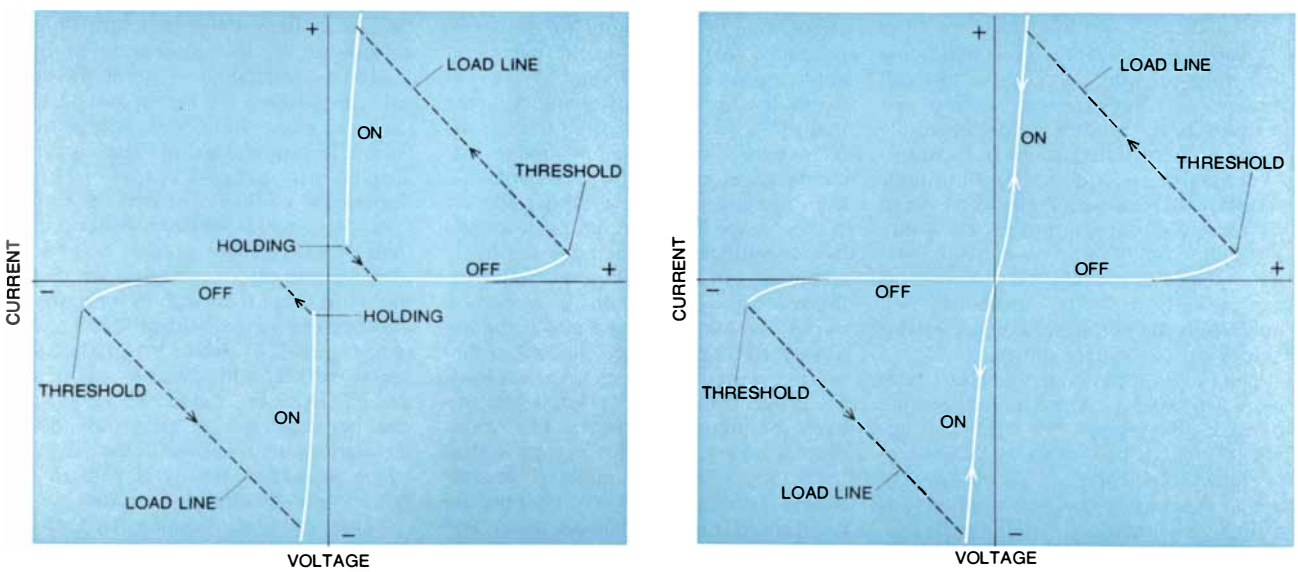
In addition to devices based entirely on amorphous solids, hybrid electronic structures consisting of amorphous/crystalline heterojunctions have recently been developed. In 1974 Don Reinhard, Floyd O. Arntz and I derived an energy-band model that explains the low-field electronic and photoconductive characteristics of such heterojunctions. Petersen and I later found that the model could be extended to the case in which the amorphous semiconductor is switched into the conductive state.

The hybrid heterojunctions can be extremely useful. They can readily be made to act as a rectifier, and their rectification capability persists even after the glass is switched. Furthermore, since amorphous solids do not have a rigid crystal structure, the hybrid devices have a significantly lower density of energy states along the interface between the materials of the heterojunction than all-crystalline heterojunctions do. Since interface states tend to degrade the performance of heterojunctions, their sharp reduction in glass-hybrid devices leads to near-ideal performance. Petersen and I also found that the chalcogenide glass in its conducting state can be used to inject high-energy electrons directly into the conduction band of the crystalline semiconductor.

This last finding immediately suggested that a chalcogenide glass could serve as the emitter of a novel transistor. Recently Petersen, Shaw and I fabricated such a device, employing crystalline silicon as the base and collector. When the emitter of this transistor is in the resistive state, current gains are on the order of .05. When the glass emitter is switched with a small pulse, however, the gain increases by a factor of about 200, to about 10. Depending on the ratio of the collector voltage to the base voltage, the transistor can be left in its high-gain state or returned to the low-gain state. In the low-gain state the single device acts as a threshold amplifier; in the high-gain state it is a "threshold latching" amplifier (a system that conventionally requires three subunits: a threshold detector, a latching circuit and an amplifier). The device can also be used to convert a system from an oscillator to an amplifier. Furthermore, since the device provides an additional state (the low-gain condition) that is clearly distinguishable from the saturation and cutoff conditions characteristic of an ordinary transistor, the glass-hybrid device can be used as the basis for ternary logic. Such logic can specify three conditions: minus, zero and plus. It provides the most efficient means known for storing and transmitting information.

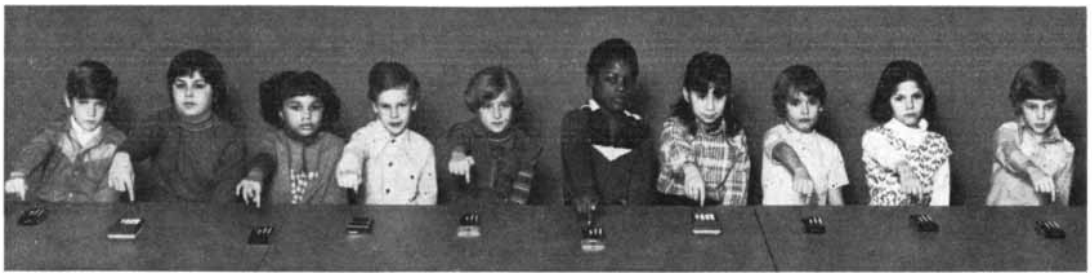
Work is just beginning on heterojunctions between the chalcogenide glasses of the threshold type and crystalline semiconductors other than silicon or germanium, but there is no question of the potential importance of these glasses as efficient electron injectors. No crystalline material yet developed can duplicate this function.

Amorphous-semiconductor devices are now at a critical stage of development. Extraordinary progress has been made in the past year or so. For the first time we seem to understand the basic reasons for the unique properties of chalcogenide glasses, we have been able to modulate the electronic properties of both chalcogenide glasses and tetrahedral amorphous semiconductors and we can quantitatively understand both the nature of the conducting state and the recovery process in threshold switches. In the past the development of an understanding of the physics behind previously obscure phenomena has always led to a period of rapid technological advance. An important example is the field of crystalline transistors. The original devices were point-contact transistors, the theory of which was a mystery. As the theory of semiconductors became clearer a quantitative model for the junction transistor was developed, although no such device had yet been fabricated. Because the physics of point-contact transistors remained obscure, within two years junction transistors completely dominated the field. Although one cannot predict the future with any degree of certainty, it does seem safe to conclude that over the next few years amorphous-semiconductor devices will develop rapidly and be widely accepted.



TWO KINDS OF SWITCHING have been realized with chalcogenide glasses. These diagrams show current as a function of voltage; the slope of the curve (*white*) corresponds to the conductance of the material. In threshold switching (*left*) the application of a voltage greater than the threshold value increases the conductivity of the glass by a factor of about a million. If the voltage in the conducting state is reduced below the holding point, the device switches back to the low-conductance, or "off," state. Memory switching (*right*) requires a glass with different properties. If the glass is held in the high-conductance, or "on," state for a certain length of time, the glass will remain conducting even after applied voltage has been completely removed. In this state, which is now known to be due to a filament of crystalline material, the off state can be restored by applying a suitable pulse of current, which dissolves the filament and re-forms the glass.

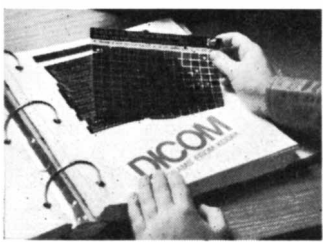
Reading, writing, and what?



Just in case on some grim day in their future they push the buttons and nothing happens, "math skills" are still taught in the third grade.

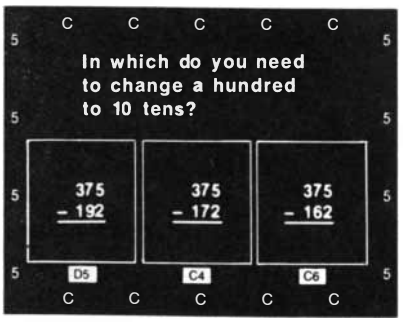
Right or wrong, modern pedagogy refrains from recommending flogging for incorrect answers in arithmetic. The modern pedagogue, a kindly person who first eliminates boredom, inattention, and plain naughtiness as explanations of the error, seeks to spot some misconception of principle. Just where is the break in the chain of concepts?

To help answer, a diagnostic method worked out with Educational Testing Service* of Princeton, N.J. is being introduced across the country. Now join the third grade and pay attention:

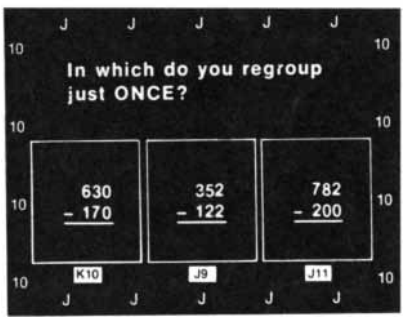


Your friend Sylvester is initiating you. Sylvester claims he breezed right through the 4th blue microfiche. He explains that when you finish you see a frog with the number 7, which you write down and hand in to Mrs. Brown. So you take out the 4th blue microfiche, put it in the reader, turn on the light inside, and shift the microfiche to Row A, Column 1.

- At A-1 you are told to shift to B-1.
- There, if you are quite sure that 1 hundred = 10 tens, you are directed to C-1.
- Easy choice. You know that 1 hundred, 3 tens = 13 tens, so you go to C-2.
- This one is tricky. 4 hundreds, 2 tens does not = 42, nor is it the same as 5 hundreds, 12 tens. Deciding that it is the same as 3 hundreds, 12 tens, you now go to G-2 and from there to a lot of easy subtractions.
- Now at C-5 you are confronted with this:



- You are getting a little bored and make a quick decision for C-4, which leads you to more subtraction. They get harder, but you are paying good attention now. Then at J-10 comes this horrible question:



- Is it J-11?

At J-11, which is also about regrouping, you are guessing again. You guess K-8 which turns out to be the frog. You're done! When you report the frog's number to Mrs. Brown her key will tell her that you understand enough about subtraction to regroup tens as ones, that that's more than Sylvester understands, but that you fail to recognize whether regrouping is required once or twice. Tomorrow, during individual-project time, she will enlighten him and then get you out of your own, slightly higher-level hangup.**

*Which also writes and administers the Scholastic Aptitude Tests (SAT).

**Never mind that your grandfather would probably also exit at Frog 8. He didn't learn to regroup either. He called it "borrowing," but in those days kids had to learn arithmetic instead of mathematics.

SCIENCE AND THE CITIZEN

Amended Dogma

The unraveling of the molecular biology of the gene has given rise to several "dogmas." Perhaps the most basic dogma is expressed by the phrase "One gene, one protein," which states that the information required for the synthesis of each protein is encoded in a gene unique for that protein; given the sequence of codons (groups of three coding units) that represents a gene in the linear molecule of DNA holding the organism's genetic blueprint, one can deduce the unique sequence of amino acids in the protein specified by that gene. Although the dogma has not been repealed, it has been distinctly amended by the recent work of Frederick Sanger and his colleagues at the British Medical Research Council Laboratory of Molecular Biology in Cambridge.

For some time evidence had been accumulating that the information contained in the DNA of one of the smallest bacterial viruses, designated ϕX174, was insufficient to code for the nine kinds of protein molecules the virus is able to manufacture when it invades a host cell. Although the exact roles of the various proteins have not been established, some are involved in replicating the virus's DNA and others in providing the capsid, or package, in which the DNA is tightly enfolded. The finished virus is an icosahedral structure about 25 nanometers in diameter. When estimates of the total number of amino acids in the nine protein products were compared with the number that could be encoded by the single-strand DNA molecule in the virus, it appeared that the DNA molecule was 10 to 15 percent too short. The DNA was known to consist of about 5,400 nucleotides. A sequence of three nucleotides, a codon, is needed to specify each amino acid. The codons are designated by triplet combinations of the four letters *A, T, G* and *C,* which stand for adenine, thymine, guanine and cytosine, the substituents that distinguish the four nucleotides. All together there are 64 codons, which represent all possible combinations of the four letters taken three at a time. Each codon (except for three "nonsense" codons) specifies one of the 20 amino acids that serve as the building blocks of proteins. Since there are 61 "sensible" codons, a particular amino acid is typically identified by more than one codon. Thus the amino acid valine can be specified by any one of four codons: *GTA, GTC, GTG* or *GTT.* On the other hand, methionine is represented by only one codon, *ATG,* which also serves as an initiator, signaling the start of a protein chain.

Conclusive evidence that the DNA of ϕX174, if it is translated in normal fashion, consists of too few nucleotides to account for the virus's nine proteins was provided last fall when Sanger and his co-workers published the complete nucleotide sequence of ϕX174 and simultaneously cleared up the mystery of the missing genetic information. When they compared the nucleotide sequences of the DNA and the amino acid sequences of the corresponding proteins, they found that in two cases the code for a second protein was embedded in the nucleotide sequence for a larger protein. Thus the code for gene product *B* was embedded in the code for gene product *A,* and the code for gene product *E* was embedded in the code for gene product *D.* In other words, with two genes that would normally yield only two proteins, ϕX174 is able to make four distinctly different proteins.

Moreover, in each case the smaller protein is not simply a truncated portion of the larger one; it represents a completely different reading of codon triplets achieved by a phase shift. For example, the sequence of nucleotides that is parsed as *... G, AAG, TTA, ACA...* to specify the amino acids lysine, leucine and threonine in protein *A,* is read from a point one nucleotide earlier in the synthesis of protein *B,* yielding the codons *... GAA, GTT, AAC, A...,* which specify the amino acids glutamine, valine and asparagine. It seems remarkable that a stretch of 360 nucleotides (120 codons) can be carved out of the 1,536 nucleotides (513 codons) of gene *A* and, with a phase shift of one, give rise to a viable protein *B,* totally unlike any portion of protein *A.*

Until Sanger's group discovered the existence of such phase-shift readings apparently no one had entertained the possibility. In retrospect there were several good reasons for ignoring it. In virtually every organism for which one can compare DNA length and total length of protein product there is usually an apparent excess of DNA. What has puzzled investigators is not the lack of DNA but rather the presence of so much "silent" DNA. The exception presented by ϕX174 did not seem serious as long as precise information was lacking.

Also militating against the possibility of dual reading was the existence of the three nonsense codons, *TAG, TAA* and *TGA,* which, having no amino acid counterparts, often act as chain terminators. Hence in any completely random stretch of 64 codons one would expect to find three nonsense codons. In the phase-shifted reading of genes *A* and *B* in ϕX174 there are no nonsense codons in 633 overlapping triplets. Even if one had imagined such a possibility, it would seem even more improbable that a single phase-shift reading would give rise to a second viable protein when a single mutation in a gene (the change of one nucleotide in several hundred) commonly gives rise to a protein so defective that it cannot function.

In spite of the amendment of the one-gene, one-protein hypothesis it is not believed the production of more than one protein by one gene is a common occurrence. It is suspected that it may have arisen in the case of ϕX174, and perhaps in similar small viruses, because the protein capsid places an absolute limit on the virus's DNA content. If the virus is to develop any new functions in the course of its evolution, it can do so only if additional information is extracted from a fixed amount of DNA.

American Dilemma

It has been clear for some years to those who stopped to think about it (although few did stop to think about it) that U.S. higher education could not continue indefinitely to grow at a substantially faster rate than the U.S. population and economy as a whole. The growth accelerated during the 1960's, both to accommodate the postwar-baby generation and to take advantage of increased Federal funding, particularly for the natural sciences. Large numbers of doctoral-level scientists and engineers were hired by the universities, distorting the age distribution of faculties: in 1973 some 23 percent of the Ph.D. scientists and engineers in four-year colleges and universities were between 30 and 34 years old. Today most of them hold tenured positions. The result, points out the *Public Interest Report* of the Federation of American Scientists, is the phenomenon of "tenure block": a disproportionately large group of young, tenured faculty members with a long way to go before retirement.

The combination of tenure block and decreasing higher-education enrollment has already produced a sharp decrease in the number of junior-faculty openings; for a few years in the mid-1980's, according to a study by the late Allan M. Cartter, there will be just about no openings at all. Meanwhile the output of Ph.D.'s continues to increase as the postwar generation emerges from the graduate schools. The result of all these demographic processes, the F.A.S. observes, is that the new scientist who wants a tenured position at a first-class university "will find, in the next decade or two, an extraordinary paucity of openings." On the assumption that a society benefits from advances in scientific knowledge and that a steady flow of young workers with fresh ideas is crucial, the F.A.S. urges that some kind of

The Peugeot 604 gives you the engineering of a Mercedes-Benz, the handling of a BMW, the elegance of a Jaguar, and a level of comfort superior to all of them.

It's one thing to build a car in a class with three of the world's most sophisticated sedans. It's quite another to build a car that in any way achieves superiority over them.

In building the Peugeot 604, however, we've done precisely that.

Like the Mercedes 280 E, it has fully independent suspension, a responsive six-cylinder engine (ours is a V-6), a unitized body with thousands of welds, impact-absorbing crush zones front and rear, anatomically designed seats, and meticulous attention to detail.

Like the BMW 530i, it has power-assisted rack and pinion steering, four-wheel power disc brakes, stabilizer bars, McPherson struts, and a head-spinning ability to make tight turns.

Like the Jaguar XJ6L, it has elegant styling (ours is by Pininfarina), air conditioning, AM/FM stereo, power windows, and fully-reclining seats (we also offer an electric sunroof and genuine leather upholstery).

But unlike all of them, the Peugeot 604 has been engineered for a superior level of comfort. With oversized shock absorbers patented by Peugeot. With large coil springs and long vertical wheel travel to keep the car from practically ever bottoming out.

With a "floating" differential and a steel torque tube enclosing the drive shaft to reduce noise and vibration—a major source of passenger fatigue. And with seats that are tuned to the suspension.

All these components have been engineered to work in harmony, so the car is obviously comfortable when you first ride in it. And incredibly comfortable when you've been riding in it all day.

But comfort isn't the only thing separating the 604 from its competition. There's also its suggested retail price, which starts at around $11,000 P.O.E. And that may be its most comforting feature of all.

PEUGEOT

No one builds cars the way we build cars.

For information on buying, leasing, or overseas delivery, see your Peugeot dealer. Or call 800-243-6000 toll-free (in Conn., call 1-800-882-6500). Peugeot Motors of America, Inc., Lyndhurst, N.J. 07071.

"holding pattern" is needed to keep the very best young scientists at work during the 1980's, after which the tenured professors in the "demographic bulge" will begin to move toward retirement.

One possibility, the F.A.S. suggests, would be a system of highly competitive five-year fellowships (renewable once) for junior faculty members. A fellowship-holder might negotiate for a tenured position in about his sixth year at a university—approximately when tenure is ordinarily achieved. If tenure was granted, the university would still not have to pay the new faculty member for another four years; if it was not granted, the scientist would at least have a four-year cushion against being disemployed. Another proposal, said to be under consideration by the National Science Foundation, contemplates giving research fellowships to senior faculty members that would free them from teaching and remove them from the university's payroll—provided the university applied the money thereby saved to the hiring of young Ph.D.'s. Neither of these schemes, the F.A.S. points out, would significantly affect the large oversupply of scientists and engineers who will be receiving doctorates in the 1980's: "What we are talking about ... is ... the most scientifically talented 100 or 200 persons [entering the job market each year]."

Armed Camp

Even with the effects of inflation removed, the amount of money the nations of the world spent for military purposes rose from $285 billion in 1966 to $345 billion in 1975. At the end of the decade about 5.5 percent of the total of the world's gross national product was devoted to military expenditures. Moreover, the international trade in arms (mainly in the form of exports from the developed countries and imports by the developing countries) rose from $7.1 billion in 1966 to a peak of $10.5 billion in 1973, declining thereafter to $9 billion in 1975. These data appear in *World Military Expenditures and Arms Transfers 1966–1975,* published by the Arms Control and Disarmament Agency.

To eliminate the effect of inflation the agency made its calculations in terms of prices as they stood in 1974. Examining the trend of military expenditures by region and by country, the agency found that "the rate of growth appears to have been decelerating since 1969 in the developed countries and to have been accelerating in the developing countries." Expenditures in the North Atlantic Treaty Organization (NATO) countries have declined since reaching a peak in 1968; the change is mainly attributable to the decrease over the same period of U.S. military spending (measured at 1974 price levels) from $114 billion in 1968 to $83.3 billion in 1975. Expendi-

tures by the Warsaw Pact nations "have continued their steady climb." Members of the Organization of Petroleum Exporting Countries (OPEC) "have been spending at an increasing rate."

Expressed in terms of the percentage of gross national product that is devoted to military expenditures, the major nation with the highest figure in 1975 was Israel (34.6 percent) and the one with the lowest was Japan (.94 percent). For the U.S. the figure in 1975 was 6 percent. No firm figure is given for the U.S.S.R. because of the difficulty of making fully comparable calculations, but the agency estimated that the Russian military burden as a fraction of the gross national product is "in the range of 11 to 13 percent." The agency also made a comparison of the "relative burden of military expenditures" in 1975 by grouping nations according to their per capita gross national product [*see illustration below*].

As for the arms trade, the agency found that in general worldwide deliveries of arms have decreased since the peak years of 1972 and 1973. In 1975 the nation with the largest volume of arms exports, expressed in dollars held constant at the 1974 level, was the U.S. ($4.4 billion), and the nation with the highest volume of arms imports was Iran ($1.02 billion). The agency noted, however, that one should interpret arms-trade figures with caution, since "the relative economic value of arms to supplier and recipient may be considerably different" and much of the trade involves financial arrangements other than direct payment.

Genetic Disarmament

A large number of the "worst-case scenarios" put forward to exemplify the concerns of some observers about the potential hazards presented by research involving the propagation of recombinant-DNA molecules in bacterial hosts focus on the possibility that a new pathogenic form of the host bacterium might be created inadvertently by experimenters working with the novel gene-splicing technique, that such a "rogue" strain might escape accidentally from the laboratory, that it might manage to survive and multiply in the natural environment (or find a way to transfer its pathogenicity to another microorganism), that the newly created pathogen might prove to have acquired somewhere along the way an extraordinary resistance to standard procedures available for its eradication and that it might therefore lead to an uncontrollable epidemic among human beings or other organisms.

Such hypothetical sequences of events, considered by most epidemiologists and other experts on infectious diseases to be exceedingly improbable, have been loosely characterized as *Andromeda Strain* scenarios, after the popular science-fiction novel of a few years back. In spite of the very low probability that the conjectures will ever be realized, they have proved difficult to dismiss entirely, partly because of certain unanswered questions about the biology and genetics of the host cell that has been generally adopted for such research: a strain of the ubiquitous intestinal bacterium *Escherichia coli*.

It was because of these uncertainties—and the public anxiety they might reasonably be expected to engender—that investigators engaged in the early stages of this line of research, meeting at the International Conference on Recombinant DNA Molecules held at Asilomar in California in February, 1975, agreed that most (but not all) of the then current work on recombinant DNA should proceed only if "appropriate safeguards, principally biological and physical barriers adequate to contain the newly created organisms, are employed." The specific containment procedures recommended by the participants in the Asilomar conference formed the basis of the more detailed and somewhat stricter set of guidelines

Military outlays as percentage of gross national product	Per capita gross national product							
	Less than $100	$100 to $199	$200 to $299	$300 to $499	$500 to $999	$1,000 to $1,999	$2,000 to $2,999	More than $3,000
10 or more	Kampuchea (Cambodia) North Vietnam	South Vietnam	China Egypt	Jordan North Korea	Syria	Iran Iraq	Israel Oman	U.S.S.R
5 to 10	Chad Laos Somalia	Pakistan Yemen (Sana)	Yemen (Aden)	Albania South Korea	China (Taiwan) Cuba Malaysia Mongolia	Bulgaria Portugal Romania	Greece Hungary Poland Singapore	Czechoslovakia East Germany Qatar Saudi Arabia U.S.
2 to 4.9	Burundi Ethiopia Mali Rwanda Upper Volta	Burma Central African Rep. Guinea India Indonesia Tanzania Zaïre	Equatorial Guinea Mauritania Sudan Uganda	Bolivia Congo Morocco Nigeria Philippines Rhodesia Thailand	Algeria Angola Brazil Chile Guyana Nicaragua Peru Turkey Uruguay Zambia	Argentina Cyprus Lebanon South Africa Yugoslavia	Bahrain Italy Spain Venezuela	Australia Belgium Canada Denmark France Kuwait Netherlands Norway Sweden U.K. West Germany
1 to 1.9	Afghanistan	Benin Haiti	Cameroon Honduras Kenya Madagascar Togo	Colombia El Salvador Ghana Senegal	Dominican Republic Ecuador Guatemala Ivory Coast Paraguay Tunisia		Gabon Ireland	Austria Finland Libya Luxembourg New Zealand Switzerland
Less than 1		Bangladesh Gambia Lesotho Malawi Nepal Niger Sierra Leone	Sri Lanka	Botswana Liberia Mozambique Swaziland	Costa Rica Fiji Mauritius Mexico Panama	Barbados Jamaica Malta Surinam	Trinidad and Tobago	Iceland Japan United Arab Emirates

The relative burden of military expenditures around the world, 1975

issued last June by the National Institutes of Health to govern the conduct of all NIH-supported research on recombinant-DNA molecules. The NIH guidelines, now in effect for all Federally supported experimentation in this area, are in turn expected to be transcribed soon into Federal legislation extending the dual containment requirements to all recombinant-DNA research.

In their present form the NIH guidelines call for four levels of physical containment, designated *P*1, *P*2, *P*3 and *P*4, which are matched to various prospective experiments estimated to represent four corresponding levels of risk. In addition the NIH established rules for biological containment, a new concept in laboratory safety based largely (in the words of the organizing committee of the Asilomar conference) on the "realization that special bacteria and vectors which have a restricted capacity to multiply outside the laboratory can be constructed genetically, and that the use of these organisms could enhance the safety of recombinant-DNA experiments by many orders of magnitude." The NIH guidelines define the specific criteria for biological containment in terms of three different classes of host-vector systems, designated *EK*1, *EK*2 and *EK*3, which have progressively diminishing probabilities of survival outside the laboratory.

The *EK*1 host-vector system is simply the standard laboratory-adapted strain of *E. coli,* designated *K*-12, used in conjunction with either of two recombinant-DNA cloning vectors: the bacterial virus known as bacteriophage lambda or certain nonconjugative plasmids, small extrachromosomal loops of bacterial DNA that are considered incapable of transferring genes between cells by the process known as conjugation.

The *K*-12 strain of *E. coli* has been cultivated exclusively in microbiological laboratories since 1922, when it was isolated from a human patient at the Stanford University Medical Center. The strain has been widely used for experiments in genetics and molecular biology since the 1940's, and as a result more is known about its biological and genetic properties than about any other living organism. In the course of its long sojourn in the laboratory the *K*-12 strain has undergone numerous mutations, becoming "semiaddicted" to the special conditions and nutrients of its artificial environment. Accordingly *K*-12 has become quite dissimilar from the dozens of "wild" strains of *E. coli* that normally inhabit the intestinal tract of warm-blooded animals. For example, in experiments in which *K*-12 has been fed to healthy, well-nourished animals, including human beings, it has not been possible to show that the laboratory strain has successfully colonized the intestinal tract in competition with its more robust wild relatives.

Because of the special properties of the *K*-12 strain of *E. coli* and of its prescribed cloning vectors, these systems have been judged to provide an adequate measure of biological containment for recombinant-DNA experiments generally considered to present little or no potential danger. (The designation *EK*1 applied to such host-vector systems is derived from the *E* in *Escherichia* and the *K* in *K*-12.)

For experiments in which the potential hazard is estimated to be higher the NIH guidelines stipulate higher levels of biological containment than those afforded by the *EK*1 host-vector systems. Thus an *EK*2 host-vector system is one in which genetic modifications of the host and/or the vector are shown in a wide range of laboratory tests to reduce the survival probability of a recombinant-DNA molecule that has escaped to the outside world to one in 100 million.

Work on the development of suitable *EK*2 host-vector systems began almost immediately after the Asilomar meeting in early 1975. By January, 1976, a group of experimenters led by Roy Curtiss III of the University of Alabama completed the construction of a new *E. coli* strain, derived from *K*-12, that incorporated 15 separate genetic defects de-

signed to make it a safer and more useful host cell for recombinant-DNA experiments. According to Curtiss, his team's strategy was three-pronged: "First, we introduced a constellation of mutations that resulted in the destruction of the bacterial cell and its genetic information if it should attempt to grow outside its carefully controlled test-tube environment. Second, we introduced another constellation of mutations that caused the strain to be extremely sensitive to various environmentally encountered substances or physical environments that would cause cell death independent of the attempts of the organism to grow. Third, we introduced a constellation of mutations that reduced the likelihood of transmission of recombinant DNA to other microorganisms that could be encountered in nature."

The new disarmed *E. coli* strain fabricated by Curtiss and his co-workers was named χ1776 "in celebration of the bicentennial." Last November, NIH director Donald S. Fredrickson certified the χ1776 strain as meeting the *EK*2 host-vector requirements when it is used in conjunction with a set of improved nonconjugative plasmid cloning vectors developed by Herbert W. Boyer of the University of California at San Francisco and Donald R. Helinski of the University of California at San Diego.

More recently Frederick R. Blattner of the University of Wisconsin, Philip Leder of the NIH and Philip Sharp of the Massachusetts Institute of Technology and their colleagues have independently designed and constructed safer bacteriophage-lambda vectors, and Curtiss' group has made another *E. coli* host strain (designated χ1953) for use with several of these lambda vectors. Some of these newer host-vector systems have already been approved by the NIH as meeting the *EK*2 standards and others are pending approval.

In testifying before a Congressional subcommittee last month on the effectiveness of biological containment in recombinant-DNA research Curtiss remarked: "χ1776 and a number of the other components of *EK*2 host-vector systems have been criticized by some members of the scientific community as not being sufficiently perfect.... It should be pointed out that the development of biological containment systems is in its infancy and that improvements are continually being made as new information is obtained. As for the safety tests not yet done, these are part of *EK*3 testing, which is just commencing."

Although the results of the ongoing *EK*3 tests may serve to dispel some of the concerns embodied in the current proliferation of *Andromeda Strain* scenarios, they are unlikely to satisfy those critics of recombinant-DNA research who view such experiments as the "thin end of the wedge" leading to the deliberate misuse of recombinant-DNA techniques for evil purposes, principally in the area of human genetic engineering. Indeed, it is difficult to see what evidence of a scientific character can be brought to bear on such fears, which (to borrow again from science fiction) might be grouped under the heading of *Brave New World* scenarios.

Ambivalent Leptons

The electron and the muon have long seemed an odd couple. They are so much alike—differing only in mass—that it is hard to understand why two independent particles should exist. Either one alone would seem to suffice. On the other hand, since there are two particles there is no apparent reason why there should not be more. Some recent rumors and speculations offer hope that these peculiarities will be explained. They suggest that the distinction between the electron and the muon may not be absolute. They also suggest that the electron and the muon may be not an isolated pair of particles but the first members of an extensive family.

Both the electron and the muon are leptons: particles with no measurable size and no evidence of an internal structure. The only other known leptons are

two types of neutrino (one each for the electron and the muon). In addition to lacking all the properties the electron and the muon lack, the neutrinos also lack electric charge and all (or almost all) mass.

The muon is more than 200 times more massive than the electron, but otherwise the two particles are indistinguishable. They are not, however, interchangeable: the muon cannot simply give up its extra mass and become an electron; some indefinable essence ("mu-ness") must first be carried off. The only confirmed mode of decay for the muon is transformation into an electron by the emission of a muon-type neutrino (carrying off the mu-ness) and an electron-type antineutrino. Other possible modes of decay, widely discussed in recent months and perhaps observed experimentally, imply a violation of the conservation of mu-ness.

The possible observation of an anomalous muon decay was reported by a group of seven investigators at the Swiss Institute for Nuclear Research (SIN) near Zurich. With the 590-million-electron-volt cyclotron at SIN they detected half a dozen events (out of several billion) in which a muon seemed to decay into an electron and a gamma ray, or high-energy photon. The gamma ray can carry off the mass of the muon but not the property of mu-ness.

The Swiss investigators deferred publication or public announcement of these preliminary findings until the experiment could be repeated and more data could be collected. Nevertheless, news of the possible discovery prompted immediate theoretical speculation on the experiment's significance and consequences. It now seems far from certain that anomalous muon decays have actually been observed at SIN, and attempts to reproduce the result elsewhere have so far been unsuccessful, but interest in the topic remains strong. The Swiss investigators still have not published a report of their findings, but theoretical interpretations of them have begun to appear. Those interpretations involve radical adjustments to the physicist's model of the universe. They introduce the possibility of a new fundamental force in nature, they attribute mass to the neutrinos and they postulate new leptons.

One of the interpretations is that of Steven Weinberg of Harvard University and James D. Bjorken of the Stanford Linear Accelerator Center (SLAC). Writing in *Physical Review Letters,* they consider a model of the leptons in which the conservation of mu-ness is not a basic law of nature but instead arises as an almost accidental feature of a large class of physical theories. Events such as the decay of the muon into an electron and a gamma ray would not be forbidden, but they would be suppressed; they could not proceed through any of the four known basic forces of nature (the strong, electromagnetic, weak and gravitational forces). The observation of such events might then imply the existence of a fifth force, unaffected by the suppression, which has been given the name "superweak." An interesting feature of this model is that the suppression appears only if at least one type of neutrino has a mass greater than zero. That is certainly possible; experimental upper limits on the mass of the neutrinos are not very stringent.

Another approach to the problem of decays that change mu-ness has been taken by T. P. Cheng of the University of Missouri and Ling-Fong Li of Carnegie-Mellon University. Their comments, also published in *Physical Review Letters,* were written before the SIN findings became known. They point out that the probability of events such as the decay of a muon into an electron and a photon would be greatly enhanced by the existence of additional leptons. Such new leptons would probably include particles much more massive than the electron and the muon, and new neutrinos as well.

Evidence suggesting the possibility of additional leptons has been accumulating for more than a year in experiments at several laboratories. At SLAC, for example, a small but consistent discrepancy that may signal a new lepton has been observed in the products of collisions between electrons and positrons. The discrepancy consists of a few extra muons that cannot readily be accounted for by other mechanisms. The interpretation of these events is uncertain, but a hypothesis of growing plausibility is that the muons arise from the decay of an electrically charged lepton with a mass equivalent to between 1.6 and two billion electron volts. At other laboratories quite different experiments have suggested new leptons with other masses, perhaps including a massive electrically neutral lepton.

If the leptons should turn out to be a large family of particles, counting and cataloguing them could become a major preoccupation of physicists over the next decade. An experiment that might provide an important early clue to their number has been proposed by A. K. Mann and Henry Primakoff of the University of Pennyslvania. The experiment might also place a new upper bound on the masses of the neutrinos. It is an ambitious plan for physics on a continental scale.

Mann and Primakoff propose to search for "neutrino oscillations," spontaneous conversions of one neutrino type into another. Such oscillations are not decays of one particle yielding the other but are reversible changes of quantum-mechanical state. In a sense they are manifestations of an underlying ambiguity in the identity of the neutrinos. If the oscillations are possible, then a pure beam of muon-type neutrinos

must become increasingly contaminated with neutrinos of the electron type. No matter what the original composition of the beam, it must eventually come to be an equal mixture of both neutrino types (or of all neutrino types if there are more than two).

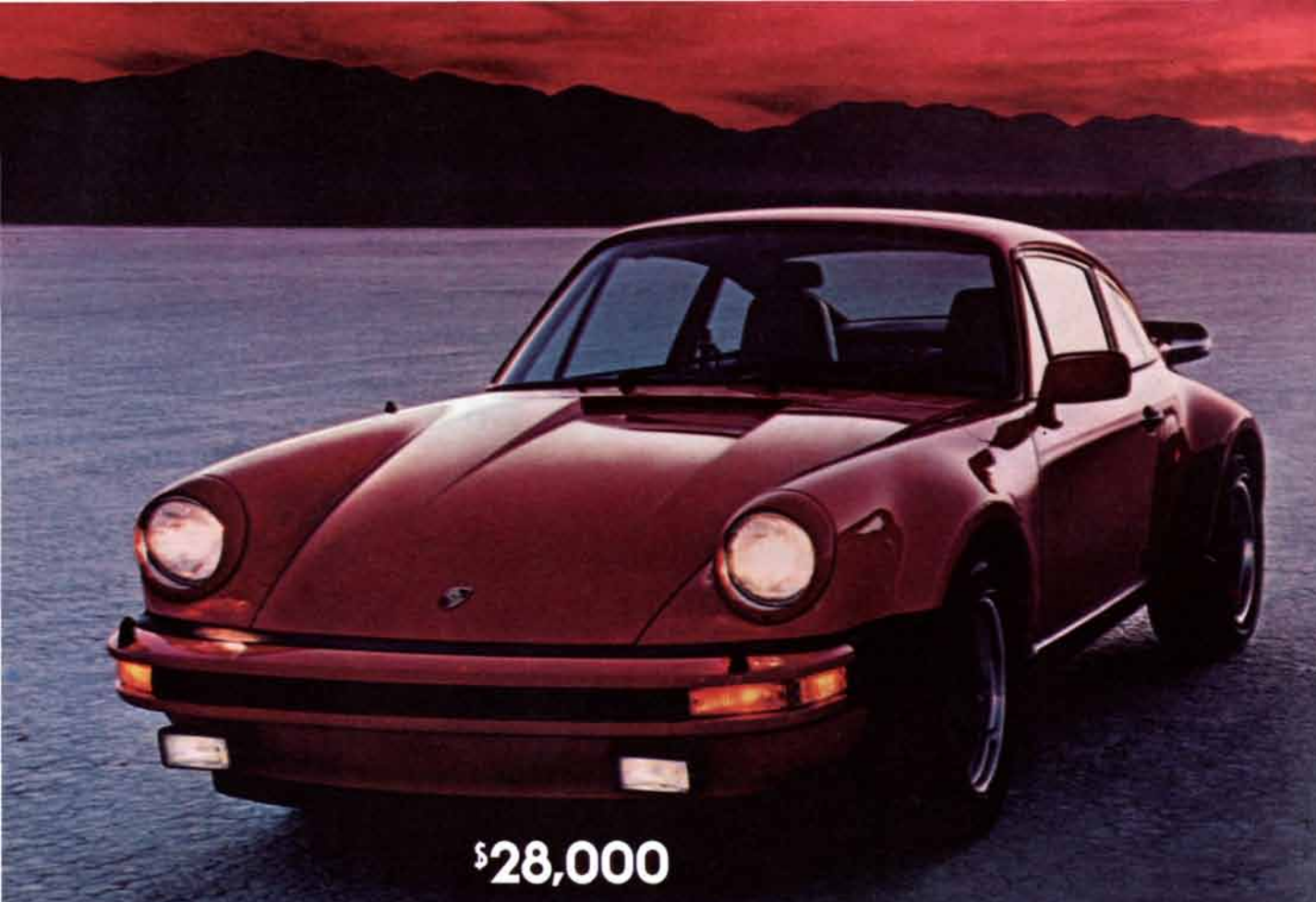
Neutrino oscillations are expected only if certain constraints on the theory of leptons are met; to that extent they test the theory. The oscillations should be possible only if the conservation of mu-ness is not absolute. In other words, the muon-type neutrino can be freely converted into an electron-type neutrino only if a muon can also be converted into an electron. Another condition is that the masses of the neutrinos cannot all be zero, and if they are greater than zero, they cannot all be the same.

If neutrino oscillations could be detected, the proportion of each type of neutrino in a beam that has reached equilibrium would provide a simple measure of the number of neutrino types. For example, if the muon type makes up half of the beam, then there are just the two known types; if it makes up a third, there is another kind of undiscovered neutrino. Since neutrinos and massive leptons seem to come in pairs, the proportion would also estimate the total number of leptons.

The difficulty of measuring neutrino oscillations is one of scale. The oscillations are expected to be rather slow, and a neutrino, moving at virtually the speed of light, would travel a considerable distance before the probability of an oscillation became significant. An appropriate length for measuring the oscillations might be 1,000 kilometers. Because of the neutrino's distinctive lack of properties and its reluctance to interact with matter such a grandiose experiment might be feasible. Mann and Primakoff propose to direct a high-energy neutrino beam, such as the one generated at the Fermi National Accelerator Laboratory, into the ground at a shallow angle. The neutrinos would then pass unhindered through the solid earth. A detector, discriminating between muon-type and electron-type neutrinos, would be placed 1,000 kilometers away, where the beam emerged from the ground at the same shallow angle.

The Thorium Option

It is generally acknowledged that if nuclear fission is to play a major role in supplying the world's long-term energy needs, a new generation of nuclear reactors capable of "breeding" more fissionable fuel than they consume must be developed and put into widespread commercial operation by about the end of this century. The problem is that present-day power reactors are designed to operate in a "once through" fuel cycle that extracts only a small fraction of the energy available in natural uranium ore,



THE DIFFERENCE BETWEEN OUR TWO CARS IS APPARENT. BUT LET US TELL YOU ABOUT THE SIMILARITIES.

Similarities between a $28,000* sports car and a $5,445* sedan?

It's not as far-fetched as it sounds, really. Not when you consider that both our cars are designed to achieve the highest degree of performance, efficiency and comfort. Which results in similar solutions to many automotive engineering challenges:

They both have CIS fuel injection systems. This provides optimum combustion, resulting in smoother performance, easier cold weather starts and, most important of all, better mileage. Indeed, according to EPA estimates, no 5 passenger sedan gets better mileage than the Audi Fox —36 mpg on the highway and 24 in the city. The Porsche Turbo Carrera

*Suggested 1977 retail price. P.O.E. Transportation, local taxes, and dealer charges additional.

gets 24 mpg on the highway and 14 in the city which is truly extraordinary when you consider it's the fastest production car Porsche ever built. (EPA estimates for standard shift. Mileage may vary depending on condition of the car, optional equipment, as well as how you drive.)

Also, both cars utilize a unique transaxle system with the engine mounted over the drive wheels. This not only reduces the weight (because there's no long drive shaft), but it keeps the transmission and drive train in one place for even power transfer and it concentrates the weight over the drive wheels for added traction.

We insist, of course, that both cars have such high performance features as rack-and-pinion steering, overhead cam engines, and high strength, lightweight construction.

But we feel that what's more important than any specific feature is a common dedication to engineering excellence and innovation. A dedication that has been a hallmark of Porsches and Audis through the years.

A dedication that has a lot to do with pride. Regardless of price.

DRIVE ONE OF OUR LEGENDS

PORSCHE + AUDI

and continued reliance on reactors of this type would seriously deplete the world's economically recoverable uranium reserves in a matter of decades.

The breeder reactor promises to relieve the uranium-supply situation by generating new fissionable fuel in a "breeding blanket" of naturally nonfissionable material surrounding the reactor's active core. A "fertile" atomic nucleus in the breeding material is transmuted into a fissionable nucleus by capturing one of the multitude of stray neutrons produced by the chain reactions of fissionable nuclei in the core. Thus comparatively abundant isotopes such as uranium 238 and thorium 232 can be transmuted respectively into the fissionable fuels plutonium 239 and uranium 233. The new fuels would then be available to supplement the supply of the rare isotope uranium 235, the only readily fissionable material found in nature. In this way the world's potential reserves of nuclear fuel would become virtually inexhaustible.

In a number of countries around the world plans are going forward to separate plutonium from the spent fuel of existing reactors by chemical means and to recycle the plutonium as fresh fuel first in light-water reactors and later in breeder reactors. So far the nations that also have large-scale breeder-development programs under way (the U.S., the U.S.S.R., France, Britain, West Germany and Japan) have concentrated their efforts almost exclusively on a single approach: the breeding of fissionable plutonium 239 from fertile uranium 238 in liquid-metal-cooled fast breeder reactors (LMFBR's).

The commitment to a future "plutonium economy" implied by these trends has been questioned on various grounds, among which perhaps the most serious is the concern that with large amounts of reprocessed and freshly bred plutonium coming into worldwide commercial circulation certain nations or terrorist groups could divert some of this readily fissionable material for the purpose of building nuclear weapons. At the very least, it is argued, attempts to safeguard weapons-grade plutonium against national diversion or criminal theft will require much more elaborate and expensive international controls than those now in force for protecting the smaller amounts of plutonium mixed in with highly radioactive fission products in the spent fuel from today's reactors.

The concern that the spread of plutonium-breeding power reactors might therefore encourage the proliferation of nuclear weapons has led two Princeton University researchers, Harold A. Feiveson and Theodore B. Taylor, to propose the consideration of a more "proliferation resistant" alternative fuel cycle for the breeder reactors of the future. In their scheme thorium 232 (the only form of thorium found in nature)

would be substituted for most of the uranium 238 as the principal fertile isotope in the breeder. The resulting material would consist of the fissionable isotope uranium 233, mixed with enough uranium 238 to "denature" it, that is, to make it unsuitable for use in a nuclear weapon without further isotope-separation, a complex and expensive physical process currently within the means of only a few countries. No such denaturing isotope exists in significant quantities for plutonium.

Feiveson, a political scientist formerly associated with the Arms Control and Disarmament Agency, and Taylor, a physicist and former nuclear-weapons designer, have outlined their proposed thorium-based breeding cycle in some detail in a study prepared for the 1980's Project of the Council on Foreign Relations. According to their approach, the fresh fuel supplied to various national power reactors would contain no weapons-grade material. This fuel would consist of a mixture of roughly one part uranium 233, eight parts uranium 238 and between 50 and 80 parts thorium. Present types of reactors, they state, could be converted to operate on this fuel mixture. The fabrication of the fresh fuel assemblies and any necessary denaturing of the uranium 233 could be done at internationally controlled, regional centers. Such centers would also undertake the reprocessing of spent fuel from the thorium reactors. The regional reprocessing centers, they add, "would, of course, require physical security against theft of plutonium by criminals or terrorists. But fresh fuels shipped to national power plants and spent fuel shipped back to the regional centers would not be attractive targets for theft: the fresh fuel is not of weapons grade, and the spent fuel is so highly radioactive and so diluted as to be essentially self-protecting against theft and the subsequent processing required to make nuclear explosives."

Assuming a fuel cycle in which recycled uranium 233 is denatured with uranium 238, it seems unlikely that current types of nuclear power plants could breed enough uranium 233 to be self-sustaining on a thorium cycle. Nevertheless, Feiveson and Taylor maintain, "they could be part of an overall system that *is* self-sustaining if some additional source of uranium 233 could be provided at the same internationally controlled regional centers at which the fuel is reprocessed." Two such possibilities for supplementing the uranium 233 in the proposed thorium fuel cycle are discussed in the study; both would involve converting thorium 232 into uranium 233 at the regional site, either in plutonium-burning fast breeder reactors or in the intense neutron environment of fusion "converter" reactors.

The preliminary study of the thorium option conducted by Feiveson and Taylor indicates that its characteristics depend very strongly on the assumed growth rate of nuclear power. In general, they say, "the economics and room for flexibility in the cycle look more attractive for low projections of nuclear-power growth rates than for high ones."

In arguing for a thorough reassessment of the available nuclear-fuel-cycle options, Feiveson and Taylor are careful to point out that "we do not view the thorium cycle or even a complete phase-out of civilian nuclear power as alternatives to the plutonium economy that would remove the risks of nuclear-weapons proliferation. Huge quantities of nuclear materials suitable for military purposes, and of course nuclear weapons themselves, will still have to be securely protected from theft. Furthermore, pressures for nations to acquire nuclear weapons will persist at least as long as the nuclear-weapons states continue to behave as though they feel more secure with nuclear weapons than without them. We see no practical way technically to *prevent* any nation that really wants nuclear weapons from acquiring them."

In the case of their own proposed "alternative fission future" the Princeton authors state: "There are a great number of technical and institutional issues involved in the implementation of a thorium cycle that are unresolved; we do not presume that upon further analysis the thorium option will necessarily look very attractive, either from a proliferation perspective or from other perspectives. Rather, we wish to make the case that, given certain attractive reactor engineering characteristics of a thorium cycle and the fact that uranium 233 and uranium 235 can be denatured, whereas plutonium cannot, the present worldwide momentum toward a uranium-plutonium breeding cycle in preference to a thorium option should at least be thoroughly reexamined before irreversible commitments are made to the plutonium economy."

From Bronze to Steel

How did the Bronze Age end and the Iron Age begin? A widely received hypothesis connects the transition with shadowy disorders in the eastern Mediterranean, beginning in about 1200 B.C., that saw, among other violent events, the overthrow of the Mycenaean city-states of Greece by mysterious raiders known as the "Sea People." Deprived of access to tin, and perhaps also copper, the hypothesis goes, the smiths of the period were forced to make greater use of iron, a familiar and abundant metal that had previously been smelted for ornamental purposes rather than utilitarian ones. Working with iron because they had no choice, the smiths soon discovered its many advantages over bronze. As a result, although the ingredients for bronze again became available in about 900 B.C., the older metal never regained its former position.

This picture cannot be correct, argues Robert Maddin, professor of metallurgy at the University of Pennsylvania. First of all, bronze is superior to iron as a material for tools and weapons. Iron will deform under stresses of between 20,000 and 30,000 pounds per square inch, whereas bronze does not until the stress is between 60,000 and 70,000 p.s.i. Both metals can be made stronger by cold forging; the process increases the tensile strength of iron to about 100,000 p.s.i. and that of bronze to more than 120,000 p.s.i. Iron does not melt, and so cannot be cast, at temperatures below 1,500 degrees Celsius (a level not easily achieved by early smiths), whereas bronze melts at a little higher than 1,000 degrees C., or about the temperature used in firing pottery. Finally, iron corrodes rapidly and bronze oxidizes very slowly.

Why is it, then, that iron suddenly eclipsed bronze at about the end of the second millennium B.C.? Because, Maddin says, it was not really iron that was used but steel. Speaking at a recent University of Pennsylvania symposium on laboratory analysis in archaeology, he reported the results of metallurgical studies of early Iron Age tools and weapons from sites in Israel. With the cooperation of museum officials he cut several corroded iron artifacts in half, thereby exposing uncorroded metal. Microscopic examination revealed that the interior metal contained from .3 to .5 percent carbon. The metal on the surface of the objects, Maddin calculated, would have included about 1 percent carbon. Iron that has been "steeled" with that much carbon will not deform under stresses of less than 140,000 p.s.i.; forging increases the metal's tensile strength to 200,000 p.s.i., or almost twice that of bronze.

Maddin suggests that the early producers of carbon steel did not consciously attempt to carburize the iron but may have believed prolonged and intense heating in the charcoal fires of the smithy in some way purified the metal. Iron absorbs carbon at a rate related to the temperature of the fire; carburization at 1,100 degrees C. is more than twice as fast as it is at 1,000 degrees. Quenching the incandescent metal in water, a process the ancient smiths presumably invented for their own safety when they began to work in iron (a cast bronze artifact could be left in its mold until it was cold), hardens carbon steel more than forging does, although the process makes the metal brittle. In any event, those who used the steeled-iron axes, adzes, picks, shovels, swords, daggers, spear points and arrowheads were satisfied thereafter to let bronze assume the ornamental role previously played by iron.

Cancer Immunology

Cancer cells have "foreign" labels, yet they can escape destruction by the immune system. Efforts are being made to learn how they do so and to utilize the immune response for treatment of the disease

by Lloyd J. Old

In the past decade there has been a great resurgence of interest in approaches to cancer based on immunology: the study of the body's mechanisms of defense against foreign invaders. Such approaches have been based on the belief that there is something unique about a cancer cell that distinguishes it from normal cells, and that this difference can be recognized by the body's immune system. The investigation of the immunological reaction against cancer cells and the paradoxical ability of many tumors to persist in spite of it is now an active field of cancer research.

It is my purpose to set down some of the principles of cancer immunology as we now know them, to illustrate them by generally accepted examples and to describe current directions and questions. The object of the most intensive and precise study in this field has been the mouse, and so my discussion will be largely devoted to results obtained with this experimental animal.

Cell-Surface Antigens

Of all the parts of the living cell that might be altered when it becomes cancerous the surface of the cell is the most suspect. Many of the regulatory signals that control the growth and multiplication of normal cells have been found to have their primary site of action at the cell surface, and it is easy to visualize how subtle changes in the surface structure of cells may have profound effects on their behavior. It is well known that when normal cells are placed on a glass surface, the cells will stop migrating and growing when they begin to touch one another, a phenomenon known as contact inhibition. If an equal number of cancer cells are placed on a glass slide, however, they will not stop moving and dividing when they meet but will instead grow into a multilayered mass. It therefore seems possible that an aberration in the normal structure of the cell surface could account for the loss of contact inhibition and hence for the uncontrolled growth of malignant cells.

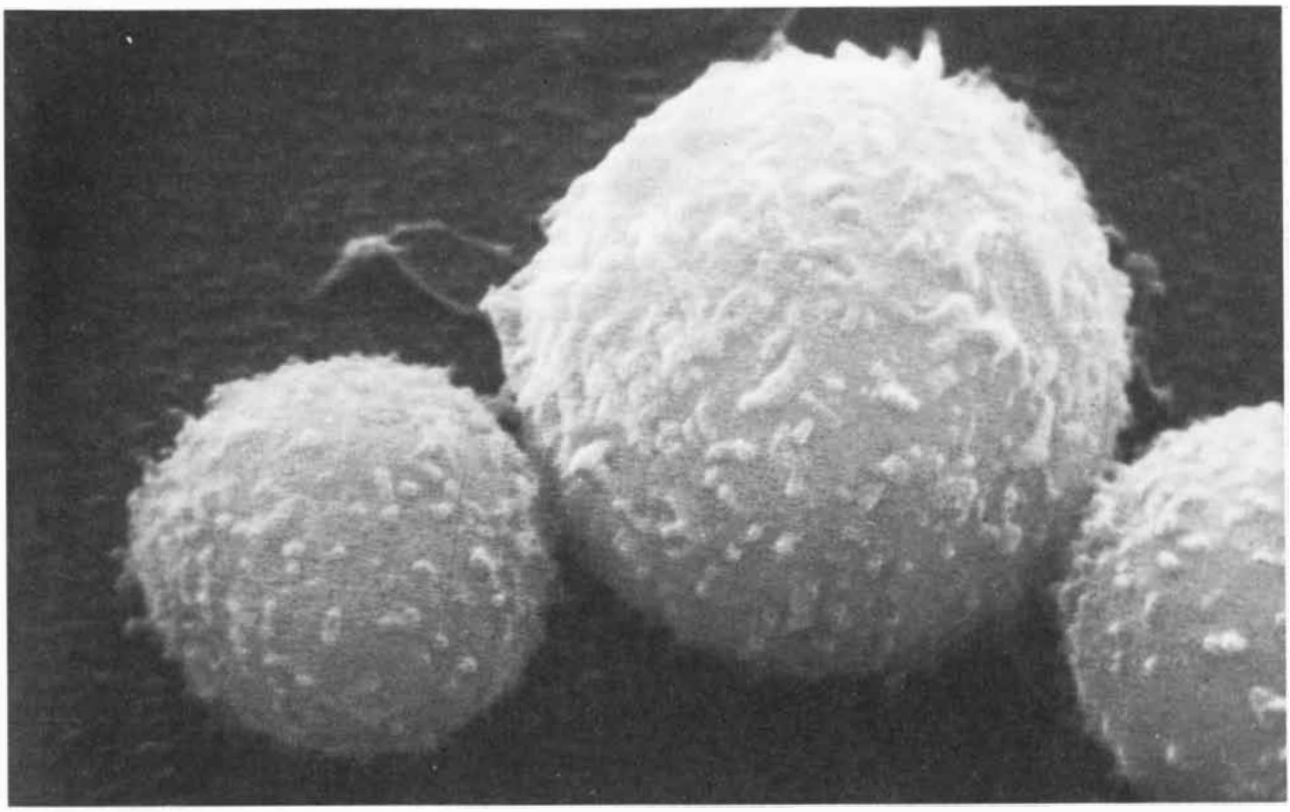
This belief has been supported by the recognition that both normal and malignant cells possess a complex array of molecular markers on their outer surface, consisting primarily of protein with a small amount of carbohydrate, and that the markers differ from one cell type to another. Moreover, when cells from one animal are transplanted to an unrelated animal of the same species, the surface markers elicit an immunological reaction resulting in an attack by the host's lymphocytes: the white blood cells that play a crucial role in the body's defenses against infection. For this reason these molecular labels are called antigens, although it must be realized that the fact they are capable of eliciting an immune response in an unrelated recipient animal probably has little to do with their primary biological function on the cell surface. In the context of cancer immunology the discovery of cell-surface antigens that distinguish cancer cells from normal cells has been the starting point of attempts to explain the malignant transformation in terms of cell-surface changes and of attempts to control cancer by immunological means.

The initial demonstration of cancer-specific antigens came from observations made in mice belonging to highly inbred strains. In these strains, as a result of many generations of successive brother-to-sister matings, each mouse is a genetic (and therefore antigenic) replica of every other. Grafts of normal tissue from a mouse of one inbred strain will generally be accepted by other mice of the same strain but will be rejected by mice of different strains. This means that the antigens on the grafted cells are identical with those of genetically related individuals but differ from those of genetically unrelated ones. The discovery of these "laws" of transplantation immunity explained the long-standing riddle of why grafts of organs or skin from one human being to another, attempted many times in premodern medicine, failed in almost every case to survive more than a short period: the antigens on the surface of the grafted cells had elicited an immune response in graft recipients that lacked these antigens, resulting in the destruction of the grafted tissue. Transplantation antigens, as they are called, are thought to be present on all cells of the body. In man the strongest and best-studied transplantation antigens are the *H-LA* antigens and in the mouse they are the *H*-2 antigens.

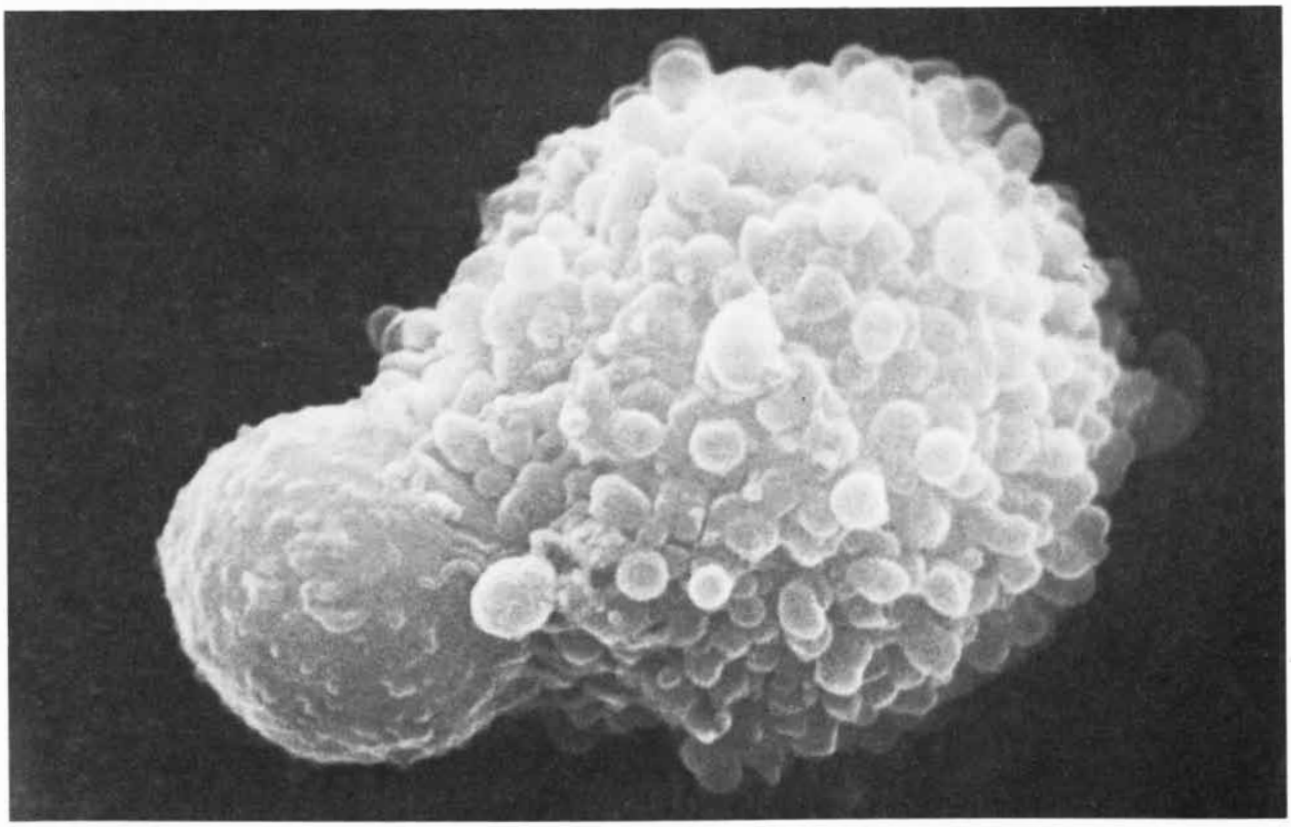
A False Start

When the study of tumor immunology began at the turn of the century, inbred strains of mice had not yet been developed. In retrospect this accounts for the great confusion that originally plagued the field. Early investigators found that when tumors arising in mice were transplanted to other mice, the grafted tumor first began to grow and then shrank and disappeared under the fierce attack of the recipient's lymphocytes. Phenomena such as this one encouraged the erroneous belief that the cures represented immune reactions that were specifically directed against cancer-specific antigens.

The subsequent discovery of transplantation antigens, which exist on both normal and cancerous cells, shattered these illusions. It was soon realized that since the early tumor-transplantation experiments had been done with noninbred animals, the reports of specific tumor immunity were in fact describing an immunity directed against genetically foreign grafts generally and not immunity directed against the tumor as such. This was shown to be the case when it was found that resistance against tumor grafts could also be elicited by immunization with grafts of normal tissue. Investigators then realized that the demonstration of tumor-specific immunity would require the study of tumors arising in mice of the same highly inbred strain that possessed identical transplantation antigens on their cells. If a tumor from one such individual could be shown to behave like a graft of foreign tissue and be rejected by an otherwise genetically identical mouse, it would constitute convincing evidence for the existence of tumor-specific anti-



LYMPHOCYTE ATTACKS A CANCER CELL in this scanning electron micrograph made by Andrejs Liepins of the Sloan-Kettering Institute for Cancer Research. The cancer cell (*large spheroid at right*) has molecular labels (antigens) on its surface that enable the sensitized lymphocyte (*smaller cell at left*) to selectively attack and kill it. The magnification of the micrograph is some 14,500 diameters.



DEATH OF A CANCER CELL is indicated by the blebs, or deep folds, that have appeared on its surface membrane. The mechanism by which lymphocytes kill tumor cells is unclear but most probably involves the release of a toxic factor that disrupts the cell membrane.

gens that were distinct from the transplantation antigens found on both malignant and normal cells.

Chemically Induced Tumors

The first real milestone in the search for tumor-specific antigens was reached in 1943, when Ludwik Gross, now at the Veterans Administration Hospital in the Bronx, discovered that inbred mice could be immunized against a tumor that developed in a mouse of the same inbred strain. Gross's work received little notice at the time because of the prevailing pessimism about immunological studies of cancer, and 10 years passed before Edward J. Foley of the Schering Corporation provided further evidence for tumor-specific antigens in a series of mouse tumors. In his experiments Foley used tumors induced by methylcholanthrene, a multiring hydrocarbon capable of causing cancer in mice two to four months after its injection under the skin. The tumors were induced in inbred mice, so that they could be maintained by the serial transplantation of tumor cells into mice of the same strain, where they multiplied until the mouse died. Foley found that the growth of a tumor transplant, followed by its removal, increased the mouse's resistance to a second transplant of the same tumor.

In 1957 Richmond T. Prehn and Joan M. Main of the National Cancer Institute enlarged on Foley's work by showing that an inbred mouse that rejected a tumor graft from a genetically identical donor would accept a skin graft from the same animal in which the tumor had originated. The grafts of normal skin survived permanently, so that the rejection of the transplanted tumor could be considered convincing evidence for new tumor-specific antigens that had arisen during the malignant transformation of a normal cell into a cancer cell.

Yet even these results were not universally accepted, because the possibility continued to be raised that there were residual undetected genetic differences among members of the same inbred strain. This criticism was met by George and Eva Klein of the Royal Caroline Institute in Stockholm, who showed that even a mouse in which the tumor was originally induced could be immunized against cells from its own tumor, eliminating any doubt that tumor resistance might be due to transplantation antigens alone. These initial findings have since been confirmed in many laboratories around the world. Studies with cancer-causing chemicals other than methylcholanthrene have shown that they too induce tumors that are antigenic in the inbred strain of origin.

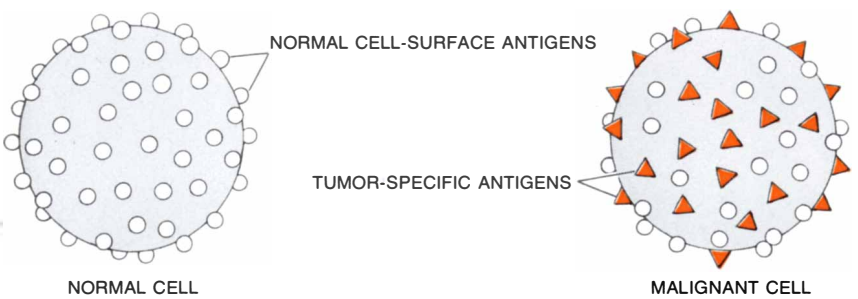
A consistent and remarkable feature of chemically induced tumors is that each tumor elicits immunity to itself but not to any other tumor. Thus tumors induced by the same chemical agent in the same inbred strain exhibit unique antigens. Even when two tumors are induced in the same animal, each can be shown to have distinct antigens. Although the subject of much speculation, the genetic origin of these new antigens is not known. The two possibilities that have been considered most seriously are that the information giving rise to them resides in genes that have mutated (either as a direct result of the interaction of the cancer-causing chemical with the cell's DNA or as an indirect result of the malignant transformation) or in genes that are ordinarily "silent" in adult life but are active at other times, possibly in fetal life. There are other explanations, but until these antigens have been characterized in more precise terms than transplantation techniques can provide their origin will remain obscure.

Virus-induced Tumors

In addition to the tumors that can be artificially induced by chemical substances and by physical agents such as X rays, mice naturally develop an array of malignancies that arise with increasing frequency with age. Each inbred mouse strain has a characteristic resistance or susceptibility to the development of a particular type of cancer: some have a high incidence of leukemia, others develop mammary cancer and still others remain cancer-free. The naturally occurring tumors that do develop can be placed in two categories. One category includes tumors that arise without known cause. The other includes tumors that are elicited by viral agents. Those agents are the RNA leukemia virus originally described by Gross and the RNA mammary-tumor virus discovered by John J. Bittner of the University of Minnesota Medical School. A number of other viruses have also been identified that cause leukemia or other types of cancer in laboratory mice, although they do not do so under normal circumstances. Nevertheless, they have proved to be invaluable laboratory tools for dissecting the way viruses cause cancer and for uncovering the immunological factors that influence tumor development and growth.

One such virus, the polyoma DNA virus, which was discovered independently by Gross and by Sarah E. Stewart and Bernice E. Eddy at the National Institutes of Health, is widespread in certain mouse populations, but tumors caused by the virus arise only rarely, if at all, under natural conditions in normal mice. Adult mice injected with polyoma virus do not develop tumors. Tumors can be induced, however, if the virus is injected into immature mice (which have a weak immune system) or adult mice that have been immunologically crippled by such procedures as removal of the thymus gland at birth or whole-body exposure to X rays. The polyoma tumors that arise can then be successfully transplanted to normal adult mice of the same inbred strain.

Karl Habel of the National Institutes of Health and Hans O. Sjögren, working with George Klein at the Royal Caroline Institute, made the surprising observation that such transplanted tumors would not grow if the recipient mouse had been previously injected with polyoma virus as an adult. This finding—that immunization with a tumor-inducing virus conferred resistance to transplants of tumor cells elicited by that virus—can best be understood by assuming that tumor cells do arise in adult mice following the initial injection of polyoma virus, but that these cells are recognized as being foreign by the immune system and are rejected, leaving the animal tumor-free. When polyoma tumor cells from another animal are later transplanted to the same mouse, the tumor cells are swiftly eliminated because the mouse's immune system has been properly alerted by the earlier growth and rejection of its own tumor cells. In contrast to the antigenic uniqueness of each chemically induced tumor, transplantation studies have shown that all tumors induced by polyoma virus have the same virus-specified cell-sur-



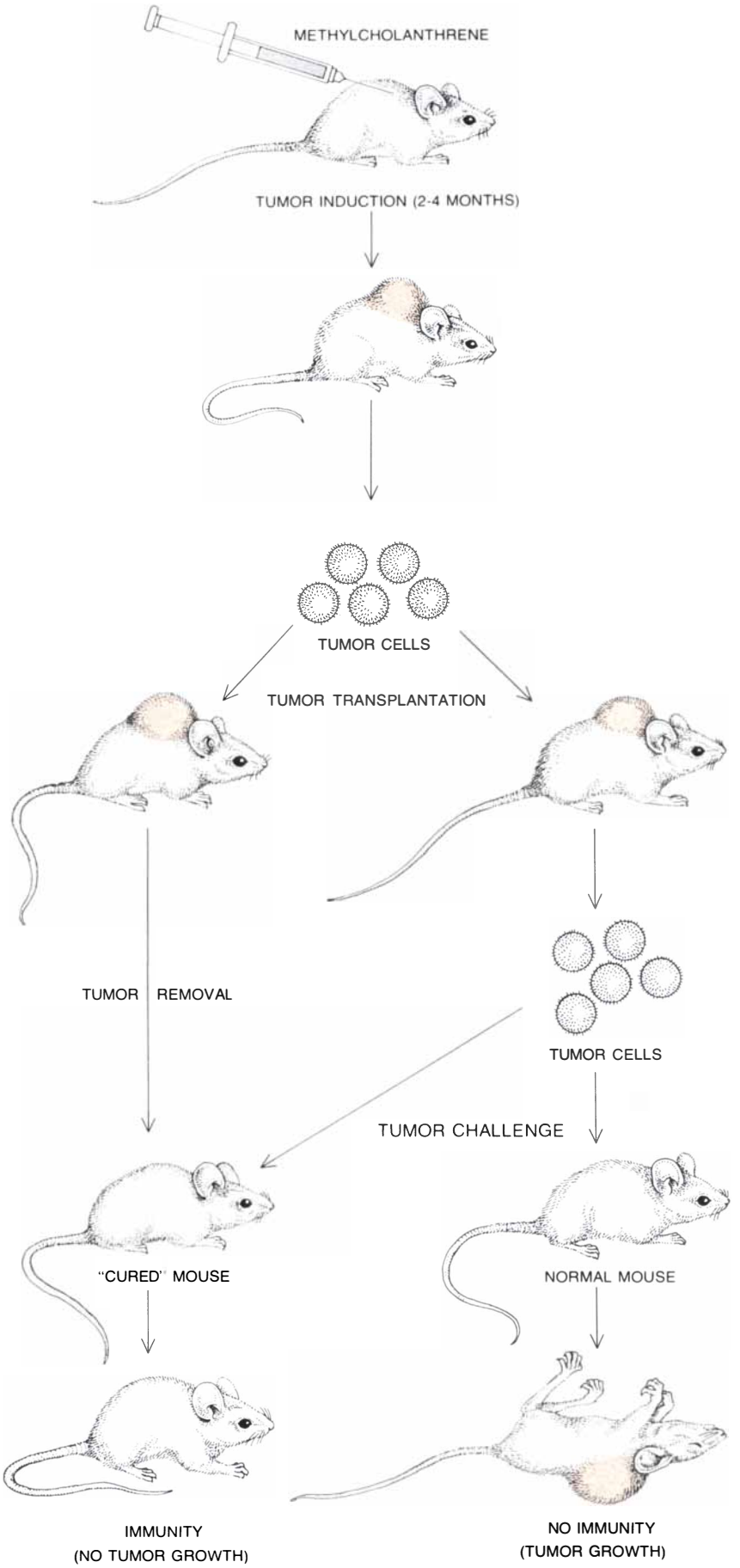
TUMOR-SPECIFIC ANTIGENS appear on the surface of most and perhaps all cancer cells as a consequence of the malignant transformation. Like the transplantation antigens that exist on both normal and cancer cells, the tumor-specific antigens are thought to be complexes of protein and carbohydrate that have been synthesized within the cell and inserted into cell membrane. Since tumor antigens differ from native ones, they should elicit an immune response.

face antigen, so that immunization with any one polyoma tumor confers resistance to any other. The study of tumors induced by other cancer-causing viruses, such as SV40 or adenovirus, has led to the same conclusion: Identical antigens appear in different tumors induced by the same virus (although each cancer virus induces a different antigen).

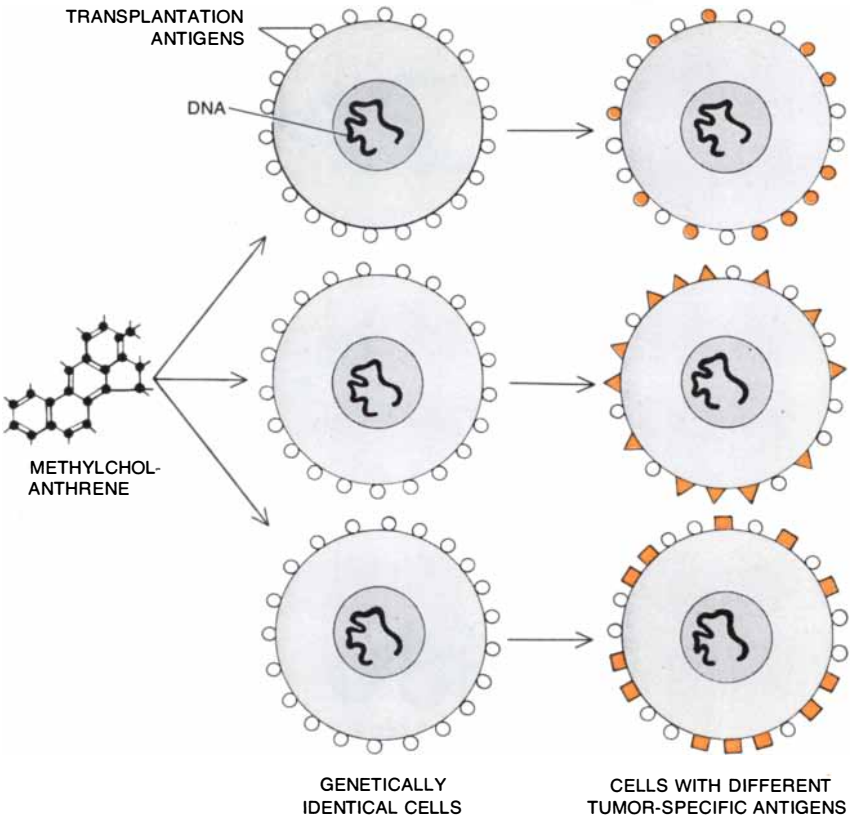
Tumors caused by any one agent, whether it is chemical or viral, vary greatly in their immunizing capacity. Tumor-specific antigens are described as being "weak" or "strong," depending on how effectively they render mice immune to a subsequent challenge with the same tumor. Naturally occurring tumors of mice have generally been found to be only weakly antigenic, and in certain cases they do not appear to be antigenic at all. In some instances we know why antigens escape detection, and this will be discussed below. In others the technique used for detecting antigen is not sufficiently sensitive. Thus it is not known whether tumor-specific antigens are characteristic of all cancers without exception or whether cells can become cancerous without gaining these distinctive markers. On the basis of current evidence it would be premature to form an opinion on this crucial question.

Tumor Immunity

Up to this point I have been discussing the detection of cell-surface antigens and have said little about the nature of the immunological response of the host animal to such antigens, which is clearly of prime concern if the response is to be manipulated to the host's advantage. The rejection of tumors, like the rejection of grafted tissues from a nonrelated donor, is primarily mediated by lymphocytes. It has been known since the early 1960's that there are two different classes of lymphocytes, both of which arise from primitive stem cells in the bone marrow and which migrate to organs where they differentiate into cells that can interact specifically with antigens. One route of differentiation results in the development of *B* lymphocytes, or *B* cells. These cells interact with antigen and subsequently synthesize and secrete antibodies: complex proteins with active sites capable of binding specifically to the stimulating antigen. The



IMMUNITY to chemically induced tumors can be demonstrated in mice. Tumors appear two to four months after an injection of methylcholanthrene under the skin. Cells from the induced tumor can then be successfully transplanted to other mice of the same inbred strain. If the tumor is removed from a recipient mouse, that mouse will be resistant to subsequent injections of cells from the same tumor. Tumor cells injected into a normal mouse of the same strain, however, will proliferate and kill it.



UNIQUE ANTIGENS appear in each tumor induced in mice by the injection of cancer-causing hydrocarbons such as methylcholanthrene or benzpyrene. The mechanism by which this enormous antigenic variety is generated may involve genetic mutation or the activation of genes that are not normally expressed in adult life but that serve some function in fetal life.

other route is through the thymus gland, where the stem cells differentiate into different classes of *T* lymphocytes, or *T* cells, that become responsible for the cellular immune response. *T* cells that have been exposed to antigen appear to synthesize antibodylike molecules at their surface with which they recognize that antigen on subsequent challenge, initiating a series of reactions that can result in the destruction of cells having foreign transplantation antigens or tumor-specific antigens on their surface. *T* cells also release factors that mediate complex interactions between *T* cells, *B* cells and other cells of the immune system, particularly the large scavenger cells called macrophages. In short, both the humoral (antibody) and cellular (*T* cell and macrophage) aspects of the immune response appear to be involved in tumor immunity, although *T*-cell immunity may be the more essential.

With the recognition of tumor-specific antigens cancer immunologists sought analytical methods that would be faster and less cumbersome than the transplantation techniques where tumor rejection by the inbred mouse is the immunological end point. Over the past 50 years an array of powerful techniques have been devised to demonstrate immune reactions not in animals but in the test tube. These techniques have been widely, if not always critically, applied to the study of cancer. In recent years much effort has gone into developing methods that measure the reactivity of cellular components of the immune system, particularly the *T* lymphocyte. Although considerable progress has been made in this area, further understanding of the conditions required for the function of *T* cells in the test tube is needed before the value of methods to detect cellular immunity to cancer can finally be assessed.

Serological Methods

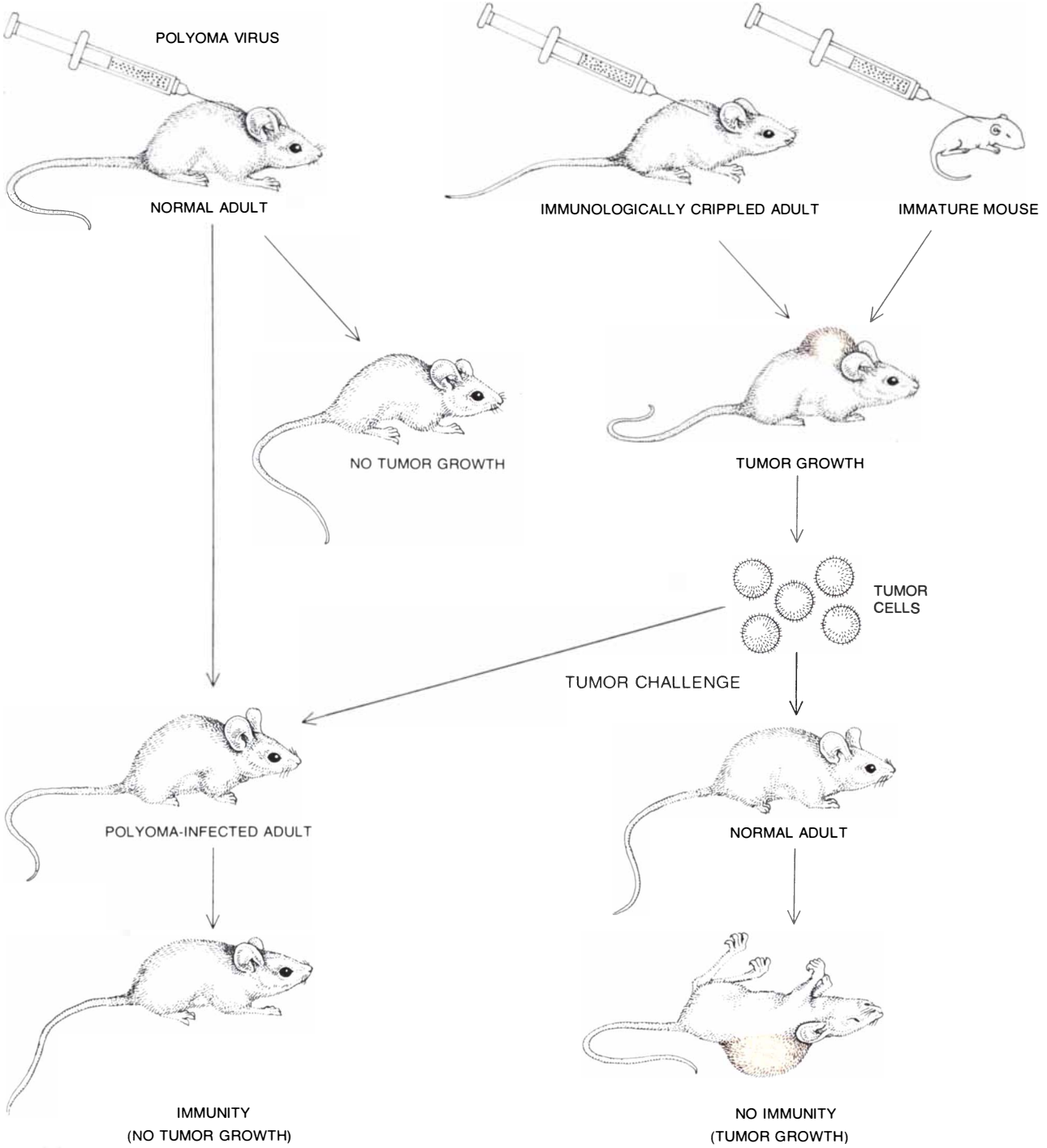
By far the most advanced and versatile methods at the immunologist's disposal are those, known collectively as serological methods, that use antibody as the analytical probe. The value of these methods in laboratory research and clinical practice cannot be overemphasized; two examples that come to mind are the diagnosis of infectious diseases and the typing of blood for transfusion. An enormous effort has gone into attempts to detect cancer-specific antigens by immunizing horses, rabbits and goats with human cancer tissue and then analyzing the antibodies the animal manufactures to see if any can be shown to react specifically with cancer cells but not with normal cells. (Much of this work was inspired by the still unfulfilled hope that a cancer-specific serum would be valuable in human cancer therapy.) In general such efforts have been disappointing, with initial claims for cancer specificity yielding on further analysis to the more realistic assessment that the "cancer antigen" is found in some normal tissue.

The potential in the proper application of this approach is still great, however, and can be illustrated by the discovery of Garri I. Abelev of the U.S.S.R. Academy of Medical Sciences. Abelev immunized rabbits with the blood of mice growing a liver tumor. After all the antibodies that reacted with the blood of normal mice had been removed from the rabbit serum, antibody still remained that reacted with some component in the blood of cancerous mice. Further study showed that the same component was present in the blood of the normal mouse fetus, and Abelev and others subsequently made comparable findings in human beings. This antigen, called alpha-fetoprotein, belongs to a class of substances that are manufactured in fetal life and then fall to undetectable levels shortly after birth, presumably because the genes controlling their production have been inactivated. As a consequence of the malignant transformation the genes controlling the production of these fetal proteins may become reactivated, and the antigen may reappear. The detection of alpha-fetoprotein in the blood has proved to be useful in the early diagnosis of liver cancer in human beings. There are other fetal antigens, and their detection and analysis has become an active area of cancer immunology.

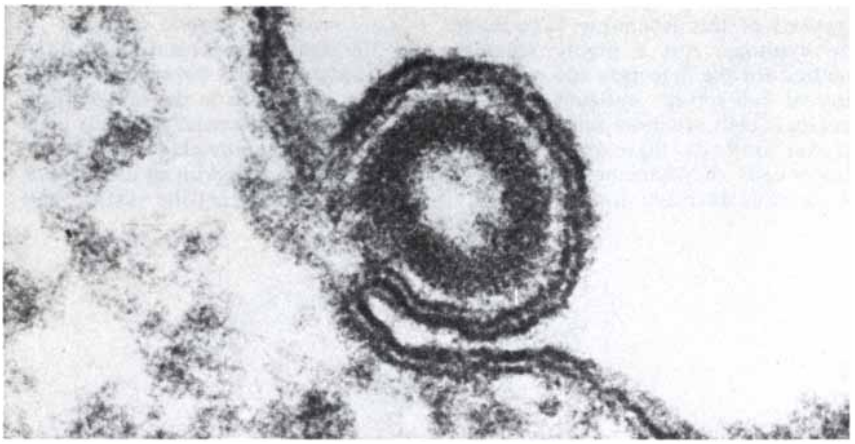
One of the most useful serological methods for the study of cell-surface antigens of cancer cells has been the cytotoxic test, originally developed by Peter Gorer of Guy's Hospital in London. In this technique cells bearing a particular surface antigen are killed when they are incubated with antibody specific to that antigen in the presence of the blood-serum factor known as complement. Complement is actually a group of proteins that, when antibody binds to antigen, are activated sequentially in a "cascade" of enzymatic reactions. The final active complex appears to weaken the membrane of the antigenic cell, leading to its lysis, or dissolution. The sequence of complement reactions serves to amplify the original antigen-antibody reaction several thousandfold, so that the binding of only a few antibody molecules can result in the destruction of the antigenic cell. Complement has been implicated in many immunological phenomena, but its most important role in

tumor immunity is to bring about the lysis of cells bearing tumor-specific antigens against which antibody has been made. In the cytotoxic test devised by Gorer tumor cells that have been killed by antibody are identified by adding a dye to the solution; the lysed cells take up the dye but intact cells do not. Modifications of this technique have made the cytotoxic test a highly sensitive method for the detection and quantitation of cell-surface antigens. Because leukemia cells are more sensitive to cytotoxic antibody than other types of cancer cells, they have been the subject of the most intensive study.

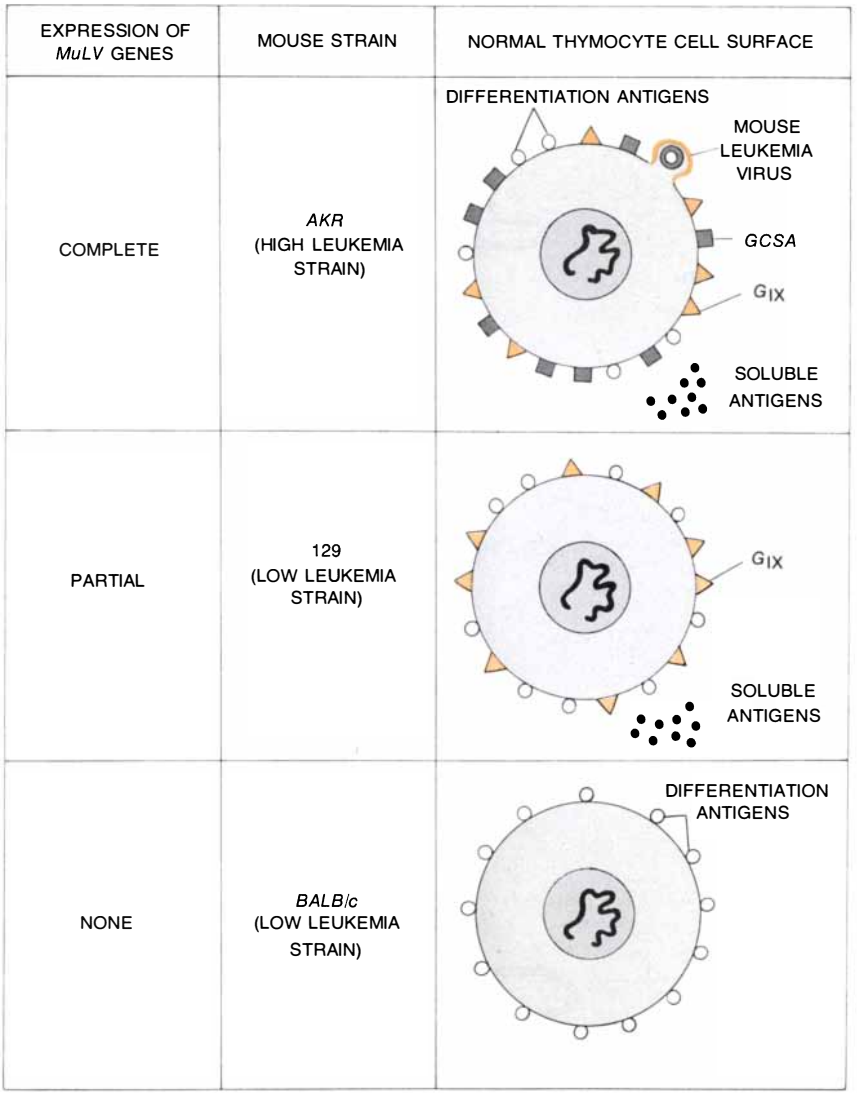
Leukemia is a general term for the proliferation of abnormal white blood cells, either as solid tumors called lymphomas or as cells in the bloodstream. In the mouse this cancer generally originates in the thymus gland, which provides the serologist with an ideal opportunity for comparing the surface anti-



POLYOMA VIRUS INDUCES TUMORS when injected into either newborn mice, which lack a strong immune system, or adult mice whose immune system has been crippled by irradiation or removal of the thymus gland at birth. Healthy adult mice are resistant to the induction of tumors by polyoma virus but will accept transplants of polyoma-induced tumors from other mice of the same inbred strain. Infection with polyoma virus before such a transplant protects the recipient against proliferation of the tumor cells, presumably because its immune system has been alerted by the growth and rejection of its own polyoma-induced tumor cells. Such cross-immunity occurs because, unlike chemically induced tumors, which express unique antigens, all polyoma-induced tumors express same surface antigen.



MOUSE LEUKEMIA VIRUS buds from surface of a cell in this electron micrograph made by Etienne de Harven of the Sloan-Kettering Institute. Magnification is 310,000 diameters.



VIRAL LEUKEMIA ANTIGENS are found on the surface of leukemia cells, on normal thymocytes of mice bred for a high incidence of leukemia and on cells intentionally infected with mouse leukemia virus (*MuLV*). The two major antigens, *GCSA* and G_{IX}, are both protein components of the *MuLV* particle. When they are complexed with carbohydrate and inserted into the cell membrane, they become cell-surface antigens. *GCSA* appears only when virus is being actively produced, but G_{IX} may also appear on the normal thymocytes of certain strains in the absence of virus production. This observation and others suggest that the *MuLV* genes are incorporated into the genetic material of all mice but are expressed to varying degrees.

gens of the leukemia cell with those of the normal thymic cell, or thymocyte. In 1961 Edward A. Boyse, Elisabeth Stockert and I at the Sloan-Kettering Institute for Cancer Research began our serological study of the surface antigens of leukemia cells and of normal thymocytes. At the time only one antigen could be demonstrated on lymphoid cells; it was the major transplantation antigen *H*-2, which is present on most, if not all, mouse cells. Over the years we and our colleagues, particularly Tadao Aoki and Toshitada Takahashi, have defined 10 additional antigens, so that more is known about the surface antigens of these cells than about those of any other cell type in the mouse.

Leukemia Antigens

Three major categories of surface antigens in addition to *H*-2 have now been recognized on thymocytes and leukemia cells. The first can be illustrated by a series of antigens, called *Ly* antigens, that mark all lymphoid cells derived from the thymus. These antigens have been named differentiation antigens because the genes controlling their expression are activated only in cells following a particular pathway of differentiation from embryonic cells into specialized cells. Thus leukemia cells of thymic origin have *Ly* antigen as a direct result of their cellular ancestry. Genes that determine *Ly* antigens follow the rules of conventional Mendelian genetics, with mice having either one *Ly* type or the other. These antigens have been of considerable interest to immunologists since the discovery by Hiroshi Shiku, Pawel Kisielow and their colleagues at the Sloan-Kettering Institute that three classes of *T* cells with different immunological functions can be distinguished by the patterns of *Ly* antigens expressed on their surface.

The second category of surface antigens on leukemia cells, and under certain circumstances on normal cells, are viral antigens that can be traced to the murine (mouse) leukemia virus, abbreviated *MuLV*. Much that we now know about the biology of this virus and the leukemias it causes was learned through the application of serological techniques for detecting viral antigens. In fact, as we shall see, the expression of a viral antigen may be the only evidence for the presence of *MuLV*. A range of *MuLV* antigens have now been identified; they include structural components that go to make up the virus particle, viral antigens that become incorporated into the cell surface and soluble viral antigens that slough off *MuLV*-infected cells into the body fluids.

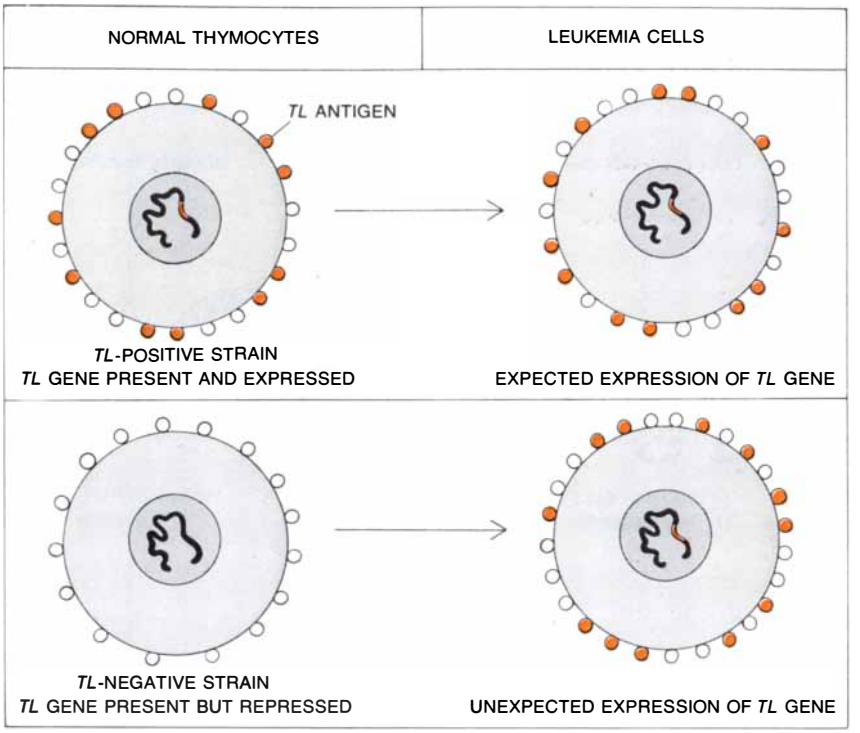
The first opportunity to follow this virus in its natural host came through the description of what is known as the Gross cell-surface antigen (*GCSA*), which is expressed whenever a cell pro-

duces *MuLV*. We found *GCSA* in a variety of normal and leukemic tissues from mouse strains that have been bred for a high incidence of leukemia, such as the strains designated *AKR* and *C*58, but it was not detected in any normal tissue from strains with a low incidence of leukemia. What was unexpected was our finding that *GCSA* could occur in the solid tumors and leukemias of these low-incidence strains. This was one of the first indications that all mice might be infected with *MuLV* and that other factors determine whether the virus is expressed or not, a view that is now widely held.

From the work of Wallace P. Rowe and his colleagues at the National Institutes of Health we know that the genetic information for *MuLV* is ubiquitous in mice as an integral part of the genome (the full set of mouse genes), and that these viral genes are transmitted from generation to generation along with the rest of the mouse genes. Whether or not *MuLV* genes are fully expressed, resulting in an "infected" mouse and the manufacture of virus particles, is determined by other genes that differ from one inbred mouse strain to the next. Strains have also been found where the viral genes are partially expressed, with only low levels of virus or viral proteins produced, or not expressed at all. Hormones, cancer-causing chemicals and exposure to X rays may also activate the normally silent *MuLV* genes, and in certain strains of mice these agents may cause leukemia as well.

Full *MuLV* expression alone, however, does not appear sufficient to render a cell malignant. If one calculates the number of proteins the genetic information carried by this virus could code for, the viral products one knows about can probably account for all the viral genes. Since infected cells can express the full range of viral proteins without showing any evidence of malignancy, their presence per se does not cause leukemia. The fact that leukemia usually originates in the thymus, even though *MuLV* genes are probably incorporated in the DNA of every other nucleated cell of the mouse, indicates that only those cells following this particular pathway of differentiation are susceptible to the leukemia-producing influence of the virus. How the forces of differentiation control the expression and function of the *MuLV* genes is a central question in cancer research.

Analysis of G_{IX}, another antigen specified by *MuLV*, clearly illustrates the effect of cellular differentiation on the expression of the viral genes. G_{IX} was originally recognized as a surface antigen of normal thymocytes, being detected in some mouse strains but not in others. Its relation to *MuLV* was established by the fact that it could be induced in cells lacking the antigen by deliberate infection with the virus. Like *GCSA*, G_{IX} is expressed whenever *MuLV* is expressed and may appear on the leukemia cells of mice that do not normally express G_{IX} on their thymocytes. Unlike *GCSA*, however, G_{IX} is also present on normal thymocytes in certain mouse strains that show no evidence of *MuLV* infection. In these strains G_{IX} is found only on thymocytes and not on any other lymphoid tissue.



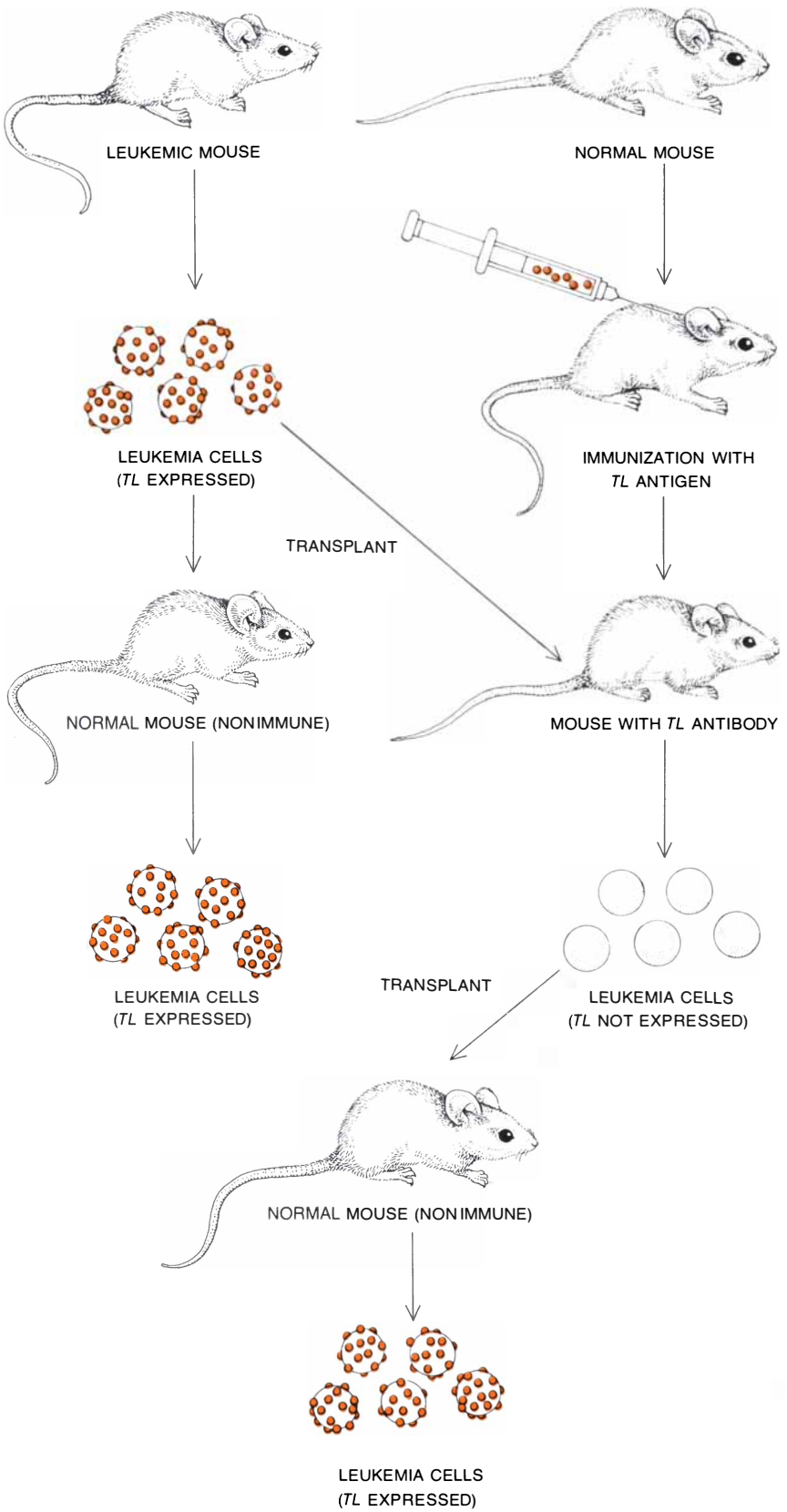
THYMUS-LEUKEMIA ANTIGEN (*TL*) appears on leukemia cells and also on the normal thymocytes of certain inbred mouse strains, designated *TL*-positive. Other mouse strains lack *TL* on their normal thymocytes (*TL*-negative strains), but when these mice develop leukemia, the leukemia cells may nonetheless express *TL*. This observation suggests that all mouse strains contain the genetic information for *TL* but that in *TL*-negative strains it exists in a silent state that can be activated in the course of the transformation of a normal cell into a malignant one.

Genetic analysis has revealed additional unique features of G_{IX}. Unlike *H*-2 and other common cell-surface antigens, which require only a single gene for their expression, G_{IX} appears to require two genes. Moreover, whereas other cell-surface antigens are the product of one or another form of the same gene, mouse strains that lack G_{IX} reveal no alternative gene product. Another peculiarity of the G_{IX} system is the fact that mouse strains can be placed in one of four categories according to the relative amounts of G_{IX} antigen they express on their thymocytes. This unusual feature may have its origin in the different numbers of copies of *MuLV* genes incorporated into the genetic material of different strains.

The relation between these leukemia antigens and *MuLV* has been further clarified by the finding that G_{IX} is a constituent of the principal glycoprotein in the envelope of the *MuLV* virus particle, designated *gp*70. Moreover, Erwin Fleissner and his associates at the Sloan-Kettering Institute have recently found that *GCSA* is immunologically related to two of the virus particle's internal protein components, *p*30 and *p*15.

Three important points stand out in our analysis of G_{IX}. First, cellular differentiation is of crucial importance in the expression of *MuLV* genes. In normal mice of certain low-leukemia strains the *MuLV* gene for G_{IX} is activated only in lymphoid cells undergoing differentiation in the thymic environment. In this sense G_{IX} can be considered a differentiation antigen comparable to the *Ly* antigens, since it is expressed only on a highly specialized cell type.

Second, the relation between *MuLV* and its host is an intimate one, both at the genetic level and at the level of the cell surface. G_{IX} is a viral gene product that has been incorporated as a cell-surface constituent of a normal cell, the thymocyte. Another way to think of this finding is that in the course of the mouse's evolution viral and host genes collaborated closely to produce the surface structure of certain cells. If the origin of G_{IX} from *MuLV* were not known, it would be considered the product of a normal mouse gene. For a variety of reasons, particularly the widespread occurrence of *MuLV* in mice, there is a growing suspicion that *MuLV* performs a critical function in the normal life of



ANTIGENIC MODULATION is a well-documented "escape mechanism" that enables antigenic cancer cells to evade the immune response of the host. When leukemia cells bearing *TL* are injected into mice that have been previously immunized against the antigen (and hence have synthesized antibody to it), *TL* disappears from cell surface. The leukemia cells are then functionally "naked" and cannot be recognized by *TL* antibody or killer lymphocytes. When modulated cells are injected back into a nonimmune mouse, *TL* reappears on the cell surface.

the host animal, although what that function is remains an open question.

Third, the appearance of G_{IX} and *GCSA* in leukemias and other tumors of mice that do not express these antigens during normal life bears significantly on the origin of tumor-specific antigens. I have mentioned the possibility that activation of normally silent genes could account for the individually unique antigens of chemically induced tumors. For the appearance of G_{IX} or *GCSA* in the leukemias of low-incidence strains the genetic origin would be *MuLV*.

The *TL* System

The other example of a tumor-specific antigen that owes its origin to genetic activation came from our discovery in 1963 of the *TL* system, the third category of cell-surface antigens I shall discuss. *TL* has many features in common with G_{IX}, but we have no direct evidence linking it to a virus, even though we suspect there is some kind of association. The antigen is called *TL* because the only types of cells on which it is found are normal mouse thymocytes and leukemia cells. Normal mouse strains can be classified as *TL*-positive or *TL*-negative on the basis of typing the thymus for *TL* antigens. The unusual feature of *TL*, which it shares with G_{IX}, is that when leukemias arise spontaneously or are induced by physical or chemical agents in mice that lack *TL* on their thymocytes, the leukemia cell may express it. As with G_{IX}, this has been explained on the grounds that all mice have the genetic information for *TL* but that other genes control whether or not it will be expressed on normal thymocytes. Leukemia brings about a change in this genetic control, resulting in the anomalous appearance of *TL* on the surface of the leukemia cell.

These findings raise an interesting question about the antigens cancer immunologists classify as being tumor-specific. *GCSA*, G_{IX} and *TL* can each be considered tumor-specific under conditions where they appear in mice that in normal life never express them. Yet these antigens are not truly tumor-specific in the sense that they occur only on cancer cells, since they can also occur as components of normal cells in other strains of mice. In fact, the only truly leukemia-specific antigen we have found in the mouse is a component of the *TL* system designated *TL*-4, which has never been detected on a normal cell.

Although a good start has been made, we still have only a superficial understanding of the enormous variety of surface antigens on normal and malignant cells in the mouse. Other antigens have been identified, but they have not been analyzed sufficiently to be placed in one category or another. Of course, the ultimate aim of this research is not simply

to compile an inventory of cell-surface antigens but to use these antigenic markers to understand how the cell surface is constructed, how malignancy changes that structure and how the immune system responds to such changes.

Escape Mechanisms

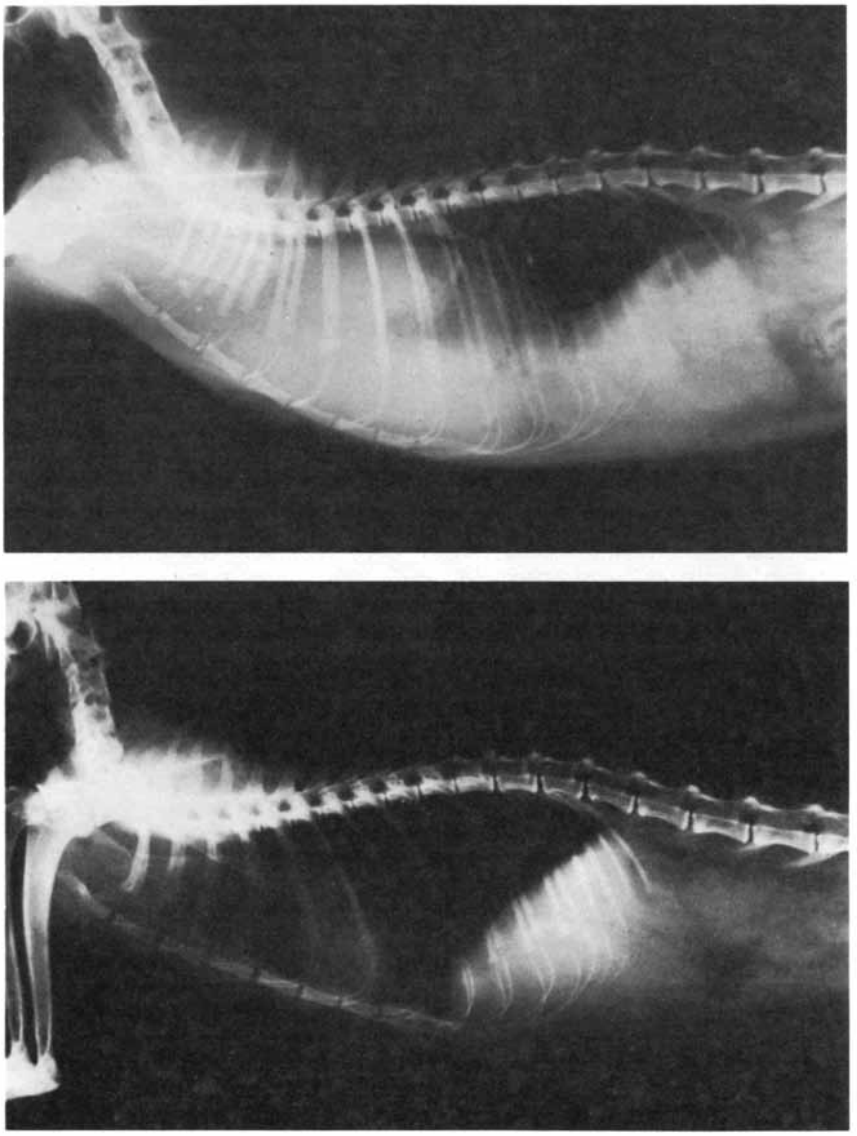
The specific transplantation antigens of tumors induced by chemical agents or polyoma virus, and the G_{IX} and *TL* antigens occurring on leukemia cells of mice that lack these antigens on their normal cells, illustrate the range of cancer-specific antigens that have been detected so far. In fact, when sufficiently sensitive techniques are used, most cancer cells in the mouse have been shown to bear such specific surface antigens. How, then, can these tumors arise without eliciting an effective immune response that suppresses their growth?

There is no single satisfactory explanation for the success of tumors in the face of an immune response. A variety of mechanisms have been proposed, however, that would enable tumor cells to escape the consequences of their antigenicity. Some are deceptively simple, such as the one that has come to be known as "sneaking through." Cancer starting as individual cells may not excite immunological recognition during the early phase of its growth. By the time the immune system is alerted the cancer is established and too large for the body to deal with.

Perhaps the most surprising and best-studied escape mechanism is antigenic modulation: the ability of cancer cells to mask or lose antigen in the face of immunological attack. It was discovered by our group when we were faced with the paradox that the immunization of mice with *TL* antigens did not lead to resistance to the growth of leukemia cells with *TL* on their surface. Such leukemia cells grew as well in mice with high levels of *TL* antibody as they did in nonimmunized mice, although that much antibody should have given the mouse good protection. Boyse and I found that leukemia cells from the *TL*-immunized mice had lost *TL* antigen. When these leukemia cells were transplanted back into mice lacking antibody, however, the antigen reappeared on the cell surface. Obviously not all tumor antigens undergo antigenic modulation, as otherwise it would not have been possible to detect the antigens of chemically or virally induced tumors with transplantation techniques.

Another way cancers might escape immunological destruction is through a flooding of the body with tumor antigens. The antigens would then bind to specific antibodies or to specific receptors on lymphocytes and prevent them from recognizing and destroying the cancer cells. We refer to this escape route as "antigenic blindfolding."

The three mechanisms I have been describing represent strategies on the part of the tumor cell that enable it to escape destruction. Deficiencies and other abnormalities in general or specific immunological reactivity on the part of the host may also play an important role in the development and progressive growth of antigenic tumors. As I have mentioned, polyoma virus cannot induce tumors in adult mice unless their immunity is artificially lowered. Robert A. Good and his colleagues, who at that time were working at the University of Minnesota Medical School, discovered that human patients with genetic diseases impairing the normal immune response show a far greater incidence of cancer. Cancer is also known to arise more frequently in patients who have received an organ transplant and who must be given immunosuppressive drugs to minimize graft rejection. The fact remains, however, that cancer can arise in experimental animals and human beings whose general immune response is, as far as we can tell, perfectly normal. Since it is now clear that immunological reactions are under genetic control, one possibility for the emergence of antigenic tumors in individuals with otherwise normal immunity would be that they lack the immune-response



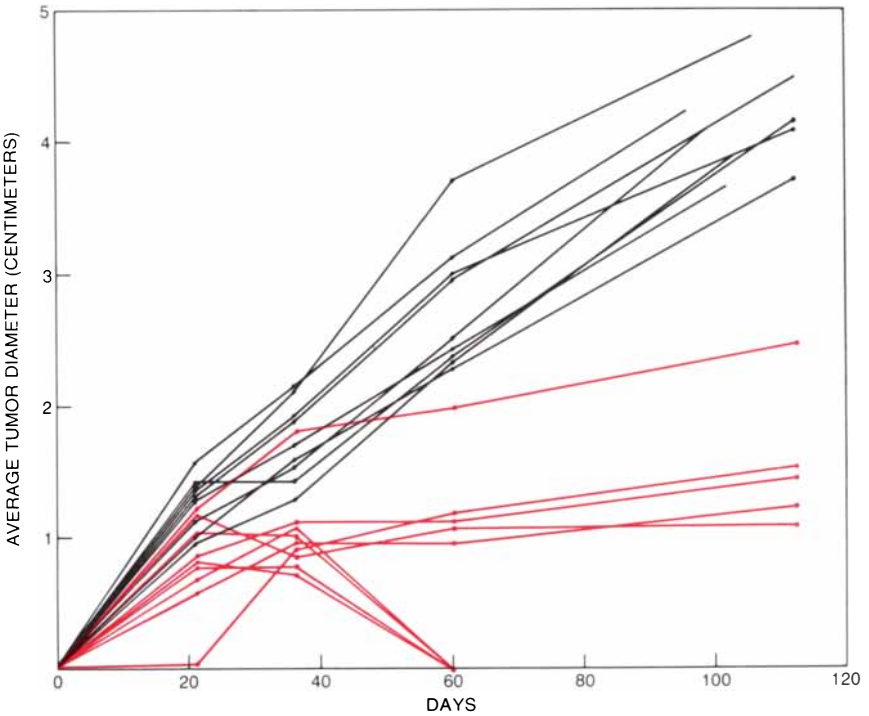
DRAMATIC TUMOR REGRESSION in a cat was produced by infusing the animal with blood plasma containing complement, a series of proteins that are involved in the destruction of antigenic cells in the presence of specific antibody. X ray at top shows how the chest cavity of the cat is entirely filled with an advanced lymphoma (*cloudy area*). X ray at bottom, made 12 days after infusion of plasma containing complement, shows that the tumor has completely regressed. The evidence indicates that the deficiency of secondary immunological factors such as complement may be responsible for the failure of the immune system to destroy certain antigenic tumors even when high levels of specific antibody are present in the host animal.

(*Ir*) genes that are required for the recognition of certain kinds of tumor-specific antigens.

Another mechanism for the specific damping of tumor immunity in the presence of a normal immune system is known from work on mammary tumors of the mouse done by Donald L. Morton at the National Cancer Institute and David W. Weiss at the University of California at Berkeley. The great majority of these tumors are known to be caused by the Bittner virus, which is commonly passed from generation to generation through the milk: infant mice acquire the infection by suckling on infected mothers. If the infant mice are nursed by foster mothers that are not infected with the virus, the infants do not become infected and do not develop mammary tumors. Transplantation tests, like those for the antigens of chemically induced tumors, show that mammary tumors are far more antigenic in mice that lack the virus than in mice that acquired it at birth. Mice infected at birth apparently become immunologically unresponsive ("tolerant") to certain antigens related to the virus and the tumor it induces, and so they are not able to mount an effective immunity to the tumor when it appears in later life.

Most discussions of escape mechanisms stress, in the final analysis, the absence of specific immunity mediated by antibody or immune cells. Experiments conducted by Robert L. Kassel, Elizabeth A. Carswell and me indicate, however, that secondary factors required for the action of antibody might critically limit certain immune reactions to cancer. We have found extensive destruction of leukemia cells after leukemic mice have been infused with blood plasma from healthy mice. The active factor in the plasma has been identified as one component of the complement system, which, as I have mentioned, is a series of blood proteins that are involved in a wide range of immune reactions, including the destruction of cells by antibody. Current evidence points to the likelihood that some leukemic animals make adequate levels of specific antibody, but because they lack sufficient complement the antibody fails to destroy the leukemia cells. When complement is provided from an external source (in the infused plasma), the leukemia cells are killed. As my colleague William D. Hardy, Jr., has shown, plasma-mediated leukemia-cell destruction occurs in cats as well.

What is clear from this discussion of escape mechanisms (and several other mechanisms have been suggested) is that the progressive growth of a tumor in no way indicates the absence of an immune response or the absence of tumor-specific antigens. Of course, the ultimate escape mechanism in the immunological sense is the total lack of tumor-specific antigens. We do know of a process, named immunoselection, in which the most antigenic cells in a tumor are destroyed by immunological reactions, leaving the tumor cells with the least antigenicity to continue multiplying, but no antigenic tumor has yet been found in which a permanent loss of tumor-specific antigens has been demonstrated. Given the ways tumors can escape immunological destruction, however, we cannot know whether any tumor in the mouse lacks tumor-specific antigens. There is also no way at present to assess the strength of tumor-specific antigens in human beings, since the transplantation techniques required for such assessment in the mouse cannot be applied in man. Until the techniques are developed to obtain the necessary information one can only speculate whether the proper models for human cancer are mouse tumors with the strongest antigens or mouse tumors with the weakest ones.



ANTITUMOR EFFECT OF BCG (bacillus Calmette-Guérin), an attenuated strain of the tuberculosis microorganism, is shown graphically. Test mice were injected with BCG and given a tumor transplant one week later (day zero on graph); controls received only the transplant. Tumors in the BCG-infected mice showed a slower rate of growth or regressed completely (*colored curves*) whereas those in control mice continued to grow, eventually killing all of them (*black curves*). BCG appears to work by generally stimulating the host's immune system.

Immunotherapy of Cancer

The discovery of tumor-specific antigens has naturally generated much interest in possible immunological approaches to the therapy and prevention of cancer. It is popular to speak of immunotherapy as being a fourth kind of cancer treatment (to be added to surgery, radiation and chemotherapy), but it would be wrong to infer that there is at the moment any comparable regime of immunotherapy with predictable benefit to the cancer patient. In general there are only hints of what the future will bring.

The two cancer therapies most widely used, surgery and radiation, are maximally effective only when the cancer is localized. If cancer cells are disseminated throughout the body, therapy is required that will reach and destroy them wherever they are, and this is the chief theoretical advantage of chemotherapy. Although there has been significant progress in the control of certain forms of cancer by chemotherapy, all cancer chemotherapeutic agents injure normal cells to some degree, and this frequently places limits on their ultimate effectiveness. The principal attractiveness of immunotherapy lies in the extraordinary specificity of immunological reactions: the destruction of cancer cells marked by specific antigens should leave normal cells that lack these antigens uninjured. For this reason a vast amount of work is being done in laboratories and clinics around the world to apply immunological principles to the treatment of malignant disease.

It is now generally believed that experiments of nature represented the first attempts at immunotherapy. Physicians

The photographer is Marie Cosindas. The medium is Polaroid's Polacolor® Land film. The result is the remarkable photograph at the right, "Dolls," a work of art recently acquired by the Museum of Fine Arts in Boston. Now, Polacolor 2 film has the same unique metallized dyes found in Polaroid's SX-70® film. It has the same exceptional clarity and stability. And its brilliant colors are among the most permanent and fade resistant ever developed in photography. Polacolor film is used by amateur, professional and scientific photographers throughout the world. Polaroid, the choice of the artist in the creation of her art.

This Polaroid Polacolor photograph was acquired by Boston's Museum of Fine Arts for its permanent collection.

have long observed and recorded cases in which advanced cancers regressed after an acute bacterial infection. William B. Coley, a New York surgeon born in 1862, spent his life developing a therapy for cancer based on such observations. Coley started by attempting to infect cancer patients with bacteria, but because that turned out to be surprisingly difficult to accomplish he began to inject the patients with mixtures of killed bacteria. These mixed bacterial vaccines, which came to be known as Coley's toxins, reproduced certain features of acute bacterial infection, such as fever and chills, and in a substantial number of cases the tumors in his patients partly or completely regressed.

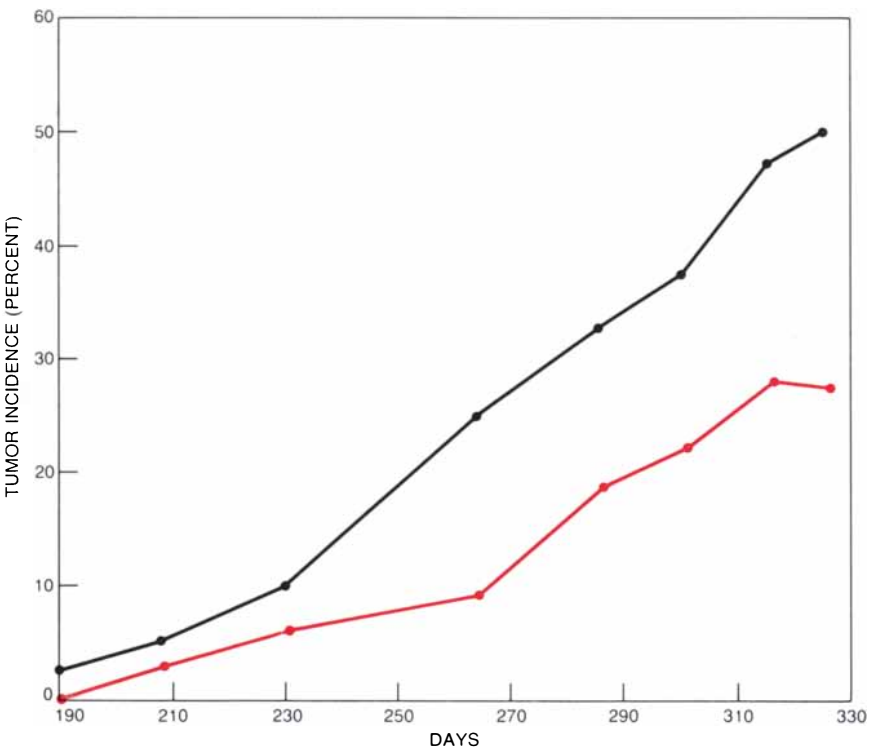
Helen Coley Nauts (Coley's daughter) of the Cancer Research Institute has now collected and analyzed nearly 1,000 case histories of patients treated in this way, and there can be no doubt that for some of the patients the treatment was highly effective. Why, then, did it fall into disuse and for a time even into disrepute? Part of the answer lies in the great hopes that attended the emergence of radiation therapy in the 1920's and 1930's. Another reason was that no scientific rationale could be put forward to account for the activity of Coley's toxins, and physicians are loath to adopt a therapy whose basis they do not understand and whose effectiveness they cannot predict. As the field of immunology grew, knowledge caught up with the Coley phenomenon, and several findings began to shed light on how Coley's toxins exerted their antitumor effect. For one thing bacterial products were found to have profound effects on immunological reactions. When animals are injected with an antigen along with killed tuberculosis bacilli, they manufacture far more antibody than they do if they receive antigen alone. With the discovery of tumor-specific antigens in animals, it seemed likely that human tumors might also be antigenic, and that the bacterial products in Coley's toxins were potentiating an otherwise inadequate immune response against the cancer.

Immunopotentiators

A final link between Coley's work and the modern era came when Baruj Benacerraf, who was then working at the New York University School of Medicine, Donald A. Clarke and I found that mice infected with bacillus Calmette-Guérin (an attenuated strain of tuberculosis bacilli used to vaccinate human beings against the disease) showed increased resistance to a variety of tumors. Bacillus Calmette-Guérin (BCG) produces a transient infection in mice, and in the course of the animal's reaction to the infection its immune system becomes maximally stimulated. In addition to being more resistant to cancer the BCG-infected mouse has a greater ability to manufacture antibody, to reject grafts of foreign normal tissue and to survive certain bacterial and viral infections that are normally lethal. Considerable progress has been made, particularly by Edgar Lederer of the Institute for the Chemistry of Natural Substances in France, in extracting and identifying the immunologically active components of BCG. Other agents, some of microbial origin and some synthetic, have been found to have BCG-like effects on the resistance of the animal host. Since these agents heighten immunological reactivity under defined conditions, they have been named immunopotentiators.

A major challenge to cancer immunologists is to determine how immunopotentiators work and to sort out the roles of the humoral and cellular components of the immune system. Much interest has been focused on the macrophage as the key cell in mediating the effect of BCG and other immunopotentiators. The macrophage is called a scavenger cell because of its capacity to phagocytize (engulf) and digest bacteria, dead cells and other foreign matter. Since the macrophage was recognized by Élie Metchnikoff late in the 19th century, it has been considered essential to health, particularly through its role in resistance to infection. BCG and other immunopotentiators have striking effects on macrophages, causing them to become more phagocytic, to synthesize more digestive enzymes and to divide. Such activated macrophages have been found by Peter Alexander of the Chester Beatty Research Institute in England and by Jack S. Remington of the Palo Alto Medical Research Foundation to inhibit and sometimes kill cancer cells in tissue culture under conditions where normal cells are not harmed.

Some understanding of how immunopotentiators may bring about tumor-cell destruction has come from our analysis of a phenomenon known to cancer researchers for more than 50 years and referred to as tumor hemorrhagic necrosis. Certain tumors of mice and other rodents undergo acute hemorrhage and destruction following the injection of gram-negative bacteria or a factor from the cell wall of such bacteria. This reaction has long been viewed as the experimental counterpart of Coley's observations on human cancer. The cell-wall factor that causes tumor hemorrhagic necrosis and is also responsible for many of the toxic effects of these bacteria is called endotoxin; it has been chemically identified as a lipopolysaccharide, a complex of lipid (fatty) and carbohydrate molecules. Unlike BCG, which in mice is most effective if it is adminis-



INHIBITION OF SPONTANEOUS TUMOR DEVELOPMENT follows infection of young mice with BCG. The spontaneous mammary tumors in certain mouse strains are known to be caused by the Bittner virus, which is passed from mother to offspring in the milk. Tumors begin to appear when the mice reach six months of age and continue to develop at a steady rate, with more than half of the mice having tumors after a year (*black curve*). If mice are infected with BCG two months after birth, they have a lower rate of tumor development (*colored curve*).

DC-10.
When you have a choice...

To many people, "all airplanes look alike," but more and more discerning travelers are seeing and choosing the McDonnell Douglas DC-10. More than 100,000 passengers board DC-10s each day.

In addition to its size, the most distinguishing characteristic of the DC-10 is the engine perched high in the tail, where it is superbly efficient.

DC-10s are nearly everywhere. Thirty-four airlines fly more than 4,900 flights a week to 150 cities on six continents. In fact, DC-10s fly to more places, more often, than any other wide-cabin jet.

When you fly, don't settle for less. Whether your journey is a short trip, or a long flight across an ocean, you'll find the DC-10 as gracious inside as it is graceful outside. Tell your airline or travel agent, "When I have a choice, I prefer the DC-10."

The DC-10

MCDONNELL DOUGLAS

EQUAL OPPORTUNITY IN PROFESSIONAL CAREERS. SEND RESUME: BOX 14526, ST. LOUIS, MO. 63178

The new $2984* Colt. Isn't a Datsun. Isn't a Toyota. It's a lot of little Dodge.

The new Dodge Colt is such a lot of car, it's got Mr. D. and Mr. T. confused. Because Colt offers you the value you'd expect from this import, plus Dodge Colt sales and service coast to coast.

47 MPG highway, 30 MPG city. Colt will give you great mileage, according to EPA estimates.** Your mileage may vary, according to your car's condition, equipment, and your driving habits. And Dodge Colt runs on either regular or unleaded gas.

Looking for a long list of standard features? Well, you get it on all of our '77 Colt models. Even our lowest priced two-door coupe gives you whitewall tires, two reclining bucket seats, tinted glass in all windows, carpeting, adjustable steering column, simulated wood-grained instrument panel, four-speed manual transmission, quiet sound insulation, trip odometer, locking gas cap, and electric rear window defroster.

And we offer you an optional automatic transmission to go with the standard 1.6 liter engine.

So if you're thinking "import," think about Dodge Colt. It's not a Datsun. Not a Toyota. For $2984, it's a lot of little Dodge. See it at your Dodge Colt Dealer's.

*Manufacturer's suggested retail price, not including destination charge, taxes, title, and options. California prices higher.

**Equipped with standard 1.6 liter engine, four-speed manual transmission, and 3.31 rear axle ratio. California mileage lower.

"Very nice, Mr. D."

"I thought it was yours, Mr. T"

Dodge

A PRODUCT OF CHRYSLER CORPORATION

Dodge Colt. It's a lot of little Dodge.

tered before a tumor transplant, endotoxin has its strongest effect on established tumors.

The action of endotoxin on tumor cells is not direct, as is shown by the fact that endotoxin is not lethal for tumor cells in tissue culture. For many years tumor hemorrhagic necrosis was believed to result from an action of endotoxin on the circulatory system, leading to the decreased supply of blood to the tumor and the death of tumor cells for lack of oxygen. The work of our group at the Sloan-Kettering Institute has provided another explanation for this effect of endotoxin. We have detected a factor in the blood of animals injected with endotoxin that has the capacity to kill tumor cells directly. Our current findings lead us to believe endotoxin causes macrophages to release this factor, which is in turn responsible for the tumor-cell destruction.

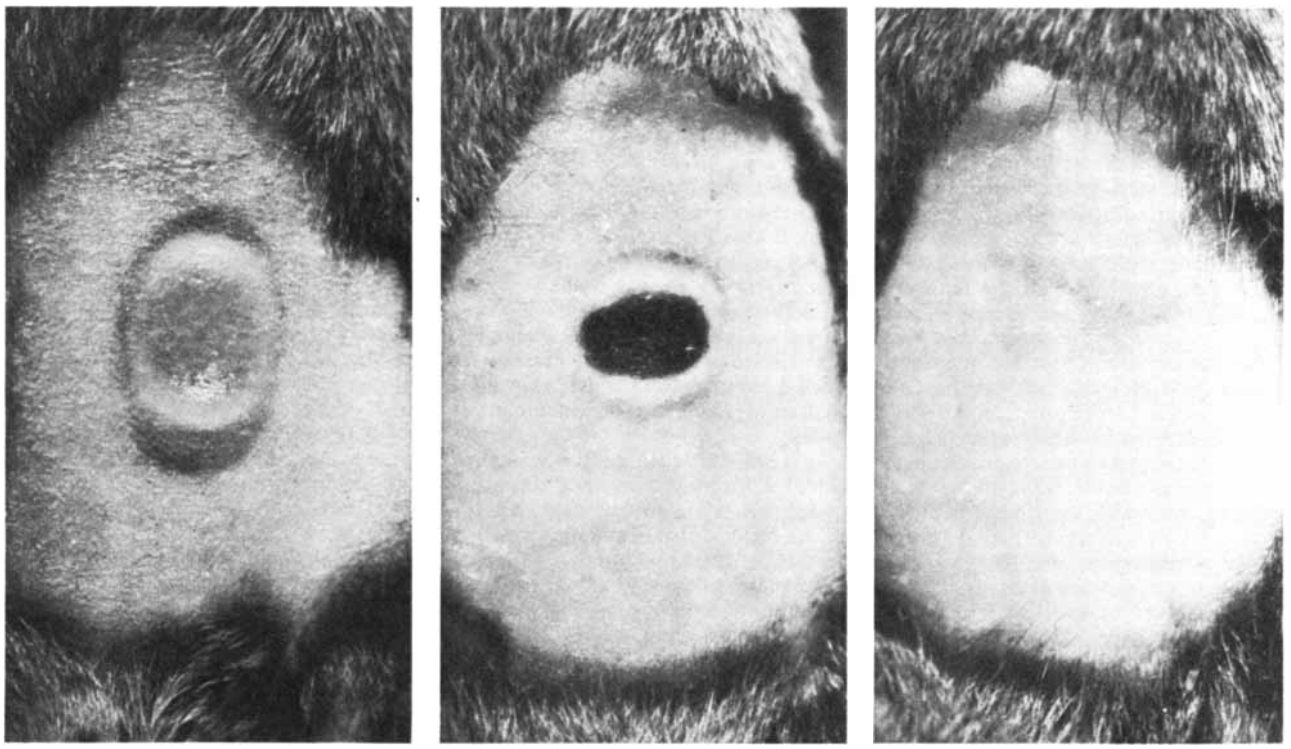
Immunopotentiators are now being discovered that act selectively on components of the immune system, in contrast to BCG and other microbial substances, which act more broadly. As we learn to measure and understand the specific immunological defects associated with cancer and other diseases, these more specific immunopotentiators may become valuable therapeutic agents.

Immunopotentiation, whether general or selective, is only one of the many ways that immunological principles are being applied to cancer therapy. The idea of a specific cancer vaccine is as old as immunological thought itself. Time and time again physicians have attempted to alter the course of cancer by injecting patients with an extract of the patient's own cancer cells or of another patient's cancer cells. Considering the fact that most of the patients in these clinical studies had far-advanced cancer, any beneficial result would have been surprising, and indeed there was none. With the newer concepts and findings of cancer immunology, however, specific immunization with cancer vaccines is still as attractive a possibility as ever and one with a sound scientific rationale. Before we can expect this approach to work we need to know more about the best source and form of tumor antigen and how to combine it most effectively with immunopotentiators. As always, work with the mouse will undoubtedly lead the way in providing the necessary information.

Future Directions

The greatest triumph of immunological research has been its success in preventing formerly devastating infectious diseases in human beings. What its ultimate contribution to the understanding and control of cancer will be remains for the future to determine. The discipline of cancer immunology is a new one, however, and to expect major practical applications of its principles and findings at this point would be unrealistic. What we can expect over the next few years is a growing understanding of the nature of tumor-specific antigens in experimental animals and a detailed knowledge of the immune response they elicit. The study of immunology has been profoundly influenced by the realization that genes exert a tight control over both the nature and the intensity of immunological reactions, and the genetic control of immune responses to cancer and cancer viruses now needs definition. Learning how to modify or overcome genetic restrictions on immune responses may provide the most powerful way of preventing tumors from escaping immunological destruction. In research on human cancer the task is to apply the findings derived from the study of animal tumors to the analysis of human cancer. Among other things this means carefully controlled assessments of the value of immunopotentiators in the treatment of human cancer and rigorous laboratory studies to define human cancer antigens as precisely as we can define them in experimental animals.

As we look to the immediate future neither uncritical optimism nor new waves of doubt and pessimism appear to be justified. Substantial advances have been made, and we can expect that immunological investigations will continue to make notable contributions to our understanding of malignant disease.



HEMORRHAGIC NECROSIS of a transplanted tumor occurs after the injection of endotoxin, a component of the cell wall of gram-negative bacteria. Within 12 hours extensive hemorrhage and cell death have turned the tumor black (*middle*). Regression follows several days later (*right*). Effect of endotoxin is indirect and appears to be mediated by macrophages, which release a tumor-killing factor.

The Case of the Missing Sunspots

Old records indicate that between 1645 and 1715 there were virtually no spots on the sun. It seems likely that the activity of the sun varies considerably, and that the present period is an unusually active one

by John A. Eddy

In 1893 E. Walter Maunder, superintendent of the solar division of the Royal Greenwich Observatory in London, searched through old books and journals and could hardly believe what he was finding. It seemed likely that for many years astronomers had missed an all-important truth: the sun was not the regular and predictable star that all had believed it to be. If Maunder could accept what he was reading, the sun had changed in major ways in fairly recent times. Specifically, the old accounts revealed that for a period of 70 years ending in about 1715 sunspots and other solar activity had all but vanished from the sun. If that had really happened, Maunder knew the implications could be profound, not only for astronomy but perhaps also for climate and thus for the future conditions of life on the earth.

Sunspots are the best-known feature of the sun, and the cycle of about 11 years in which they appear and disappear is among the best-documented facts in astronomy. Although for centuries astronomers in the Orient had recorded large sunspots with the unaided eye, in the Western world the dark spots on the sun were largely ignored until they were observed with the telescope in 1611 by Galileo and several others. From that time to the present they have been under continuous telescopic scrutiny. In 1843 Heinrich Schwabe, a German amateur astronomer, noted from his own observations that when he plotted the average number of spots seen per year, they seemed to come and go in an obvious cycle with a period of about 10 years. His discovery surprised professional astronomers, who had long held that there was nothing at all cyclic in the appearance of sunspots or in other solar activity. Soon after Schwabe's announcement, however, other observers confirmed the existence of the cycle, refining it to 11.2 years. From old observatory records the Swiss astronomer Rudolf Wolf also established that the cycle had been running since at least 1700, which seemed to him to be the limit of reliability of the available records. By 1893 it was well known that the sunspot cycle was associated with other signs of solar activity and with recurrent terrestrial effects such as auroras. The curve of the annual number of sunspots from 1700 on looked like a sample of a wave that continued unchanged both forward and backward in time. Few doubted that sunspots and the 11-year cycle were enduring features of the sun. Then as now those features were accepted as evidence that activity on the sun was regular, which in turn meant that the sun was constant and predictable.

In 1893 the sun was at a maximum of the sunspot cycle and hundreds of sunspots were visible, as Maunder was well aware. Even in years when the cycle is at a minimum at least a few sunspots are usually found; for an entire month to pass without any spots appearing on the sun is rare. Yet in the musty records of the 17th century, just before the start of the familiar sunspot curve, Maunder had found first-hand accounts of the passage of year after year without the appearance of spots. For 32 years not one spot had been seen on the sun's northern hemisphere. For 65 years no more than a single small group of spots was seen at any one time. Several periods lasting for as long as 10 years had passed when no one found a spot anywhere on the sun. Maunder discovered that the total number of sunspots reported between 1645 and 1715 was less than what is seen in a single average year today.

In 1894 Maunder published a paper titled "A Prolonged Sunspot Minimum" giving the details of that strange period in the history of the sun and calling attention to its implications. If the apparent paucity of spots was real, it should shake the foundations of solar astronomy. Maunder pointed out that such an unusual period could serve as a powerful test of relations between the earth and the sun: if the normal ups and downs of the 11-year sunspot cycle could be detected in changes in the magnetic field of the earth or perhaps in the weather, then a prolonged change in the behavior of the sun should be accompanied by major effects on the earth.

It is not clear whether anyone listened to Maunder. Not much attention had been paid to an earlier paper he wrote on the same subject in 1890, or to one published a year before that by the German astronomer Gustav Spörer, who had first brought the sunspot-deficient period to Maunder's attention. Maunder did not give up. In 1922 he tried again with another paper, again titled "A Prolonged Sunspot Minimum," again pointing to the significance of the 70-year period to solar astronomy and terrestrial physics. Six years later Maunder died, and the sunspot cycle went on and on as though to mock him. His papers either were forgotten or were ascribed to his being an enthusiast who put too much faith in old and sketchy records.

Several years ago I felt it was time that we clear up the case of the missing sunspots, which had hung too long like a skeleton in the closet of solar physics. I was annoyed by occasional references to it in connection with a coincident change in the world climate. As a solar astronomer I felt certain that it never could have happened, and my interest in history made the prospect of cross-examining Maunder's assertions an appealing one.

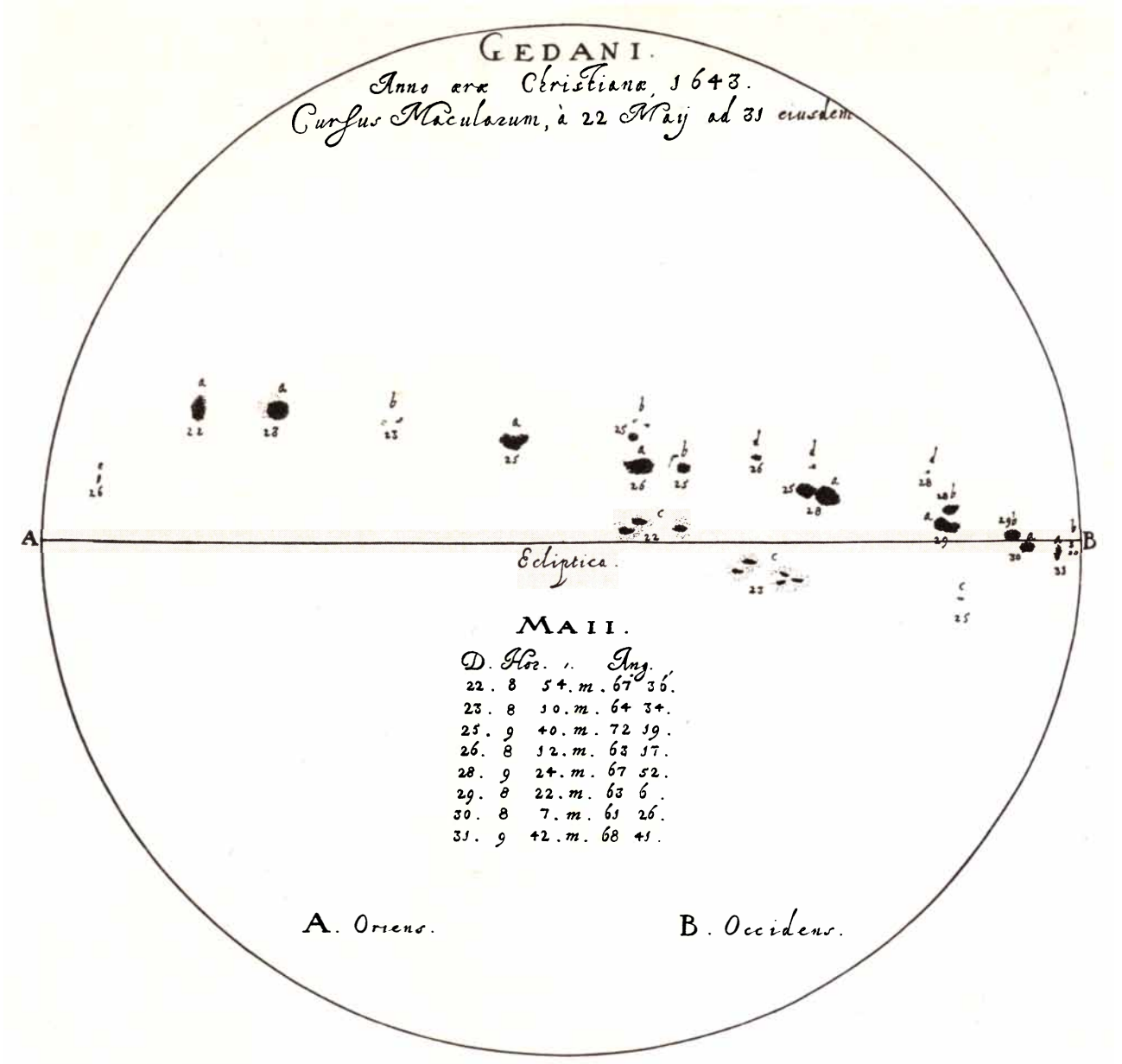
The task had the ring of a detective story: a crime, serious for astronomy and perhaps for the earth itself, had been reported as having been committed far in the past. Did it really happen? Maunder's original clues in the case were now more than 250 years old, yet they were still intact in libraries that held journals of the 17th and 18th centuries. Even more encouraging, in the half century since Maunder's death new evidence had come to light in the advance of solar physics. That evidence included catalogues of historical observations of auroras, compilations from the Orient of sunspots observed with the unaided eye and a fuller understanding of how a completely inactive sun might appear during a total solar eclipse. Best of all, I had at my disposal a powerful tool in the

modern analysis of the annual growth rings of trees. What was important for my work was not the width of the rings, which gauges only the local climate, but their chemical content, which holds an indirect record of changes on the sun. None of these clues had been known to Maunder. None of them was in itself conclusive. All of them combined, however, along with the original historical records, could possibly resolve the perplexing case.

In studying the astronomers' accounts from the period I found to my surprise that they were exactly as Maunder had represented them, and so I began calling the sunspot-deficient period the "Maunder minimum." When a new sunspot was reported in 1671, in the middle of the Maunder minimum, the editor of *Philosophical Transactions of the Royal Society of London* had this to say: "At Paris the excellent Signior Cassini hath lately detected again Spots on the Sun, of which none have been seen these many years that we know of."

The editor went on to describe the last sunspot seen, 11 years earlier, for the benefit of readers who had forgotten what one looked like. Cassini himself (G. D. Cassini, the founder and first director of the Paris Observatory) wrote of the same event: "It is now about 20 years since astronomers have seen any considerable spots on the sun, though before that time since the invention of the telescope they have from time to time observed them." Could those words have been written in a time of normal solar behavior?

John Flamsteed, the first Astronomer Royal of England and the first director of the Royal Greenwich Observatory, recorded how he searched the sun for seven years before he could find a single



SEVENTEENTH-CENTURY DRAWING OF SUNSPOTS was published by Johannes Hevelius of Danzig in his book *Selenographia* (1647). It shows the progress of several groups of spots across the face of the rotating sun from May 22, 1643, through May 31 of the same year. From a series of such drawings the author and his coworkers have deduced the rotation rate of the sun in the middle of the 17th century, just at the beginning of the 70-year period when sunspots were nearly absent. The form of the spots in the drawing and the details of their dark umbra, surrounded by the lighter penumbra, demonstrate that the quality of the telescopes available in the 17th century was high enough for contemporary astronomers to observe sunspots in much the same detail that astronomers can now.

SUNSPOT CYCLE IS NOT REGULAR in either frequency or amplitude, as can be seen in this curve of the mean annual sunspot number, a measure of the number of spots visible on the face of the sun at any one time from 1610 through 1976. The interval between neighboring maximums of the sunspot cycle is not a regular 11 years; it has been as short as eight years and as long as 17. Moreover, some sunspot maximums, such as the one in 1959, are far more intense than others, such as the weak ones early in the 19th century.

spot. The sunspot-deficient period was commonly referred to in books written well after the end of the period. Significantly, it was not until the discovery of the 11-year sunspot cycle by Schwabe that references to the long absence of sunspots in the 17th century began to disappear. Recollections of the sunspot absence could explain why astronomers were first surprised by Schwabe's announcement of the cycle. A suppressed or interrupted sunspot cycle could explain why more than 230 years elapsed between the first telescopic observation of sunspots and the belated discovery that they come and go in a pronounced cycle.

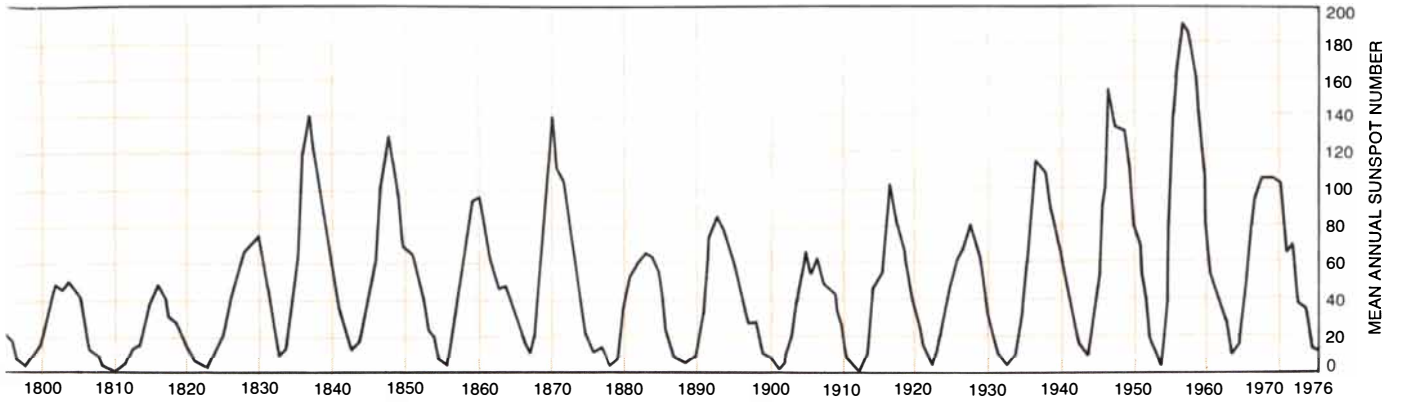
But could we trust the old observers? How good were their telescopes? How carefully did they watch for sunspots? The 17th century was a long time ago: it was the age of Louis XIV, and people wore quaint clothes and wrote in labored ways. In the same period, however, Cassini discovered the major dark division of Saturn's rings and Saturn was found to have at least five satellites. The Maunder minimum began 35 years after Galileo constructed his first small telescope. In the intervening years both optics and astronomy flourished. The 17th century was the age of suspended telescopes that had a focal length as long as 200 feet. It was the age of Newton's first reflecting telescope and of many other innovations. Astronomers observed and counted spots on the sun in much the same way that we do today, and their instruments were little different from those employed for the purpose over the next two centuries. Their drawings of sunspots, preserved in journals and books, are scarcely inferior to the drawings of observers in A.D. 1977. I am convinced that astronomers in the time of Louis XIV possessed both the instruments and the ability to see all but the tiniest of sunspots—if there had been spots on the sun to see. I believe that these observers were every bit as capable as we are, that they most likely had the same drives and professional interests and that they probably had more time to spend at the telescope.

Did they keep a continuous watch on the sun? Or did Maunder take absence of evidence for evidence of absence? Two facts suggest to me that absence of evidence is not the problem. The absence of sunspots was repeatedly noted in the period, and if one accepts that the motives of the 17th-century observers were much like ours, I suspect that a particularly close watch was kept to find new spots to challenge what even at the time was perceived as being strange. Moreover, articles in the contemporary journals establish that when a new sunspot was found, it provided the occasion for writing a paper. Today even at sunspot minimum there are so many spots to be seen that if a paper were written for each one, no journal could publish them all.

Could the weather have interfered with the observations? Could Europe have had an unusually large number of overcast days for 70 years that would have kept astronomers away from their telescopes? It was indeed a time of unusual cold in Europe, but it was not a time of total overcast. If it had been, we would have heard about it from complaints in the journals by astronomers, who have never been known to be either long-suffering or taciturn. Furthermore, in the 17th century nighttime astronomy was vigorous and active: comets were regularly sighted, and contemporary advances in knowledge of the planets required not only clear skies but also a steady atmosphere.

Historical reports of the aurora borealis—the "northern lights"—leave even less room for doubt that the Maunder minimum is real. Displays of the aurora are linked to the level of solar activity. Below the Arctic Circle the number of nights when the aurora is seen correlates well with the number of spots on the sun. In general how often the aurora is seen also depends on the observer's distance from the earth's magnetic poles. Auroras are more frequent at higher latitudes and are rare near the Equator, because at low latitudes the geometry of the lines of force in the earth's magnetic field shields the atmosphere against the incoming solar particles that cause the auroras. Under ordinary conditions of normal solar activity at least 500 and perhaps as many as 1,000 auroras are seen in the populous areas of Europe over a period of 70 years. Yet between 1645 and 1715 hardly any auroras were reported in Europe. Even in Scandinavia, where today auroras are visible almost nightly, so few were seen that they were considered portents. During the Maunder minimum there was one span of 37 years when not a single aurora was recorded anywhere on the earth. When one was finally seen in England in March, 1716, at the end of the Maunder minimum, the astronomer Edmund Halley, who was then the Astronomer Royal, was moved to write a paper explaining the phenomenon. He confessed that he had never seen an aurora before, although he was then 60 and had always watched for one. Although Halley did not know it, his life had spanned most of the Maunder minimum.

The 70-year sunspot minimum is a particularly clear-cut feature if one plots the number of reports of historical auroras by year. If Maunder had looked at nothing but that record, he could have written his papers. In the counts of historical auroras, however, there is also another effect that demands explanation. Very few auroras were reported in ancient times compared with the number seen today. Why are such reports so scarce even before 1645? The records show that the number of auroras reported begins to rise steeply in about 1550, is interrupted by the Maunder minimum,



The author calls the period between 1645 and 1715 the Maunder sunspot minimum, after the British solar physicist E. Walter Maunder, who first pointed out the period's possible significance for conditions on the earth. The record of sunspots observed before 1650 is sketchy. The first isolated peak around 1612 (*far left*) is derived from the observations of Galileo; the second peak is derived from the observations of Christoph Scheiner, recorded in his book *Rosa ursina;* the third peak is derived from the observations of Hevelius.

and then jumps by a factor of almost 20 after 1716. How much of that steep rise in the number of auroras reported after 1550 is a social effect, say a reflection of the interest in astronomy in the Renaissance or, in a later time, the impact of Halley's paper?

I suspect that much of the rise in the number of auroras reported since medieval times is indeed a social effect. Other evidence, however, leads me to believe that at least a part of the rise is a physical effect—the effect of a changing sun. There is evidence that there were other prolonged periods like the Maunder minimum in earlier times. They show up clearly in early aurora records and the frequency of unaided-eye sunspot reports. I have come to believe that in the long view the present-day frequency of sunspots and auroras is probably unusual, and that since the 17th century the activity of the sun has risen steadily to a very high level—a level perhaps unequaled in this millennium.

Reports of sunspots observed without the aid of a telescope provide a check on the reports of auroras and on the reality of the Maunder minimum. Spots on the sun were reported at least as early as the fifth century B.C., and were recorded fairly regularly thereafter, particularly in the Orient. Large spots or groups of spots can be seen fairly easily with the unaided eye as the sun is rising or setting, or when it is heavily obscured and colored by smoke. In 1933 the Japanese astronomer Siguru Kanda compiled a list of known unaided-eye sunspot observations from Japan, China and Korea. He found that in the Christian Era an average of between five and 10 such spots have been seen per century, including a few periods when the spots were seen more frequently and several gaps when no spots were seen. One of the gaps runs from A.D. 1584 to 1770, spanning the Maunder minimum.

This evidence is sketchy, and there are social effects that could explain the gaps, particularly considering that the normal number of sightings is so small. I would be inclined to dismiss the gaps in sunspot observations in the Orient as coincidence if it were not for the fact that the reports of unaided-eye sunspots match very well with the frequency of the auroras reported in Europe for a period of more than 2,000 years.

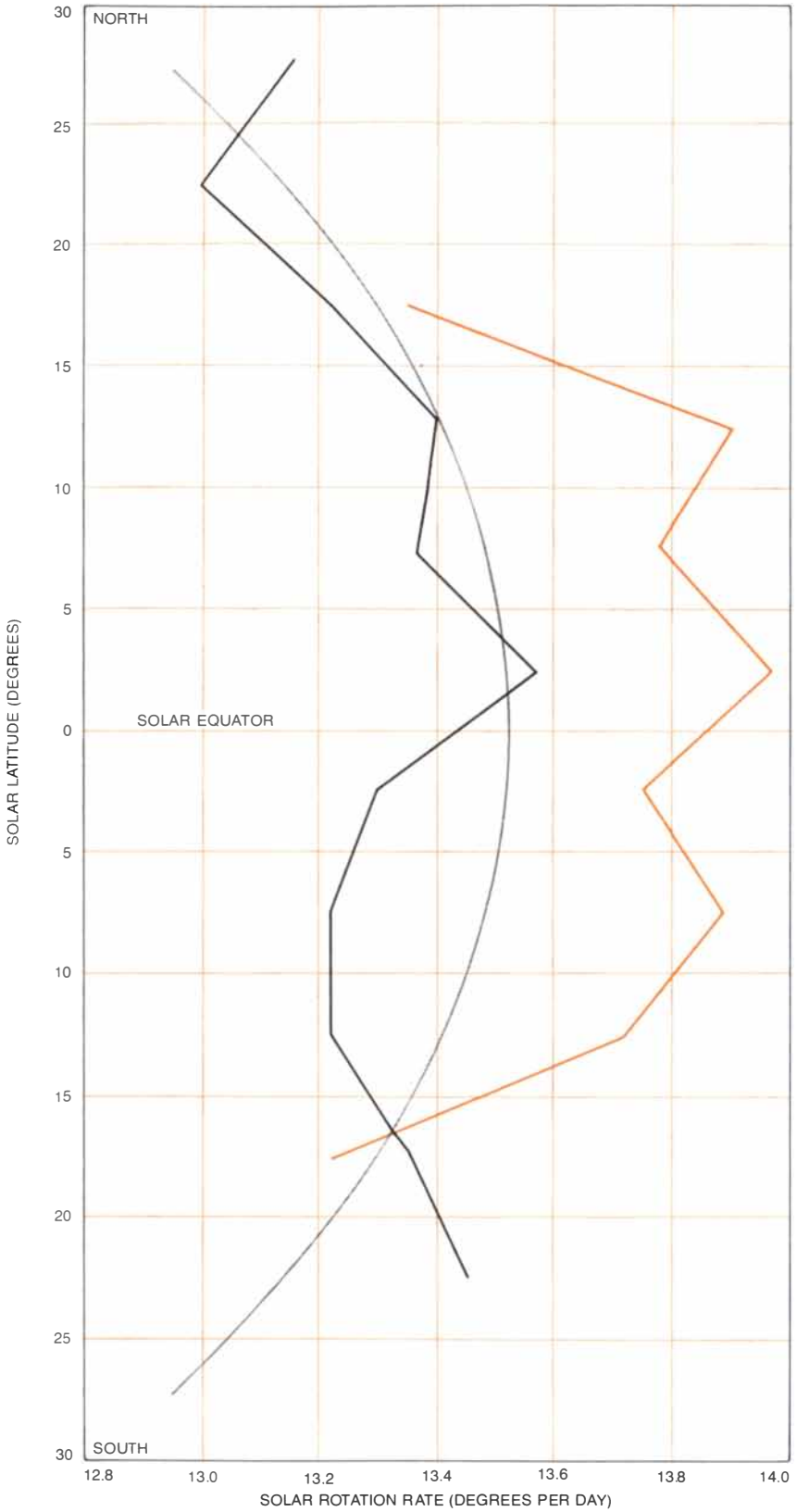
In addition to reports of auroras and sunspots, descriptions of the appearance of the solar corona during total eclipses provide another kind of evidence for checking on past levels of solar activity. The shape of the corona, the sun's tenuous outer atmosphere, when there are many spots on the sun is quite different from the shape when there are few spots. Sunspots are the locus of strong magnetic fields on the surface of the sun, and those magnetic fields shape the pale white coronal streamers that are seen around the sun during a total eclipse. When there are numerous sunspots and therefore concentrated solar magnetic fields, the corona is filled with streamers that extend from the sun like the petals of a dahlia. When there are fewer sunspots, the number of coronal streamers declines as though someone had plucked the petals off. At sunspot minimum the corona seen during a total eclipse is dim and limited to a few streamers near the equator.

What would happen to the corona if there were no sunspots, and none had appeared for years or decades? An observer would still see the zodiacal light, or "false corona," around the black disk of the moon, a narrow ring of pale reddish light without any streamers. Unlike the true corona, the zodiacal light is not a part of the sun's atmosphere; it is simply sunlight scattered by dust in interplanetary space. During a total solar eclipse the zodiacal light is ordinarily overpowered by the brighter corona, but it is always there.

Between 1645 and 1715 there were 63 opportunities to see the sun totally eclipsed. I have endeavored to track down observations of all of them to find out how the corona was described at the time of the Maunder minimum. Most of the eclipses could be observed only from inaccessible regions, and it was not the custom in the 17th and 18th centuries, as it is today, to send eclipse expeditions to follow the shadow of the moon to the ends of the earth. A number of the eclipses were visible near observatories in Europe, however, and they were particularly well observed. Not a single European eclipse was missed, and several additional ones were observed by scholars who happened to be in the New World or in Asia.

Most of the professional astronomers observing an eclipse did not look up at the sun in the sky but spent the precious moments watching the image of the eclipse projected through a telescope onto a screen. On the image they measured the details of the sun's obscuration by the moon, which was then a matter of pressing interest. Those who did look up when the eclipse was total, however, are consistent in what they describe: around the moon was a pale ring of dull light, of uniform breadth, reddish and narrow. Not one observer described the whitish structured corona that is so familiar and so striking today. On the basis of such reports I am convinced that in the time of the Maunder minimum the streamered corona was either entirely absent or so dim that all one could see was the zodiacal light.

On the face of it the eclipse reports would seem to be clear evidence that the Maunder minimum was real. As with the reports of the auroras, however, we must also consider what was seen before that time. I can find no description of a structured corona during any eclipse in any era before the Maunder minimum! Total eclipses of the sun have captured the attention and awe of man for millenniums; how, in thousands of accounts of



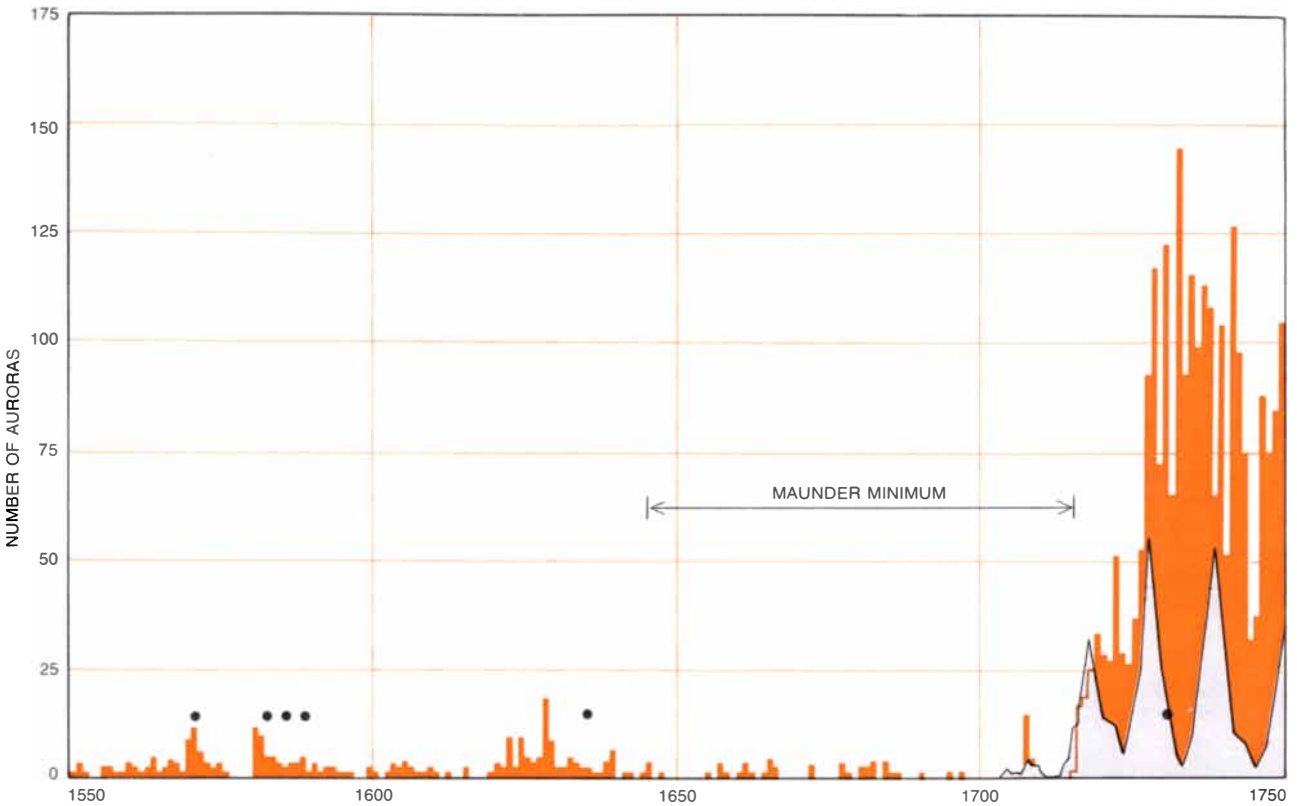
ROTATION OF SUN'S EQUATOR SPED UP just before the onset of the Maunder minimum. The sun does not rotate as a solid body; the equatorial regions spin faster than the regions at higher latitudes. Here the modern rate of the sun's rotation at different latitudes is plotted in degrees per day (*gray*). Scheiner's observations show that the sun's rotation rate in the 1620's was much the same as it is today (*black*). Hevelius' observations in 1640's (*color*), however, show that sun's equatorial rotation rate compared with polar rotation rate had increased three times. It is not known whether this was a cause or an effect of Maunder minimum.

hundreds of total solar eclipses, could the most beautiful and moving spectacle of all have escaped attention? Perhaps we see only what we look for, no matter how striking the scene, and it is true that the corona was not known to be a part of the sun until late in the 19th century. Perhaps the observers were temporarily dazzled by the last rays of the sun before the eclipse reached totality and they therefore missed the ethereal corona. Perhaps until the Enlightenment of the 18th century no one felt moved to describe the corona.

I suspect that anyone who has seen the breathtaking beauty of the corona with the unaided eye will find these excuses entirely inadequate, as I do. It seems quite possible that the corona was actually absent from the sun for a period much longer than the Maunder minimum, possibly for several hundred years. Its absence would thus coincide with an earlier extended minimum in solar activity that I call the Spörer minimum, which is apparent in the counts of auroras, in the absence of unaided-eye sunspot observations and in the evidence from the tree-ring record. Such an absence implies that the corona of today might be a transient, modern feature of our sun.

Another independent clue regarding the Maunder minimum has come from work Dorothy E. Trotter, Peter A. Gilman and I have done toward reconstructing the pattern of the sun's rotation in the 17th century. Modern observations demonstrate that the sun does not rotate like a solid body; it lags at higher latitudes, so that the sun's equator rotates once in 27 days and its polar regions rotate once in 31 days. Modern solar theory contends that the observed differential rotation of the sun produces the sunspot cycle through the action of a solar dynamo, in which deep-seated solar magnetic fields interact with the differential rotation of the surface to generate electric currents, giving rise to the magnetic fields of the sunspots. Therefore if there were a marked anomaly in the number of sunspots, one might expect to find a concurrent anomaly either in the sun's magnetic field or in the rotation of its surface.

The rotation of the sun can be measured quite accurately from series of drawings in which the daily positions of sunspots are shown. The sunspots serve as tracers of the surface rotation, as leaves floating on a stream trace the flow of the water. Trotter, Gilman and I have reconstructed the patterns of solar rotation in the 17th century by examining original drawings of the sun and sunspots from two old books: the *Rosa ursina,* by Christoph Scheiner, published in 1630, before the start of the Maunder minimum, and the *Selenographia,* by Johannes Hevelius, which fortuitously ap-



NUMBER OF AURORAS OBSERVED each year between 1550 and 1750 (*color spikes*) shows that displays of the "northern lights," which are related to the level of solar activity, nearly disappeared during the Maunder minimum. Moreover, early in the 18th century, when the modern sunspot cycle (*gray area*) seems to have begun, number of reported auroras increased by 20 times. The times when greater numbers of auroras were seen also correspond to times when sunspots seen with unaided eye (*dots*) were reported in the Orient.

peared in 1647, at the beginning of the minimum.

In each book daily drawings of the sun are shown for a nearly continuous period of about two years. The solar rotation we derive from Scheiner's drawings in *Rosa ursina* for the years 1625 and 1626 looks much like the pattern of today: the rotation of the sun differs with latitude by the same amount. Hevelius' drawings, made from 1642 through 1644, just as the last spots faded before the long deficiency, show a remarkable and significant change. The sun's equator had speeded up, completing a rotation one full day faster than it had in 1625! In fact, the rate of rotation of the equator compared with the rate at higher solar latitudes was accelerated by a factor of three. In capturing this change in the pattern of the sun's surface rotation just as the spots began to vanish did we catch the perpetrator of the crime? Was the change in the sun's differential rotation the cause of the Maunder sunspot minimum or an effect?

At this point in my detective work the complexion of the case had of course completely changed. There was now little doubt that Maunder and the old accounts were entirely right. A last and telling clue remained in the annual growth rings of old trees.

The radioactive isotope carbon 14 is formed in the upper atmosphere of the earth through the action of galactic cosmic rays. The flux of those cosmic rays is in turn modulated by the activity of the sun, which alters the sun's extended magnetic field. When the sun is very active, its extended magnetic field shields the earth from some of the galactic cosmic rays; fewer strike the upper atmosphere and less carbon 14 is formed. When the sun is less active, its extended magnetic field weakens; the earth receives a higher dose of cosmic rays and the carbon-14 content of the atmosphere increases. It is true that influences other than the sun affect the production of carbon 14. Nevertheless, if we had an exact record of how much carbon 14 was in the atmosphere in the past we would also have an indication of the activity of the sun in the past.

Trees provide just such a record, nicely separated into one-year divisions by the annual growth rings. The carbon 14 formed in the upper atmosphere eventually finds its way into trees through the intake of carbon dioxide in photosynthesis. The ratio of the carbon 14 in the carbon dioxide to the common isotope of carbon at the time each ring formed can be determined by analyzing the wood in the ring. This painstaking analysis has been carried out for many years at a number of tree-ring laboratories, chiefly because a record of the past abundance of carbon 14 is needed for accurate carbon-14 dating in archaeology and other disciplines. And several specialists in that work have pointed out the potential value of the carbon-14 record as an indicator of the past activity of the sun.

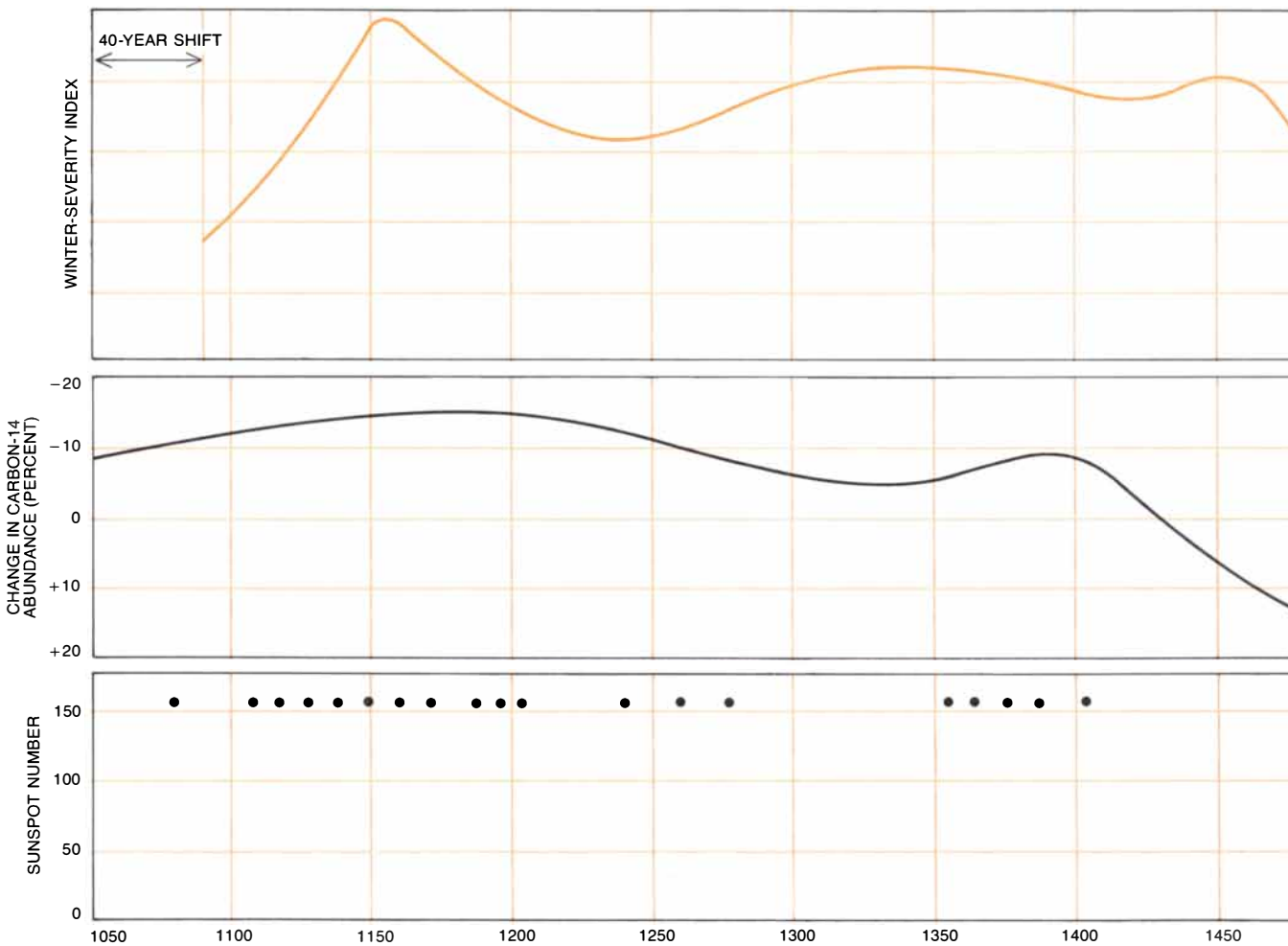
As early as 1958 the Dutch investigator Hessel DeVries called attention to a striking anomaly in the carbon-14 abundance in tree rings from the second half of the 17th century and the first part of the 18th. In those years the abundance of carbon 14 rose sharply, exactly as it would have if the sun had been anomalously inactive during the time. The DeVries effect has been confirmed in tree-ring data from around the world. Its association with a concurrent anomaly in the behavior of the sun was not fully appreciated, however, mainly because the history of solar activity was so poorly known. Now it seems certain that the marked excursion in the carbon-14 record first found by DeVries strongly confirms the existence of the Maunder minimum.

The historical verification of the Maunder minimum has thus made possible an exciting breakthrough in the recovery of solar history. Through the analysis of the rings of the bristlecone

pine, the oldest living thing on the earth, the carbon-14 record available today extends back to at least 5000 B.C. The dominant feature in the long carbon-14 record is a slow modulation of the abundance of carbon 14 exhibiting an apparent period of about 10,000 years. That modulation is induced by changes in the strength of the magnetic field of the earth, which acts as another shield against galactic cosmic rays. Numerous other, shorter-term excursions are also present in the carbon-14 record. Now I had found that the most recent of these shorter-term carbon-14 excursions corresponded to a solar anomaly that seemed well established by direct and indirect historical evidence. Although that possible association had been pointed out before, for the first time I knew precisely what kind of change on the sun produced an excursion of a known amount in the carbon-14 record. At last we could calibrate the carbon-14 record in terms of solar change. With the Maunder minimum as our key we could decode the long-hidden history of the sun, not just to the time of Louis XIV or Galileo but back to before the beginning of the Bronze Age some 7,000 years ago. As sometimes happens in crime detective work, the solution of one case had broken open a larger and possibly more important one.

At least 12 features as prominent as the Maunder minimum stand out in the carbon-14 record of the past 5,000 years. Each lasts between 50 and several hundred years. Some are in the direction of severely depressed solar activity, like the Maunder minimum. One of these is the Spörer minimum, lasting from about A.D. 1400 to 1510, when again few auroras and unaided-eye sunspots were reported and when there are no reports of a structured corona from any observer of a total solar eclipse. Other major excursions go in the opposite direction and therefore correspond to eras of anomalously high solar activity, possibly even higher than what we have known in modern times. One of those excursions, which I call the medieval maximum, falls between about 1100 and 1300. Another, I think, is the one we live in now, characterized by the rising level of solar activity that followed the Maunder minimum and that may or may not have reached its peak in the highest recorded sunspot maximum in 1959. Only during this modern maximum do we find accounts of the structured corona and of the high incidence of auroras.

There is uncertainty in interpreting the present era of solar activity from carbon-14 evidence because of the Suess effect: a marked dilution of the abundance of carbon 14 in tree rings formed since late in the 19th century. It has been attributed to the introduction of large



CHANGES IN WORLD CLIMATE coincide with long-term variations in the activity of the sun. The sunspot cycle is shown from 1610 to 1950 (*gray*) along with Oriental observations (to 1750) of large sunspots seen with the unaided eye (*dots*). The relative abundance of carbon 14 in the atmosphere, determined from the analysis of the annual growth rings in trees, is plotted from A.D. 1050 to 1900 (*black*). The ratio of carbon 14 to the common isotope of carbon varies with the level of solar activity: more carbon 14 is produced in the atmosphere when the sun is quiet and less of it is produced in the atmosphere when the sun is active. Here, for the purpose of direct comparison with the curve of solar activity, the curve of the abundance of carbon 14 is inverted, with the abundance increasing down-

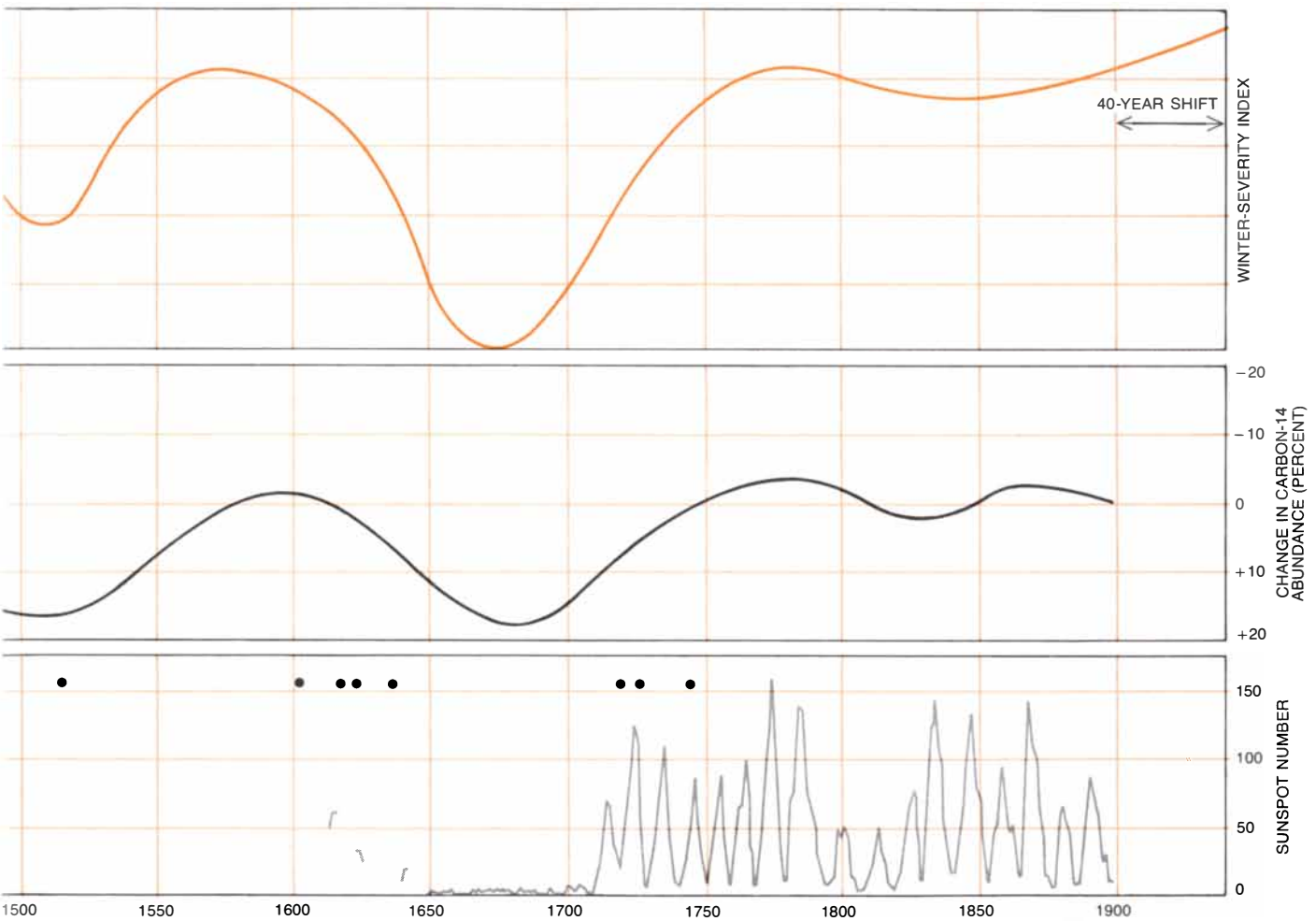
amounts of carbon with low carbon-14 content into the atmosphere through the worldwide increase in the combustion of fossil fuels during the past century. The Suess effect has apparently drawn a curtain on what we can learn from naturally induced carbon-14 changes in the modern era, precluding our using the present state of solar activity to calibrate past changes in the carbon-14 record.

One final comparison remained before me, a comparison that, as Maunder once suggested, may link the long-term solar changes with important effects on the earth. The Maunder minimum corresponds almost precisely with the coldest excursion of the "little ice age," a period of unusual cold in Europe from the 16th century through the early 19th century. In the coldest extremes of that period the average temperature was about one degree Celsius colder than it is now, according to the British climatologist Hubert H. Lamb. In that period the Alpine glaciers advanced farther than they had since the last major glaciation 15,000 years ago. In that period too the Norse colony in southwestern Greenland perished to a man, cut off from the rest of the world by pack ice that year after year failed to thaw. Could that anomaly in the climate—the severest of the past millennium—be related to the long absence of sunspots? Could the disappearance of sunspots and the distorted pattern of solar rotation be indications that the radiation flowing from the sun was slightly reduced? Modern models of the climate show that global cold periods as cold as the little ice age can be brought about by a decrease of no more than 1 percent in the total solar radiation—a subtle change that, if it were drawn out over decades, would hardly be detectable by direct measurement.

On the other hand, the coincidence could be simply that: the Maunder minimum and the little ice age could be unrelated anomalies. A colleague once warned me of the hazard of making such a simple association; he pointed out that one could equally well postulate a connection between the Maunder minimum and the coincident reign of Louis XIV. Could we say that a prolonged sunspot minimum produces a Sun King?

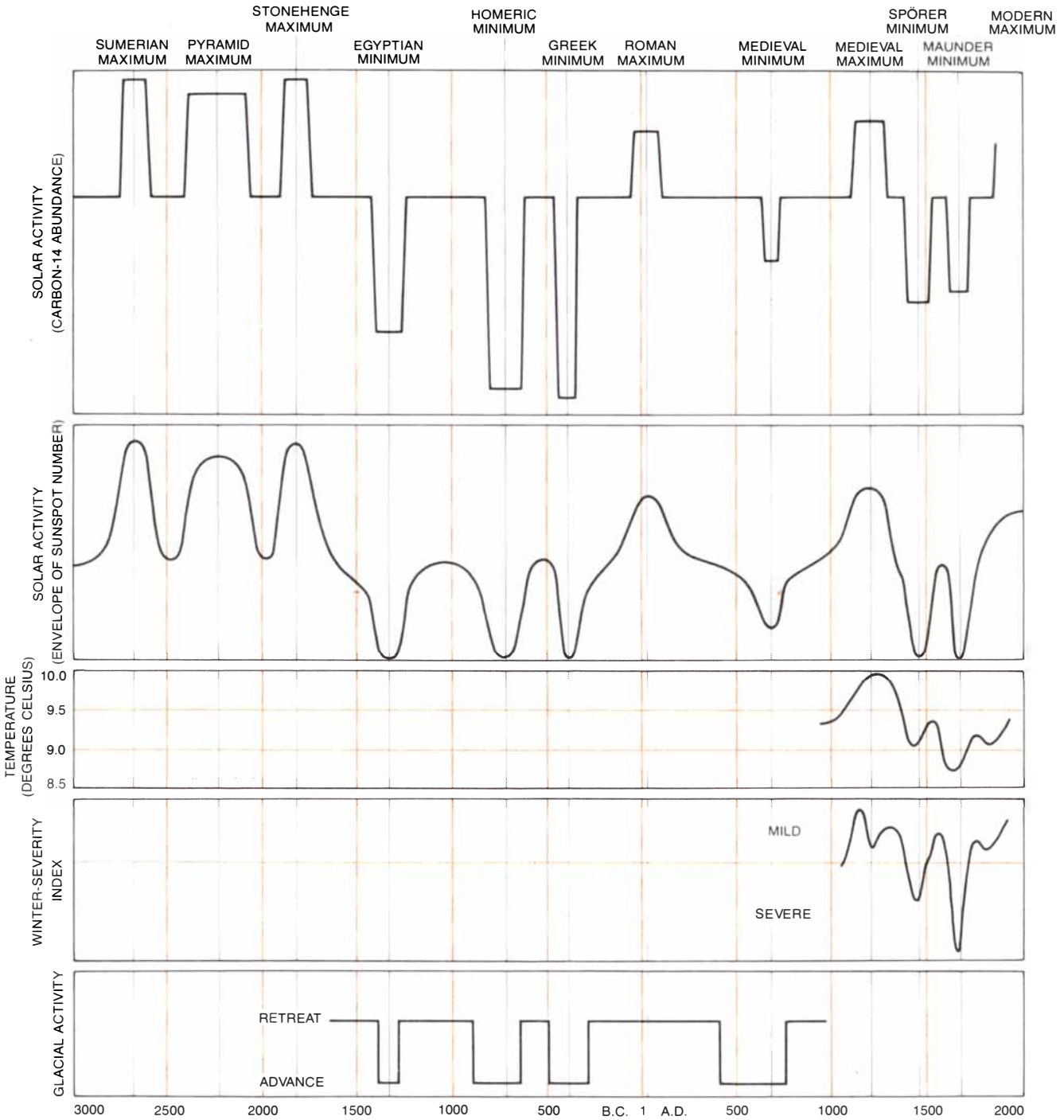
Now that the more extended record of solar history is available in the form of the carbon-14 record, however, we can test for a significant connection between solar change and climate by comparing each of the major solar excursions in the carbon-14 record with the history of the climate at the corresponding time. In making the comparison we shall be limited most by the uncertainty in the climate record: we may now know the history of the sun better than the history of our own planet!

I have compared the carbon-14-de-



ward. The carbon-14 curve and the curve of the general number of observed sunspots fit well. Peaks and troughs in the level of solar activity also coincide with peaks and troughs in the severity of winters in Paris and London (*color*), as determined from historical records by the British climatologist Hubert H. Lamb. Winter-severity index curve is shifted in this illustration to allow for the fact that there is a lag of about 40 years between the production of carbon 14 in the upper atmosphere and the intake of carbon 14 by trees. After about 1900, when both solar activity and world temperature increased, the carbon-14 record is no longer valid as a representation of solar activity; the discrepancy is due to the Suess effect, the modern dilution of carbon 14 in the atmosphere because of the combustion of fossil fuels.

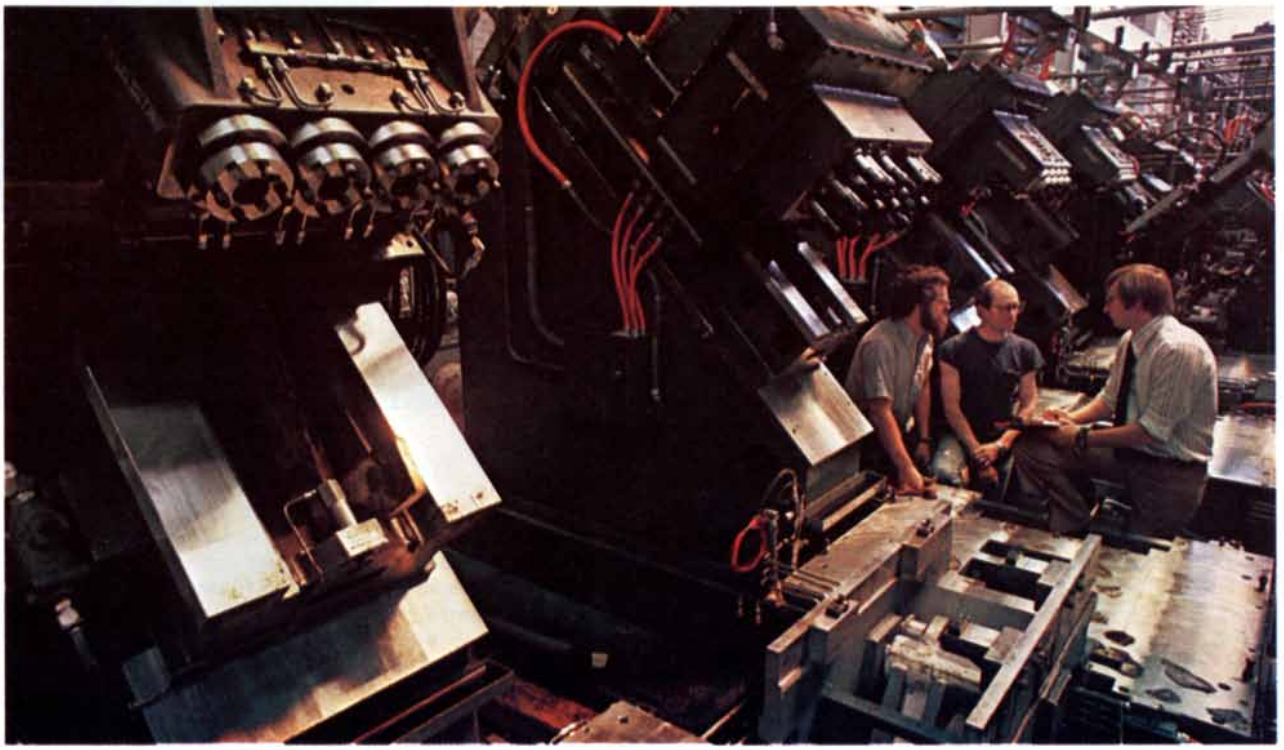
rived history of the sun with the history of the world climate derived by Lamb and others from historical records and from the advance and retreat of the Alpine glaciers. The fit is almost that of a key in a lock. Every decrease in solar activity such as the Maunder sunspot minimum matches a time of glacier advance in Europe; every rise in solar activity such as the medieval maximum matches a time of glacier retreat. The Spörer sunspot minimum corresponds in depth and duration with the first extreme dip in temperature in the little ice age. The solar medieval maximum corresponds with the well-studied medieval warm epoch when the average world temperature was as warm as or warmer than it is now. These early results in comparing solar history with climate make it appear that changes on the sun



SUN'S VARIATIONS SINCE THE BRONZE AGE can be deduced from the abundance of carbon 14 in the growth rings of the bristlecone pine, using the Maunder minimum to calibrate the carbon-14 record in terms of solar change. Pronounced excursions in the carbon-14 record are shown as a square-shouldered curve (*top*) whose heights and durations were derived from the carbon-14 record. Below this curve the author has derived the smoothed historical curve of solar activity (*second from top*), which can be interpreted as a long-term envelope of the amplitude of a possible sunspot cycle. From both curves it is evident that over the past 5,000 years there have been at least 12 solar excursions as prominent as the Maunder minimum; names suggested for earlier features have been assigned according to historical epoch. Estimate of mean annual temperature in England from about A.D. 1000 is shown (*middle*) as well as winter-severity curve from Paris and London (*second from bottom*). Black step curve at the bottom shows the times when Alpine glaciers advanced or retreated. It can be seen that for 5,000 years all climatological curves seem to rise and fall in response to long-term level of solar activity.

DP Dialogue

Notes and observations from IBM that may prove of interest to the engineering community.



Work stations of an Ingersoll transfer machine. The transfer slides are at the bottom of the picture. Visible here is part of one section of a large machine which will make transmission cases.

Simulation Makes Giant Transfer Machines More Efficient

Computer simulation is helping huge transfer machines achieve as much as 15 percent higher productivity.

Robert Callahan, president of Ingersoll Manufacturing Consultants, attributes recent gains to an IBM computer program called General Purpose Simulation System V (GPSS V). It can be used in modeling a broad range of business activities, including manufacturing, physical distribution and transportation.

Callahan's group is a subsidiary of Ingersoll Milling Machine Company, Rockford, Ill., and is separate from that company's machine tool business. It helps manufacturing clients around the world to increase their return on investment and reduce costs.

Transfer machines finish rough castings into complex pieces such as engine blocks. They are named for the transfer mechanisms which automatically move the workpieces through successive 'stations' where cutting is done.

"If a line were built as a single straight-through train of work stations," Callahan says, "every stoppage would quickly bring the entire machine to a halt. So we lay it out in several sections, each supplied from a 'bank' of workpieces which can be drawn upon when upstream sections are stopped. How close a line comes to its production potential is determined by such factors as the number of sections and the size and placement of the banks.

"It is vital," he adds, "to design the section split and the banks correctly before the line is built. But there is no straightforward analytical way to calculate performance in advance. Simulation by computer is the ideal tool for this kind of problem; with a GPSS model of the entire line, we can test proposed layouts quickly and easily."

Callahan's group runs its simulations on an IBM System/370 Model 145 operated by the parent company.

"In designing a line," he adds, "we manipulate the computer model to work our way toward the best configuration. We try different layouts, section splits, banking arrangements and tool placements. GPSS lets us model the characteristics of each tool in detail; the resulting model behaves remarkably like the real transfer line."

Advertisement

Computing Power for Engineers

Design engineers can now access the computer directly through terminals in their offices to test tentative structural configurations, run simulations and develop improved designs. A growing number of companies are raising engineering productivity with IBM interactive computing facilities.

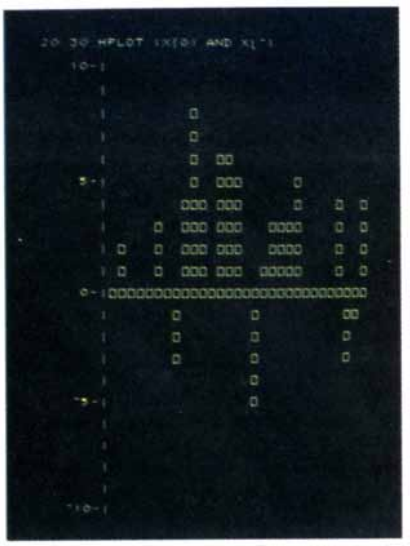
A user at a terminal can activate a previously prepared program stored in the computer. Or he can create a problem-solving routine to meet the need of the moment, using one of the simple-to-learn but powerful programming languages.

Facilities exist for presenting a computation as a curve, bar chart, histogram or frequency distribution on the screen of an IBM 3277 Display Station. In lengthy computations, intermediate results can be displayed, allowing the user to watch the trend and terminate an unpromising trial.

Often, the user of the computer finds that a calculated result suggests further trials with new parameter values, approaching an optimum solution iteratively. With interactive computing, the user can obtain multiple job "turnarounds" in a short time, rather than one or two a day.

Programs can be written for one-time use or to be kept in direct-access storage and invoked whenever needed. Such programs can be as large as a complex system simulation or as small as the evaluation of a simple expression. The computing power is always on tap, whenever it can be useful.

Facilities available from IBM bring interactive computing capability to any System/370 installation.



Simple instructions can specify graphic display (above, bar chart) of interactive computing results.

Computer Cuts Water Use More Than 30% for Farmers

Irrigation is vital to crop production in the Great Plains. The more water applied, the greater the production—up to a point. After that, water and energy are wasted and nutrients are leached out of the soil. Intuitive methods of scheduling irrigation usually lead to over-watering—or, in some cases, underwatering.

"With the help of a computer, a growing number of farmers in this area are achieving water and energy savings of 30 to 40 percent," says Paul Fischbach, extension irrigationist at the University of Nebraska.

These farmers are using one of more than 100 programs available in AGNET (Agricultural Computer Network), a remote-access computer service developed at the university.

They obtain assistance through terminals in many locations throughout the state. When a farmer enters the word 'Irrigate' an IBM computer in Lincoln responds by requesting current information on his field. He enters daily temperatures, rainfall, amount of irrigation applied and soil moisture readings since the previous update.

When the farmer has keyed in all the requested weekly data, the computer uses up-to-date weather statistics for the region and the stored characteristics of the farmer's field to determine a suggested irrigation schedule.

Nebraska farmer Kenneth Bruns takes a reading of soil moisture, one of the variables considered by AGNET in the calculation of an optimum irrigation schedule.

At the university's Institute of Agriculture and Natural Resources, James G. Kendrick, professor of agriculture economics and Thomas L. Thompson, professor of agricultural engineering, lead the continuing development of the online AGNET system, implemented on an IBM System/370 Model 158 at the Nebraska State Department of Administrative Services.

Another AGNET program helps livestock growers calculate optimum feed mixes or formulas. For maximum growth rate, the nutrition requirements are different for various species and change as the animal grows. For example, AGNET identifies 13 different nutritional balances for beef cattle.

To develop a minimum-cost mix of cattle feed that meets nutritional and palatability requirements, the computer asks what ingredients are to be considered, and at what prices. The farmer enters his own prices or uses those in the AGNET data base.

He can ask the computer "what if" questions, test his decisions against hypothetical price and cost fluctuations, and calculate the total costs of crop production under different management techniques. A financial program analyzes investment in capital equipment, using any desired cost and performance assumptions. Other programs recommend pest control schedules, make fertilizer recommendations based on soil analyses and simulate the growth of livestock.

"The computer," Kendrick notes, "is becoming an important tool for improving the economics of farming in this area."

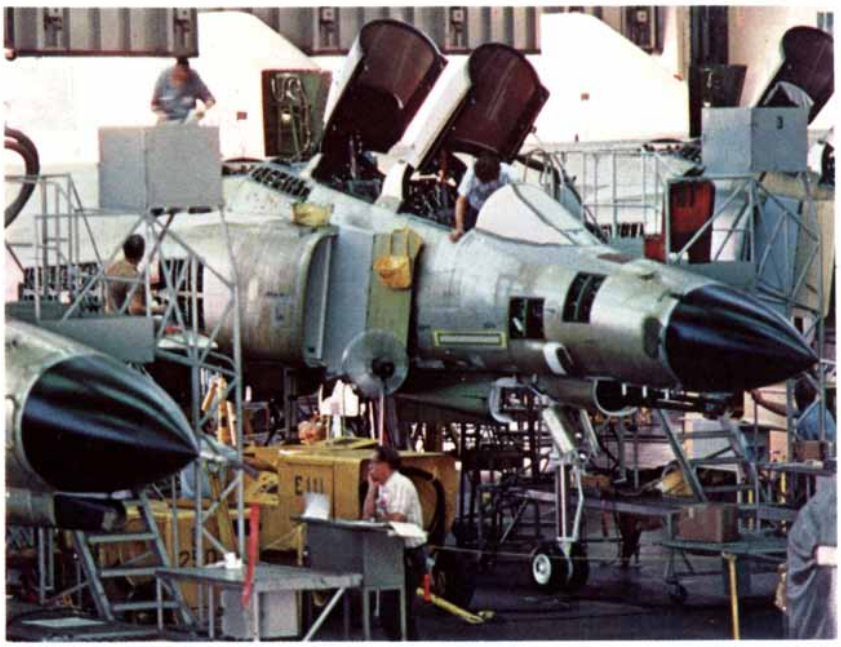
Designing Supersonic Aircraft with a Light Pen

The complex shape of a supersonic aircraft fuselage appears in crisp white lines on a CRT screen. The design engineer seated at the terminal presses a few keys and the image of the craft's landing gear unfolds into the extended position.

But the engineer observes some interference between the landing gear and the fuselage. Swiftly he touches keys and moves a light pen across the surface of the screen. As he works, the shape of the fuselage alters slightly, and when the landing gear descends again it is clear of all obstructions.

The scene is McDonnell Aircraft Company, where the engineer and his colleagues design high-performance aircraft. He is seated at one of 30 IBM 2250 Graphic Display Terminals in the St. Louis headquarters of the McDonnell Douglas Corp. subsidiary, using a graphics processing computer program called Computer Aided Design and Drafting (CADD, pronounced "caddy"). Created by McDonnell, CADD runs on two IBM System/370 Model 168 computers.

"We can sometimes identify engineering productivity gains of ten to one or more," says Stanley LaFavor, director of computer-aided technology. "In one instance, our engineers solved in two days a problem in landing gear placement which we couldn't have solved manually in six months.



McDonnell Douglas engineers design structural components of these high-performance aircraft, working interactively with the computer at graphic terminals.

"Overall, we see an average engineering productivity gain of six to one," LaFavor adds, "and we are accomplishing many engineering tasks with CADD today that couldn't be done any other way. Formerly, with these difficult design problems, we had to take the first solution that would work; now we can find the best one."

A remarkably sophisticated graphics processor, CADD permits the engineer to display sight lines (indicating, for example, the unobstructed view from the pilot's seat), or to display any desired cross section of any geometric shape he has defined to the computer. Not only landing gear but any articulated part—hinged doors in compound-curved surfaces or an in-flight missile launching system—can be displayed and its motion in space delineated and checked for interference.

CADD displays three-dimensional shapes in trimetric projection, a type of perspective frequently used in drafting. The system can rotate a projection around any selected axis, move it in any direction or change its scale. The user defines a geometry and commands system functions by means of the light pen, the keyboard of the 2250, and a set of pushbuttons which control powerful graphics functions in the CADD program.

CADD permits structural components of the aircraft to be designed in detail at the terminal. To configure a fuselage bulkhead, for example, the engineer asks the computer to display the desired cross-section of the fuselage and to subtract the skin thickness to arrive at the exterior shape of the bulkhead. He then works out the machined shape—webs, flanges, lightening holes—using the light pen and the keyboards.

In addition to its graphics capabilities, CADD serves as the core component of a related set of computer programs supporting all aspects of structural design of an aircraft. These subsystems are linked to one another by the CADD data base, in which all details of the aircraft are accumulated as design progresses. One subsystem generates loft lines; others perform structural analysis and calculate the weight and structural dynamics as the detailed components of the craft take shape. A third calculates the operational performance of the hypothetical aircraft.

"We've eliminated much that is time-consuming and repetitive in engineering," says LaFavor. "Our major thrust has been to improve the flow of data from engineering to manufacturing, saving downstream dollars."

IBM Interactive Languages for Engineers

Three IBM programming languages are designed specifically for problem-solving by engineers and other non-data processing professionals:

1. **VS APL** A broadly applicable interactive language, simple to learn, yet uniquely powerful for scientific and mathematical problems.

2. **VS BASIC** Powerful for a wide range of problem-solving, and flexible without sacrificing simplicity.

3. **VSPC FORTRAN** Permits problem-solver to create and invoke FORTRAN programs directly. The user enters data and receives results at the terminal.

For more information on these languages, write to the address on the right.

DP Dialogue is designed to provide you with useful information about data processing applications, concepts and techniques. For more information about IBM products or services, contact your local IBM branch office, or write Editor, DP Dialogue, IBM Data Processing Division, White Plains, N.Y. 10604.

IBM®
Data Processing Division

are the dominant agent of climatic changes lasting between 50 and several hundred years.

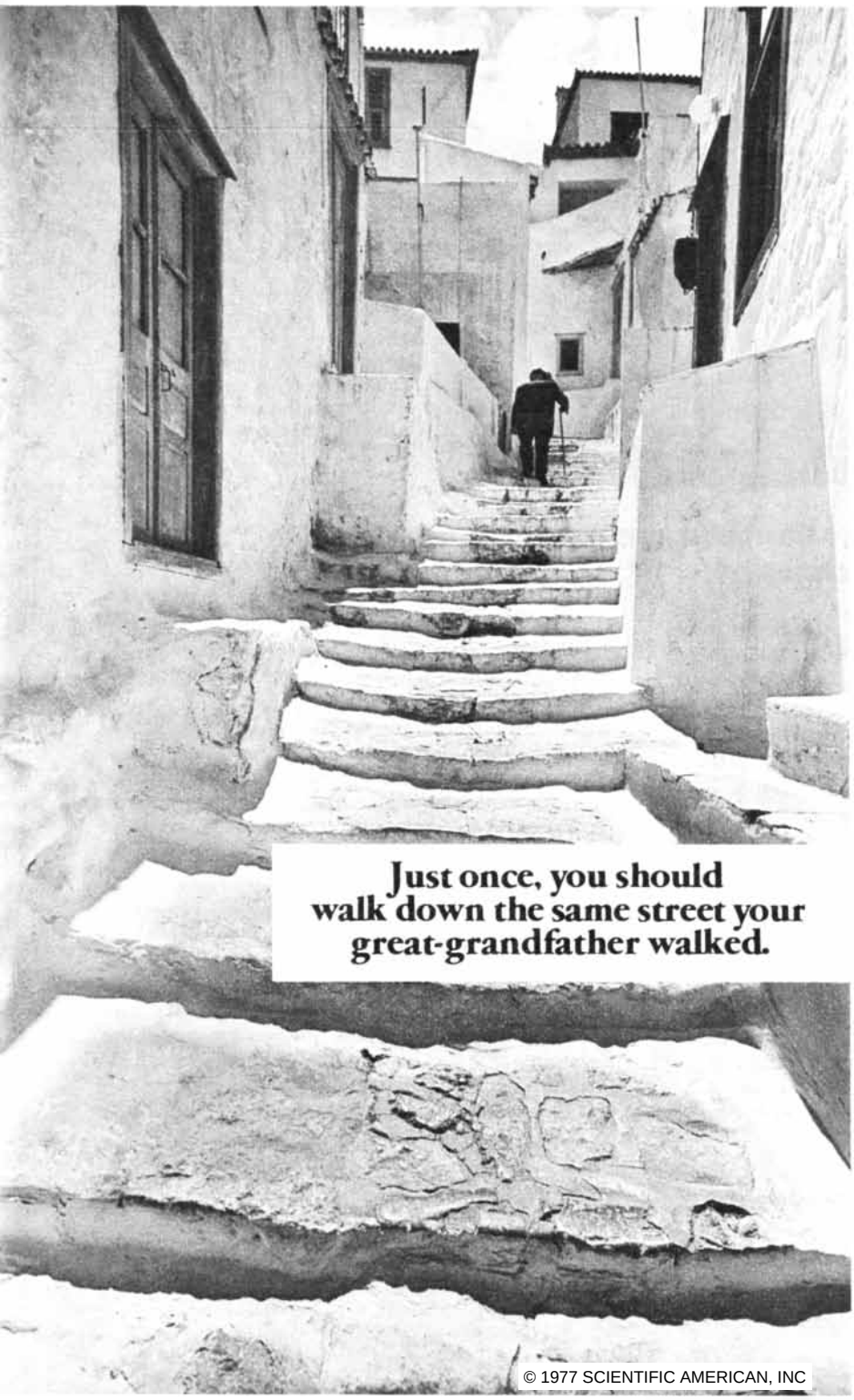
Curiously, the apparent link between solar change and climatic change may reveal little of whether short-term changes on the sun such as the 11-year sunspot cycle have any major effect on short-term weather. In the Maunder minimum or the medieval maximum we are studying not the individual ups and downs of the sunspot cycle but the longer-term envelope connecting the peaks of many cycles. It seems possible to me that the slowly varying envelope could reflect slow changes of a few percent in the total energy output of the sun, changes that could be quite independent of whether the 11-year cycle was at its maximum or at its minimum. A changing solar output could modulate the amplitude or intensity of a continuous series of 11-year sunspot cycles, each of which would still show a peak and a valley. As with an amplitude-modulated radio signal, the message would be carried not in the individual cycles of the continuous wave but in their varying amplitude, which is most recognizable in the envelope of the peaks.

The intensity of the sunspot cycle could by modulated by means of the solar dynamo, when slow changes in the flow of solar energy alter the structure of the sun's convective zone and hence the pattern of circulation at the surface. The radiative output of the sun could be almost entirely independent of the phase of the 11-year sunspot cycle. Such a mechanism might explain why investigations of the connection between the sun and the weather have been so fraught with frustration when correlations have been sought with the 11-year sunspot cycle, which may be only the carrier frequency.

It would seem that Maunder and Spörer were right and that most of the rest of us have been wrong. As is often the case in the onrush of modern science, we had too quickly forgotten the past, forgotten the less-than-perfect pedigree of the sunspot cycle and the fact that it too once came as a surprise. We had adopted a kind of solar uniformitarianism, contending that the modern behavior of the sun represented the normal behavior of the sun over a much longer span of time.

As people and as scientists we have always wanted the sun to be better than the other stars and better than it really is. Long ago we held that the sun was perfect, and when the telescope showed that it had spots on it, we took comfort in the thought that it was at least regular in its behavior. It now appears to be neither of these, and it is probably inconstant as well. Knowing this, however, opens the way to a much fuller understanding of the sun and its important influences on the earth.

All of us come from someplace else.



Picture this if you will.

A man who's spent all his life in the United States gets on a plane, crosses a great ocean, lands.

He walks the same streets his family walked centuries ago.

He sees his name, which is rare in America, filling three pages in a phone book.

He speaks haltingly the language he wishes he had learned better as a child.

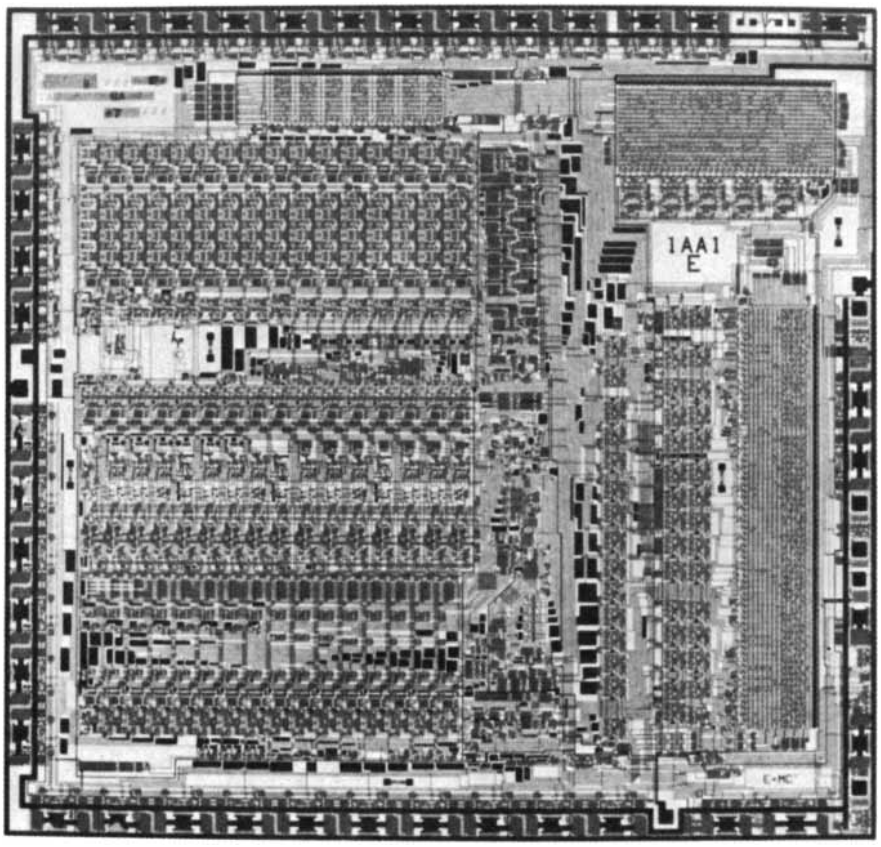
As America's airline to the world, Pan Am does a lot of things.

We help business travelers make meetings on the other side of the world. Our planes take goods to and from six continents. We take vacationers just about anywhere they want to go.

But nothing we do seems to have as much meaning as when we help somebody discover the second heritage that every American has.

PAN AM®
America's airline to the world.

See your travel agent.



Actual size of the MC² chip is ■, with 10,000 transistors in its 34 mm² area. One thing the enlargement at left cannot reveal is the chip's translucence which, as the connoisseur of LSI knows, is a property peculiar to the sapphire substrate.

Hewlett-Packard announces the first microprocessor chip from its silicon-on-sapphire LSI technology.

Ordinarily, we prefer to reveal our technological advances through the end products that embody them. An exception is our new MC² chip. HP has developed a silicon-on-sapphire technology with complementary metal-oxide-semiconductor processing (CMOS/SOS) that makes possible a range of dependable, high-speed, large-scale integrated circuits. These circuits, of which the MC² (for Micro CPU Chip) is our first, are destined to play significant roles in products that will appear from time to time in these pages.

Two long-standing criteria in HP product development are invention and producibility: providing our customers with advanced products that offer dependable, high performance at a fair price. In these days and times, performance rests heavily on large-scale integrated circuits.

A case in point is the MC², the first chip produced by our new CMOS/SOS process.

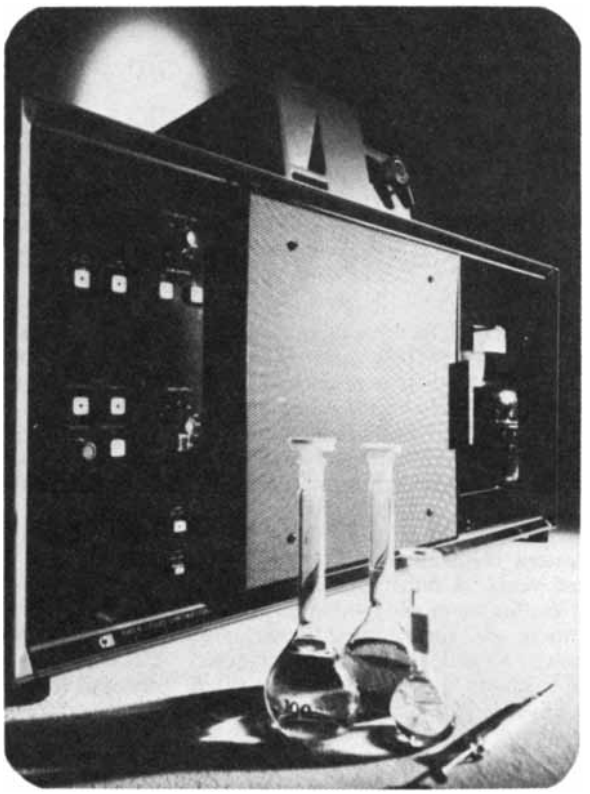
The major distinction between SOS and other integrated circuit technologies is that instead of being formed in a wafer of bulk silicon, the circuitry begins with a thin layer of silicon on a sapphire substrate. Because CMOS/SOS provides superior dielectric isolation, circuit elements can be very closely spaced. MC², for example, is a complete 16-bit parallel microprocessor that contains 10,000 transistors in a 34 mm² area.

Low parasitic capacitance is a further advantage provided by the sapphire substrate, greatly improving the speed of LSI circuits. Thus our little jewel, which executes 34 classes of 16-bit instructions, can perform a full register-to-register addition in 875 nanoseconds, yet power consumption is typically only 350 milliwatts. Instructions tailored for maximum flexibility in handling a range of peripherals make the chip easy to incorporate in electronic or mechanical devices as a direct, high-speed controller.

An article on this new, high-density chip appears in the April issue of the *Hewlett-Packard Journal*. If you'd like to delve further into the properties of MC², mail the coupon. We'll gladly send you a copy.

High-performance liquid chromatograph improves speed and accuracy of hydrocarbon group analysis in gasoline.

The HP 1084 integrates a high-performance liquid chromatograph (HPLC), digital processor, and printer/plotter in a single instrument. It automates the analysis from injection to final report of volume percent in a tenth of the time required by the traditional, standard method—and with greater accuracy.



One standard method used by petroleum refiners to characterize gasoline distillates by hydrocarbon groups is fluorescent indication adsorption (FIA). This method requires the analyst's constant attention throughout the analysis, which can run up to four hours for a single sample. And quantitative results can be biased by the presence of large amounts of olefins, diolefins, and light hydrocarbons.

A few years ago, Mr. J.C. Suatoni and some of his associates at Gulf Research in Pittsburgh, Pa., first applied high-performance liquid chromatography as an alternative approach to hydrocarbon group analysis. Now, with the microprocessor-based HP 1084, the method developed by Gulf Research scientists can be completely automated, making the analysis simpler, faster, and more precise than the FIA method.

The HP 1084 accepts all operating instructions on its control keyboard before the analysis begins. After the analyst pushes the *start* button, the instrument's microprocessor takes over full control: it automatically injects the gasoline sample; controls the operation of all aspects of the liquid chromatograph; plots the chromatogram; computes the retention time and area of every peak; computes the normalized volume percent for each hydrocarbon group; and prints a detailed analytical report. When an automatic sampler is used, the HP 1084 can analyze a whole series of samples completely unattended.

At a price of $23,500*, the HP 1084 brings to high-performance liquid chromatography the same kind of automatic operation, speed, and precision that the microprocessor-controlled HP 5840 brought to gas chromatography two years ago. Between them, the two instruments offer a choice of advanced analytical instrumentation for virtually any kind of organic material regardless of its separation characteristics, molecular weight, boiling range, and heat sensitivity.

HEWLETT hp PACKARD

1504 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, 1504 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) MC² micro CPU chip
( ) HP 1084 high-performance liquid chromatograph

Name _____
Company _____
Address _____
City _____ State _____ Zip _____

*Domestic U.S. prices only.

00744

Exploring the Herbarium

Valuable information on little-known plants that might provide new foods or drugs can be gathered by examining labels attached to the millions of specimens in the world's botanical collections

by Siri von Reis Altschul

In view of the fact that there may be as many as 800,000 species of plants on the earth it is remarkable how few of them have been directly exploited by man. Only the three major cereals and perhaps 10 other widely cultivated species stand between famine and survival for the world's human population, and a handful of drug plants has served Western civilization for several thousand years. A rather obvious question arises: Are we missing something? Surely there are, among the 250,000 or so known, named and described species, other plants that could be cultivated to enhance the world's supply of food or whose active chemical compounds could be isolated for medicinal purposes. How is one to find them?

Some years ago it occurred to me that the world's herbaria constitute an accessible and immensely rich source of information on plants with potential food or medicinal value. There are about 1,800 public herbaria in the world, housing upward of 175 million dried and pressed plant specimens that represent the approximately 250,000 different known species and some others not yet identified. I had seen enough of such collections to know that in some instances the field notes attached to a specimen included ethnobotanical information: data on local lore pertaining to the plant and on the uses to which it is put by the indigenous people of the region. I recognized that for the taxonomists and students of plant evolution who traditionally work in herbaria such information must often appear to be peripheral and incidental, and so it would for the most part have been overlooked. I decided to pursue the matter. A five-year exploration of the collections in Harvard University's Gray Herbarium and Arnold Arboretum convinced me that valuable data, gathered and recorded over several hundred years and from all parts of the world, are there on herbarium labels, waiting to be brought to light.

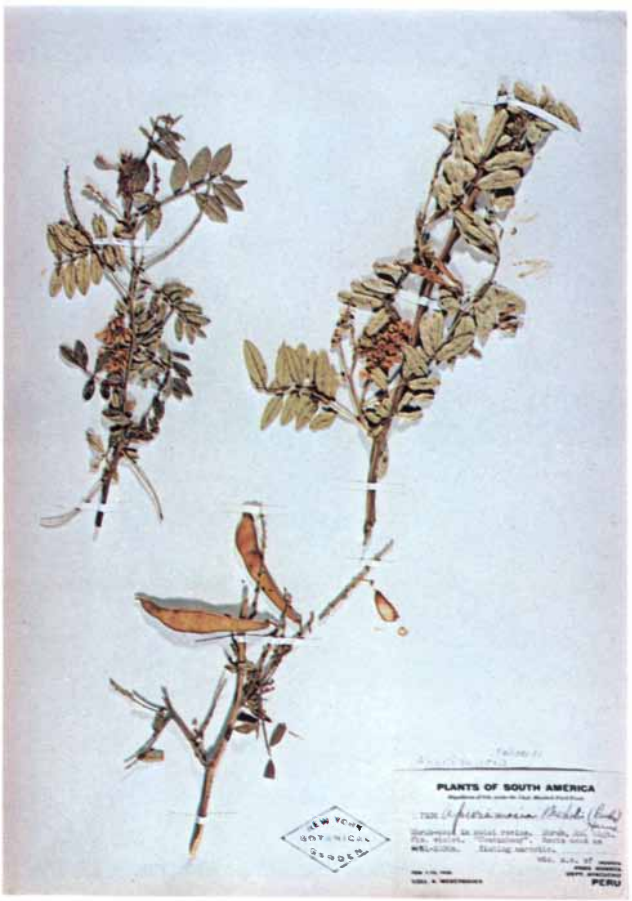
It is not really surprising that herbaria should be repositories of clues to the uses of little-known plants. From its beginnings the study of plants was deeply concerned with the uses of plants, and in particular their uses as drugs; until rather recently, in fact, botany and medicine were closely intertwined. Primitive peoples in all ages have had extensive knowledge of the vegetation around them. This is as true of the world's remaining aboriginal peoples, whose folk taxonomies are based largely on utilitarian considerations, as it was of ancient civilizations. As early as 5,000 B.C. the Chinese had well-developed floral pharmacopoeias, or sets of accepted plant-derived drugs; so did the Babylonians, the Assyrians and the Hebrews later. The works of Aristotle and Hippocrates reveal that the Greeks knew many of our currently used drugs. Pliny and Galen wrote about drug plants, and Pliny's contemporary Dioscorides produced the great *De materia medica,* which dealt with all the medicinal substances known at the time, including some 600 plants, and remained the authoritative treatise on the subject for 15 centuries.

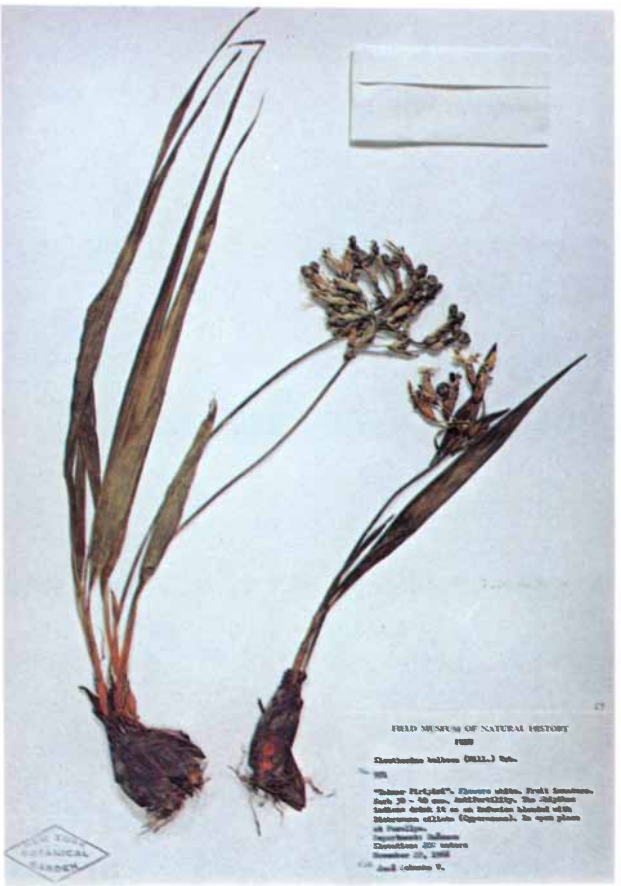
The invention of movable type in about 1440 heralded the age of herbals: illustrated compendiums of data (some true and some false) about plants, with emphasis on their medicinal or food value. Herbals stimulated a new interest in botany, both in the plants that grew north of the Alps and in those that had been known to the ancient world. Because now the written word and drawings could be duplicated easily and distributed widely, plants found in different places could be compared. As a result efforts to classify the various species increased. There does not seem to have been any attempt to create systematized collections of dried and pressed plants, however, until the middle of the 16th century, when the Italian botanist Andrea Cesalpino and his associates began to preserve specimens for future reference and authentication. A few botanists in central Europe followed suit. Soon the advantages of maintaining collections of preserved specimens were widely appreciated, and herbaria as we know them were born.

Herbals, herbaria, the changing intellectual climate of Europe and the discovery of the New World and its vast natural riches all contributed to a new trend where medicine retained its central interest in plant-derived drugs but botanists were increasingly fascinated not just by the uses of plants but by the plants themselves. It soon became apparent that the various systems by which plants were classified were cumbersome and inadequate to the growing task. Order was imposed on confusion by the great 18th-century Swedish botanist and zoologist Carl von Linné, or Linnaeus. He invented an admittedly artificial but serviceable system for classifying plants based mainly on stable and obvious flower characteristics: the number, position or relative length of the stamens. More important, he established a uniform way to designate species by two Latin names, and he saw the method accepted internationally within his lifetime. Linnaeus himself named some 10,000 species; his *Species plantarum* of 1753, which described and gave two-word names to all the known species, and his *Genera plantarum,* which described all the known genera, are still living classics to which taxonomic botanists resort regularly. It is remarkable that books written two centuries ago remain the cornerstones of a contemporary discipline.

Linnaeus lived in a time of high hopes

HERBARIUM SHEET displays a dried plant specimen together with detailed field notes, which may include ethnobotanical data suggesting food or medicinal uses. This specimen, photographed by Fritz Goro in the herbarium of the New York Botanical Garden, is a bamboo classified in the genus *Guadua* of the grass family. It was collected near a river in what was then British Guiana. The note of interest states: "Said to be poisonous. Used in arrow points."





for the accumulation of encyclopedic knowledge, hopes animated by faith in the existence of a fixed natural world endowed with immutable order in a single moment of creation; it seemed that careful observation of all the parts of that world would surely reveal its underlying coherence. Thus inspired, botanical exploration was pursued with florid zeal, notably by a remarkable group of Linnaeus' "apostles," as he called them, who embarked on numerous expeditions to the Americas, the South Seas and the Far East. Other investigators followed, and specimens poured into the herbaria of Europe. Thanks to Linnaeus' system, these masses of plants could be classified and filed conveniently, new species could be described concisely and their descriptions could be disseminated and understood.

In the 19th century botanical collecting received a different kind of impetus from the opening of the mind of Western man to new ideas about himself and his relations with the rest of the natural world. As the concept of evolution moved into human consciousness another wave of exploration was set into motion, this time in pursuit of an endlessly changing nature. Collectors were eager to find out not only what species there were but also how they had come to be, and to learn which plant groups were related and how. Specimens were collected and filed in herbaria for later investigation by new techniques ranging from cytology and physiology to plant geography and ecology. The evolutionary rationale required the revision of many artificial groupings that had been established according to other considerations; Linnaeus' classification based largely on stamens was an early casualty. The new investigations required extensive herbarium and laboratory study, often followed by further fieldwork and more purposeful collecting. Herbarium-making became a major feature of late-19th-century systematic botany, which combined taxonomy (classification as an expression of evolutionary relationships) and nomenclature. Indeed, systematics became almost an obsession in the botany departments of American and European universities, and eventually there was a reaction against collecting, classifying and herbarium-making; some collections were stored away and even dispersed. Today the collecting of new specimens and their maintenance in herbaria are recognized as being central to the basic studies of plant biology and evolution that largely occupy botanists. And systematics continues to fulfill a unique role: it absorbs and integrates knowledge from an increasing number of disciplines, and in turn serves those disciplines by identifying new species and shedding light on relations within the plant kingdom.

Throughout the world several thousand plants are valued for their medicinal attributes, most of them by primitive peoples or by folk practitioners. Modern Western pharmacology has recognized that many of the traditional drug plants contain in their tissues particular compounds that have specific effects on the human body. There is a broad spectrum of these biodynamic chemicals, which can be classified as alkaloids, glycosides, essential oils, fatty oils, resins, mucilages, tannins, gums and other substances. It has been estimated that more than half of the prescriptions written by American physicians today contain a plant-derived drug: a drug that actually comes from a plant or one that has been synthesized to duplicate (or to improve on) a plant substance.

For example, the meadow saffron, *Colchicum autumnale,* supplies us with the alkaloid colchicine, which is a specific against gout. Ipecac, from the root of *Cephaëlis ipecacuanha,* was an old remedy of South American Indians when it was discovered by the 18th-century privateer-physician Thomas Dover; ipecac became the main ingredient of Dover's powder and is given today to treat amoebic dysentery. The snakeroot, *Rauwolfia serpentina,* was sold 3,000 years ago in the bazaars of India for snakebite and as a calmative for psychic frenzies; since the 1940's its active principle, reserpine, has been recognized as an effective drug, first as a tranquilizer and now primarily as an agent to lower blood pressure. Curare is a quick-acting arrow poison still made by South American Indians from a mixture of up to 30 different kinds of plants. One of its active alkaloids is the basis of drugs that are important in modern surgery as muscle relaxants. Quinine, a widely used and inexpensive antimalarial drug, comes from the bark of a species of *Cinchona,* a shrub in the coffee family that was known to the Incas before the Spanish conquest. The Jesuits became familiar with it (hence one of its names, Jesuits' bark) and introduced it into European medicine.

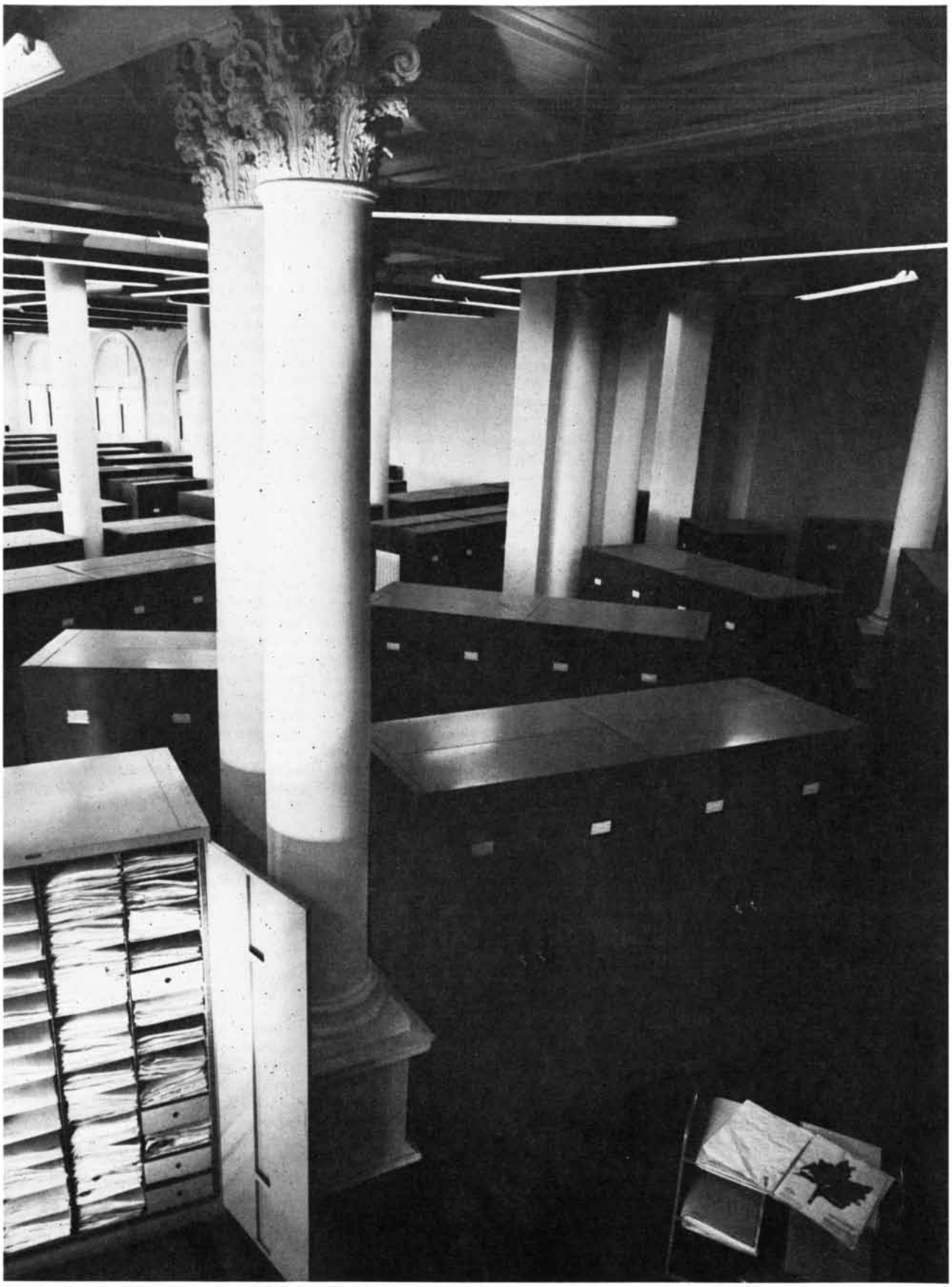
One of the oldest drug plants that are still much with us is *Ephedra,* a shrub known in China for more than 5,000 years, whose alkaloid ephedrine serves today as a nasal decongestant and a stimulant for the central nervous system. The deadly nightshade or belladonna, *Atropa belladonna,* is another old medicinal plant, from which is derived a powerful alkaloid, atropine, that dilates the pupil of the eye. Cocaine, now mainly synthesized, was discovered in *Erythroxylon coca* a shrub whose leaves are still chewed by Andean and Amazonian Indians to alleviate hunger and fatigue and as a stimulant narcotic. The leaves of *Digitalis purpurea,* the purple foxglove, contain a glycoside with specific therapeutic effects on cardiac muscle that were recognized in folk medicine even before the English physician William Withering demonstrated the plant's value late in the 18th century; digitalis is still a major drug for heart failure.

Ergot is a poisonous substance in the fungus *Claviceps purpurea,* which infects rye and other grains. In A.D. 944, 40,000 people died in France from an epidemic of ergot-contaminated bread, and such plagues recurred until their cause was identified in the 17th century. Midwives had long known, however, that ergot would speed labor and inhibit postpartum bleeding. Today a derivative of the parasite provides the most effective relief from migraine headaches that is available, and alkaloids derived from ergot are administered to contract uterine muscle in cases of difficult childbirth. From time immemorial the Japanese have eaten a red seaweed called *Digenea simplex* to rid themselves of intestinal parasites. The active principle was found not long ago to be kainic acid, a potent substance that now has become a standard treatment for intestinal parasites in the Far East.

All human cultures have their pharmacopoeias, and it has been estimated that from 25 to 50 percent of the drugs in non-Western pharmacopoeias are empirically effective. In other words, up to half of the drugs in native medicinal repertories actually may cure or provide relief—if not necessarily for the reasons given by local tradition. According to Richard Evans Schultes of the Botanical Museum of Harvard University, the screening of the plant kingdom for substances of pharmacological value is so incomplete that probably fewer than half of the compounds produced by plants have been identified, much less assessed for effectiveness.

Now, how do investigators go about locating and evaluating plants and their constituents? The usual source of leads is the published literature: classical writings, early descriptions of plants and

MEDICINAL APPLICATIONS are suggested by the notes on the sheets shown on the opposite page. *Cymbopogon densiflorus* (*top left*), a grass from Tanganyika (now Tanzania), has flowers that "are smoked alone or with tobacco by witch-doctors and cause dreams. It is said that these dreams foretell the future." *Billia columbiana* (*top right*), from Venezuela, has a "rusty brown" fruit whose dried seeds are "placed in infusion, used as drink to help mothers in child birth." The roots of *Apurimacia michelii* (*bottom left*), a shrubby legume from Peru, are "used as fishing narcotic." *Eleutherine bulbosa* specimen (*bottom right*), found in Peru by a later expedition, carries the notation: "Antifertility. The Shipibas indians drink it as an infusion."



HERBARIUM of the New York Botanical Garden contains some four million specimens, filed phylogenetically by family and alphabetically by genus and species, in 1,250 steel cases. (Capitals of modified Corinthian columns are based on the leaf of *Acanthus spinosus*.)

their effects, and later accounts by explorers and anthropologists. Even if one goes to primary sources, however, one finds that specimens usually are not cited, so that it is impossible to authenticate the identity of the plant to which the text alludes. A 16th-century Spaniard may have told of New World Indians who gather a "plant with broad leaves and small seeds, which they drink through their noses and which produces the most remarkable reactions, wherein these *Indios* throw themselves into a divine frenzy accompanied by marvelous visions," but we can only guess the identity of the plant. Even modern accounts of anthropological fieldwork among primitive societies usually have ignored the critical botanical identifications; only recently have such reports begun to cite "voucher," or reference, specimens.

Alternatively one might turn to floras, which are compiled by botanists and list the species that grow in a particular region, but such lists usually omit ethnobotanical data! Again, the seeker for new drug plants can go into the field himself and try to learn which plants are considered valuable for which purposes. It takes a long time and much preparation to become conversant with one group of people, however, let alone to become intimate enough with medicine men or elders to be allowed to share their well-guarded secrets. Substituting chemistry for local lore, the field-worker can resort to quick tests that reveal the presence or absence of certain chemicals in a plant. Unfortunately, one has to decide ahead of time what kinds of compounds to look for and then apply tests for those particular substances—alkaloids, say, or glycosides; there is no field test that can reveal all the constituents of a plant or what their potential uses might be.

Archaeology has been a source of some information on the uses of plants; from it we have learned in particular much about the origins and development of food plants. Little evidence of medicinal plants has survived from early sites of human occupation, however. Archaeological remains are usually desiccated, carbonized by fire or preserved in water and are often fragmentary and fragile. The study of coprolites (fossilized feces) reveals what foods early man ingested but not why each food was consumed.

That leaves herbaria, which were founded primarily in the interest of medicine and now are available to serve medicine again, and perhaps also to serve food science and nutrition. As herbaria have multiplied and have enlarged their collections over the centuries, collectors have persisted in making field notations about the uses and reputed effects of the plants they gathered. As late as the 19th century many collectors were physicians—including, for example, Asa Gray and John Torrey in the U.S. The firsthand observations they made were set down on labels attached to their specimens, so that there is no question about the plant to which a note refers. Because these field notes indicate where a plant was collected, investigators can return to the original locality to obtain material in bulk, if necessary, for chemical analysis. Most of these notes have never been published.

PLANT SPECIMENS are taped or glued to standard 12-by-17-inch sheets of heavy paper and filed in folders. Here a herbarium assistant is refiling sheets that have been out on loan.

It was in the course of a taxonomic study in the Harvard collections as a graduate student during the late 1950's that I began to be impressed by the generally ignored information on herbarium sheets. I was working on a rare South American narcotic and hallucinogenic genus, and I was looking for any scrap of ethnobotanical information. The published literature, much of it based on reports from the days of the Spanish conquest, rarely cited voucher specimens, and so I was overjoyed to find any data that actually accompanied a specimen. To what extent, I wondered, might the vast Harvard collections contain valuable unpublished information on the medical folklore of other plants? Would it be worthwhile to look at every one of the two and a half million herbarium sheets in order to extract such information? I felt it would be important to make the most of the bits of information that hundreds of collectors had gathered and recorded in the hope that they might thereby contribute not only to taxonomy but also to the advancement of general knowledge. Once my doctoral work was completed I began the search with the help of two part-time assistants.

What we looked for and recorded was any note suggesting biodynamic constituents: those that would have effects on living tissue, from skin irritants and poisons through medications of any kind to foods, beverages and spices. We also in-

CARL VON LINNÉ (LINNAEUS) developed the system of nomenclature that facilitated the recording and classifying of species. His first major expedition was in Lapland in 1732; five years later he posed for a portrait in Lapp dress, holding *Linnaea borealis*, renamed in his honor.

cluded species associated with magic or ritual or that were said to frighten people or affect them in some other way and plants whose applications suggested they had some unreported chemical activity. In addition to plants affecting man we recorded those that were said to interact in some way with animals, including fishes or insects; plants with local names that were suggestive of biological activity ("fruta del diablo" or "café de montaña," for example), and plants that had been collected expressly for chemical assessment.

As we went along, however, we eliminated any species whose uses already were well known—perhaps a fifth of the notes of interest. The largest group we excluded were plants that had been listed in J. C. Th. Uphof's *Dictionary of Economic Plants,* against which we checked each prospective entry. We also eliminated what are called "type specimens," which by definition are reported in professional journals when a new species is named, and also any species whose Latin names alone called attention to medicinal or nutritional properties. And we omitted duplicate reports on a single species in the same geographic area.

We listed the notes in the order in which their species were found in the collections, family by family in what was a more or less evolutionary sequence. Such an arrangement would serve to call attention to similar characteristics in related species and to genera and families having biodynamic properties in common, which might deserve special attention. The entries begin with the Latin name of the specimen, cite the country in which it was collected, the collector's name and his field number for the plant and then quote the ethnobotanical or other note of interest.

It took us four and a half years to go through all two and a half million specimens of higher plants mounted on 12-by-17-inch herbarium sheets. We ended up with almost 6,000 notes of interest referring to more than 5,000 species in nearly 2,000 genera; of some 300 families in the herbaria, more than half were represented in the notes. The families that were most heavily represented were the composites, which accounted for 15 percent of our species; the spurges (8 percent); the legumes (7 percent), and the mints (6 percent). I should point out that quite different families might be heavily represented in a similar search in another herbarium, since collections tend to be specialized in certain areas; the Gray Herbarium and Arnold Arboretum are particularly rich in material from eastern Asia, North America and South America and the South Pacific.

We had undertaken the project as a search for medicinal plants, but it soon became clear that one could not always distinguish between drugs and foods; a "tea," for example, might be simply a

beverage or it might be therapeutic. Moreover, in order not to miss any biodynamic substance we had decided to include all notes that stated a species was edible or even suggested that it was ingested in some way. In any case there was no reason to try to draw a line between medicine and nutrition; nutrition affects health, and new food plants could be of great potential value. As a result what began as a search for drug plants produced a windfall of possible food plants.

Since the notes had been made by different collectors with very different interests and approaches, there was no uniformity in the kind of ethnobotanical information that was included; it ranged from general to specific—from indications that a plant was edible (1,255) to one report of a treatment for typhoid fever. In between we catalogued 922 aromatic plants, 407 said to be medicinal, 284 poisonous plants, 241 plants given for gastrointestinal disorders, 190 reported to have analgesic properties and 116 that were suitable for treating injuries. Other labels described plants providing beverages, teas, tonics, astringents or antidotes, or that are administered to treat specific illnesses or conditions.

The food plants fall into a number of categories. Some species are in genera that have not been known to include any economically valuable plants. For example, *Burckella cocco* from the Solomon Islands is said to bear "a delicious fruit, with a deep white flesh, similar to an apple, but with a strong distinct flavour of its own." In many instances what we found were little-known species that are members of genera having other species we do know as edible—for example a new member, from Peru, of the genus (*Rubus*) that includes raspberries and blackberries. The genus *Salvia* includes many beverage plants; *Salvia hispanica* is not known as one of them, but we found many notes from Central America and the West Indies referring to its edibility, one collector mentioning that "very large crops can be readily produced." *Cyphomandra* includes two South American species with edible fruits resembling the tomato; our notes include a third species, *C. splendens,* reported to be sold as food in the marketplace at Cuzco in Peru. Sometimes a mere resemblance to a food may be a clue, as in the case of *Uvaria confertiflora* from North Borneo, with a "fruit like bananas."

Occasionally there are indications that a plant may contain rare minerals or vitamins, enzymes or antibiotic agents. For example: "The natives have a superstition that if bits of this tree are given with the food of pigs, they will fatten quickly." There are also "emergency" food plants of man and other

animals, preferred grazing materials and other plants said to be eaten avidly by animals, plants that are not consumed directly but have a role in food preparation and plants that are consumed in extraordinarily large amounts. Other species of interest are those said to be good-tasting, fragrant or aromatic even if they are not reported as being edible; those that are chewed but not swallowed, and those that are associated with agricultural rituals. Such species might be developed as food adjuncts.

In the vast field of potential drug plants I shall cite some examples from one randomly selected page of our report. *Cologania angustifolia* var. *stricta,* from Mexico, has "roots said to be purgative." *Erythrina rubinervia,* from El Salvador, has flowers that "cause sleep if much eaten." In Mexico *E. lanata* is "used as a source of poison by Indians." The sap of *E. variegata* var. *orientalis* "is used for making a cough medicine" in the Solomon Islands, and the bark of *E. subumbrans* is boiled in the Philippines "and drunk for splenomegaly" (enlargement of the spleen). The leaves of *Strongylodon lucidus* from the Solomon Islands "are heated then rubbed on boils." *Mucuna monosperma* from Bombay is taken "medicinally with other ingredients to relieve acute spasms." And *Calopogonium mucunoides* of the Caroline Islands makes a "medicine for general debility; . . . the young leaves and shoots are pounded . . . squeezed into a betel nut bowl, . . . coconut water is added and the mixture is drunk."

As further examples consider potential antifertility agents and other plants associated with reproduction. We found a note about *Vochysia lomatophylla,* from Peru, which is "perhaps used by Campa tribe as contraceptive." There are other plants that are administered for the regulation of menstruation but that may act also as regulators of fertility. Several species are cited as aiding in childbirth, including *Ocimum sanctum* in the Philippines. One wonders whether or not this plant of the mint family may have a psychopharmacological constituent, as do the related genera *Nepeta* (catnip) and *Salvia,* according to recent reports. *Ocimum* may have an effect like that of the plant-derived hypnotic alkaloid scopolamine.

Herbarium specimens are long-lasting if they are well cared for. (One story has it that when the Linnaean collections were being photographed in 1939, a photographer was stung on the arm by a nettle that had been mounted almost 200 years earlier; he developed the same kind of blister a living plant would produce.) As an adjunct to reading labels, therefore, one can make "spot tests" for particular chemicals on minute bits of plant material carefully selected to preserve the specimen's integrity. By this method one can test enough representative specimens in a large herbarium to determine the approximate occurrence of a chemical in particular groups of plants. Stable alkaloids can yield a positive test after 125 years in a herbarium. Recently an alkaloid-positive result was obtained from a test on a coca leaf entombed with a Peruvian mummy in about A.D. 500.

In addition to the fields of nutrition and medicine, at least two other disciplines may benefit from herbarium searches: plant taxonomy and entomology. One of the newer approaches to taxonomy is chemotaxonomy, which seeks to determine relations among plants by means of their comparative chemistry. Information about folk practices from field notes attached to herbarium specimens often reflects plant chemistry and may serve to point out species in which particular chemicals are likely to be found. Students of ethnotaxonomy, who investigate the systems by which primitive peoples classify their plants, can also be served by our notes, which include a large number of indigenous plant names.

Current studies in entomology have made progress in illuminating the chemical interrelations of plants and insects. It now appears that certain insects take chemicals from plants they eat, store the chemicals and later emit them for defense, as "aggregation pheromones" or perhaps as sex attractants. Certain tropical American orchids produce biologically active fragrance compounds that are attractive to male bees, which do not eat the compounds but collect and store them and may possibly metabolize them. These bees live longer than other male bees, and there is some evidence that they die if they are deprived of the compounds. Might there be some clue here to aging processes in general? Herbarium notes recording unusual relationships between insects and plants could be valuable in this broad area of investigation.

Ultimately, I am convinced, explorations of the kind we conducted should be carried out in all the world's herbaria. One similar search has since been conducted at the Erbario Tropicale in Florence by Teresa Fossi Innamorati, and my colleagues and I are now engaged in another search, at the New York Botanical Garden. Most plant species, including those long put to practical use by primitive societies, have never been scrutinized from the viewpoint and with the techniques of modern science. There is reason to make haste. Many species are in danger of becoming extinct—as are the peoples and the folklore that harbor information about them—as advanced societies expand and disturb natural environments.

8 sound reasons to buy our new receiver. Plus its sound.

Sony's new, more powerful STR-6800SD receiver should get a warm reception. Because it not only looks different from other receivers, it *is* different.

It has some features found in more expensive separate components—and other features found nowhere else at all.

1. The most-used controls all in one place. The level control, muting switch, tuning knob and input and tape selectors are all in the upper right-hand corner.

2. A dial pointer that doubles in length when it's close to a station. Together with the signal strength meter and the center channel meter, this Sony exclusive helps you tune more accurately.

3. A stepped level control to keep both channels equal. It guarantees unprecedented accuracy—to within ½ db over the whole volume range.

4. MOS FET front end electronics unitized tuning. Because it's unitized, the receiver tunes the same whether it's cold or warmed up. And MOS FET gives it a very wide dynamic range.

5. Dolby noise reduction system. So you can benefit from Dolby broadcasting. Instead of being an extra, it's built in—operated from the front panel.

6. Phase locked loop. It gives you better stereo separation and less distortion.

7. LEC (low emitter concentration) transistor. This Sony exclusive in the preamp phono stage yields tight RIAA equalization, low noise, low distortion and a wide dynamic range.

8. Sony's most powerful receiver. It delivers 80 watts minimum RMS continuous power per channel at 8 ohms from 20 Hz to 20,000 Hz with no more than 0.15% total harmonic distortion. It has a direct-coupled power amplifier with true complementary symmetry output stages.

And more. To these specifications (remember, we state them conservatively), add Sony's proven reliability. And you get a receiver that produces a sound that'll make you understand why you have ears.

That's the STR-6800SD at $600. Or, for less power and a few less features—but no loss of fidelity—the STR-5800SD at $500 and the STR-4800SD at $400 (all suggested retail prices).

A sound investment.

SONY

Rat Societies

Norway rats survive in a hostile human environment by means of complex social mechanisms that ensure communal peace, equal opportunity between the sexes and early learning of vital information about the environment

by Richard Lore and Kevin Flannelly

In spite of the universal revulsion that *Rattus norvegicus* (the common Norway, or brown, rat) elicits in human beings and the centuries of active warfare mankind has waged against these little rodents, they continue to thrive on all continents inhabited by man. No other species has been quite so effective in exploiting human territorial expansion or human resources. Indeed, their remarkable ability to survive in close proximity to man has given rise to the notion that they have become a functionally parasitic species that is almost entirely dependent on the by-products of human activities. Although the geographic extension of the brown rat's range in historic times was certainly facilitated by human migration and trade, these animals also flourish in areas not inhabited by man. Thus rats can evidently survive without man. It remains true that no other animal has managed to fare so well in open competition with *Homo sapiens*. How do the rats do it?

Some of the reasons are obvious. *Rattus norvegicus* is omnivorous and can thrive on anything and everything human beings eat. Moreover, the rats' small size and gnawing persistence (both literal and figurative) make it difficult to prevent them from gaining access to human food supplies. Their nocturnal habits, coupled with the fact that they live in subterranean burrows or other equally inaccessible sites, enable them to survive quite severe environments and to avoid direct confrontations with human beings. Finally, rats have a comparatively high reproductive rate, and their ability to coexist with man provides them with a margin of safety against many natural predators such as snakes, owls and a variety of small carnivorous mammals.

Still, other species endowed with similar adaptations or even more effective ones have done poorly in the fight for survival engendered by human population growth, and one has to look elsewhere to understand the unique success of *Rattus norvegicus*. Even closely related species have not done as well. The genus *Rattus* is remarkably large, with more than 400 species, but apart from *Rattus rattus* (the Alexandrine, or black, rat) no other member of the genus begins to match the worldwide distribution of *Rattus norvegicus*. Presumably the behavior of many other species of the same genus is similar to that of *Rattus norvegicus*, but the available data on most members of the genus are insufficient to provide any clear indication of why the latter species is so widespread. Of course, with so many closely related species the term "rat" is zoologically meaningless, and here the term is restricted to the ubiquitous member of the genus, *Rattus norvegicus*.

Until quite recently not much was known about the social behavior of rats. To be sure, psychologists and biologists have worked with the laboratory variety of *Rattus norvegicus* for a long time, but interest was always focused on the behavior of individual rats raised and tested in artificial environments with little or no exposure to other members of their own species. Rats in the laboratory have logged endless miles in runways, mazes and activity wheels, pressed little bars countless times, been injected with every concoction imaginable and have had all possible combinations of their anatomical organs removed, but they go through it with little or no contact with other rats. Part of the problem involves our own perspective: we like to think that fluid and complex social behavior reaches its pinnacle in man and other primates. Consequently in the past most workers interested in social behavior chose to work with primates because the social behavior of rats was considered to be primitive and uninteresting, with the commonest response of one rat to another being a vicious attack.

More recent work indicates that the social interactions of rats are surprisingly complex. At the same time rats' social systems are fluid and therefore capable of rapid change when they are confronted with the characteristic human tendency to alter environments in a radical fashion. It seems that rats exhibit a curious combination of primitive social mechanisms coupled with a repertory of social abilities normally expected only of "higher mammals." Indeed, few other mammals seem so well adapted to group living; the critical element in the ability of rats to compete so successfully with man might well be their ability to develop an efficient social organization.

A basic question involves the role of prior experience in the social behavior of the species. If normal social behavior is readily exhibited by animals that lack previous social experience with other members of their species, one can assume that the social interactions characteristic of that species are genetically fixed at birth and hence are quite resistant to modification by the environment. Earlier work suggested that this was the case with rats. Unlike adult primates that have been reared in isolation from birth, adult rats raised in isolation in small barren cages readily breed and maintain their young in perfectly adequate fashion. Actually their ability to thrive in the highly artificial laboratory environment might just as easily be seen as an indication of their adaptiveness rather than a demonstration that reproduction and maternal care are not influenced by previous experience.

In order to obtain more information on how social history influences the subsequent social interactions of laboratory rats, David P. Luciano and one of us (Lore) established small "colonies" of adult animals in our laboratory at Rutgers University. Half of the colonies consisted of adult laboratory rats raised from weaning in small groups of the same sex. The animals in the remaining colonies were from the same litters, but they had been isolated at weaning and thus had no social experience until they were put together as adults. We observed virtually no fighting among the members of either type of colony. The adjustment to group living by the rats reared in isolation, however, was clearly more difficult than it was for the socially

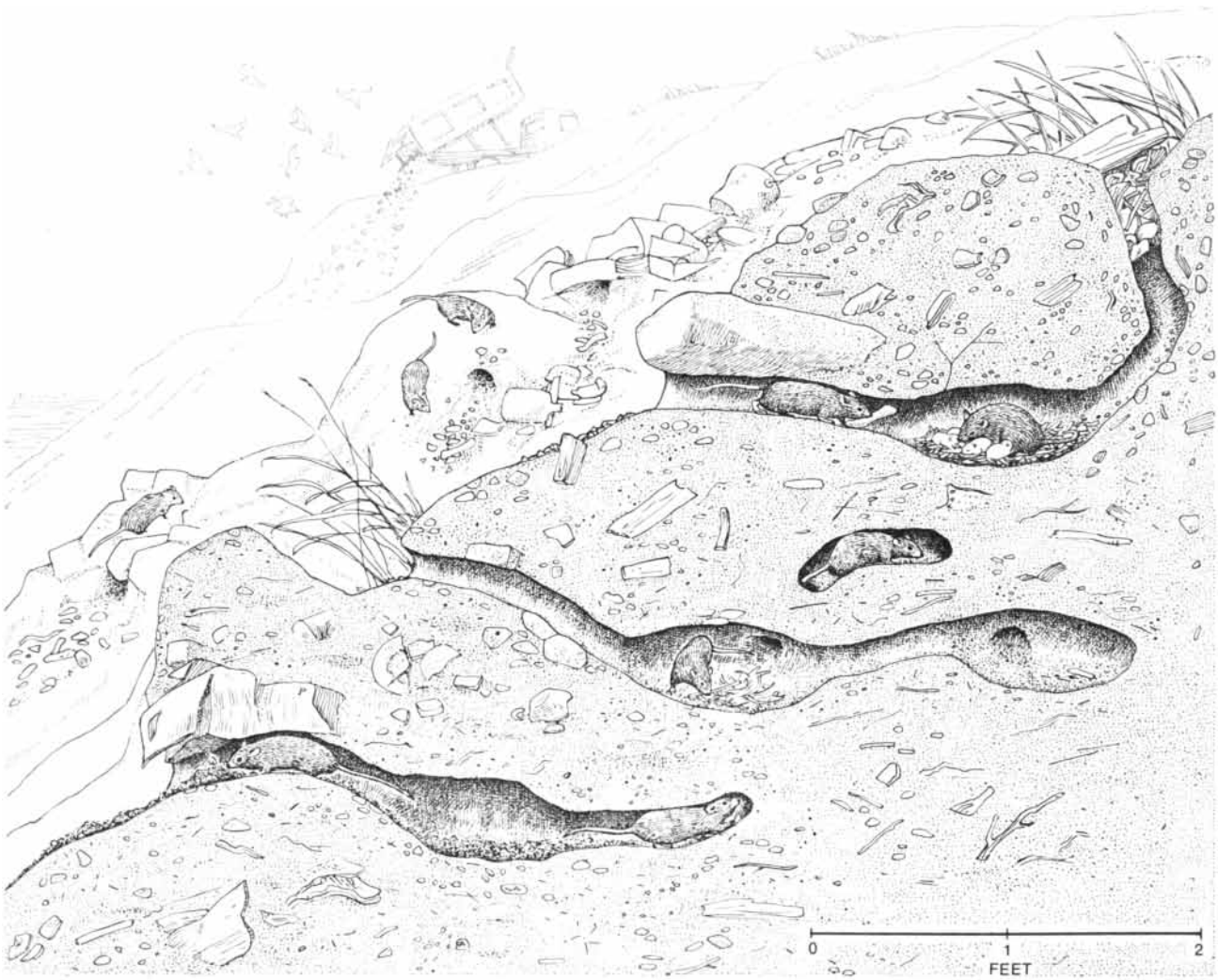
experienced animals. Although all the females gained weight and all the males lost it, the isolation-reared males lost more than twice as much as the socially experienced males, and the isolation-reared females gained only half as much as the socially experienced females.

After several days we placed unfamiliar "intruder" rats, with or without previous social experience, in the colonies for a period of 21 hours. Behavioral observations were made during the first hour of an intruder's stay in a colony, and a detailed assessment of the intruder's physical condition was made after its removal. The tests indicated that some intruders suffered serious physical injuries and large weight losses, but only when an isolation-reared intruder was exposed to a colony of socially experienced rats. Our socially experienced colonies responded peacefully to animals with a similar social history, but they were violently intolerant of a seemingly identical animal that lacked postweaning social experience. The colonies of isolation-reared rats were indifferent to all intruders.

Hence aggressive intolerance was not caused just by a disparity in the social history of the colony and the intruder. It appears that instead isolation-rearing produces an adult rat that behaves in a highly abnormal fashion, provoking an attack by the normal rats. Surprisingly there were no appreciable differences in the duration of fighting during the first hour an intruder was in the colony's cage. It seems that rats seldom initiate an immediate attack against a strange rat. Rather, fighting occurs only after a fairly long period of investigation, threat-posturing and "aggressive grooming."

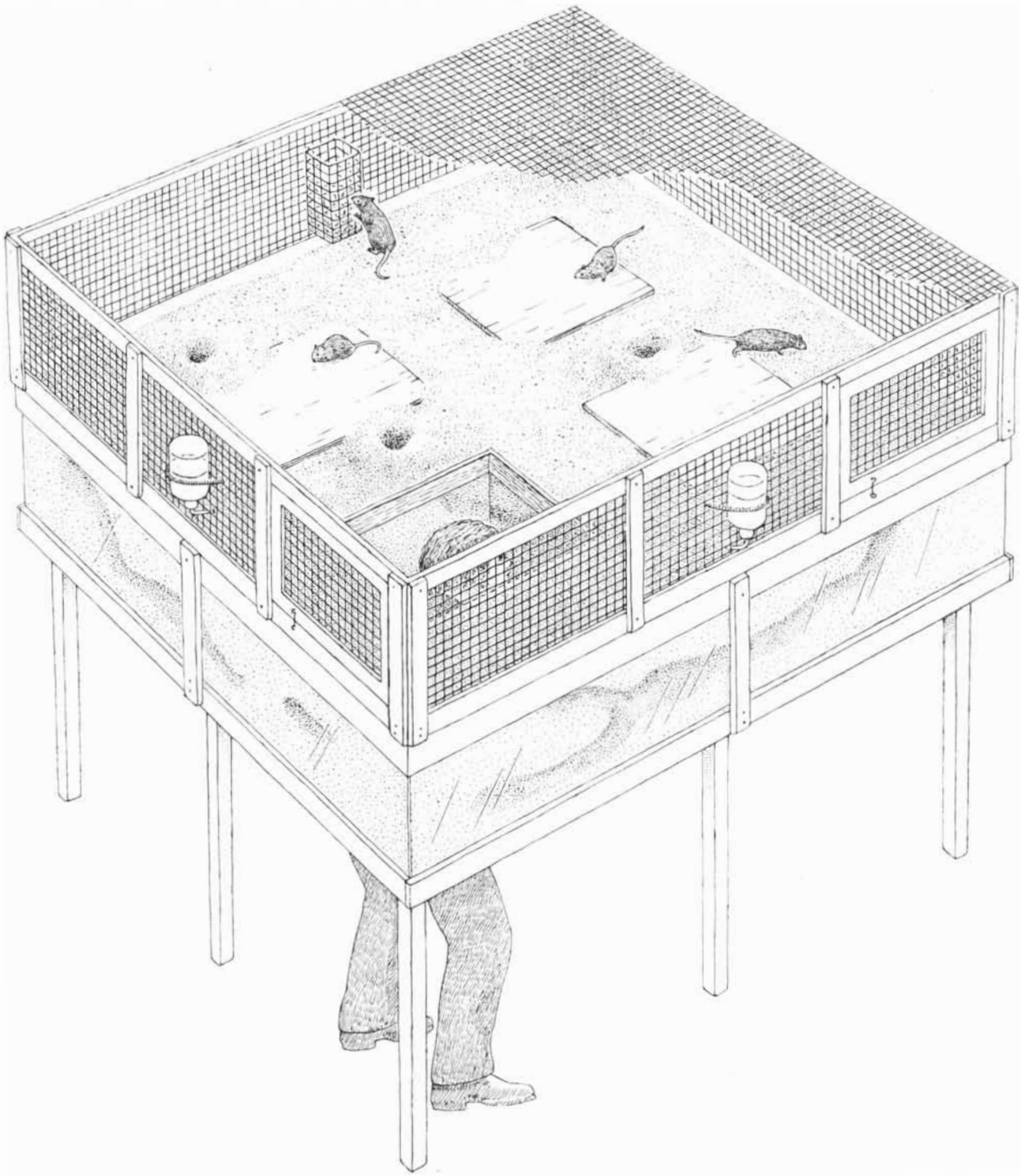
Aggressive grooming, unlike the grooming activities that frequently take place between established members of a rat colony, is usually limited to the head region and is readily distinguished from normal grooming by the firmness of the grooming animal's forepaw grasp and the vigor of its tugging at the opponent animal's fur. Hence aggressive grooming is an important precursor activity that communicates aggressive intent. In our study the amount of aggressive grooming observed during the first hour of an intruder rat's stay was an excellent indicator of the intruder's weight loss and injuries when it was removed from the colony 21 hours later. Under natural conditions a rat that is subjected to such preliminary aggressive activities would probably flee the colony's territory and remain unhurt.

We do not know precisely what aspects of the isolation-reared intruder rat's behavior prompted such intoler-



CLOSELY SPACED BURROWS characteristic of a location with a dense population of rats were found by the authors on excavating a site near an active landfill operation in central New Jersey. Desirable locations of this type are usually well drained and provide ready access to food and water. The absence of large and elaborately interconnected burrow systems in such sites suggests that under field conditions very large rat colonies might actually represent an aggregation of smaller but separate social units that limit their defense of territory to a single burrow and its immediate surroundings. Recent studies indicate that the members of an established group might know one another as individuals and not by means of a common colony odor; under such circumstances alien rats would be readily accepted.

ance, but later work indicated that socially experienced rats often produce ultrasonic signals that appear to inhibit attack. In an experiment conducted with Philip Farina we put the same male intruder rat into the cage of an established resident rat on two occasions separated by a one-week interval. On the first encounter the resident fought the intruder, and both the resident and the intruder lost a considerable amount of weight during the 24-hour test. When the now experienced intruder rat was put into the cage of a second resident rat a week later, however, no serious fighting ensued and both animals maintained their body weight. Clearly the experience of a single defeat was enough to modify the intruder rat's behavior so that it was able to inhibit the attack of the second resident rat.



ARTIFICIAL ENVIRONMENT designed for investigating the burrowing behavior of rats was constructed by the authors in Lore's backyard. The large square cage, measuring approximately six feet on a side, was equipped with clear Plexiglas panels extending below ground level on all four sides. In addition four observation chambers, also made with clear Plexiglas walls on all sides, were built into the belowground portion of the cage. The cover of one of the observation chambers has been removed to reveal an observer crouched under it.

A detailed analysis of the intruder rat's behavior during the first 30 minutes of its stay in the resident rat's cage revealed that the intruder's visible social interactions with each resident were not appreciably different in the two encounters. For example, the intruder paid the same amount of attention (sniffing, following, approaching and so on) to the residents on both occasions. Rats also exhibit a rich variety of stereotyped postures in their interactions with one another, but in this study the number of defensive and submissive postures exhibited by intruder rats was actually higher in the first encounter. The latter finding suggests that visually submissive postures do not communicate complete submission but only a temporary truce, since fighting is soon resumed.

By far the most marked difference in the intruder rats' behavior on the second encounter was their readiness to emit long trains of ultrasonic cries, principally in the range between 20 and 25 kilohertz. The cries were produced much earlier and oftener in the initial 30 minutes of the second test. Obviously one thing the intruder had learned during its disastrous encounter with the first resident was to submit completely very early when it was confronted with the same situation again. The calls sometimes fall within the audible range; to the human ear they sound remarkably like the whimpering of an injured dog.

Of course, the communication of complete submission might involve other sensory pathways. The cries were usually produced only by a comparatively immobile intruder rat while it was breathing deeply, but no single posture was associated with the emission of the calls. It is also possible that the submissive rat might emit a distinctly different odor. Bill Carr of Beaver College has demonstrated that rats can distinguish between the odors of victorious and of defeated fellow animals.

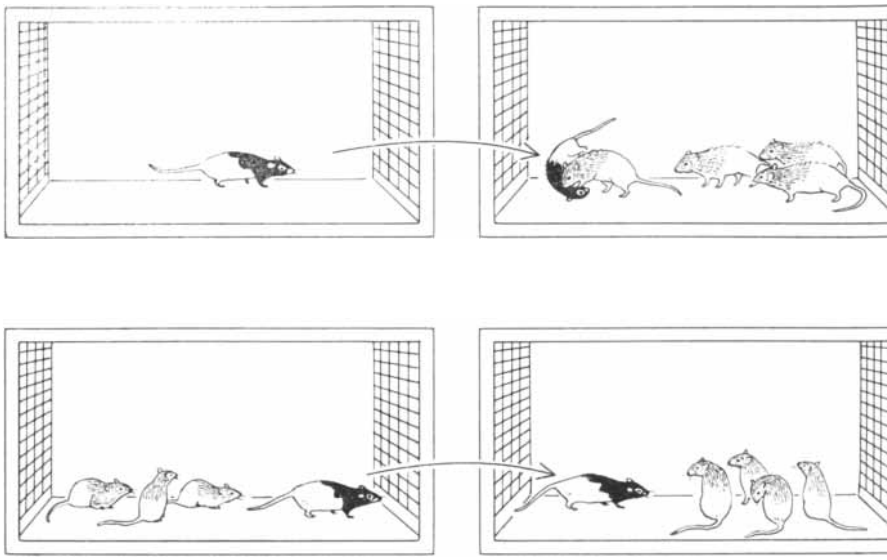
As it happens, the male rat emits a similar ultrasonic cry just after copulation. Ron Barfield of Rutgers, who discovered this curious phenomenon, has suggested that the signal functions to ensure the continued presence of the female for additional copulation. The vocalizations might also serve to prevent the female (or other males) from attacking during a period when the sexually exhausted male is obviously quiescent and in no mood for a fight. In many mammalian species aggressive interchanges between sexual partners are common, but in rats sexual activity is remarkably peaceful.

Comparable ultrasonic vocalizations are produced whenever a rat is badly frightened or in pain. In our laboratory a few brief experiences with a mild electric shock are enough to condition an animal to emit these calls as soon as it is placed in the shock apparatus and be-

STEREOTYPED POSTURES observed in rats include normal grooming (*top*), aggressive grooming (*middle*) and fighting (*bottom*). Normal grooming takes place frequently among established colony members. Aggressive grooming, which is sometimes observed when an unfamiliar "intruder" rat is placed in a cage with a socially experienced resident rat, is usually limited to the head region and is easily distinguished from normal grooming by the firmness of the resident animal's forepaw grasp and the vigor of its tugs at the intruder's fur. Under field conditions an intruder that is subjected to such preliminary aggressive activities would probably flee. In a typical fight resident rat attacks while intruder offers little or no resistance.

fore the onset of shock. It seems likely that the calls also serve as alarm or warning signals for nearby rats under field conditions, but to our knowledge this possibility has not been explored.

The production of the ultrasonic cries is almost certainly an unlearned response to stress, but the important point here is that these animals can rapidly learn to produce the cries in situations previously associated with stress. In other words, this most effective submission signal is readily conditionable and can be elicited in anticipation of a potentially dangerous situation. The motor responses associated with the expression of aggression or appeasement are doubtless innate, but the particular stimulus that evokes this response is determined in large part by an individual animal's experience. For example, a rat with a past history of successful aggression against other rats would not emit such auditory signals of appeasement when it was put into a cage containing a resident animal. A fight would be much more likely to ensure. The prolonged and often acrimonious controversy over whether aggression is an innate characteristic of animals is based in large part on the failure to make a distinction between the motor patterns by which aggression is expressed and the stimuli that elicit aggression. The motor sequence might be innate for any given species in the sense that it is similar in all members of that species and does not require learning. For any given animal, however, the probability that a particular stimulus or situation will elicit aggression is quite variable and is clearly dependent on previous learning.

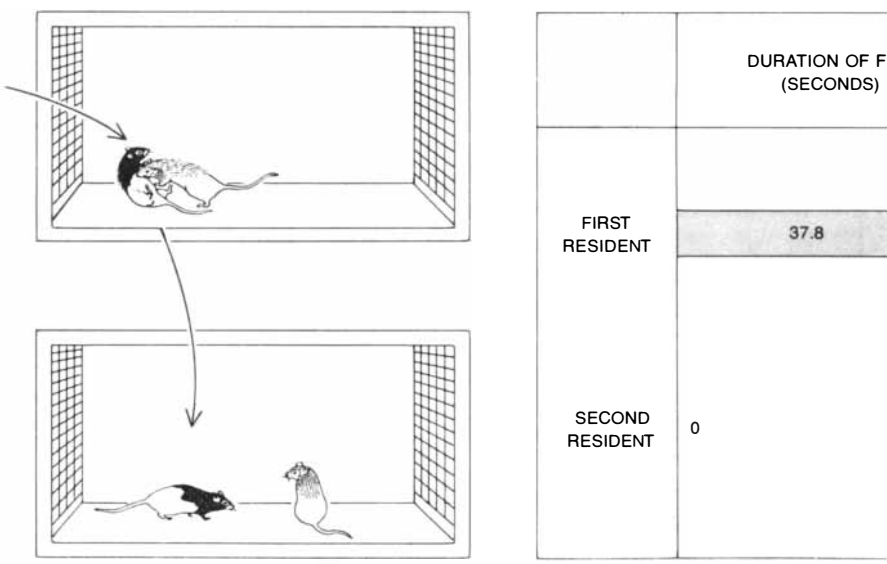


EARLY SOCIAL EXPERIENCE (or lack of it) appears to have a strong influence on the reception an intruder rat is likely to meet on being placed in a cage with a small colony of other rats, whether the latter were themselves raised in isolation or in a group. Most intruders are accepted peacefully, but colonies of socially experienced animals were particularly intolerant

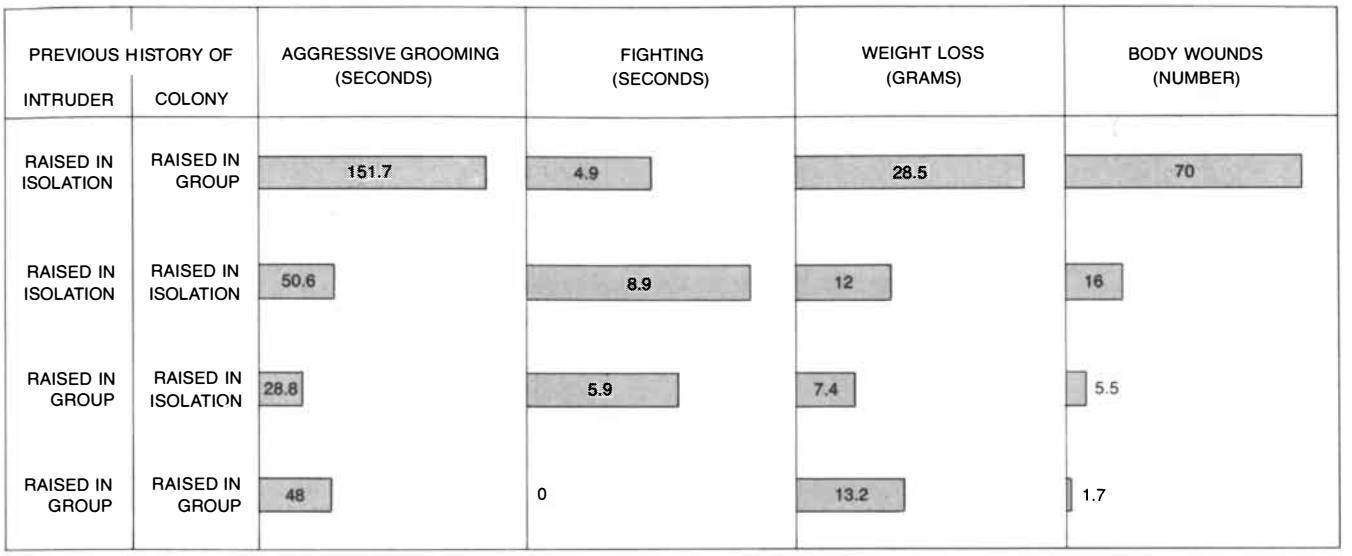
Discussions of the social behavior of rats invariably focus on their reputation for unbridled aggression. For example, Konrad Lorenz in his widely read book *On Aggression* considers the rat and man to be uniquely similar in their propensity to attack and kill their own kind. Apart from the fact that many other species kill members of their own species, Lorenz' conclusions are questionable on the grounds that they are based on studies of captive wild rats kept in small enclosures. Fighting and deaths do occur in groups of captive rats, but captivity almost certainly distorts and accentuates aggressiveness in virtually every species. Under field conditions an alien rat may be attacked by an established group, but the fight seldom lasts long enough to result in serious injury, because the alien simply flees the area.

Moreover, there is evidence that alien rats are not always attacked by an established group. The German ethologist Hans-Joachim Telle trapped rats in one place and introduced them to different areas containing established rat populations. Vigorous and sustained territorial defense was observed only when the existing population had fewer than 20 members. When the size of the group was slightly larger, the alien animal was initially attacked but soon won acceptance. Much larger colonies did not defend their territory at all and readily accepted alien rats. Telle's findings suggest that the members of an established colony might know one another as individuals and not by means of a common colony odor, as most other workers have assumed. In large groups alien rats are not recognized as such because the information required for the recognition of a great many animals is beyond the individual rat's processing capacity.

Alternatively the large colonies observed by Telle under field conditions might actually represent an aggregation of smaller but separate social units that limit their defense of territory to a single burrow and its immediate surroundings in densely populated sites. When we excavated a number of burrow systems at



SINGLE DEFEAT EXPERIENCE inflicted on an intruder rat by an established resident rat appears to modify the intruder's behavior sufficiently so that it is able in some way to inhibit the attack of a second resident rat. In this series of experiments, conducted by the authors in collaboration with Philip Farina, an adult male rat was placed in the cage of a resident rat on two occasions separated by a one-week interval. On the first encounter the resident usually

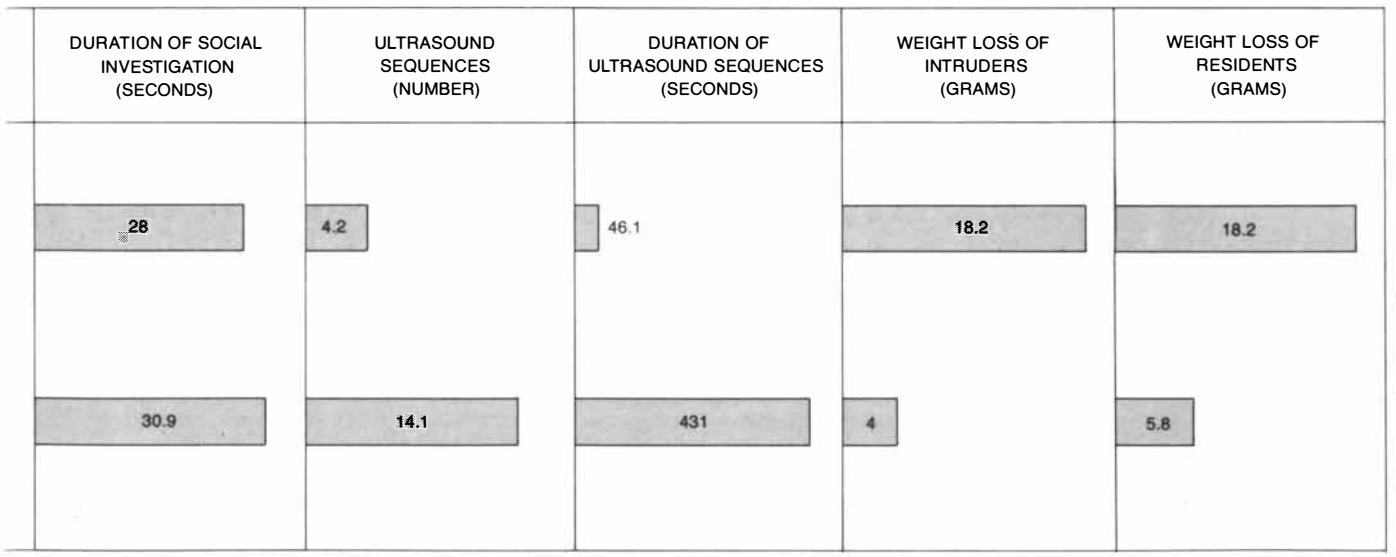


of isolation-reared intruders. Each of the eight rat colonies tested contained two males and two females. A total of 24 males were used as intruders. Each test lasted for 21 hours. Figure for total wounds was arrived at by adding the number of wounds smaller than a half centimeter plus two times the number of wounds larger than a half centimeter but smaller than one centimeter plus four times the number of wounds larger than one centimeter. Tests were conducted by David P. Luciano and one of authors (Lore) at Rutgers University.

a local landfill site in central New Jersey, we found that desirable locations (proximity to food and water, a slope for good drainage and so on) contained burrows virtually the same in design and size as those in less desirable areas. The only difference is that burrows in the best locations are quite close to one another. The absence of large and elaborately interconnected systems in good locations suggests that the basic social unit might remain fairly constant but that the amount of territory defended by each group varies inversely with the size of the area's entire population.

Whatever the mechanism, Telle's findings demonstrate a crucially important way in which territorialism can be short-circuited so that the rat population can increase rapidly in particularly desirable habitats. In any event the reputed aggression of wild rats is exaggerated. After 12 years of fieldwork Telle reports that serious conflict was seen only when an intruder attempted to join a small colony of 20 or fewer rats.

Another misconception concerning the socially primitive and aggressive nature of rats involves the persistent notion that adult males readily kill infant or juvenile rats. This belief is widely held in spite of the fact that many commercial breeders of laboratory rats regularly maintain breeding pairs together throughout pregnancy and lactation. When we compared laboratory rats born and raised with and without the breeding male present in the maternity cage during the preweaning period, we



attacked the intruder, and both the resident and the intruder lost a considerable amount of weight during the 24-hour test. When the intruder was subsequently put into the cage of a second resident rat, the resident appeared to accept the presence of the now experienced intruder peacefully and both animals maintained their body weight. The most marked difference in the intruders' behavior on the second encounter was their increased tendency to emit long trains of ultrasonic cries at a frequency of about 22 kilohertz. The cries, which apparently have the effect of inhibiting attack, were produced much earlier and more often during the initial 30 minutes of the second test.

Digital Watch Breakthrough

There are several big changes taking place in digital watches. Here are all of them in one product.

Your digital watch is either too thick, uses up batteries quickly, has just a few functions or is hard to read under certain lighting conditions. You still have to press a button, flick your wrist, or hold your watch at just the right angle to read the time.

The digital watch industry has gone through four years of rapid change, but the disadvantages cited above have finally been resolved in one totally new product—the Sensor Laser 440 Digital. The Laser 440 is so different that it represents a dramatic departure from conventional digital watches.

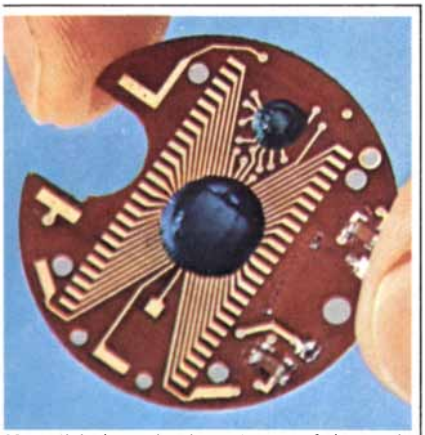
NO BUTTON TO PRESS

There is no button to press since the display glows in the dark. A glass ampoule, charged with tritium and phosphor and sealed by a laser beam, is placed behind the new CDR (crystal diffusion reflection) display. When room lights dim, the self-contained tritium light source will compensate for the absence of light by glowing brightly and illuminating the display.

No matter when you wear your watch—day or night—just a glance will give you the correct time. There's no button to press, no special viewing angle required, and most important, you don't need two hands to read the time.

CHANGE YOUR OWN BATTERIES

The Laser 440 is only 8 millimeters thick—thinner than many of the so called thin digitals being advertised today. The new CDR display draws 100,000 times less current than an LED watch when displaying the time so your single commercially-available Union Carbide battery lasts years longer. In fact, part of our warranty includes all the batteries you'll ever need, free of charge, for five full years. To replace a battery, simply open up the battery hatch on the back of your watch, tap out the old battery and drop in the new one.



Most digital watches have dozens of electronic components. The Laser 440 has only six—two integrated circuits, a crystal and three microcapacitors. All components are bonded directly on the printed circuit board. By hermetically sealing the integrated circuits and using fewer components, the Laser 440 is considerably more reliable than other watches that do not yet have this complete integration.

The new Sensor Laser 440 digital watch glows in the dark so there's never a button to press to read the time in darkness or in sunlight.

THE ULTIMATE ACHIEVEMENT

Other manufacturers have devised unique ways to produce a watch you can read at a glance. The $300 LED Pulsar requires a snap of the wrist to turn on the display, but the Pulsar cannot be read in sunlight and its display uses 100,000 times the current of the Sensor display. The $400 Longine's Gemini combines both an LED and liquid crystal display. (Press a button at night for the LED display, and view it easily in sunlight with the liquid crystal display.) But you must still press a button to read the time. All these applications of existing technology still fail to produce the ultimate digital watch: one you can read under all light conditions without using two hands. Until the new Laser 440.

And if you've owned a digital watch for a year, chances are you've had it in for repair more than once—a very common consumer complaint. The Laser 440 is so service-free and has such high quality that it should rarely, if ever, require service. It is backed by a solid five-year warranty—your assurance of our commitment to this outstanding new product.

The Laser 440 has both time and stop watch functions. Six digits are on display—four large digits and two small ones. You choose between hours, minutes and seconds or hours, minutes and date by pressing a button. The Laser memory remembers the number of days in a month and resets automatically on the first day of the new month.

The 440 is also available in an 11 function chronograph (stop watch) and is truly the ultimate Laser timepiece. You can time two separate laps of a multilap race keeping one lap in memory. You can accumulate time; you can view the time of one lap while continuing to time a lap stored in memory. As a business executive, you can time long distance phone calls and interviews. Lawyers can keep track of their services, and doctors can time the vital signs of their patients. Even while the chronograph is functioning, you can still view the time—something even many of the expensive digital chronographs cannot do.

BUILT DIFFERENTLY

All wires have been replaced with circuitry printed on one single thin surface. On this same surface are two integrated circuits which use gold contacts and are hermetically sealed to protect their several thousand micro components. The American-made Laser is shock resistant and uses a tough mineral glass crystal to protect the rugged electronics from the everyday water and humidity tortures normally given any watch.

NEW QUARTZ BREAKTHROUGH

Digital watch accuracy depends on the quartz crystal. Even the best crystals change frequency with shock or age (especially when first produced). The Laser 440 uses the new and very expensive, tuning fork crystal. It is first aged to not shift frequency more than five parts per million per year (more accurate than most radio or TV time signals) so the extreme accuracy you expect is built into your watch from the first day you wear it. The crystal is cushioned and solidly bonded to the crystal carrier eliminating all fine wires that may break from shock. In short, the advanced design of the crystal will assure guaranteed accuracy greater than 5 seconds per month—year after year after year.

The Laser 440 is ideal for pilots because of its cockpit visibility and chronograph functions, perfect for the businessman who depends on his watch for split-second accuracy and the ultimate watch for anybody who wants unquestionably the finest digital watch ever offered at any price.

The Laser 440 is not only thin but is designed to conform to the contour of your wrist. A bulky digital watch can become annoying to wear—especially if you have a thin wrist.

HOW WE PROVIDE THE FINEST SERVICE

Can our company provide better service than even your local jeweler? We think so. If your Laser malfunctions during its unprecedented five-year warranty, just call us on our toll-free line. We have made arrangements with United Parcel Service to pick up your Laser at your door, at our expense, and we give you a loaner watch to use while your Laser is repaired. You pay nothing to have your watch serviced during its five-year warranty (that is if service is ever required) and we are as close as your phone or door.

WHO WILL BACK YOUR 5 YEAR LIMITED WARRANTY

Two solid companies are behind your new Laser. JS&A is America's largest single source of space-age products—a substantial company and a leader in electronics for over a decade. Our commitment to the consumer and to service is a matter of record. Check with the Better Business Bureau in your very own community, the Northbrook, Illinois Chamber of Commerce (312) 498-5555, or any of the 100 national magazines and newspapers in

found that fewer pups died and that the growth rate of the surviving pups was greater when the male was present. Twice during lactation the female was removed for two-hour periods with no effect on the pups. Most males behaved quite maternally toward the young animals. In a related study alien adult males were placed in the maternity cage with the mothers and their litters. At first the female's threats kept the male from approaching the litter, but after the first hour or so the mother's protectiveness diminished. No pups were killed by any of the strange males.

Further evidence for the inhibition of aggression directed toward young rats has been provided by a study conducted at the E. R. Johnstone Training and Research Center by Don Thor and one of us (Flannelly). It was found that neither young and sexually immature intruders nor adult females are attacked by established colonies even though these small colonies are viciously intolerant of adult males. Tolerance of alien young and females provides a second means by which migration to new groups can be accomplished. Certainly young rats can be threatened and intimidated by adults, but physical injury is rare. It seems likely that adult animals perform important socializing functions in established colonies; as it happens, however, virtually all investigators have worked only with groups of animals that are the same age.

Male rats are rarely seen to attack female rats, but a female will readily attack an alien male that enters the immediate nesting area while the female is nursing. The female invariably succeeds in driving the male away. In mixed-sex colonies, however, alien animals are usually first attacked by the resident males, and often females are never observed fighting. There are exceptions, of course, and when females do fight intruders, the attack is likely to be quite vicious and is not preceded by the usual preliminary threat-posturing and aggressive grooming of the male. One prime determinant of whether females fight intruders is the sexual composition of the group. A male intruder placed in an established colony consisting only of females fares about as badly as one placed in an all-male colony. Living in mixed-sex groups appears to increase aggressiveness in males and to inhibit it in females.

The reciprocal nature of the relation between sex and aggression is also illustrated in mixed-sex colonies where the males have been deprived of their sense of smell. In this case the males will not fight an intruder, but the resident females will attack the intruder far more frequently than they normally would in mixed-sex groups. Under field conditions the presence of aggressive males provides the stable environment required for females to devote the bulk of their energies to reproduction. If the males are incapacitated for some reason, the females are capable of performing the defense role quite adequately, but reproductive effectiveness would be impaired.

In its conflict with man *Rattus norvegicus* has developed admirably efficient feeding strategies that enable the members of a colony to avoid poisoned baits and to adjust to sudden changes in the available food supply. At every step in the complex feeding sequence group membership plays a crucial role. Both laboratory rats and wild rats tend to avoid any contact with novel objects in their environment. Typically a new food is completely avoided for several days, and it may never even be sampled if the existing diet is nutritionally adequate. Eventually small sublethal quantities of the new food may be ingested. If feeding animals get sick, the new food is thereafter avoided by the entire colony. This phenomenon, called conditioned food aversion, has forced psychologists to revise some basic principles of learning theory and has generated a lively interest in how animals acquire biologically meaningful information. For example, it takes many trials to teach a hungry rat to press a bar for a food reward, and the rat never learns the trick if the food is not presented immediately after the bar is pressed. In contrast a rat learns to avoid poisoned food in a single trial, in spite of the fact that it may not get sick until six hours after it has eaten. A medicinal effect has also been demonstrated. Here the animal is first made sick (usually with lithium chloride), and then just before it recovers it is allowed to eat a novel food. Thereafter the animal will prefer the novel food associated with recovery.

Under certain circumstances a rat will avoid even a highly preferred food that it has eaten regularly before. A small colony of wild rats living in an outdoor enclosure behind the Lore home became addicted to eating marshmallows placed in one location. Marshmallows in other locations were initially avoided but finally came to be accepted and eaten wherever they were found. Marshmallows placed in a live trap like those used to catch the animals months before, however, were never touched.

Does each rat in an existing population learn to avoid poison from direct experience or is there a social mechanism whereby one individual's unpleasant gastrointestinal experience can be communicated to others in the same population that have not been poisoned? A series of experiments by Bennett G. Galef, Jr., of McMaster University provides clear evidence that information of this kind is readily communicated to other rats. In one of his tests a small colony of adult wild rats had

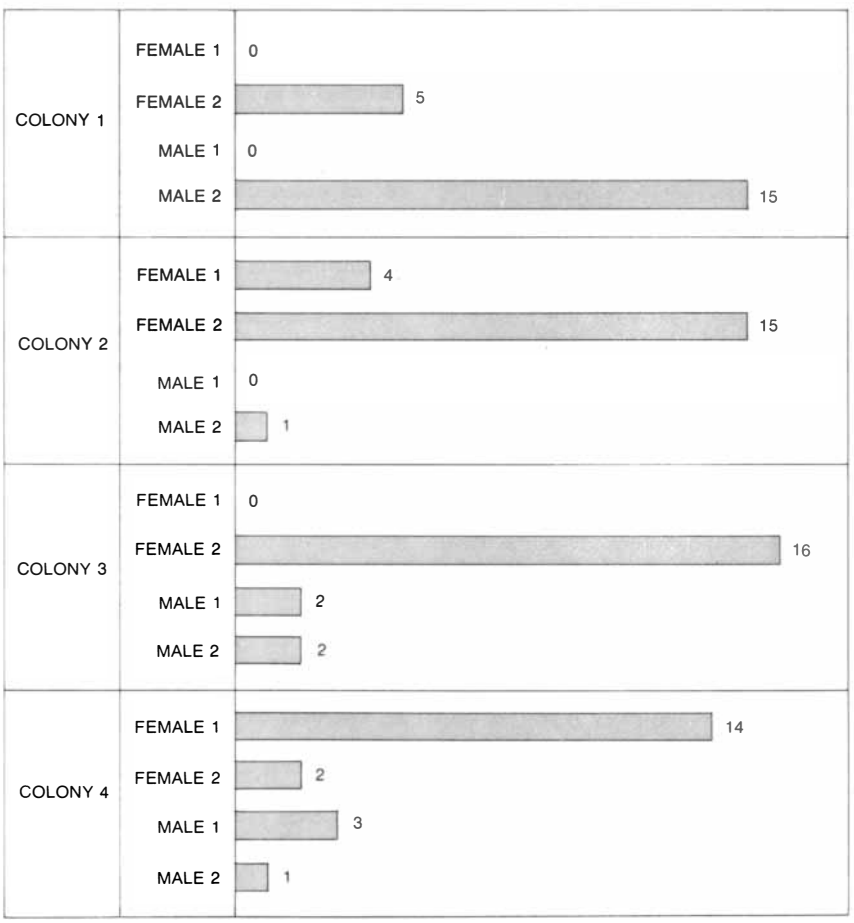
equal access to two nutritionally adequate diets differing in color, texture, taste and smell. To facilitate the observation of all the rats' feeding activities both diets were available in separate bowls for the same three-hour period every day. First small amounts of lithium chloride were added to one of the diets, a laboratory mixture called Purina Rat Chow, which under normal circumstances was not the rats' preferred food. The colony quickly learned to feed exclusively on the preferred diet, called Turtox, and consistently ignored the nonpreferred diet even when it later contained no lithium chloride.

Galef was particularly interested in the feeding preferences of the first litter born in the colony; observations during the three-hour feeding periods indicated that those pups also fed exclusively on the Turtox. Moreover, when the pups were transferred to a separate cage, they continued to eat only the preferred food. Lithium chloride was then added to the Turtox diet of the same colony, and the adults promptly switched to eating only the Purina Rat Chow, normally the nonpreferred food. Furthermore, the next litter ate only the Purina Rat Chow! Clearly parental food preferences had been transmitted to the young.

Later work by Galef provides a good account of how simple the mechanism for transmitting food preferences between generations actually is. Young rats regularly follow adult animals during their initial ventures outside the maternal nest, and thus their first solid food is likely to be ingested when the adult animal stops to feed. Thereafter familiarity with the food they first ate, coupled with a tendency to eat in groups and a negative reaction to new foods, gives rise to a strong preference for the colony's preferred diet. Curiously, Galef could find no evidence that the young animals were learning to actively avoid the poisoned food. Similarly, his work does not support the notion that rats mark poisoned food with some kind of alarm pheromone. The practice of actively trailing adult animals, however, is not the only social mechanism whereby feeding sites are located. Young animals tend to explore and feed only in an area that contains the residual olfactory cues left by adults. Presumably the odor of anal excreta and fermented urine is an attractant for hungry rats.

In more recent work Galef discovered a second independent mechanism by which parental food preferences are transmitted to the young at an even earlier age. In one study a group of lactating females was fed only Turtox and a second group ate only Purina Rat Chow. All mothers were removed from the maternity cage for feeding, so that the pups could not come in direct contact with the solid food eaten by their mothers. When the pups were between 17 and 23 days old, a series of preference tests revealed that both groups of pups had developed a strong preference for the food their mothers were eating.

Subsequent control experiments ruled out the possibility that the young animals could have acquired this preference by ingesting maternal feces or bits of food that might have clung to the mother's fur. Again the mechanism is deceptively simple. The mother's diet apparently influences the taste and odor of her milk, and her offspring are sufficiently sensitive to these clues for them to later prefer solid foods with the same taste and odor. It is known that many substances ingested by lactating females pass intact into the nursing infant. Anyone who has experienced the strong odor and taste of onion in cow's milk from a herd in a pasture infested with wild onions can appreciate the potential effectiveness of this mechanism. Rat infestations probably begin most often with the offspring of one adult pair or of a single pregnant female. Therefore either or both of the food-preference mechanisms demonstrated by Galef would protect the colony from the ingestion of potentially dangerous foods.



EQUAL ACCESS BY BOTH SEXES to a limited food supply, a measure of degree of socialization of a species, was observed in this experiment, conducted by the authors with Christine Galbreth. Mixed-sex colonies of adult rats with their normal laboratory food continuously available were given a small piece of milk chocolate, a highly prized food among rats, twice a day for 10 days. In each of the four colonies tested one animal monopolized the chocolate on most trials. Females were more than twice as likely to succeed in eating the preferred food in this particular series of tests, a situation that contrasts sharply with strong male dominance in feeding behavior observed in less socialized species such as lions and nonhuman primates.

Successful reproduction in a social species is contingent on females' having free access to the group's food supply. During food shortages it is crucial that the females, which are smaller and less aggressive than the males, receive adequate nourishment if the group is to survive as a breeding unit. According to the criterion of free access to available nutrients, nonhuman primates are imperfectly socialized, since dominant males usually eat first and monopolize any choice or preferred foods. Occasionally the dominant male will tolerate the cautious approach of a favored female or a juvenile that snatches food and runs away, but often the offender receives a vicious clout for its efforts. Similarly, in

the lion pride females eat only after the adult males have fed, even when a lioness was responsible for the kill. Little is known about male-female feeding patterns in most species, but females in other vertebrate groups seem to fare no better. For example, a recent study of weaverbirds in a small flock disclosed that weight loss following food restriction was confined to females.

In the rat aggressive dominance, which is usually enjoyed by the largest and strongest males, is not well correlated with other measures of dominance. For example, Louise Baenninger of Temple University found that aggressive dominance (based on winning fights) in small, all-male groups of rats is not related to competitive success at gaining access to limited amounts of either food or water. Her results indicate that in this species the concept of dominance must be far more narrowly defined. Furthermore, if feeding priorities are not related to aggressive dominance, then female rats in mixed-sex groups may enjoy equal access to the colony's food resources.

To test the latter proposition we were joined by Christine Galbreth in observing the feeding behavior of four small colonies of rats, each consisting of two males and two females, after regular intervals of food deprivation. A recessed barrier in the food hopper allowed only one animal at a time to eat. During the first 30 minutes of each feeding session an observer noted the feeding order and the total time spent in feeding by each animal. In three of the colonies a particular male-female pair consistently monopolized the hopper during the daily feeding sessions; in the fourth colony the two females controlled the hopper. In all four colonies the eight males controlled the hopper only 41 percent of the time; the eight females were observed eating 59 percent of the time. Feeding order was closely related to control of the hopper. Hence those rats that fed first or second within each colony also tended to feed longer. At the end of the study the eight animals (two from each colony) that monopolized the food source had gained significantly more weight than their colony mates. In addition three of the four females that succeeded in raising litters during the food-deprivation period were ones that had dominated the food hopper.

In a second study we examined the comparative success of male and female rats in groups competing for a highly preferred food. Before we established these small colonies we gave a larger number of individually housed rats a small square of milk chocolate twice a day for 12 days. We selected for the colonies only those rats that ate the chocolate ration within five minutes on the last two days. Most of the animals easily met the five-minute criterion, because

rats have a well-developed sweet tooth; thus every member of the colony was an avid chocolate eater before the colony was established. Starting on the day after the colonies were established a single piece of chocolate was dropped into the colony cage twice a day for 10 days, and an observer noted which animal or animals ate the chocolate. Purina Rat Chow was continuously available, so that none of the animals went hungry.

The results of this study indicated that the chocolate was eaten almost immediately by a single animal in most trials. In eight of the 80 trials two rats shared the tidbit. In only one trial did three animals consume at least some of it. In these nine trials the animal that fed for the longest time was designated the winner. One rat in each colony succeeded in eating the preferred food in more than 70 percent of all the trials. In all four colonies females ate the chocolate in 56 of the 80 trials. In these experiments the females obviously had equal access to the group's food supply. Indeed, the females' access to the food supply was somewhat better than equal. In the first study the females were more effective in monopolizing the limited food ration, and in the second study they clearly ate most of the chocolate. The results of both feeding studies indicate the existence of a well-defined system of feeding priorities in groups of rats that increases the likelihood that some members of both sexes of a local population of animals will survive a period of food shortage in good breeding condition.

In spite of the fairly rigid feeding order within our colonies we found no evidence that the rats were aggressively competing for food, although occasionally one animal would push another aside at the food hopper. These attempts were much more likely to be successful if they were initiated by an animal with a past history of monopolizing the hopper. In the second study the successful animal simply ate the chocolate where it was dropped while the other colony members remained at rest. A casual observer would think that chocolate eating was randomly determined, with the winning animal being either the one that happened to be awake and alert at the time or the one closest to the chocolate when it was dropped into the cage. Nevertheless, only one animal consistently ate the chocolate, and all members of the colony were probably aware of its presence in their cage, considering its strong and distinctive odor.

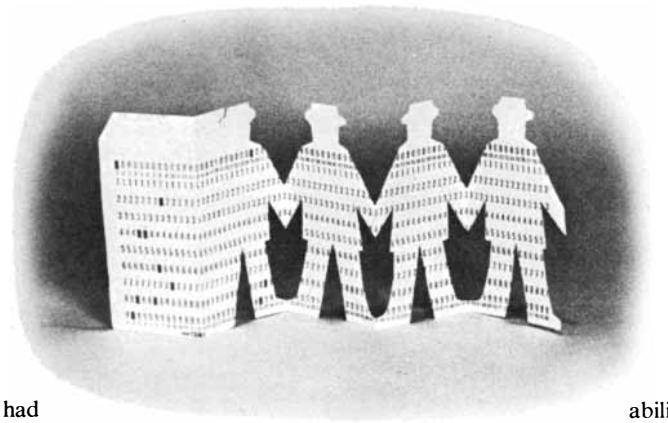
In summing up our findings a few words of caution are in order. We have attempted to demonstrate that rats of the species *Rattus norvegicus* are well adapted for group living and that these "primitive" animals are capable of a richer and more flexible social organization than has been thought possible by most animal behaviorists. Learning and early social experience play important roles in the normal social development of the species, and the animals' capacity to communicate information between generations represents the crucial first step in the acquisition of culture. Contrary to a widespread belief, among rats serious fights are rare, and the one-sided and uneasy truce between the sexes that is characteristic of many other mammals is replaced by an easy reciprocity and interchange of roles that are beneficial to both sexes.

It is important to note, however, that probably none of the social mechanisms we have discussed is unique to rats. For example, many mammals (including man) can develop an aversion to foods they ate just before they got sick, and some of them probably have the capacity to communicate these aversions to their offspring. Moreover, other species have certainly evolved the capacity to develop cooperative and long-term relations between the sexes; witness the social canids, among which pair bonds are unusually strong and permanent. To be sure, the world is not overrun by wolves and foxes; the important point here is that no single aspect of the behavior of *Rattus norvegicus* can account for its remarkable competitive success. The survival formula of these animals involves the fact that their exceptional social capacity is coupled with the physical resilience enjoyed by all members of the ancient and large order of rodents.

Finally, our work is by no means complete. We lack basic information on the social structure of these animals in the wild. For example, do large aggregations of rats represent a single functional social unit or many smaller units of approximately the same size that tolerate one another's physical proximity by defending much smaller territories? Or do adult males tolerate one another under field conditions without the necessity of dominance orders that invariably develop in captivity? Precisely what is the typical age and sex composition of free-living groups of rats? So far the results of field studies of rats contain imprecise and conflicting definitions of terms such as "clans," "colonies" or "bands" of animals, and thus they offer little more than anecdotal evidence on the social organization of rats. Furthermore, we know almost nothing about the behavior of old animals, in the laboratory or outside it, and even less about the behavior of older animals in groups. We are convinced that imaginative research on established groups of these highly socialized animals could provide a helpful perspective in our attempt to develop a psychology equipped to deal with the social problems of normal adult members of the human species.

Technology:

not enough information, too much information and the information democracy.



It's hard to imagine life without the almost instantaneous communication of new information. When man first set foot on the moon a half-billion people were watching.

Yet during the War of 1812, the Battle of New Orleans took 2600 lives, sixteen days after a peace treaty had been signed. The news from Ghent, Belgium had not reached New Orleans.

Tremendous strides in communication technology make this kind of mistake seem impossible today. The question now is whether we have progressed from not knowing enough to knowing too much.

Our communications systems have evolved with us.
Man is the paramount communicator. The very communication skills we were born with give us the capacity to devise technologies that in turn enhance our ability to effectively communicate.

First, perhaps, came burnt-stick ceremonial drawings on cave walls. Then the stylus and brush, enabling us to write. With writing, more complex information could be communicated and built upon. The limitations of memory were overcome.

Then came movable type, ushering in a printing technology that produced a quantum jump in communication. Information became available to the many instead of the few. Mass education began.

The problem of knowing too much.
The total sum of human knowledge has doubled in the last twenty years. And every year the increase accelerates. This expansion of knowledge is a great benefit. But it is also a great burden. Information is of no value unless it is controllable, accessible and containable.

The probable answer is the electronic computer. It is revolutionizing society in the same way as the printing press, but far more rapidly.

The Computer Age is still in its infancy.
Electronic data processing lets us store and tap vast networks of information. We see it at work in business forecasts and weather forecasts, in government reports and traffic reports and countless other applications.

But we've also experienced a less favorable effect: computer errors, even crime by computer. Some consider these problems to be the fault of technology, irritants that outweigh the benefits. More often than not they are the result of improper programming and human shortcomings.

A dehumanizing force or a "humanizing" one?
While skeptics argue that computerized information processing threatens a loss of individuality, the reverse is more likely true. The drudgery of routine information processing will be shifted from people to machines. People will be free to become more innovative, more productive and better informed.

Sophisticated information storage and communication will make more knowledge more readily available to more people. Future students may enjoy the luxury of truly individualized education, programmed to meet personal needs and to match personal abilities. Financial transactions, such as taxation, may be custom-tailored for a new fairness. Preventive medicine could be personalized on a broad scale, allowing up-to-the-minute readings on each individual's physical condition.

How communication technology can help bring about an "information democracy."
Computerization might produce the first totally democratic system of government. Each person's opinions and needs could be registered in a perpetual referendum, enabling governments to become more responsive to their citizens.

And imagine the impact of making available the mass product of human thinking to every person on earth through a working world library. Anyone would be able to study or enjoy any existing piece of information. The end of ignorance would become an achievable goal.

Impossible dreams? No more so than the printing press. Technology makes the impossible possible. And while new technology has always raised new questions and brought new risks, we need not fear it. Technology is man's vehicle for improving the quality of life. If we use it wisely, it will take us wherever we want to go.

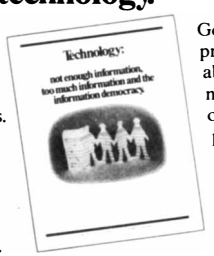
Science and technology can solve many problems. If they don't, what else will?

We believe in the promise of technology.

Gould is sponsoring a series of "white papers" to foster a clearer understanding of technology. Subjects include:

1. Technology and the Rise of Man.
2. Technology and the Rise of the United States.
3. Technology and the Prophets of Doom.
4. Technology and the World Trade Gap.
5. Technology and Energy.
6. Technology and Free Enterprise.
7. Technology and Communications.

Excerpts from number seven are published above. For the complete text of this or any other one in the series, specify by number and write: "Dialogue on Technology," Dept. SA.



Gould is an electrical and industrial products company, committed to above-average earnings, growth, and new product development. Technology is important to us because we put it to work in products like batteries, instruments, electric motors and controls, electrical switchgear and substations. For a copy of our annual report, write Gould Inc., "Annual Report," Dept. SA, 10 Gould Center, Rolling Meadows, Illinois 60008.

GOULD

Minolta by Minolta by Irving Penn.

The easy-to-use used camera with easy-to-change lenses.

More than three million people have bought Minolta SR-T 35mm single lens reflex cameras. And now they cost substantially less than ever before.

Why is the Minolta SR-T such a well-used camera?

It's easy to use, for one thing. With Minolta's patented through-the-lens metering system, you get perfect exposures just by lining up two needles in the big, bright viewfinder.

Interchangeable lenses, for another. All Minolta SR-T models accept all of the nearly forty Rokkor-X and Celtic lenses made by Minolta. So you can get all kinds of interesting photographs—from fisheye wide-angle shots to close-up pictures of faraway subjects with a super-telephoto lens—using just one camera.

And the lenses are easy to change, with the unique Minolta bayonet mount that requires only a quick turn to put a lens on or take one off.

Ask for the easy-to-use Minolta SR-T at your photo dealer. Or for more information, write Minolta Corporation, 101 Williams Drive, Ramsey, N.J. 07446.

MINOLTA

More than three million people own a Minolta SR-T.

Stein's Paradox in Statistics

The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average

by Bradley Efron and Carl Morris

Sometimes a mathematical result is strikingly contrary to generally held belief even though an obviously valid proof is given. Charles Stein of Stanford University discovered such a paradox in statistics in 1955. His result undermined a century and a half of work on estimation theory, going back to Karl Friedrich Gauss and Adrien Marie Legendre. After a long period of resistance to Stein's ideas, punctuated by frequent and sometimes angry debate, the sense of paradox has diminished and Stein's ideas are being incorporated into applied and theoretical statistics.

Stein's paradox concerns the use of observed averages to estimate unobservable quantities. Averaging is the second most basic process in statistics, the first being the simple act of counting. A baseball player who gets seven hits in 20 official times at bat is said to have a batting average of .350. In computing this statistic we are forming an estimate of the player's true batting ability in terms of his observed average rate of success. Asked how well the player will do in his next 100 times at bat, we would probably predict 35 more hits. In traditional statistical theory it can be proved that no other estimation rule is uniformly better than the observed average.

The paradoxical element in Stein's result is that it sometimes contradicts this elementary law of statistical theory. If we have three or more baseball players, and if we are interested in predicting future batting averages for each of them, then there is a procedure that is better than simply extrapolating from the three separate averages. Here "better" has a strong meaning. The statistician who employs Stein's method can expect to predict the future averages more accurately no matter what the true batting abilities of the players may be.

Baseball is a sport with a large and carefully compiled body of statistics, which supplies convenient material for illustrating the workings of Stein's method. As our primary data we shall consider the batting averages of 18 major-league players as they were recorded after their first 45 times at bat in the 1970 season. These were all the players who happened to have batted exactly 45 times the day the data were tabulated. A batting average is defined, of course, simply as the number of hits divided by the number of times at bat; it is always a number between 0 and 1. We shall denote each such average by the letter y.

The first step in applying Stein's method is to determine the average of the averages. Obviously this grand average, which we give the symbol $\bar{y}$, must also lie between 0 and 1. The essential process in Stein's method is the "shrinking" of all the individual averages toward this grand average. If a player's hitting record is better than the grand average, then it must be reduced; if he is not hitting as well as the grand average, then his hitting record must be increased. The resulting shrunken value for each player we designate z. This value is the James-Stein estimator of that player's batting ability, named for Stein and W. James, who together proposed a particularly simple version of the method in 1961. Stein's paradox is simply that the z values, the James-Stein estimators, give better estimates of true batting ability than the individual batting averages.

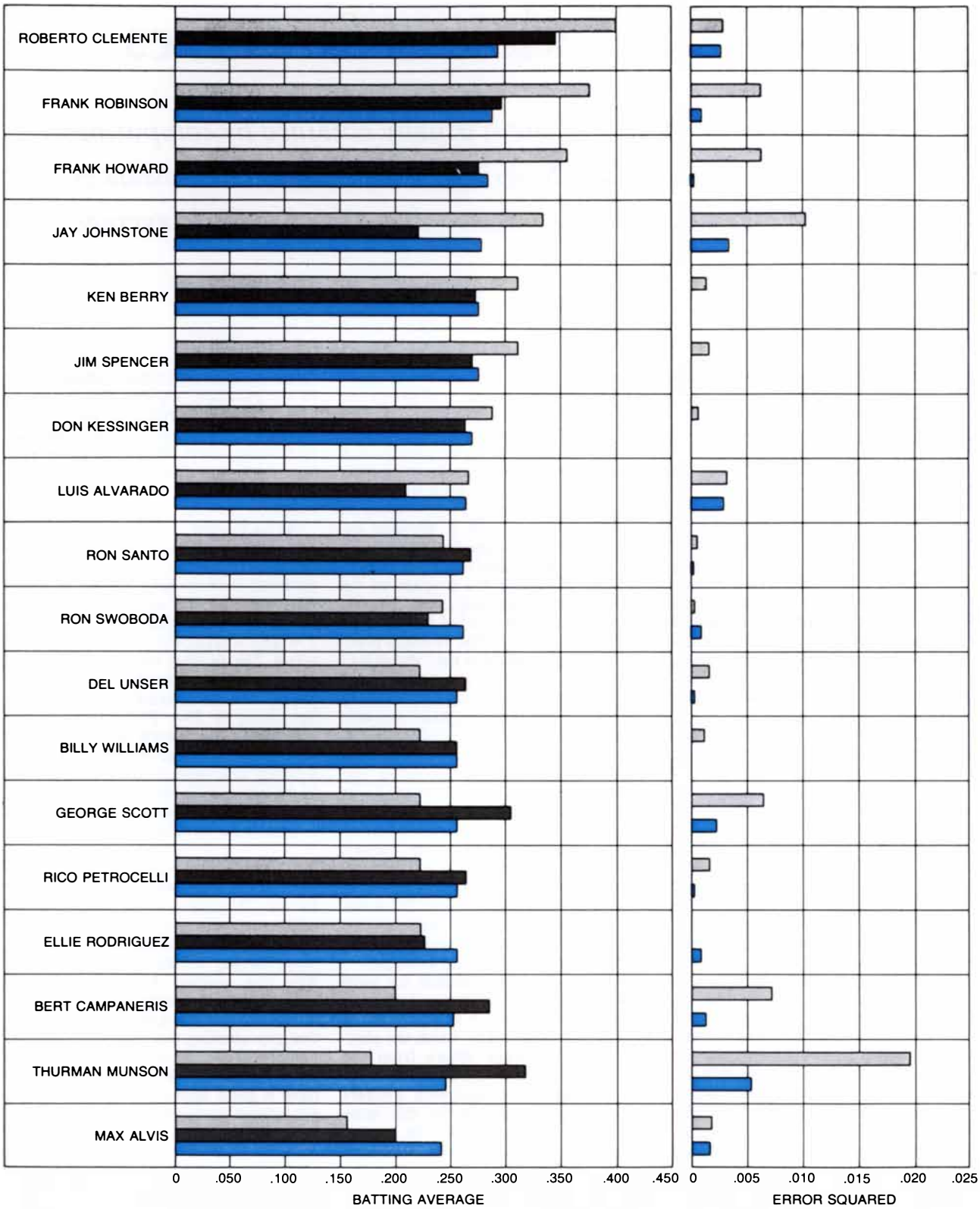
The James-Stein estimator for each player is found through the following equation: $z = \bar{y} + c(y - \bar{y})$. The quantity $(y - \bar{y})$ is the amount by which the player's batting average differs from the grand average. The equation thus states that the James-Stein estimator z differs from the grand average by this same quantity $(y - \bar{y})$ multiplied by a constant, c. The constant c is the "shrinking factor." If it were equal to 1, then the equation would state that the James-Stein estimator for a given player is identical with that player's batting average; in other words, y equals z. Stein's theorem states that the shrinking factor is always less than 1. Its actual value is determined by the collection of all the observed averages.

In the case of the baseball data, the grand average $\bar{y}$ is .265 and the shrinking factor c is .212. Substituting these values in the equation, we find that for each player z equals $.265 + .212(\bar{y} - .265)$. Because c is about .2, each average will shrink about 80 percent of the distance to the grand average, and the total spread of the averages will be reduced about 80 percent.

As an example consider the late Roberto Clemente, who was the leading batter in the major leagues when our statistics were compiled. For Clemente y is equal to .400, and z can be determined by evaluating the expression $z = .265 + .212(.400 - .265)$. The result is .294. In other words, Stein's theorem states that Clemente's true batting ability is best estimated not by .400 but lies closer to .294. Thurman Munson, in a batting slump early in the 1970 season, had an average of only .178. Substituting this value in the equation, we find that his estimated batting ability is substantially increased: the James-Stein estimator for Munson is .247.

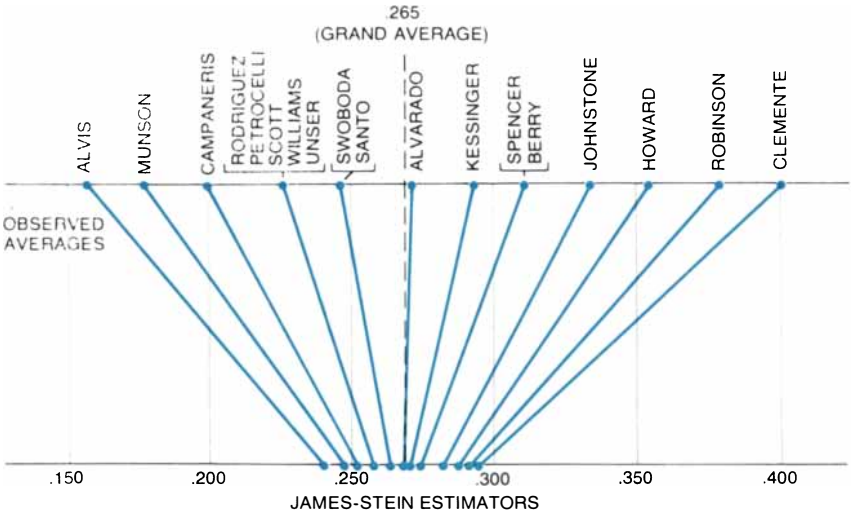
Which set of values, y or z, is the better indicator of batting ability for the 18 players in our example? In order to answer that question in a precise way one would have to know the "true batting ability" of each player. This true average we shall designate with θ (the Greek letter theta). Actually it is an unknowable quantity, an abstraction representing the probability that a player will get a hit on any given time at bat. Although θ is unobservable, we have a good approximation to it: the subsequent performance of the batters. It is sufficient to consider just the remainder of the 1970 season, which includes about nine times as much data as the preliminary averages were based on. The expected statistical error in such a sample is small enough for us to neglect it and proceed as if the seasonal average were the "true batting ability" θ of a player. That is one reason for choosing batting averages for this example. In most problems the true value of θ cannot be determined.

One method of evaluating the two es-



BATTING ABILITIES of 18 major-league baseball players are estimated more accurately by the method of Charles Stein and W. James than they are by the individual batting averages. The averages employed as estimators are those calculated after each player had had 45 times at bat in the 1970 season. The true batting ability of a player is an unobservable quantity, but it is closely approximated by his long-term average performance. Here the true ability is represented by the batting average maintained during the remainder of the 1970 season. For 16 of the players the initial average is inferior to another number, the James-Stein estimator, as a predictor of batting ability. The James-Stein estimators, considered as a group, also have the smaller total squared error.

timates is by simply counting their successes and failures. For 16 of the 18 players the James-Stein estimator z is closer than the observed average y to the "true," or seasonal, average θ. A more quantitative way of comparing the two techniques is through the total squared error of estimation. This is measured by first determining the actual error of each prediction, given by $(\theta - y)$ and $(\theta - z)$, for each player. Each of these quantities is then squared and the squared values are added up. The observed averages y have a total squared error of .077, whereas the squared error of the James-Stein estimators is only .022. By this comparison, then, Stein's method is 3.5 times as accurate. It can be shown that for the data given 3.5 is close to the expected ratio of the total squared errors of the two methods. We have not just been lucky.



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Suppose a statistician makes a random sampling of automobiles in Chicago and finds that of the first 45 recorded nine are foreign-made and the remaining 36 are domestic. We want to estimate the true proportion of imported cars in Chicago, a quantity represented by another unobservable θ. The observed average of $9/45 = .200$ is one estimate. Another can be obtained by simply lumping this problem together with that of the 18 baseball players. Substituting the value .200 in the equation used in that problem gives a James-Stein estimator of .251 for the imported-car ratio. (Actually the addition of a 19th value changes the grand average $\bar{y}$ and also slightly alters the shrinking factor c. The changes are small, however; the amended value of z is .249.)

In this case intuition argues strongly that the observed average and not the James-Stein estimator must be the better predictor. Indeed, the entire procedure seems silly: what could batting averages have to do with imported cars? It is here that the paradoxical nature of Stein's theorem is most uncomfortably apparent. The theorem applies as well to the 19 problems as it did to the original 18. There is nothing in the statement of the theorem that requires the component problems to have some sensible relation to one another.

The same disconcerting indifference to common sense can be demonstrated in another way. What does Clemente's .400 observed average have to do with Max Alvis, who was poorest in batting among the 18 players? If Alvis had had an early-season hitting streak, batting say .444 instead of his actual .156, the James-Stein estimator for Clemente's average would have been increased from .294 to .325. Why should Alvis' success or lack of it have any influence on our estimate of Clemente's ability? (They were not even in the same league.)

It is questions of this kind that have been raised by critics of Stein's method. In order to reply to them it will be necessary to describe the method rather more carefully.

Taking an average is an easy and familiar process that seems to need no justification. Actually it is not obvious why the average is so often useful in estimating the true center of gravity of a random process. The explanation lies in the distribution that the values of the random variable tend to assume.

The distribution most common in scientific work is the "normal" distribution, described by a bell-shaped curve; it was first investigated in depth by Gauss and is sometimes called the Gaussian distribution. It is constructed by assuming that the random variable can take on any value along some axis; the probability that it falls within any given interval is then made equal to the area under the same interval of the bell-shaped curve. The curve is completely specified by two parameters: the mean, θ, which lies at the peak of the curve, and the standard deviation, which measures how closely the values are distributed around the mean. It is customary to assign the standard deviation the symbol σ (sigma). The larger the standard deviation is, the more widely dispersed the data are.

In probability theory a known mean and standard deviation are employed to predict future behavior. A problem in statistics proceeds in the opposite direction: from observed data the statistician must infer the mean θ and the standard deviation σ.

Suppose, for example, the measurement of some random variable x yields the five successive values 10.0, 9.4, 10.3, 8.6 and 9.7. Suppose further the values are known to be part of a normal distribution with a standard deviation of 1. What is the value of the true mean θ? In principle the mean could have any value, but some values are more likely than others. A mean of 6.5, for example, would require that all five values be under the extreme tail of the curve and that none be found near the center. Gauss showed that among all possible choices for the mean, the average $\bar{x}$ of the observed data (which in this case has a value of 9.6) maximizes the probability of obtaining the data actually seen. In this sense the average is the most likely estimate of the mean; in fact, Gauss constructed the normal distribution just so that it would have this property.

There is a further justification, also pointed out by Gauss, for choosing the average as the best estimator of the unobservable mean θ. Gauss noted that the average of the data is an "unbiased" estimator of the mean, in the sense that it favors no selected value of θ. To be more precise, the average is unbiased because the expected value of $\bar{x}$ equals the true θ no matter what θ may be. There are infinitely many unbiased estimators of θ, none of which estimates θ perfectly. Gauss showed that the expected squared error of estimation for the average $\bar{x}$ is lower than that for any other linear, unbiased function of the observations. In the 1940's it was demonstrated that no other unbiased function of the data, whether it is linear or nonlinear, can estimate θ more accurately than the average, in terms of expected squared error. An essential contribution to that proof had been made in the 1920's by

R. A. Fisher, who showed that all the information about θ that can possibly be found in the data is contained in the average $\bar{x}$.

In the 1930's a mathematically more rigorous approach to statistical inference was undertaken by Jerzy Neyman, Egon S. Pearson and Abraham Wald; the ideas they developed are part of what is now known as statistical decision theory. They discarded the requirement of unbiased estimation and examined all functions of the data that could serve as estimators of the unknown mean θ. These estimators were compared through a risk function, defined as the expected value of the squared error for every possible value of θ.

Consider three competing estimators: the average of the data, $\bar{x}$; half that average, $\bar{x}/2$, and the median of the data, or middle value. For both the average and the median the risk function is constant; that is merely another way of saying that their expected squared error in predicting the mean θ is the same no matter what the value of θ really is. Of the two constant risk functions, the one for the average $\bar{x}$ is uniformly smaller by a factor of about two-thirds; clearly the average is the preferred estimator. In the language of decision theory the median is said to be "inadmissible" as an estimator of θ, since there is another estimator that has a smaller risk (expected squared error) no matter what θ is. (It should be mentioned, however, that when the data have a distribution other than the normal one, it is possible for the order of preference to be reversed.)

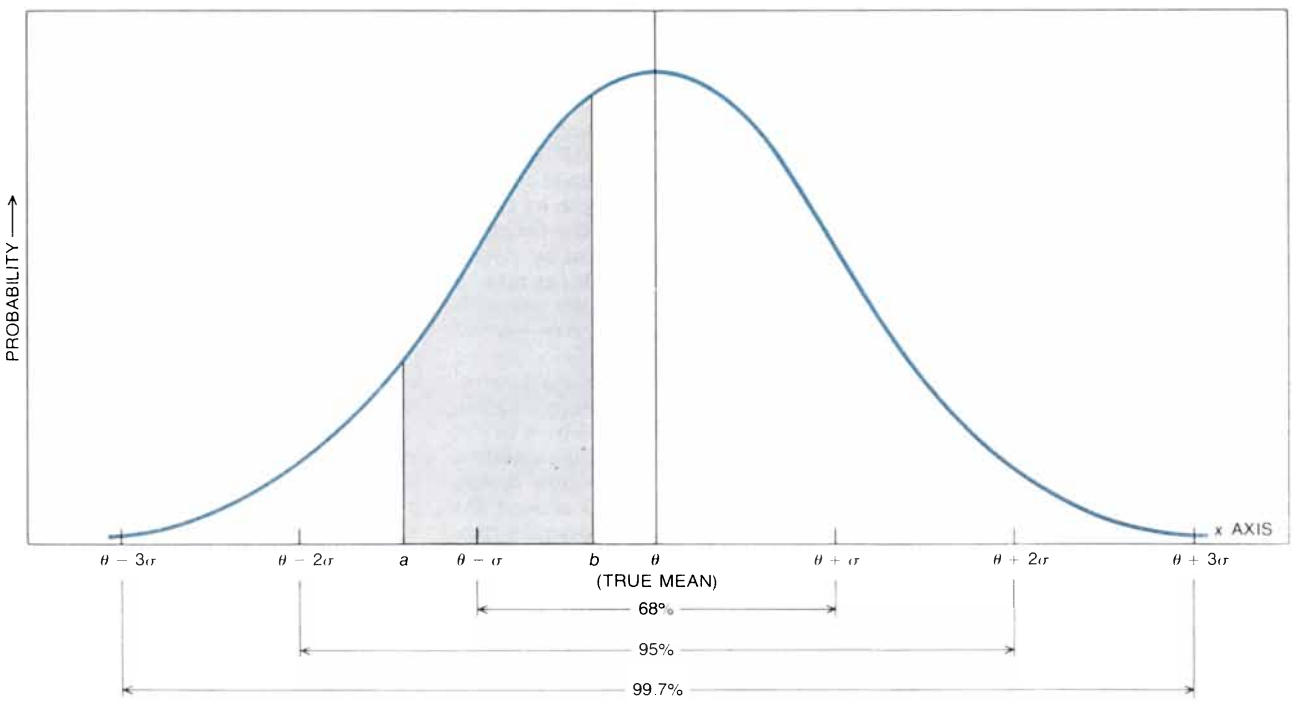
For the estimator $\bar{x}/2$, which is biased toward the value $\theta = 0$, the risk function is not constant; this estimator is accurate if θ happens be close to zero, but the expected squared error increases rapidly as the true mean departs from zero. The risk function describes a parabola, with the minimum point at $\theta = 0$; if the mean does happen to be zero, then the risk function for $\bar{x}/2$ is four times smaller than that for the average itself. At large values of the mean, however, the average $\bar{x}$ regains its superiority. With other estimators we can poke down the risk function below that of the average at any point we wish to, but it always pops up again somewhere else.

There remains the possibility that some other estimator has a risk that is uniformly lower than that of the average. In 1950 Colin R. Blyth, Erich L. Lehmann and Joseph L. Hodges, Jr., proved that no such estimator exists. In other words, the average $\bar{x}$ is admissible, at least when it is applied to one set of observations for the purpose of estimating one unknown mean.

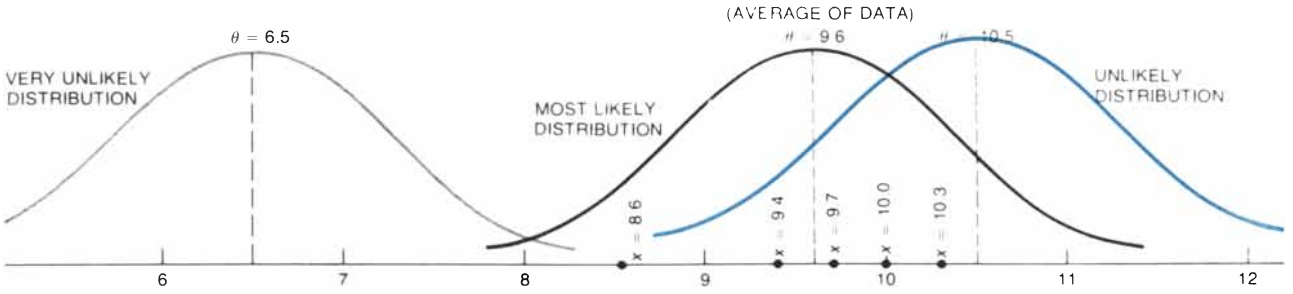
Stein's theorem is concerned with the estimation of several unknown means. No relation between the means need be assumed; they can be batting abilities or proportions of imported cars. On the other hand, the means are assumed to be independent of one another. In evaluating estimators for these means it is once again convenient to employ a risk function defined as the sum of the expected values of the squared errors of estimation for all the individual means.

The obvious first choice of an estimator for each of several means is the average of the data related to that mean. The entire historical development of statistical theory from Gauss through decision theory argues that the average is an admissible estimator as long as there is just one mean, θ, to be estimated. Stein showed in 1955 that the average is also admissible for estimating two means. Stein's paradox is simply his proof that when the number of means exceeds two, estimating each of them by its own average is an inadmissible procedure. No matter what the values of the true means, there are estimation rules with smaller total risk.

In 1955 Stein was able to prove this proposition only in those cases where the number of means, a quantity we shall designate k, was very large. Stein's 1961 paper written in collaboration with James extended the result to all values of k greater than 2; moreover, it did so in a constructive manner. Stein and James not only showed that estimators must exist that are everywhere superior to the



NORMAL DISTRIBUTION of a random variable around the mean value of that variable provides the fundamental justification for estimation by averaging. The distribution is defined by two parameters, the mean, θ, which locates the central peak of the distribution, and the standard deviation, σ, which measures how widely scattered the data points are. It is assumed in defining the distribution that the variable x can take on any value on the x axis. The most likely value of x is, by definition, the mean θ. The probability that x lies within any given interval on the axis, such as that between the points a and b, is equal to the area under the bell-shaped curve between those points.



PROBLEM IN STATISTICS is to deduce from a set of data the true mean and standard deviation of the distribution. Even when it is known that the distribution is a normal one and that the standard deviation is 1, the mean could in principle have any value. Some values, however, are more likely than others. For example, the five data points (x) given here could be described by a normal distribution with a mean of 6.5 only if all five points were more than two standard deviations above the mean. It can be shown that the data are most likely to be generated by a distribution with a mean equal to the observed average of the data, denoted $\bar{x}$. In this case the average is equal to 9.6.

averages; they were also able to provide an example of such an estimator.

The James-Stein estimator has already been defined in our investigation of batting averages. It is given by the equation $z = \bar{y} + c(y - \bar{y})$, where y is the average of a single set of data, $\bar{y}$ is the grand average of averages and c is a "shrinking factor." There are several other expressions for the James-Stein estimator, but they differ mainly in detail. All of them have in common the shrinking factor c; it is the definitive characteristic of the James-Stein estimator.

In the baseball problem c was treated as if it were a constant. Actually it is determined by the observed averages and therefore is not a constant. The shrinking factor is given by the equation

$$c = 1 - \frac{(k-3)\sigma^2}{\Sigma(y-\bar{y})^2}.$$

Here k is again the number of unknown means, σ^2 is the square of the standard deviation and $\Sigma(y - \bar{y})^2$ is the sum of the squared deviations of the individual averages y from the grand average $\bar{y}$.

Let us briefly explore the meaning of this rather forbidding equation. With k and σ^2 fixed, we find that the shrinking factor c becomes smaller (and the predicted means are more severely affected by it) as the expression $\Sigma(y - \bar{y})^2$ gets smaller. On the other hand, c increases, approaching unity, and the shrinking is less drastic as the expression $\Sigma(y - \bar{y})^2$ increases.

What do these equations mean in terms of the behavior of the estimator? In effect the James-Stein procedure makes a preliminary guess that all the unobservable means are near the grand average $\bar{y}$. If the data support that guess in the sense that the observed averages are themselves not too far from $\bar{y}$, then the estimates are all shrunk further toward the grand average. If the guess is contradicted, then not much shrinking is done. These adjustments to the shrinking factor are accomplished through the effect the distribution of averages around the grand average $\bar{y}$ has on the equation that determines c. The number of means being estimated also influences the shrinking factor, through the term $(k - 3)$ appearing in this same equation. If there are many means, the equation allows the shrinking to be more drastic, since it is then less likely that variations observed represent mere random fluctuations.

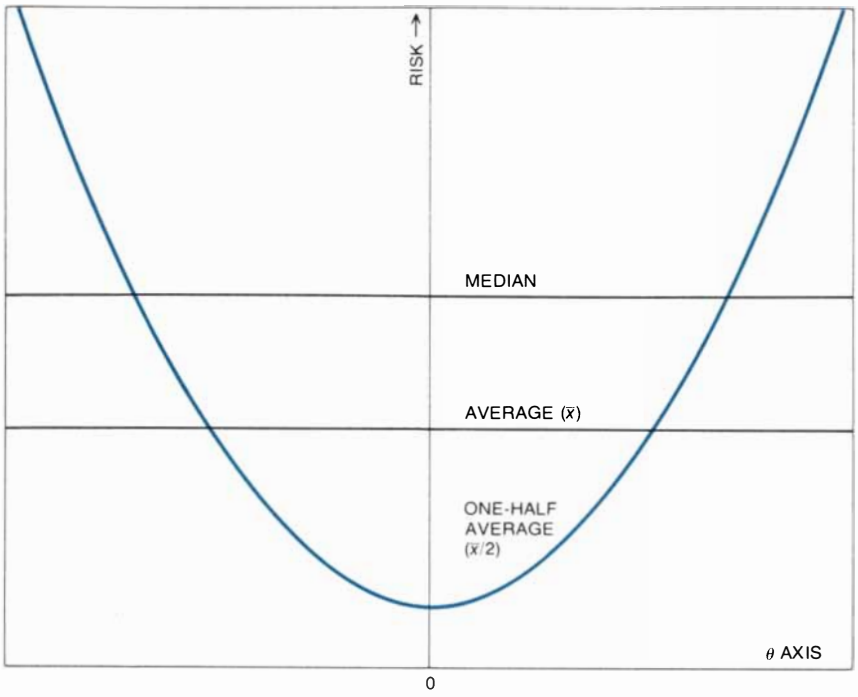
With c calculated in this manner, the risk function for the James-Stein estimator is less than that for the sample averages no matter what the true values of the means θ happen to be. The reduction of risk can be substantial, particularly when the number of means is larger than five or six. The risk function is not constant for all values of the true mean θ, as it is for the observed averages. The risk of the James-Stein estimator is smallest when all the true means are the same. As the true means depart from one another the risk of the estimator increases, approaching that of the observed averages but never quite equaling it. The James-Stein estimator does substantially better than the averages only if the true means lie near each other, so that the initial guess involved in the technique is confirmed. What is surprising is that the estimator does at least marginally better no matter what the true means are.

The expression for the James-Stein estimator that we have employed refers all observed averages to the grand average $\bar{y}$. This procedure is not the only one possible; other expressions for the estimator dispense with $\bar{y}$ entirely. What cannot be avoided is the introduction of some more or less arbitrary initial guess or point of origin for the estimator. The observed averages, it will be noted, do not depend on a choice of origin. Before Stein discovered his method it was felt that such "invariant" estimators must be preferable to those whose predictions change with each choice of an origin. The theory of invariance, to which Stein had been a principal contributor, was badly shaken by the James-Stein counterexample. From the standpoint of mathematics this is the most unsettling aspect of Stein's theorem. Indeed, the paradox was not discovered earlier largely because of a strong prejudice that the estimation problem, being stated without reference to any particular origin, should be solved in a similar way.

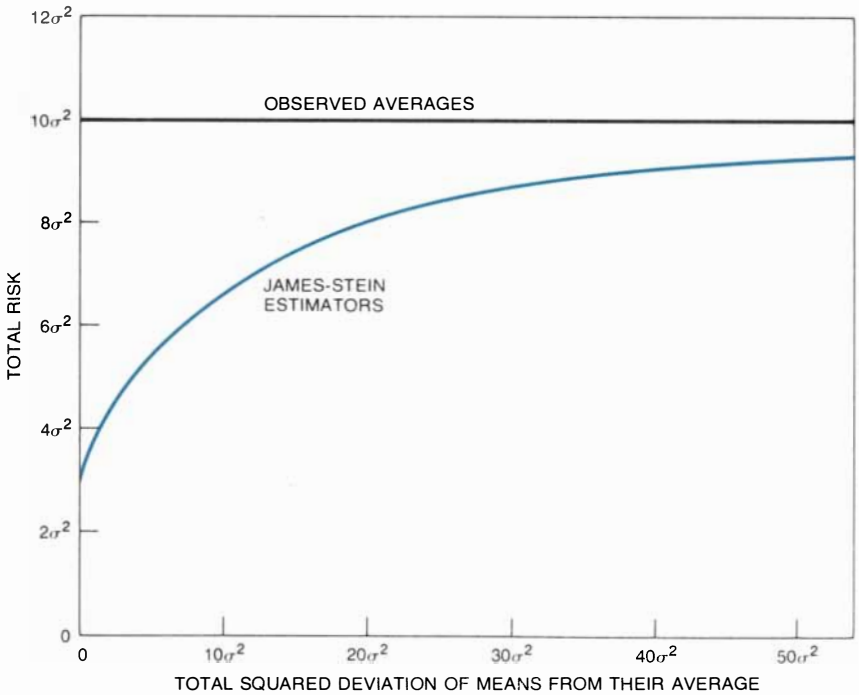
Applications of Stein's method tend to involve large sets of data with many unknown parameters. Some of the difficulties of such problems, as well as the practical potential of the method itself, can be illustrated by an example: an analysis of the distribution of the disease toxoplasmosis in the Central American country of El Salvador.

Toxoplasmosis is a disease of the blood that is endemic in much of Central America and in other regions of the Tropics. In El Salvador roughly 5,000 people drawn in varying numbers from 36 cities were tested for toxoplasmosis. The observed rate of incidence for each city can conveniently be expressed by comparison with the national rate (that is, with the grand average $\bar{y}$). A measured rate of .050, for example, denotes a city with an incidence of the disease 5 percent higher than the national average. The measured rates have an approximately normal distribution. The standard deviations of these distributions are known, but they differ from city to city, depending inversely on how large a sample population was tested in that city. It is the task of the statistician to estimate the true mean θ of the distribution for each city from the measured incidence y.

In this case the appropriate form of the James-Stein estimator is $z = cy$. The simplification, which was introduced by us, is made possible by the chosen manner of expressing the observations y. They are defined in such a way that the grand average $\bar{y}$ is zero, and terms containing $\bar{y}$ therefore drop out of the equation. On the other hand, the estimation



VARIOUS ESTIMATORS of a single true mean, θ, can be evaluated by way of a risk function. The risk is defined as the expected value of the squared error of estimation, considered as a function of the mean θ. The average of the data, $\bar{x}$, is an estimator with a constant risk function: no matter what the true mean is, the expected value of the squared error is the same. The median, or middle value, of the data also has constant risk, but it is everywhere greater (by a factor of 1.57) than the risk of the average. Half the average ($\bar{x}/2$) is an estimator whose risk depends on the actual value of the mean; the risk is smallest when the mean is near zero and increases rapidly when the mean departs from zero. For the estimation of a single mean there is no estimator with a risk function that is everywhere less than the risk function of the average $\bar{x}$.



TOTAL RISK FUNCTION for the James-Stein estimators is everywhere less than that for the individual observed averages, as long as the number of means being estimated is greater than two. In this case there are 10 unknown means. The risk is smallest when all the means are clustered at a single point. As the means depart from one another the risk of the James-Stein estimators increases, approaching that of the observed averages but never quite reaching it.

procedure is now complicated by the fact that the shrinking factor c is different for each city, varying inversely as the standard deviation of y for that city. This dependence of the shrinking factor on the standard deviation has a simple intuitive rationale. A large standard deviation implies a high degree of randomness or uncertainty in a measurement. If the measured incidence is unusually large, it can therefore be attributed more reasonably to random fluctuations within the normal distribution than to a genuinely large value of the true mean θ. It is thus proper to reduce this value drastically, that is, to apply a small shrinking factor.

The same argument can be made even more forcefully by returning for a moment to baseball. Frank O'Connor pitched for Philadelphia in the 1893 season. He batted twice in his major-league career, hitting successfully both times. His observed batting average is hence 1.000. The James-Stein rule for the 18 players considered above estimates O'Connor's true batting ability to be $.265 + .212(1.000 - .265) = .421$ (ignoring the effect of the new data on the grand average and on the shrinking factor). This is a silly estimate, although not as silly as 1.000. A perfect average after two times at bat is not at all inconsistent with a true value in the range from .242 to .294 that is estimated for the other players. The shrinking constant c applied to O'Connor's average should be severer in order to compensate for the smaller amount of data available for him.

For the El Salvador observations, most of the shrinking factors are quite gentle, between .6 and .9, but a few are in the range from .1 to .3. Which set of numbers should we prefer, the James-Stein estimators or the measured rates of incidence? That depends largely on what we want to use the numbers for.

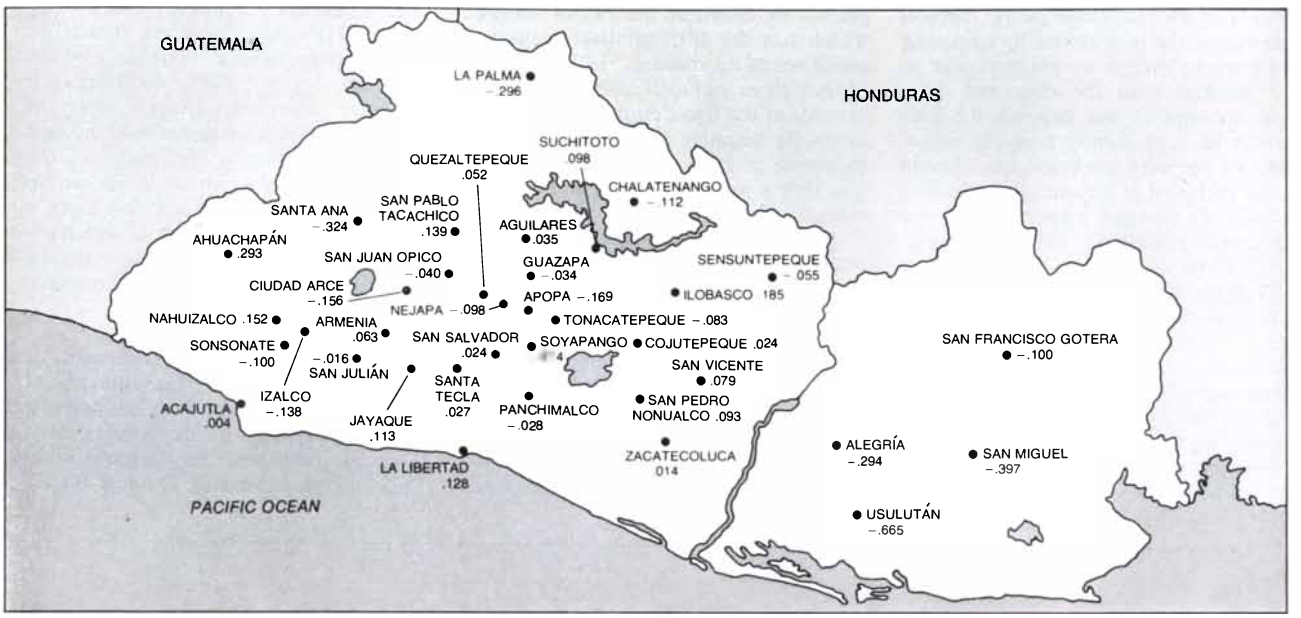
If the Minister of Health for El Salvador intends to build local hospitals for people suffering from toxoplasmosis, the James-Stein estimators probably offer the more reliable guidance. The reason is that the expected value of the total squared error is smaller for the James-Stein estimators; in fact, it is smaller by a factor of about three. The important point in this calculation is that the expected error is added up for all the cities. Any particular hospital might be the wrong size or in the wrong place, but the sum of all such mismatches would be smaller for the James-Stein estimators than for the observed rates.

The James-Stein estimators are also likely to be preferable for determining the ordering of the true means. In this regard it is notable that the city with the highest apparent incidence (according to the measured rates y) is ranked 12th according to the James-Stein estimators.

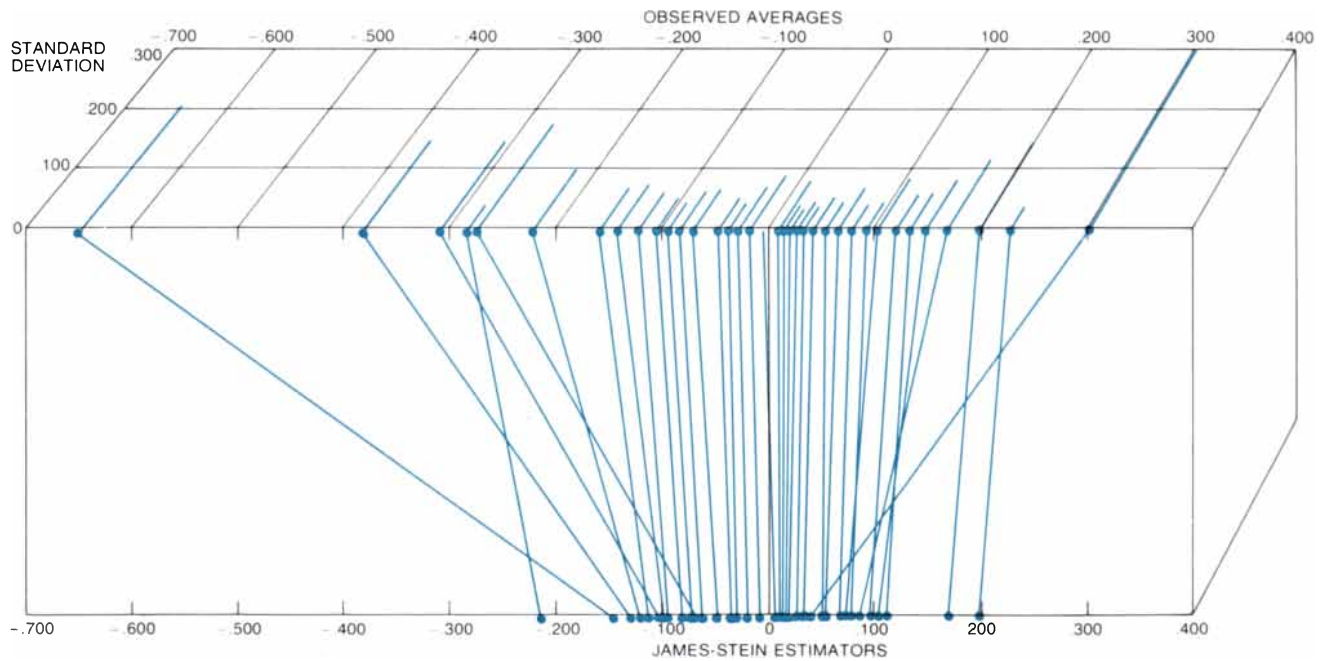
The estimate is drastically reduced because the sample was very small in that city. This information might be useful if there were funds for only one hospital.

Suppose an epidemiologist wants to investigate the correlation of the true incidence in each city with attributes such as rainfall, temperature, elevation or population? Once again the James-Stein estimators are preferred; a rough calculation shows that they would give a closer approximation in about 70 percent of the cases.

There is one purpose for which the measured incidence may well be superior to the James-Stein estimator: when a single city is considered in isolation. As we have seen, the James-Stein method gives better estimates for a majority of cities, and it reduces the total error of estimation for the sum of all cities. It



INCIDENCE OF TOXOPLASMOSIS, a disease of the blood, was surveyed in 36 cities in the Central American country El Salvador. The measured incidence in each city can be regarded as an estimator of the true incidence, which is unobservable. The measured incidence has a normal distribution whose standard deviation is determined by the number of people surveyed in that city. The measured rates are expressed in terms of deviation from the national incidence (the average of the rates observed in all the cities). Thus zero denotes exactly the national rate, and a city with a measured incidence of −.040 would have an observed rate 4 percent lower than the country as a whole.



SHRINKING of the observed toxoplasmosis rates to yield a set of James-Stein estimators substantially alters the apparent distribution of the disease. The shrinking factor is not the same for all the cities but instead depends on the standard deviation of the rate measured in that city. A large standard deviation implies that a measurement is based on a small sample and is subject to large random fluctuations; that measurement is therefore compressed more than the others are. In the El Salvador data the most extreme observations tend to be correlated with the largest standard deviations, again suggesting the unreliability of those measurements. Compared with the observed rates, the James-Stein estimators can be proved to have a smaller total error of estimation. They also provide a more accurate ranking of the cities.

cannot be demonstrated, however, that Stein's method is superior for any particular city; in fact, the James-Stein prediction can be substantially worse.

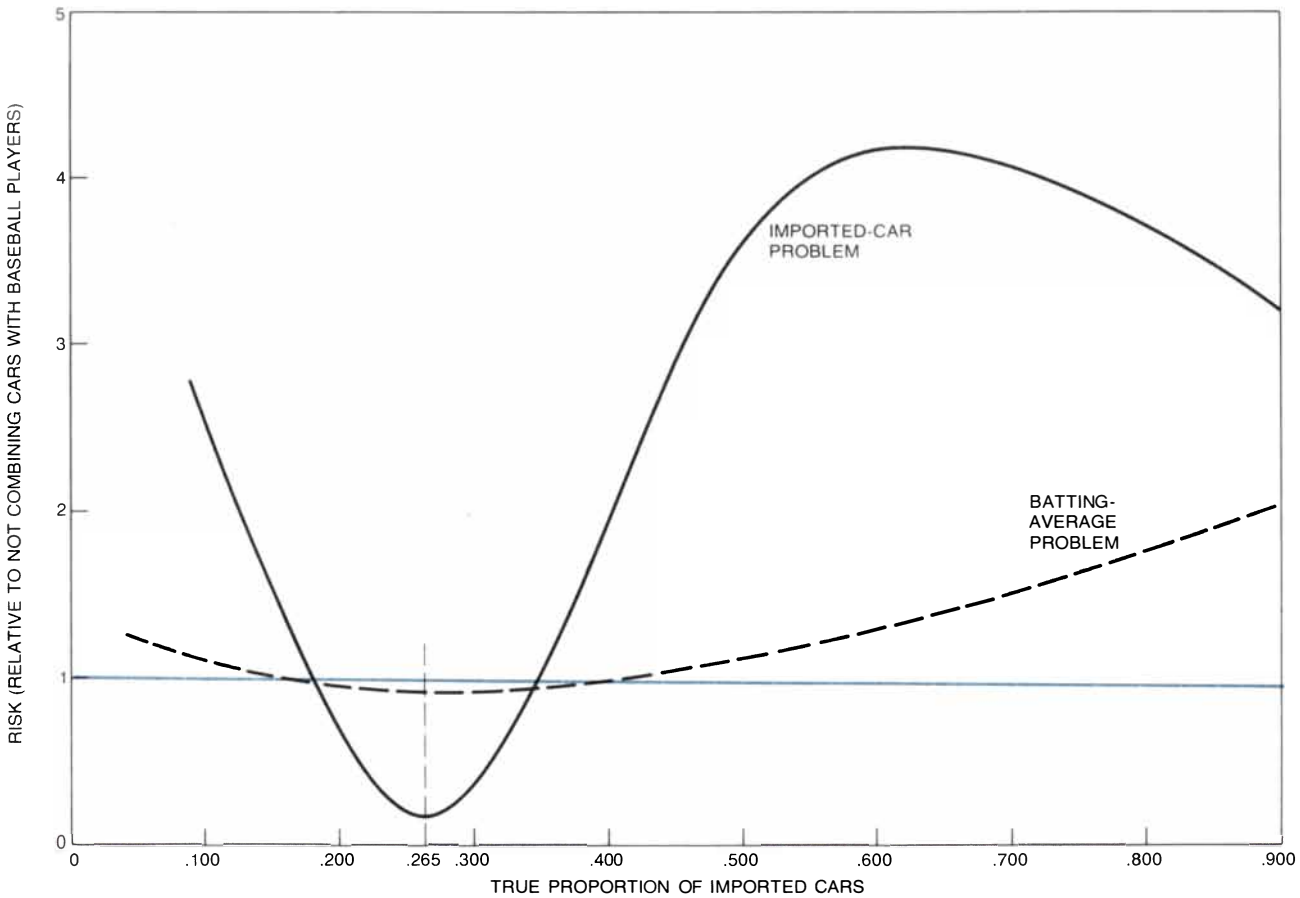
Estimating the true mean for an isolated city by Stein's method creates serious errors when that mean has an atypical value. The rationale of the method is to reduce the overall risk by assuming that the true means are more similar to one another than the observed data. That assumption can degrade the estimation of a genuinely atypical mean. Now we see why imported cars should not be included in the same calculations with the 18 baseball players. There is a substantial probability that the automobiles will be atypical.

Suppose we ignore this hazard and lump together all 19 problems; we can then calculate the total expected squared error as a function of the true percentage of imported cars. It turns out that the risk for both the baseball players and the automobiles is reduced only if the percentage of imported cars happens to lie in the same range as the estimated batting averages; otherwise the risk of error for both kinds of problem is increased.

The question of whether or not a particular mean is "typical" is a subtle one whose implications are not yet fully understood. Returning to the problem of toxoplasmosis in El Salvador, let us single out for attention the city of Alegría, which has the fifth-smallest measured incidence of the disease: $-.294$. It is one of four cities included in the survey that are east of the Rio Lempa; all four have distinctly negative values of measured incidence y. It is plausible to suppose that this is no coincidence and that the rate of toxoplasmosis east of the Lempa is genuinely lower. A James-Stein estimator that consolidates information from the entire country therefore may be less than optimal in these cities. We have developed techniques for taking advantage of extra information of this kind, but the theory underlying those techniques remains rudimentary.

An astute follower of baseball might be aware that just as each player's batting ability can be represented by a Gaussian curve, so too the true batting abilities of all major-league players have an approximately normal distribution. This distribution has a mean of .270 and a standard deviation of .015. With this valuable extra information, which statisticians call a "prior distribution," it is possible to construct a superior estimate of each player's true batting ability. This new estimator, which we shall give the label Z, is defined by the equation $Z = m + C(y - m)$. Here y is again the observed batting average of the player, but $\bar{y}$, the grand average, has been replaced by m, the mean of the prior distribution, which is known to have the value .270. In addition there is a different shrinking factor, C, which depends in a simple way on the standard deviation of the prior distribution (equal to .015).

This procedure is not a refinement of Stein's method; on the contrary, it predates Stein's method by 200 years. It is the mathematical expression of a theorem published (posthumously) in 1763 by the Reverend Thomas Bayes.



UNRELATED PROBLEMS can be lumped together for analysis by Stein's method, but only at the risk of increasing error. To the 18 batting averages computed earlier, for example, one might add a 19th number representing the proportion of imported cars observed in Chicago. New James-Stein estimators could then be calculated for both the baseball players and the automobiles, based on the grand average of all 19 numbers. Nothing in the statement of Stein's theorem prohibits such a procedure, but the evident illogic of it has justifiably been criticized. In fact, including the unrelated data can reduce the risk function only if the proportion of imported cars happens to be near the mean batting average of .265; otherwise the expected error of estimation for both the cars and the baseball players is increased.

He was able to show that this estimator minimizes the expected squared error associated with the randomness in both the observed averages (y) and in the true means (θ).

The formula for the James-Stein estimator is strikingly similar to that of Bayes's equation. Indeed, as the number of means being averaged grows very large, the two equations become identical. The two shrinking factors c and C converge on the same value, and the grand average $\bar{y}$ becomes equal to the mean m precisely when all players are included in the calculation. The James-Stein procedure, however, has one important advantage over Bayes's method. The James-Stein estimator can be employed without knowledge of the prior distribution; indeed, one need not even suppose the means being estimated are normally distributed. On the other hand, ignorance has a price, which must be paid in reduced accuracy of estimation. We have shown that the James-Stein method increases the risk function by an amount proportional to $3/k$, where k is again the number of means being estimated. The additional risk is therefore negligible when k is greater than 15 or 20, and it is tolerable for k as small as 9.

In this historical context the James-Stein estimator can be regarded as an "empirical Bayes rule," a term coined by Herbert E. Robbins of Columbia University. In work begun in about 1951 Robbins demonstrated that it is possible to achieve the same minimum risk associated with Bayes's rule without knowledge of the prior distribution, as long as the number of means being estimated is very large. Robbins' theory was immediately recognized as a fundamental breakthrough; Stein's result, which is closely related, has been much slower in gaining acceptance.

The James-Stein estimator is not the only one that is known to be better than the sample averages. Indeed, the James-Stein estimator is itself inadmissible! Its failure lies in the fact that the shrinking factor c can assume negative values, and it then pulls the means away from the grand average rather than toward it. When that happens, simply replacing c with zero produces a better estimator. This estimator in turn is also inadmissible, but no uniformly better estimator has yet been found.

The search for new estimators continues. Recent efforts have been concentrated on achieving results like those obtained with Stein's method for problems involving distributions other than the normal distribution. Several lines of work, including Stein's and Robbins' and more formal Bayesian methods seem to be converging on a powerful general theory of parameter estimation.

MATHEMATICAL GAMES

The "jump proof" and its similarity to the toppling of a row of dominoes

by Martin Gardner

"Yes," he said, "a man has no need of eyes to perceive that."

—PLATO, *The Republic*, Book V

Will the sun rise on January 1, 2000? There is no way to be absolutely sure. It is always possible that the world will end before then by divine decree or by some natural calamity. Perhaps a giant comet (as in Immanuel Velikovsky's mythology) will cause the earth to stop rotating and the sun to stand still upon Gibeon. The most we can say is that it is an excellent bet that on January 1, 2000, the sun will rise as usual. Our jump from a finite set of past sunrises to an infinite future set, or at least to a future set with a large number of elements, is an empirical induction.

Mathematicians have an analogous technique known as mathematical induction or complete induction that also supports a jump from a finite set of cases to a larger or an infinite number of cases. Unlike empirical induction, the mathematical technique is entirely deductive. A "jump proof," as it is sometimes called, is as certain as any proof can be in mathematics.

To prove something by mathematical induction we must first have a series of statements (usually an infinite series but not necessarily so) that can be put into a one-to-one correspondence with the sequence of positive integers. Second, we must establish that the statements are related to one another by what Bertrand Russell termed the "hereditary property." If any statement is true, its successor—the "next" statement—is true. Third, we must show that the first statement is true. It then follows with iron certainty that all the statements are true.

Jump proofs have been likened to a row of bricks or dominoes that are standing on end and all topple over when you unbalance the first one. Hugo Steinhaus compared mathematical induction to a pile of envelopes, each containing a note that says: "Open the next envelope, read the order and carry it out." If you are committed to obeying the order in the first envelope, you must open all the envelopes and obey all the orders.

Hundreds of classic problems in recreational mathematics are proved in the general case by mathematical induction. Into how many pieces can you cut a pie with n straight cuts? In how few moves can you transfer n disks in the Tower of Hanoi puzzle? This month we discuss a class of mind-twisting logic puzzles for which the application of induction to the general case is less well known and is fraught with curious perils.

We begin with the old puzzle of the colored hats. Three men, *A*, *B* and *C*, close their eyes while someone puts on the head of each either a black hat or a red one. They open their eyes. Each man sees the two hats not his own. If he sees a red hat, he raises his hand. As soon as he knows the color of his own hat he must say so.

Suppose all three hats are red. The three men raise their hands. After a period of time *C*, who is smarter than the others, says: "My hat is red." How does he know?

C reasons as follows. "Suppose my hat is black. *A*, seeing my black hat, will know at once that his own hat is red. Otherwise, why would *B*'s hand be raised? *B* will reason the same way and also will know at once that his hat is red. Neither *A* nor *B*, however, has said anything. Their hesitancy can only be explained if they see a red hat on me also. Therefore my hat is red."

Consider now the case of four men, all with red hats. If the fourth man, *D*, is smarter than the rest, he will reason: "Suppose my hat is black. The other three men have their hands raised because they see red hats. This is precisely the preceding case. After a suitable lapse of time, *C*, the smartest of the three, will deduce that his hat is red and say so." *D* then waits to see if *C* says anything. Because *C* says nothing, *D* knows his own hat is red.

Clearly this procedure generalizes. If there are five men, *E* will know his hat is red, because if it is black, the situation is reduced to the preceding case; after a suitable lapse of time *D* will know his hat is red. *D*'s silence proves to *E* that all hats, including his own, must be red. And so it goes for any number of men. Mathematical induction forces us to conclude that if n men all have red hats, the smartest of them will eventually deduce that his own hat is red.

This generalization usually provokes arguments because the problem demands so many fuzzy assumptions about degrees of smartness and lengthening lapses of time that the problem becomes unreal. Presumably if there are 100 men, after a few hours the smartest will know his hat is red, then after a while the second-smartest will know and so on down to the two stupidest men.

The fuzziness can be avoided by giving the same problem in a more precise form. There are three men, *A*, *B* and *C*, and five hats. Three hats are red and two are black. Each man is assumed to be honest and "rational" in the sense that he can quickly make any valid deduction no matter how complicated it is. As before, the men close their eyes, and an "umpire" puts a red hat on each man. The other two hats are hidden. Instead of being told to raise their hand if they see a red hat, the men are asked in order: "Do you know the color of your hat?"

A truthfully answers no. *B* also says no. *C* says, "Yes, my hat is red." How does he know?

A surprising aspect of this problem is that *C* can answer yes even though he is blind! Moreover, it is not necessary for *B* to see *C*'s hat. Think of the three men as being seated in a row of chairs, as is shown in the top illustration on page 131. Each man sees only the hats on the men in front of him. *C*, who is the man in the third chair, is blind in the sense that he sees no hats.

C reasons as follows: "*A* can say yes only if he sees two black hats. His saying no proves that hats *B* and *C* are not both black. Suppose my hat is black. *B* can see that it is. Therefore as soon as *B* hears *A* say no, he knows his own hat is red. (Otherwise hats *B* and *C* would be black and *A* would have said yes.) The fact that *B* also said no can be explained only if he sees my hat is red. Therefore I can answer yes."

This problem, like the preceding one, generalizes easily to n men seated in a row of chairs, a supply of n red hats and $n - 1$ black hats. Assume that a fourth man, *D*, is seated ahead of *C*. All hats are red. *D* reasons that if his hat is black, the three men behind him will see his black hat and know that only two black hats are left for themselves. Thus the problem is reduced to the preceding case, which is solved. After *A* and *B* had said no, *C* would say yes. But *C* also says no,

Saving the Main Frame.

The Harris 1600 Remote Communications Processor

Some central computers get bogged down by nickel and dime jobs from distant plants and offices. With Harris "intelligent" terminals and controllers, remote locations can handle routine processing on the spot, saving main frame time for main frame work.

Write: Harris Corporation,
55 Public Square, Cleveland, Ohio 44113.

HARRIS

COMMUNICATIONS AND INFORMATION HANDLING

Every day, we help America save more energy than it takes to light New York.

More than enough for every bulb, neon sign, streetlamp and fluorescent light New York City burns every day.

And that doesn't include *all* the energy Combustion Engineering helps save.

Only the fuel saved by the Ljungstrom® heat exchanger, just *one* of the energy-recovery products from C-E Air Preheater.

Installed on a steam generator or process furnace, the Ljungstrom captures heat energy that would be lost through the smokestack and sends it back to work.

With the Ljungstrom, a utility, refinery or process plant uses up to 25% less fuel. In fact, over the past five years, the energy saved by Ljungstrom units around the world equals 150 million tons of coal, 190 million barrels of oil and 1.4 trillion cubic feet of natural gas.

We do more than save fuel.

For over 60 years, we've helped provide the equipment, technology and research to produce energy, conserve it, squeeze more use out of it, even create new forms of it.

C-E Power Systems, for example, designed the steam generators that produce 40% of all the thermal-electric power in the Free World.

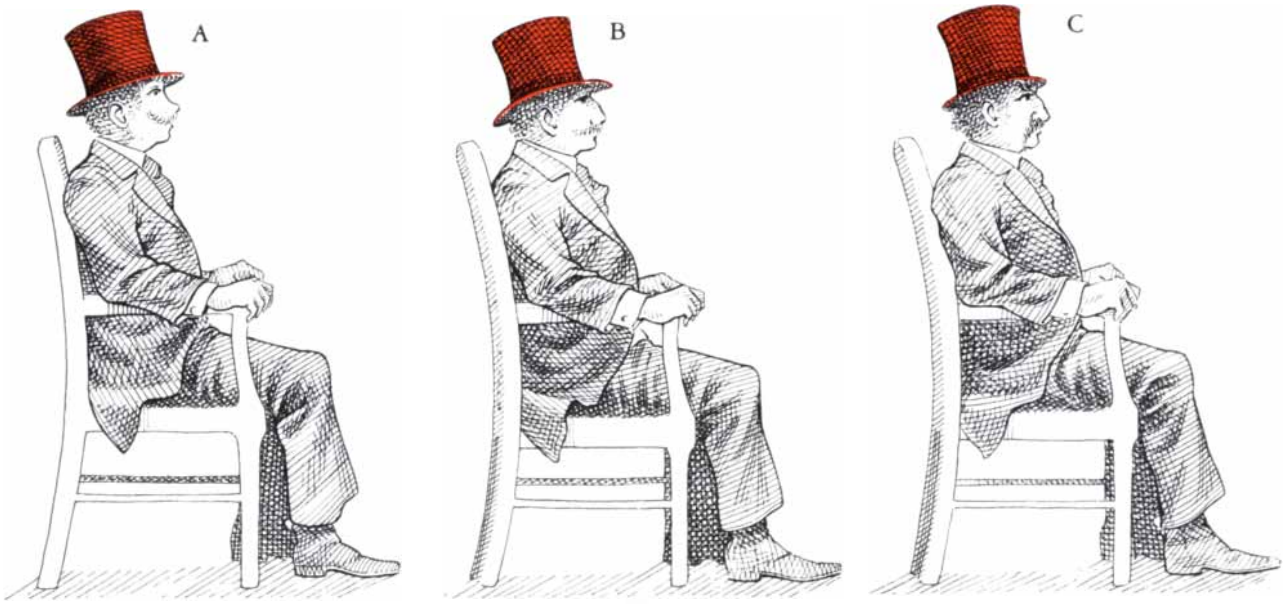
C-E Natco helps process 25% of all the oil and gas the world uses every day.

C-E Lummus has created a clean liquid fuel by removing the ash and sulfur from coal.

For more information about our activities around the world, write Combustion Engineering, Inc., Dept. 7008-27, 902 Long Ridge Road, Stamford, Connecticut 06902.

COMBUSTION ENGINEERING

The Energy Systems Company



The problem of the colored hats

which proves to *D* that his own hat must be red. Mathematical induction at once extends the solution to *n* men. If all have red hats, all will say no except the *n*th man, who will know his hat is red.

A more difficult question can now be asked. Picture again three men in a row of chairs and assume that the umpire gives them any combination of hats from the set of five. The men are questioned in order of increasing "blindness" (*A, B, C*). Will one of them always be able to answer yes? And does this situation generalize to *n* men and a set of *n* red and $n - 1$ black hats? Will there always be a yes answer on or before the *n*th question regardless of what hats are put on their heads?

In most problems of this type we encounter a curious paradox. Consider the case of three men, all hats red and each man able to see the other two. *A* and *B* answer no, *C* answers yes. Why is it necessary to question *A?* Before *A* is asked both *B* and *C* know that he must say no; *B* knows because he sees the red hat on *C,* and *C* knows because he sees the red hat on *B*. If *B* and *C* know how *A* will answer, how can asking *A* and hearing his reply add any significant new information? On the other hand, if the questioning begins with *B, C* is unable to make his deduction. Can you explain this seeming paradox?

Hats of two colors are equivalent to hats labeled 0 and 1, the integers in binary notation. There are dozens of problems, closely related to the hat ones, in which more than two colors are involved but that are easier to understand if instead of colors we use positive decimal integers. The following two-person game was sent to me last year by David Gale, a mathematician at the University of California at Berkeley.

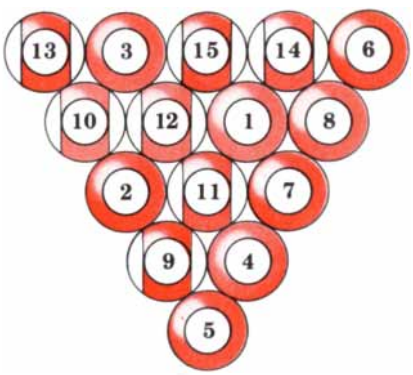
The umpire chooses any pair of consecutive positive integers. A disk with one of the numbers is stuck to the forehead of one man and a disk with the other number is stuck to the forehead of the other man. Each man is honest and rational. Each one sees the other's number but not his own. Each knows (and knows the other knows) that the two numbers are consecutive.

The umpire asks each man if he knows his number, and the questioning continues back and forth until one man says yes. It is not hard to prove by the magic of induction that eventually the man with the higher number, *n,* will be the first to say yes, and that his yes will be in reply to question *n* or $n - 1$. Readers are invited to analyze the game and to state under what conditions the high man says yes to question *n* or to question $n - 1$. Only two variables need to be taken into account: whether the high man or the low man is asked first and whether the high number is odd or even.

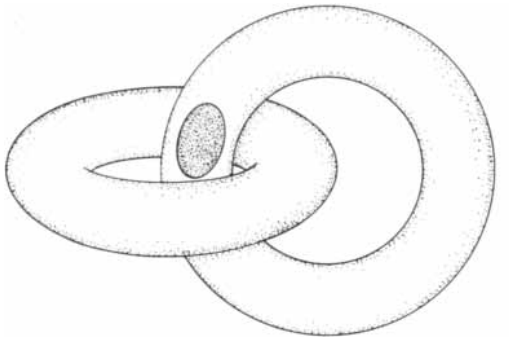
The paradox of the hat game appears here in even more striking form. I paraphrase from Gale's letter. Assume that the numbers are 99 and 100 and that the man with 100 is asked first. He will say yes to the 100th question. But why ask the first two? Each man knows before the questioning begins that the first two answers must be no. How then can asking the first two questions furnish significant information? After the first two are asked the men seemingly will know nothing they did not know before; therefore they should be no nearer to deducing their number than before, and the game will never end. How can a ritual intoning of no, which both men know must occur, shorten the number of questions required before a yes answer can be made? The argument appears impeccable.

Suppose we limit the integers to the counting numbers from 1 through 100. Each pair of consecutive numbers (1,2; 2,3; ...; 99,100) is written on a card. The umpire takes a card at random, puts its two numbers on the foreheads of two rational men and proposes the following game. The man with the lower number, *k,* must pay *k* dollars to his opponent. The umpire asks *A* if he wishes to play, and then he asks *B*. The payoff occurs only if both players say yes.

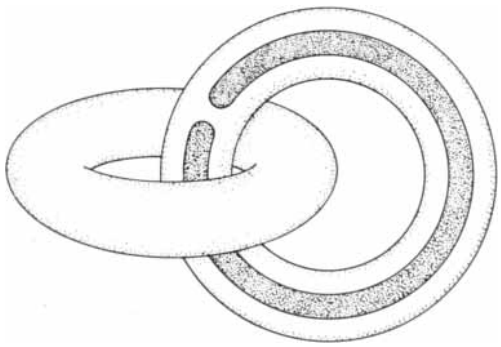
We now prove that a payoff never takes place. If *A* sees 100, he knows he has 99, and so he says no. If he sees 99, he reasons: "I am 98 or 100. If I am 100,



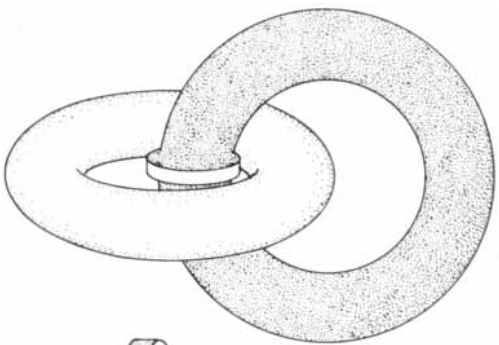
Solution of the pool-ball problem



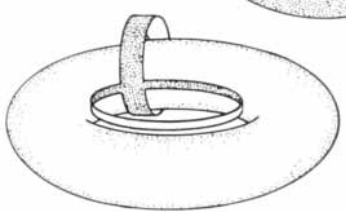
1. Mouth starts to open.



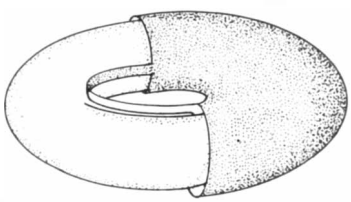
2. It lengthens to an enormous grin.



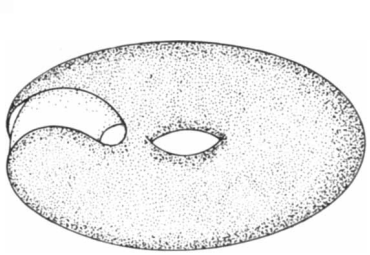
3. Grin widens until torus becomes two attached bands.



4. Horizontal band enlarges, vertical band shrinks.



5. Vertical band widens and creeps around victim.



6. Mouth closes.

How one torus eats another

then B (being rational) will say no and there will be no game. If I am 98, I surely should not play; therefore I must say no." If A sees 98, he reasons: "I am 97 or 99. If I am 99, then B will not play for the reasons given above. If I am 97, I shall lose; therefore I say no." And so on down even to seeing 1. If A sees 1, he knows he will win, but he also knows that if he says yes, B will say no.

Suppose the set of cards is infinite, with no upper bound on the integers. We now prove that both men will say yes. A reasons: "I see the number k. My number is either $k-1$ or $k+1$. If I lose, I lose $k-1$ dollars. If I win, I win k dollars. It is equally probable that I shall win or lose, and since I stand to win more than I stand to lose, the game is in my favor. Naturally I agree to play." Of course B reasons the same way. But this situation is preposterous because the game cannot favor both men.

The paradox can be magnified dramatically by taking a cue from J. E. Littlewood, who gives a version of the paradox in the first chapter of his *Mathematician's Miscellany*. Assume that there are 10^n duplicates of each card, where n is the card's lower number. Thus 1,2 is on 10 cards, 2,3 is on 100 cards, 3,4 is on 1,000 cards, and so on. The game is played as before. If either player sees number n, he knows there are 10 times as many cards with $n+1$ as there are with $n-1$. Therefore in addition to a win being a dollar more than a loss, the probability of winning seems to be for each player 10 times greater than losing! Littlewood attributes this "monstrous hypothesis" to the physicist Erwin Schrödinger.

I shall not resolve this paradox next month because I am not sure just how to do it. If readers send me their opinions, I cannot reply, but I shall report later on the response. It is not enough to prove the game fair with a playoff matrix. Obviously it is fair. The task is to explain what is wrong with the reasoning of A and B.

John Horton Conway of the University of Cambridge has given this game a confusing and deep generalization. It is easy to generalize to n men and n consecutive integers, but Conway does away with the consecutive proviso. We allow any non-negative integer (including 0) to be placed on the forehead of n men, all honest and rational. On a blackboard, which all can see, are chalked m different non-negative integers, just one of which is the sum of the numbers on the foreheads. Each man sees all the foreheads except his own. The umpire asks each man in turn, "Can you deduce the number on your head?" The questioning continues in cyclic order until a player concludes the game by saying yes.

Conway has proved the following re-

Why more people are turning to Maintenance Leasing

at Ford Authorized Leasing System dealers.

It is economical because **F**ord **A**uthorized **L**easing **S**ystem dealers are specializing in a new maintenance lease program that can possibly offer you savings.

Volume counts. As a group, **F**ord **A**uthorized **L**easing **S**ystem dealers do more leasing business than any other organization, therefore, they are well qualified to advise you on individual or fleet leasing.

It's flexible. You can choose from two different maintenance leasing plans. Pick the 24,000-mile/24-month plan or the 40,000-mile/36-month plan.

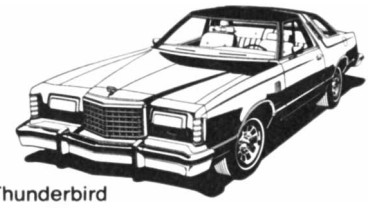
Service is available nationwide at more than 6,200 Ford Motor Company Dealers. It gives you peace of mind to know that service is always close by.

FALS
FORD AUTHORIZED LEASING SYSTEM

You can use coupons, not cash, for service. You can make a one-time payment for the convenient coupon book or have the cost prorated in your monthly rate. Insurance can be included, too.

Above all, FALS Maintenance Leasing is convenient. Your car is cared for by Ford-trained professionals using the latest diagnostic equipment and tools. And they have genuine Ford parts in stock for fast service. Remember, too, your **F**ord **A**uthorized **L**easing **S**ystem dealer offers the kind of plan you need—open-end, closed-end, cars or trucks, individual or fleet. And if low cost is important to you, he's the man you want to see.

Ford LTD



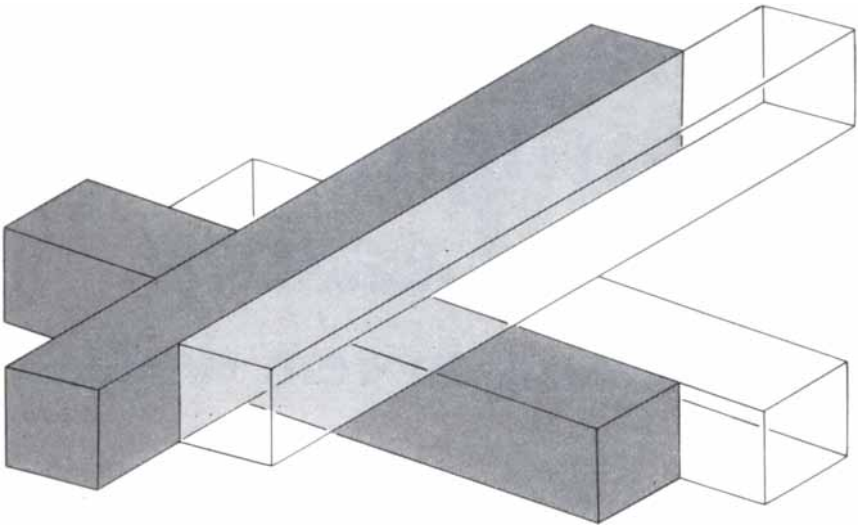
Thunderbird

Granada

FORD

FORD DIVISION

Consult your Yellow Pages for your nearest Ford Authorized Leasing System dealer.



Solution of the polycube problem

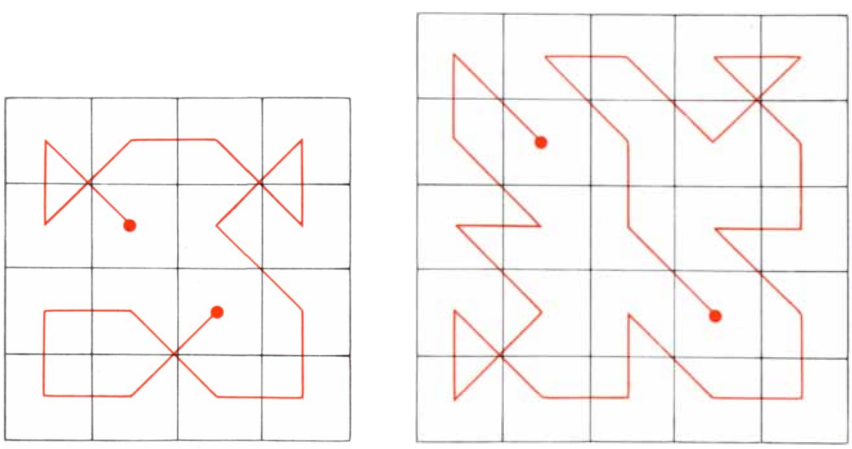
markable theorem. If m, the number of sums on the blackboard, is not greater than n, the game must terminate. For example, suppose each man has a 2 and the blackboard sums are 6, 7 and 8. Conway asserts that the game ends with a yes to the 14th question.

Formulating a general algorithm for calculating when such games will terminate is, except for certain sets of numbers, extremely difficult and far from solved. Conway writes: "One gets into an infinite regress of the form '*A* knows that *B* knows that *C* knows that *B* knows that *C* knows...' even before the first question is asked, so it is very difficult to get a measure of the information available to each player. Indeed, I felt at one time that these considerations might make the game not well defined, and that we'd get into paradox trouble. Now I don't think so. I do know it's fatally easy to make mistakes in assessing what information is available."

Conway's game presents the same paradox we have considered before. In the given example it is easy to show that before the game starts each man can predict that the first three answers will be no, and so it seems as if asking these questions can be dispensed with because after the first round the men will be no better informed than before. If the first round were eliminated, however, the same argument would apply to the next round, and the game would never end. Next month I shall explain why these paradoxes are not genuine difficulties.

Because mathematical induction often takes the form of "reducing to the preceding case," I close with an old joke. For a college freshman who cannot decide between physics and mathematics as his major subject the following two-part test has been devised. In the first part the student is taken to a room that contains a sink, a small stove with one unlighted burner and an empty kettle on the floor. The problem is to boil water. The student passes this part of the test if he fills the kettle at the sink, lights the burner and puts the kettle on the flame.

For part two the same student is taken to the same room, but now the kettle is filled and on the unlighted burner. Again the problem is to boil water. The potential physicist simply lights the burner. The potential mathematician first empties the kettle and puts it on the floor. This reduces the problem to the preceding case, which he has already solved.



Solutions of lost-king tours

Last month's problems are answered as follows:

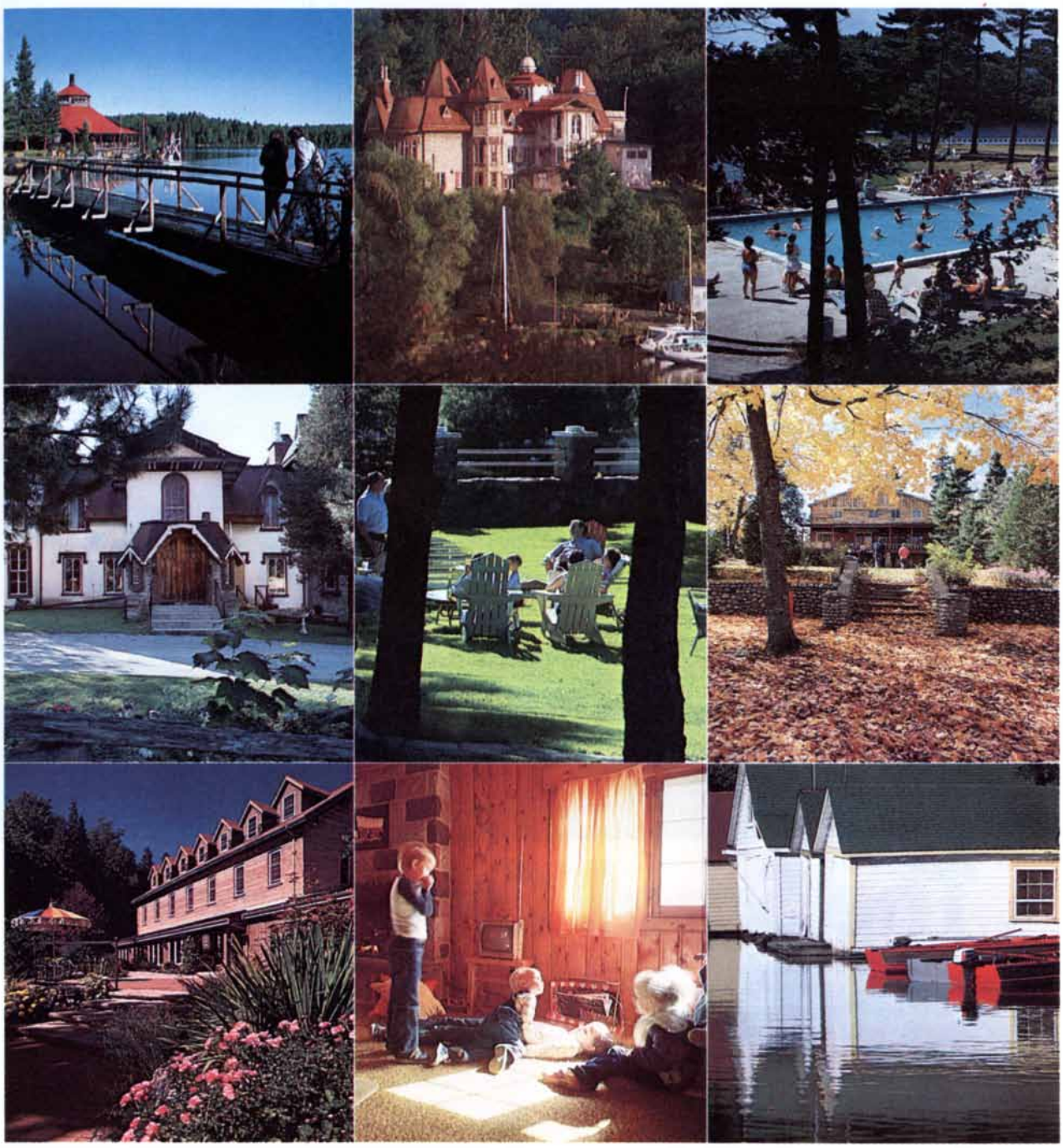
1. Except for reflection, the only solution of the problem with the 15 pool balls is shown in the bottom illustration on page 131. Col. George Sicherman of the State University of New York at Buffalo, who invented this problem, has found by computer that there are no solutions for similar triangles of orders 6, 7 and 8. Sicherman also found a simple parity proof of impossibility for all orders $2^n - 2$, where n is greater than 2.

This is how the proof goes for order 6, the smallest case. Call the triangle's top row a,b,c,d,e,f. Since addition is the same as subtraction (modulo 2) we can express the other numbers (modulo 2) by adding. The second row is $a + b$, $b + c$, $c + d$, $d + e$, $e + f$. The next row begins $a + 2b + c$, $b + 2c + d$.... Continue this way to the bottom number, which is $a + 5b + 10c + 10d + 5e + f$. The triangle contains 6 *a*'s, 20 *b*'s, 34 *c*'s, 34 *d*'s, 20 *e*'s and 6 *f*'s. All the numbers are even, therefore the triangle's parity is even. The triangle contains 11 odd and 10 even numbers, however, giving it an odd parity, so that we have a contradiction.

The above row of figures (6, 20, 34, 34, 20, 6) is the same as the seventh row of Pascal's triangle (1, 7, 21, 35, 35, 21, 7, 1) when the numbers are diminished by 1. Sicherman's general proof hinges on the well-known theorem that only rows numbered $2^n - 1$ of Pascal's triangle consist entirely of odd numbers.

Charles W. Trigg has proved, among other things, that every absolute-difference triangle of consecutive numbers must have 1 as its lowest number. He conjectures that there are no such triangles other than the 11 given last month and here. As a joke readers were asked to construct a difference triangle with 15 balls bearing the even numbers 2 through 30. The unique pattern is obtained at once by doubling each number in the solution shown.

2. One torus can be inside another in two topologically distinct ways: the in-



Still Moments

Absolute tranquility – pausing on a woodland path to listen to the myriad sweet sounds of the forest, paddling over motionless mirrored blue skies, or simply lounging, soaking up the sun, and if you tire of that, the chance to sail, hike, golf, ride, play tennis, shuffleboard, enjoy shore-lunches or barbecues, and, come nightfall, make new friends in the cozy company of a roaring fire, or be one with nature outside, beneath the Borealis, the silence broken only by the loon's lonely cry – still moments, absolute tranquility, the North begins again. For more about resorts and northern adventure vacations, call COLLECT (416) 965-4008, or write:
Ontario Travel, Dept. R., Queen's Park, Toronto M7A 2E5. We'll send you our Traveller's Encyclopaedia.

Ontario Canada

The Small Computer

Twenty-five years ago a computer as powerful as the new Processor Technology Sol-20 priced out at a cool million.

Now for only $995 in kit form or $1495 fully assembled and tested you can have your own small computer with perhaps even more power. It comes in a package about the size of a typewriter. And there's nothing like it on the market today. Not from IBM, Burroughs, DEC, HP or anybody else!

It fills a new role

If you're an engineer, scientist or businessman, the Sol-20 can help you solve many or all of your design problems, help you quantify research, and handle the books too. For not much more than the price of a good calculator, you can have high level computer power.

Use it in the office, lab, plant or home

Sol-20 is a smart terminal for distributed processing. Sol-20 is a stand alone computer for data collection, handling and analysis. Sol-20 is a text editor. In fact, Sol-20 is the key element of a full fledged computer system including hardware, software and peripheral gear. It's a computer system with a keyboard, extra memory, I/O interfaces, factory backup, service notes, users group.

It's a computer you can take home after hours to play or create sophisticated games, do your personal books and taxes, and a whole host of other tasks.

Those of you who are familiar with small computers will recognize what an advance the Sol-20 is.

Sol-20 offers all these features as standard:

8080 microprocessor — 1024 character video displa circuitry — control PROM memory — 1024 words of static lo power RAM — 1024 words of preprogrammed PROM — built cassette interface capable of controlling two recorders at 1200 bits per second — both parallel and serial standardized interfa connectors — a complete power supply including ultra quiet fan — a beautiful case with solid walnut sides — software wh includes a preprogrammed PROM personality module and a d cassette with BASIC-5 language plus two sophisticated compu video games — the ability to work with all S-100 bus product

Full expansion capability

Tailor the Sol-20 system to your applications with c complete line of peripheral products. These include the video monitor, audio cassette and digital tape systems, dual floppy disc system, expansion memories, and interfaces.

Write for our new 22 page catalog. Get all the details.

Processor Technology, Box N, 6200 Hollis St., Emeryville, CA 94608. (415) 652-8080.

Processor Technology Corporation

side torus may surround the hole of the outside torus or it may not.

If two toruses are linked and one has a "mouth," it cannot swallow the other so that the eaten torus is inside in the second sense. This result can be proved by drawing a closed curve on each torus in such a way that the two curves are linked in a simple manner. No amount of deformation can unlink the two curves. If one torus could swallow the other in the manner described, however, it could disgorge the eaten torus through its mouth and the two toruses would be unlinked. This result would also unlink the two closed curves. Because unlinking is impossible, cannibalism of this kind is also impossible.

The torus with the mouth can, however, swallow the other one so that the eaten torus is inside in the first sense explained above. The illustration on page 132 shows how it is done. In the process it is necessary for the cannibal torus to turn inside out.

A good way to understand what happens is to imagine that torus *A* is shrunk until it becomes a stripe of paint that circles *B*. Turn *A* inside out through its mouth. The painted stripe goes inside, but in doing so it ends up circling *B*'s hole. Expand the stripe back to a torus and you have the final picture of the sequence.

3. The top illustration on page 134 shows how two congruent polycubes (one shaded and one transparent) can be fitted together. By extension of the ends any finite number of such pieces can be nested in this manner so that every pair "touches" in the sense that they share a common surface and there are no interior holes. Extended to infinity, an infinite number of congruent polycubes, of order infinity, can mutually "touch."

If we drop the requirement of congruency but add the requirement of convexity, it has been known since about 1900 that an infinite number of noncongruent convex solids can mutually "touch." It is not known whether an infinite number of congruent convex solids can mutually touch, but Scott Kim has recently shown (although not published) how this arrangement can be achieved with an arbitrarily large finite number of such solids.

4. Raymond Smullyan's four logic problems are answered as follows:

(1) *A* is either telling the truth or not. Suppose he is. Then *B* is a knight and telling the truth when he says *A* is not a knight. In this case *A* is telling the truth but is not a knight.

Suppose *A* is lying. Then *B* is not a knight. *B*, however, is telling the truth when he says *A* is not a knight. Hence in this case *B* is telling the truth but is not a knight.

(2) *B* is either telling the truth or not. Suppose he is. Then *A* is a knave and must be lying when he says *B* is a knight. In this case *B* is telling the truth but is not a knight.

Suppose *B* is lying. Then *B* is surely not a knight; therefore *A* must be lying when he says *B* is a knight. Since *B* is lying, *A* is not a knave. In this case *A* is lying but is not a knave.

(3) *B* is either a knight or a knave. Suppose he is a knight. *A* and *C* must then be the same type, as *B* says. *C* is lying when he says *B* is a knave; therefore *C* is a knave. If *C* is a knave, *A* must be also.

Suppose *B* is a knave. Then *A* and *C* are different. *C* is telling the truth when he says *B* is a knave, so that *C* must be a knight. Because *A* and *C* are different *A* must be a knave. In either case *A* is a knave.

(4) Smullyan's solution of this problem is somewhat lengthy, and I shall content myself with a summary. *A* and *B* are either knight-knight, knave-knave, knight-knave or knave-knight. In each case analysis shows that *C*, whether he is a knight or a knave, must answer yes.

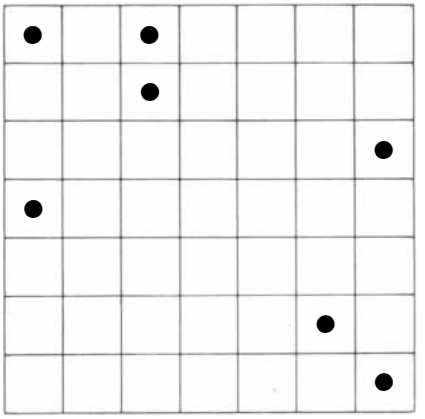
5. Solutions to the two lost-king tours are shown in the bottom illustration on page 134. The first tour is unique. It is not known whether the second one is unique or whether there is a solution with one less crossing.

6. The ellipse problem was given for a 3, 4, 5 triangle, but we shall solve it for any triangle.

The largest ellipse that can be inscribed in an equilateral triangle is a circle, and the smallest ellipse that can be circumscribed around an equilateral triangle is also a circle. By parallel projection we can transform an equilateral triangle into a triangle of any shape. When that is done, the inscribed and circumscribed circles become noncircular ellipses.

Parallel projection does not alter the ratios of the areas of the triangle and the two closed curves; therefore the ellipses that result will have the maximum and minimum areas for any triangle produced by the projection. In other words, the ratio of the area of the smallest ellipse that can be circumscribed around any triangle to the area of the triangle is the same as the ratio of a circle to an inscribed equilateral triangle. Similarly, the ratio of the area of the largest ellipse that can be inscribed in any triangle to the area of the triangle is the same as the ratio of a circle to a circumscribed equilateral triangle.

It is easy to show that the ratio of the inside circle to the triangle is $\pi/3\sqrt{3}$ and that the ratio of the outside circle to the triangle is four times that number. Applying this result to the 3, 4, 5 triangle, the area of the largest inside ellipse is $2\pi/\sqrt{3}$, and the area of the smallest outside ellipse is $8\pi/\sqrt{3}$.



Solution of the point-placement problem

Readers who would like to have a more formal proof will find it in Heinrich Dörrie's *100 Great Problems of Elementary Mathematics* (Dover, 1965), page 378 ff.

7. The illustration above shows the only way (not counting rotations and reflections) to place seven counters on an order-7 matrix so that all distances between pairs of counters are different.

8. The paradox of the fourth limerick arises when the limerick is completed in one's mind: "Whose limericks stopped at line one." To complete it is to contradict what the limerick is asserting. The four limericks presented last month prompted J. A. Lindon, the British comic versifier, to improvise the following new ones:

A most inept poet of Wendham
Wrote limericks (none would defend 'em).
"I get going," he said,
"Have ideas in my head,
Then find I just simply can't."

That things were not worse was a mercy!
You read bottom line first
Since he wrote all reversed—
He did every job arsy-versy.
A very odd poet was Percy!

Found it rather a job to impart 'em.
When asked at the time,
"Why is this? Don't they rhyme?"
Said the poet of Chartham, "Can't start 'em."

So quick a verse writer was Tuplett,
That his limerick turned out a couplet.

A three-lines-a-center was Purcett,
So when *he* penned a limerick (curse it!)
The blessed thing came out a tercet!

Absentminded, the late poet Moore,
Jaywalking, at work on line four,
Was killed by a truck.

So Clive scribbled only line five.

YOUR KIND OF BOOKS
Examine Them Free for 15 Days

TEN FACES OF THE UNIVERSE

Sir Fred Hoyle

An exciting look at the frontiers of our present knowledge of ourselves, our planet, and the universe. In the course of the book, Hoyle gives readers an autobiographical sketch of his education and development as a scientist; he presents details of his latest cosmological theory, based on a radically different interpretation of astronomical data and he poses the seeming paradox that unless civilization collapses soon, humankind may be doomed.

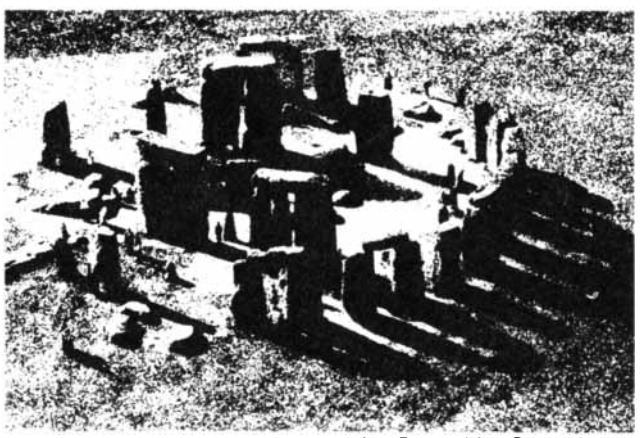
1977, 207 pp., 91 illus., softcover, $6.95

ON STONEHENGE

Sir Fred Hoyle

Hoyle presents the evidence for Stonehenge as an astronomical calculator and shows how paleolithic astronomers may have made their calculations. Along the way, he offers readers a generous helping of provocative speculation on such topics as the origins of the concept of an invisible, all-powerful god and the "unlucky" number thirteen.

1977, 160 pp., 50 illus., softcover, $6.50



Aero Films and Aero Pictorial Limited

MANAGING THE COMMONS

Edited by Garrett Hardin and John Baden

An anthology of readings that explores the implications of Garrett Hardin's modern classic of environmental literature, "The Tragedy of the Commons" for modern society. The readings trace the development of the concept of the commons and focus on ways in which the potentially destructive cultural norm of independence of individual action may be changed to promote continued human welfare and survival. The collection includes historical and recent essays, original material by the editors, and a new essay by Kenneth Boulding.

1977, 300 pp., 24 illus., softcover, $5.95

ECOLOGY AND THE POLITICS OF SCARCITY

William Ophuls

This book is a consistent political argument that modern civilization—the product of both the Enlightenment and the Industrial Revolution—has outlived its usefulness. The human and moral alternative to our present course, Ophuls demonstrates, is the creation of small-scale, simple, frugal, steady state societies.

1977, 303 pp., 12 illus., softcover, $6.95

SCIENTIFIC AMERICAN

CONTINENTS ADRIFT AND CONTINENTS AGROUND

With Introductions by J. Tuzo Wilson

The idea that the surface of the earth is broken into half a dozen enormous plates that are moving in relation to each other has revolutionized the way geologists view their subject. This book brings together a series of articles by many of the geologists who helped to bring about this conceptual revolution. J. Tuzo Wilson's introductions put the articles into perspective and provide historical background.

1977, 230 pp., 178 illus. (including many full-color plates), softcover, $5.95

NEW FRONTIERS IN ASTRONOMY

With Introductions by Owen Gingerich

Interstellar molecules. X-Ray-emitting double stars. Pulsars. Black holes. These are just a few of the astonishing new discoveries in astronomy that are described by some of the most distinguished astronomers of our time. All of them bear the stamp of authority that only practicing astronomers can give them.

1975, 369 pp., 329 illus. (151 with color), softcover, $7.50

Alan M. Eshleman

COMPUTER POWER AND HUMAN REASON

From Judgment to Calculation

Joseph Weizenbaum

Now available in softcover is this distinguished computer scientist's searching examination of the impact of scientific rationality—and of the computer in particular—on man's self-image.

"A magnificent book—the first one which really elucidates the fundamental questions of 'computers' and 'language.' "—Roman Jakobson

1977, 300 pp., 13 illus., softcover, $5.95 (Also available in hardcover)

THE THINKING COMPUTER

Mind Inside Matter

Bertram Raphael

With a minimum of technical jargon, this book discusses the capabilities of modern digital computers and how they are being used in contemporary artificial-intelligence research. It discusses the progress of AI, the goals of AI research, and the variety of current approaches to making the computer more intelligent.

1976, 322 pp., 123 illus., softcover, $6.95

HOW TO SOLVE PROBLEMS

Wayne A. Wickelgren

This book analyzes and systematizes the basic methods of solving mathematical problems. The methods are described in terms of a modern theory derived from research in computer simulation of thinking. Examples illustrating these methods include chess problems, logical puzzles, and railroad switching problems as well as problems frequently encountered in science and engineering.

1974, 262 pp., 73 illus., softcover, $6.50

THE NATURE OF HUMAN CONSCIOUSNESS

Edited by Robert Ornstein

In this monumental volume, Dr. Ornstein brings together the classic essays on consciousness and the latest reports from current scientific research. The result is a fascinating tableau of insights into human mental phenomena.

1973, 514 pp., 48 illus., 7 tables, softcover, $6.50

A PRIMER OF INFANT DEVELOPMENT

T. G. R. Bower

A vivid, authoritative description of the potential for and limits to development in the normal infant. Fascinating reading for parents, prospective parents, and indeed all those concerned with infants.

1977, 181 pp. plus index, 59 illus., softcover, $4.95

THE BIOLOGY OF ADOLESCENCE

Herant Katchadourian, M.D.

Dr. Katchadourian emphasizes sexual maturation and the endocrine processes underlying these events, but he also deals with disturbances of puberty and common physical ailments such as acne, venereal diseases, and disturbances of eating. The chapter on health hazards of the second decade of life, predominantly accidents and the use of drugs (including tobacco and alcohol), will be of special interest to adolescents and their parents.

1977, 264 pp., 120 illus., softcover, $6.95

A PRIMER OF DRUG ACTION

Robert M. Julien

Written for today's students and their parents, this book presents a medically accurate discussion of the major classes of psychoactive drugs. It is a complete sourcebook on the actions, uses, and limitations of such drugs as the sedatives, tranquilizers, stimulants, opiates, and psychedelics.

1975, 290 pp., 52 illus., 11 tables, softcover, $6.00

W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, CA 94104
58 Kings Road, Reading, England RG1 3AA

How fast would new medical discoveries travel if there were no airlines? Who knows. But we do know that today you can be almost anywhere in the U.S. within a day. And nobody knows more about the need for speed, convenience and economy in air travel than Boeing—builders of the flagships of the world.

BOEING

BOOKS

The disarmament game, slow viruses, whistled languages, carnivorous plants and Dr. Matrix

by Philip Morrison

THE GAME OF DISARMAMENT: HOW THE UNITED STATES AND RUSSIA RUN THE ARMS RACE, by Alva Myrdal. Pantheon Books ($15). Alva Myrdal writes in "near despair," with some anger and with a residue of undaunted hope. Bringing to statecraft a reputation for high and creative scholarship, she has served Sweden as minister for disarmament, as a diplomat abroad and as a tenacious and knowing leader amid the intricacies of United Nations conferences and committees over 16 years. Her verdict is clear and cold: "Those who have power have no will to disarm." The superpowers pile on the yield and the options, and the other countries share responsibility in their lesser orbits. Even the concern of the people is stilled, "as if we have all been conditioned to go on living with the threatened dangers ... as we live with the inevitable personal catastrophe of our own death.... The difference is that ... a collective nuclear suicide can be prevented."

The book is largely a kind of analytic diplomatic chronicle that displays "a history of lost opportunities," from the high-water mark of the Partial Test-Ban Treaty of 1963 to the ebbing hopes of a dozen other treaties, conferences and proposals. The acquisition of weapons continues, almost an end in itself. The two superpowers seem to be motivated by the simple, unreasonable and unattainable hope that "each must be second to none." The partial test ban, "can hardly be considered among disarmament measures." It was useful primarily for public health; indeed, the worldwide anxiety over fallout was all too easily met by placing the tests underground, which allowed the free qualitative development of nuclear weapons—an "ostrich-like solution."

The statesmen of Europe see plainly what to expect from the superpower tactical nuclear competition that is so well advanced in that theater: a scenario in which *their* countries become the sites of the nuclear explosions while the homelands of the superpowers remain sanctuaries against the mushroom cloud. Yet they choose to "buy current political stability by placing ... awful risks ... over their lives and their future!" Power reactors are now the targets of determined political movements in the main countries of Western Europe, but tactical nuclear weapons, demonstrably orders of magnitude more threatening, are outside current debate. "There is definitely no interest in public discussion" of their implications or the devastation they entail.

Event by event the story is the same. For Mrs. Myrdal this is simply unreason compounded. No doubt she is right in the last analysis, but three decades of experience need a richer explanation. It begins to appear that the leaders, made confident by that experience of their control over the weapons, see no need to forgo the political gains the race offers at every level, from internal service rivalries to the unstated sense of domination. In his final report earlier this year President Ford's Secretary of Defense Donald Rumsfeld even wrote: "However much one might wish otherwise, popular and even some governmental perceptions of the strategic nuclear balance tend to be influenced less by detailed analyses than by such static indicators of relative nuclear strengths as launchers, warheads, megatonnage, accuracy, throw-weight and the like." There the game is described; elsewhere perhaps a voice whispers that just maybe the newest guidance and a set of cunning strategies might make it possible to bring nuclear weapons to bear retail, to win smaller points, short of deterrence and overall war. This is no crude unreason but what we have long known as crackpot realism (and even longer as hubris). Mrs. Myrdal plainly implies that we are now at a transition time; that view is forming particularly in the U.S., the unquestioned leader in the race.

An engaging brief personal introduction—most readers would welcome its expansion—gives us support for the wry tone of the title. The author bears eyewitness to three events in which the superpowers stymied progress. In the first, after the near-conversion of the rivals following the Cuban missile crisis, the nonaligned delegates at the Geneva disarmament talks worked out a "quite ingenious scheme" for bridging the gap over the number of on-site inspections, which had long barred a test-ban treaty. But both big powers used diplomatic pressure in the capitals of the delegates, intimating that the delegates' initiative was jeopardizing important direct negotiations. The result was the partial test ban. Again, in 1969, the U.S. delegation at a "luxurious luncheon" blamed Mrs. Myrdal for U.S. inability to ratify the Geneva Protocol of 1925 banning chemical and biological warfare. "The pressure was enormous"; the U.S. was then using herbicides in Vietnam. Mrs. Myrdal stuck to her position, insisting that nonratification was preferable to a ratification with reservations by the U.S. Fortunately the proposal won. The third example showed "adamant negativism" on the part of the U.S.S.R. in a matter of the verification of prohibited biological weapons.

Of course, nuclear weapons are not the whole story. Guns, missiles, planes, ships and tanks are the growth industry of the world. Supersonic military jets, for example, are now flying in the colors of 39 countries of the underdeveloped world. We cannot overlook the rise of cruelty in weapons, from area bombing to gases, toxins, napalm and the light, high-velocity bullets that fragment, tumble and initiate shock fronts in the flesh they strike. The struggles are many, the victories few.

Yet hope remains. Passivity may reach an end. A world disarmament conference could come into being if the superpowers would give it life by two pledges: no first use of nuclear weapons and no nuclear attack on nuclear-weapons-free countries. The Chinese have set these as preconditions for their participation; Mrs. Myrdal too sees them as "pillars of confidence" for the gateway to any conference. Such a conference has many tasks; she lists eight, from steps toward a minimum deterrent to the elimination of foreign bases. Even the lesser states have agreements of their own to attempt.

In the end it will be we Americans who must act—and not least the technical people here, because American technology has been the chief armorer of the race, its swiftest runner. The Russian people by experience "undoubtedly" abhor war more than the American people, Mrs. Myrdal writes. But in the U.S. we can listen, we speak out, we are moved by argument; "I do not expect the book to be translated into Russian." (Yet there are those even there who will weigh its arguments.) And of America, "I know of no nation that is more capable of switching policy line." Perhaps the time is coming, as come it must for human survival, when our leaders will find enough courage and public trust to

begin in earnest the long walk back from the crevasse at our feet. Present danger is clear enough, and it does not arise far from the banks of the Potomac.

SLOW VIRUSES, by David H. Adams and Thomas M. Bell. Addison-Wesley Publishing Company, Inc. ($7.50). SLOW VIRUS DISEASES OF ANIMALS AND MAN, edited by R. H. Kimberlin. American Elsevier Publishing Company, Inc. ($47.95). Inject a mouse with a dose of Riley virus, a tumor-derived agent; after a six-hour latency the virus concentration will grow tenfold each hour for half a day. Inject instead an inoculum of the agent of the sheep disease scrapie and monitor the concentration, which is measurable only by dilution and injection into other mice; you will see nothing for three or four months, and then at last a tenfold rise in just a couple of weeks. Scrapie is the prototype of a slow-virus disease: slow to appear, slow to grow and slow to exhibit its specific and generally fatal effects.

These two books—one a personal, boldly generalizing review by two British virologists whose main interest is human disease and the other a collection of detailed review papers by 20 international contributors, most of them from animal-disease laboratories—are distinct, even complementary. They do not agree on definitions or scope; they represent overview as against survey in depth, but they mutually draw the general reader into this world of puzzling disease, where virology, immunology and pathology link at the margin of our knowledge. Adams and Bell offer easy reading, but their expert allusive account gains concreteness from the rich detail of the bigger book, even though that volume in its turn goes into detail the general reader cannot follow.

Diseases of sheep, mink and man are the chief subjects. A dozen fatal diseases of the human central nervous system stand suspect of slow-virus origin; four, perhaps, are pretty sure. The best understood of all the conditions described is in fact rejected from the class by Adams and Bell. It is the Aleutian disease of mink, so named because it is in time invariably fatal to a mutant strain of mink with a fashionable light-hued fur resembling the pelage of the Aleutian blue fox. At first the condition seemed hereditary, linked to the mutation, but it was learned that normal mink catch the infectious disease as well, although with lower mortality. The virus grows rapidly enough, but the animal dies only after protracted kidney stress, the consequence of uncontrolled production of an antibody that is useless against the virus. Only the slow-growing, nonregenerating cells of the mammalian central nervous system are the proper hosts to a slow virus in the narrow sense.

The most difficult puzzles today, the most incisive studies, center on the sheep (and goat) disease long called scrapie. Even at the time of George III scrapie was "most calamitous" for the English sheep raiser. It is a fatal central-nervous-system infection that takes its victim at three or four years of age, having made its invisible entry at or near birth. (Sheep husbandry limits the animal's average lifetime to about five years.) The name reflects the sick animal's habit of rubbing incessantly against posts or fencing until raw wounds develop over large areas of its body. In weeks or months after the first symptoms appear the animal dies; losing 10 percent of the adult sheep to scrapie is not uncommon. In 1939 scrapie broke out among sheep that a couple of years before had been vaccinated, with formaldehyde-inactivated sheep brain, against the disease known as louping ill. At once scrapie's status as an infectious disease was proved and the remarkable resistance of the agent to chemical attack, as indeed to radiation and even to heat, was first indicated.

In 1961 scrapie was first passed to mice, which invariably die of central-nervous-system lesions. The disease is slow in mice too; it takes eight months or so to diagnose mouse scrapie and thus complete the assay. Such patient work has been remarkably rewarding; the results even raise the possibility that "the central dogma" of molecular biology might founder on the unknown agent of this disagreeable disease of sheep. No nucleic acid was found in it. No virus polyhedrons have been seen under the electron microscope. Filters remove the infectious agent at a pore size that is pretty big, and yet the large dose of radiation required for inactivation puts the size of the infective target near the lower limit for viruses. Inactivation by ultraviolet radiation showed a wavelength dependence quite unlike the absorption spectrum of DNA. Perhaps, Adams once said, it is a "sugar-coated virus": a DNA tape with a polysaccharide cassette (instead of the familiar protein one), which can more tightly bind and better protect the information. Nowadays it appears that host-cell membrane structures play the protective role. Removing all brain-cell components except membrane (and recently an effort at chemical removal of membrane components as well) increases by a hundredfold the concentration of scrapie agent. There remains a long way to go to reveal a tiny DNA package, but there seems to be a good chance that the dogma will survive. Small ultraviolet-resistant plant viruses have been found in the past couple of years; it may be that, after all, the controversy "was a pure artifact . . . a failure to recognize the extent to which nucleic acid can be desensitized" by closely associated molecules.

Scrapie agent introduced into mink

produces a disease indistinguishable from an epizootic first seen in ranch mink in 1947. This "transmissible mink encephalopathy" does not spread by contact, and it is not passed from mothers to kits. The suggestion is strong that it entered with the food; mink are fed mutton. Human beings too have a scrapielike disease, rare and slow to appear, that is found in patients with an inherited predisposition and that is always fatal by progressive mental deterioration. Libyan immigrants in Israel are found to suffer from the condition, Creutzfeldt-Jakob disease, at a rate 30 times normal. These people eat sheep brain. Among the Fore people of New Guinea the women and children of certain groups suffer from kuru, a similar disease that is associated with the extraordinary and rather newly adopted ritual of rubbing on the mourning female kin the brains of a man who has died. Once thought to be genetic because of the sex linkage, kuru too has proved infectious: it can be transferred to many laboratory primates.

Behind these tales of tragedy and mystery lies a challenge. There are much commoner fatal disorders, say multiple sclerosis, that might belong to the complex and subtle class of slow viruses. Add all of them up and it may be that one person in 100 now dies of such a stealthy attack, barely recognized but excruciating in the burden on patient and society over years of wasting and incapacity. There are enigmatic signs of a connection between an RNA virus such as the measles organism and certain DNA slow viruses, as though that transcription were of major clinical effect. Indeed, the same slow virus does cause two widespread diseases of sheep first seen in Iceland, one a wasting disease of the central nervous system called visna and the other a long-delayed chronic pneumonia.

Here is a challenge to molecular biology in its own domain, important and tantalizing, yet simpler by far than the developmental issues that underlie malignant diseases.

WHISTLED LANGUAGES, by R. G. Busnel and A. Classe. Springer-Verlag ($16.40). Twenty years have passed since Professor Classe first explained in this magazine the remarkable language Silbo, with which the herders and farmers of La Gomera in the Canary Islands proudly maintain an acoustic citizen's-band surrogate among the steep, dry, canyon-cut landscapes of their island. He and his colleague have since delved deep—they are fluent in Silbo—into this remarkable cultural achievement, known to outsiders since the 17th century but poorly understood. It turns out that Gomera is not unique: in the high-Pyrenees valley of Ossau in France a single village, Aas, developed a similar scheme, which has dwindled and all but vanished as the people left their hard mountain homes. In the highlands of northeastern Turkey close to the Black Sea coast a "whole population of highlanders," some 30,000 men, women and children, depend on whistling as we depend on the telephones of the city. The three schemes are remarkably similar—almost surely by convergence rather than by diffusion, since each is tightly based on its distinct language, French, Spanish or Turkish, and no contact whatever is suggested. (A fourth whistled language, of a rather different kind, is used among the Mazateco people north of Oaxaca in Mexico.)

This book is a thin monograph that compactly summarizes what is known and how it was learned. Sound spectra, landscape photographs, careful linguistic and phonetic analyses, measures of intensity over distance, experimental tests of critical points and hints (none too easy) on how to do it, with photographs of adepts—all are offered in a work that is hard for a general reader to put down even though it is intended for the specialist in linguistics. Why Silbo and its cognates? The answer seems clear: they function adaptively. The landscapes and the work life of the three places are similar. There are no roads and the paths are winding and difficult. Crops are cultivated high in the hills, which are rather barren and, in the Turkish case, terraced. An isolated population has for a long time been given cause to think about a message channel that would carry better, with less effort, than a shout. Once built, the language becomes—as all human social accomplishments become—a source of pride and a mark of community, in these cases with the advantage of concealment from outsiders. (In the Spanish Civil War military signals were on occasion cast in Silbo, but there were Gomerans on both sides and so the measure-countermeasure drama soon ran its course.)

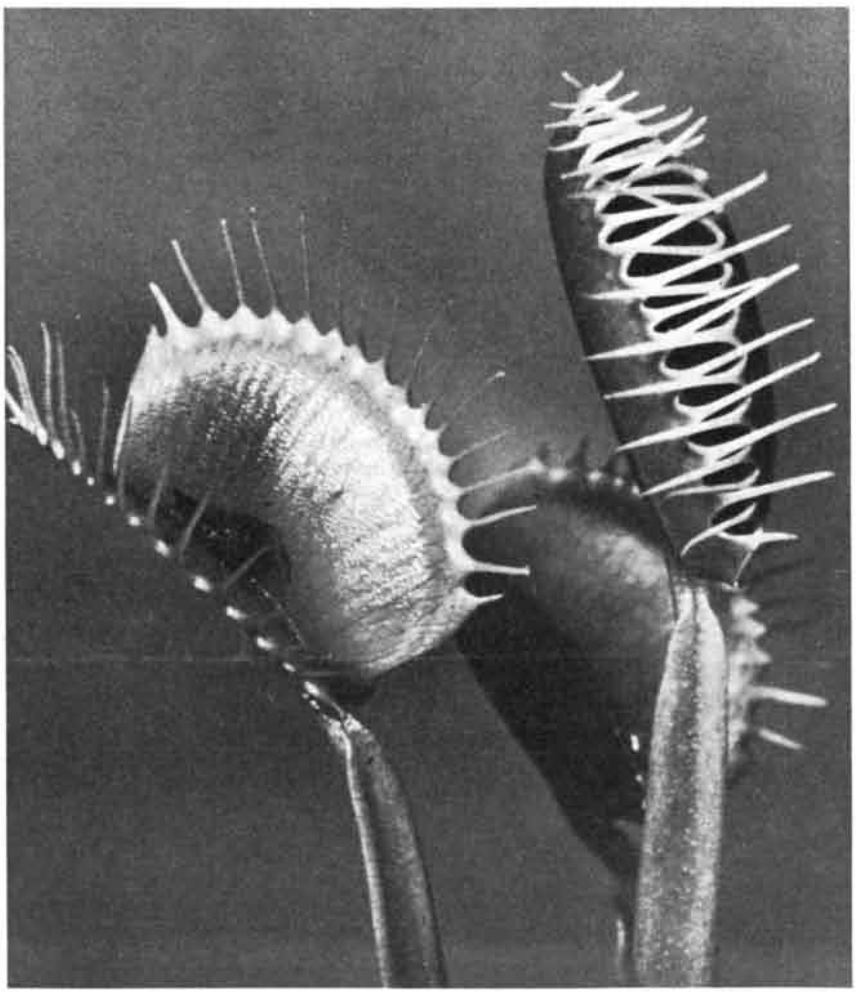
Does it work? Here are the data, both in general and from direct measurement of the propagation of sound in natural conditions. The whistlers are well aware of acoustic shadows, both topographic and thermal in origin. The noise is low; except when the trade wind is blowing La Gomera yields "the impression of profound quiet." (From time to time blackbirds sound false signals, imitating human calls!) A line of sight helps; vegetation strongly damps the signal. In wooded open valleys the useful range is one or two kilometers, but in narrow rocky valleys from eight to 10 kilometers can be reached. The whistlers do not use their lips, but with tongue on teeth or with one or two fingers in their mouth they generate rather pure piercing tones, some 120 decibels at a distance of a meter—as loud as a trumpet.

The information is placed on this car-

rier by frequency modulation as well as by all the start-stop amplitude rhythms of speech. Consonants and stops are coded into the tone by natural patterns of quick pitch change, quite parallel to the modulation of the ordinary voice during consonants, since the motions of the mouth are the source of both. Anyone can try simply whistling out an articulated phrase, particularly working with an open language such as Spanish. It is then convincing to read the study of the whistle speech in detail. It is an almost complete phonetic recoding of speech; users can recognize and repeat words from a language they do not know if they are whistled properly. Single words can be recognized, according to careful direct tests, with an error rate of some 30 percent, compared with only 4 percent for normal speech; it is the redundancy in longer messages that makes the channel fully workable.

The Mexican case is different. Children and men whistle but not grown women; this is lip whistling, much less loud and not practiced over distances greater than a few hundred meters; it is used at close quarters (say in the marketplace), unlike the other three languages. Among these people group work is more usual and there are no distant herds or flocks. The coding is quite different: their spoken language is tonal, and in the whistle language only the melodic line of speech is presented. The sentence is thus merely outlined, transposed to a high lip whistle modulated in a narrow frequency range, about half an octave around 2,000 hertz. (In Silbo the frequency range is as much as two octaves, although most of the time the carrier stays at about the same mean, which is the region of best signal detection over natural noise.) The Mazateco cannot recognize any words that lack the tonal structure of an existing word in their own language; theirs is not a full phonetic mapping. Indeed, they are reported as substituting the same whistled form for such prosodically similar names as Modesto and Gustavo.

A final chapter touches on animal signals, drum languages and the like. Maybe porpoises could learn Silbo; like canaries, they can mimic whistles. Whistle languages are almost as fast as ordinary speech and much more efficient than the



Venus's-flytraps, open and closed, from **Carnivorous Plants of the United States and Canada**

very redundant talk of drums. Here is a folk technology of telecommunication, demonstrably effective and in no way occult, mystical or telepathic.

CARNIVOROUS PLANTS OF THE UNITED STATES AND CANADA, by Donald E. Schnell. John F. Blair, Publisher, Winston-Salem, N.C. ($19.95). THE CARNIVOROUS PLANTS, by Francis Ernest Lloyd. Dover Publications, Inc. ($4.50). Golden as any host of English daffodils, a spring stand of bold *Sarracenia flava* spreads across many a South Georgia bog or wet savanna. These plants delight mind no less than eye; each bears a conspicuous flowerlike modified leaf rosette become a pitcher, which lures, entraps and digests small insects. Once in a while a frog skeleton may be found in a pitcher plant, evidence of amphibian rashness. The pigment and pattern are attractive to insects; one related species with a more tightly curled hood bears translucent "windows" on the back of the upper portion of its pitcher further to deceive the prey. There are stiff hairs to discourage climbing out, a smooth waxy lining and, deep below, the enzyme-secreting glands that maintain a "complex little ecosystem" that may include even resistant larvae; one fly larva, armed with antienzymes, feeds on the take and pupates cheerfully in the depths of the pitcher.

In the savanna around Wilmington, N.C., there is a species of active carnivore, the famous Venus's-flytrap, that is "the pet among most students of carnivorous plants." Darwin thought it "the most wonderful plant in the world." The pitcher plants have a subtle trap, but a static one. The flytrap, on the other hand, has leaf blades modified into a structure resembling a bear trap: two clamshell halves, each with strong guard hairs and sweet nectar glands along the edge. The trap stands open, its halves dished in such a way that their inner surfaces bulge inward. Touch the special trigger hairs on the surface—three hairs normally—and the halves snap closed, the dishing reversing quickly so that the bulge becomes concave on the inside and the edges come together with the guard hairs loosely meshed "like the fingers of clasped hands." Two trigger hairs must be disturbed, or one of them twice, within a fraction of a minute (but not too quickly). Once closed, the trap slowly seals during some hours to hold fluid, the lobes flatten together and the plant dines at leisure. Tease it with a piece of paper, say, and the trap will reopen well in advance of the digestive schedule.

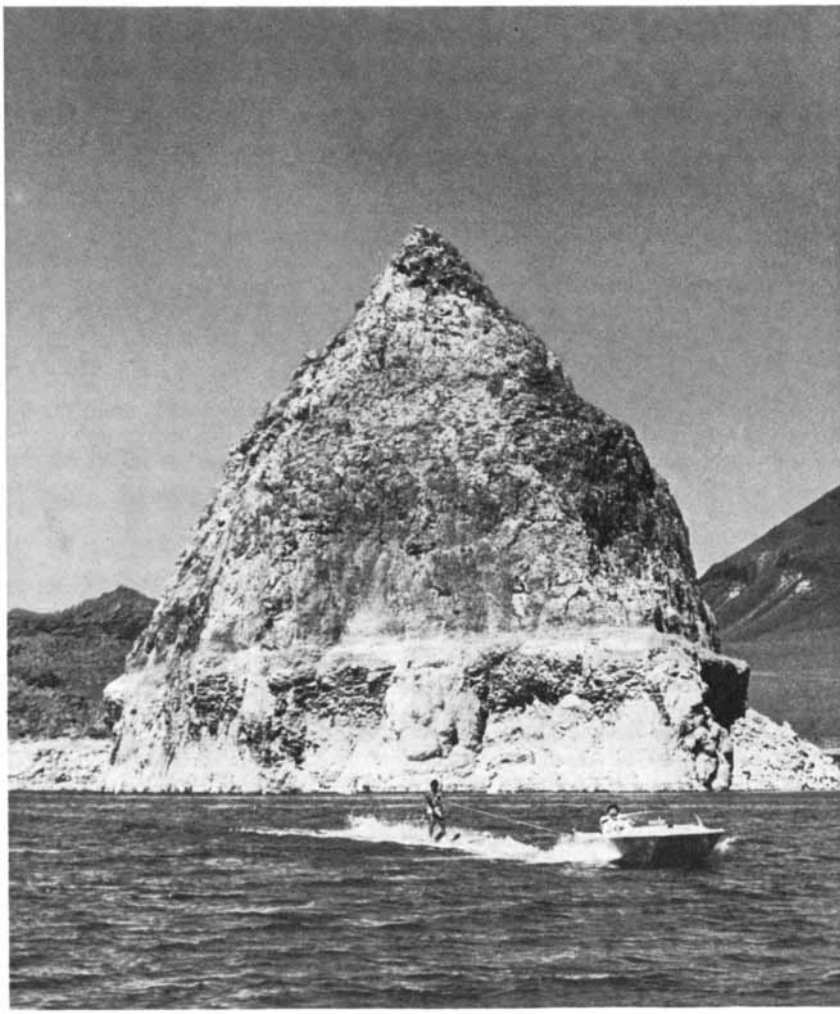
There are carnivores that simply function as flypaper. The butterworts are widespread over the northern part of the continent; they have no lures, only leaves that bear tiny glands, some of which secrete a gluey fluid and some

a digestive enzyme. Scandinavian folk biochemistry has long known that these plants help in the curding of milk. Then there are the sundews of our East Coast. Their leaves bear tiny stalks that secrete minute, sticky dewdrops of a gummy substance (which do not evaporate in sunshine and thus bestow the name). The largest species bears its droplets along the edge of erect leaves a foot or two high; a stand of enduring dewdrops glowing in the low morning sun is a remarkable sight.

The Schnell book is a model field guide to the carnivorous plants accessible to North American nature watchers. Maps, clear descriptions, literature references and beautiful photographs in color, both long shots and close-ups, support the web-footed author's report of his 19 years of fieldwork. He is a pathologist by profession and a naturalist by avocation. His work explains how and where to buy the plants and how they can be grown—even to creating an artificial bog. These are not house plants; they will not grow in a pot of "generalized dirt, treated tap water" and the temperature range of the living room. Do not be misled by advertisements into trying to grow them on the windowsill; the odds against success there are at least 1,000 to one. Never collect from the field. These plants are losing habitats to drainage, clearance flooding and eutrophication. The reputable primary dealers listed here propagate their own; of course, many secondary dealers handle plants from these nurseries.

The Lloyd book is an exact complement to the Schnell guide. It is a reprint of a hard-boiled classic of 1942, full of botanists' jargon and intricate line drawings. Its photographs are World War I gray, but its careful text, the fruit of a devoted lifetime of work by a well known Canadian botanist, discusses the state of knowledge of the carnivorous plants worldwide by taking up in order the experimental and anatomical findings of the botanists. Almost anyone who buys Schnell will soon buy Lloyd.

One must close with a good word for the widespread bladderworts. These plants are not showy; their traps are or



Pyramid Lake revisited, from **The Incredible Dr. Matrix**

the millimeter scale. "There is hardly a salubriously boggy place that does not support" one or more species of *Utricularia*. What a mechanism it has developed! Schnell shows loving photographs and Lloyd ends his treatise with a Heath Robinson kind of drawing of the mechanism turned into human hardware: double gates, motors, cams and the like. These traps are self-resetting underwater flooding chambers, kept at low pressure to sweep in the unhappy prey. For more detail you must seek the books, or better yet the bog.

THE INCREDIBLE DR. MATRIX: THE WORLD'S GREATEST NUMEROLOGIST, by Martin Gardner. Charles Scribner's Sons ($8.95). It would be redundant to identify Dr. Matrix in these pages, which each month carry the work of his friend and only chronicler, another ingenious master of words, forms and quantities. Moreover, the bulk of this book collects Dr. Matrix pieces that have appeared in *Scientific American* between 1960 and 1975. There is a certain mercantile off-flavor about noticing the well-known work of a friend and colleague, and so these book reviews have avoided mention of such volumes, highly readable though they may be.

Yet in the public good this policy is for the moment declared inoperative. In June, 1974, Dr. Matrix and his clever daughter Iva were represented in these pages as the proprietors of a pyramidal factory at Pyramid Lake, Nev., where they manufactured pyramid models of all sizes, prices and occult powers. Their doorman was a pudgy Paiute named Ree, whose grin showed only a single front tooth. ("Everyone called him One-Tooth Ree.") The final chapter of the book tells what happened after publication. The preposterous story was widely believed, Gardner reports. Hundreds wrote him for more data. Many drove to Pyramid Lake to look for the factory in vain, and one man upbraided Gardner, observing that no one at the Paiute reservation could so much as identify One-Tooth Ree. A Hawaiian reader invited Gardner to lecture on the new psi-org discoveries, all expenses paid.

"The most startling letter came from a prominent, well-known New York City publisher." The man was serious. He offered an advance of $15,000 for a quick book to be called *Pyramid Power,* which he felt he could promote into a "top best-seller." When the publisher learned that nothing Dr. Matrix had said was authentic, he was not at all put off. Use a pseudonym, he advised Gardner; even expose the book as a hoax after one year's sales! Such is the current state of the obsessive will to believe, and such is the cynicism of many who exploit it. These cults are resistant to exposure, and once the crystal ball reveals little dollar signs gleaming in the mist, let the believer beware.

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

UNDERGROUND RESERVOIRS TO CONTROL THE WATER CYCLE

PROVISIONAL INDICATIVE WORLD PLAN FOR AGRICULTURAL DEVELOPMENT: VOLS. I–II. Food and Agriculture Organization of the United Nations, 1970.

ICEBERGS AS A FRESH-WATER SOURCE: AN APPRAISAL. W. F. Weeks and W. J. Campbell in *Journal of Glaciology,* Vol. 12, No. 65, pages 207–233; 1973.

GROUND WATER IN PERSPECTIVE. Raymond L. Nace in *Water Resources Bulletin,* Vol. 9, No. 1, pages 18–24; February, 1973.

RAMAPITHECUS

NEWLY RECOGNIZED MANDIBLE OF *RAMAPITHECUS.* David Pilbeam in *Nature,* Vol. 222, No. 5198, pages 1093–1094; June 14, 1969.

RAMAPITHECUS in *Primate Evolution: An Introduction to Man's Place in Nature.* Elwyn L. Simons. The Macmillan Company, 1972.

THE PRIMATE AND OTHER FAUNA FROM FORT TERNAN, KENYA. Peter Andrews and Alan Walker in *Human Origins: Louis Leakey and the East African Evidence,* edited by Glynn Ll. Isaac and Elizabeth R. McCown. W. A. Benjamin, Inc., 1976.

AMORPHOUS-SEMICONDUCTOR DEVICES

REVERSIBLE ELECTRICAL SWITCHING PHENOMENA IN DISORDERED STRUCTURES. Stanford R. Ovshinsky in *Physical Review Letters,* Vol. 21, No. 20, pages 1450–1453; November 11, 1968.

ELECTRONIC PROCESSES IN NON-CRYSTALLINE MATERIALS. N. F. Mott and E. A. David. Oxford University Press, 1971.

AMORPHOUS SEMICONDUCTORS FOR SWITCHING, MEMORY, AND IMAGING APPLICATIONS. Stanford R. Ovshinsky and Hellmut Fritzsche in *IEEE Transactions on Electron Devices,* Vol. ED-20, No. 2, pages 91–105; February, 1973.

ON STATE OF AMORPHOUS THRESHOLD SWITCHES. Kurt E. Petersen and David Adler in *Journal of Applied Physics,* Vol. 47, No. 1, pages 256–263; January, 1976.

LOCALIZED STATES IN THE GAP OF AMORPHOUS SEMICONDUCTORS. Stanford R. Ovshinsky in *Physical Review Letters,* Vol. 36, No. 24, pages 1469–1472; June 14, 1976.

VALENCE-ALTERNATION MODEL FOR LOCALIZED GAP STATES IN LONE-PAIR SEMICONDUCTORS. Marc Kastner, David Adler and H. Fritzsche in *Physical Review Letters,* Vol. 37, No. 22, pages 1504–1507; November 29, 1976.

CANCER IMMUNOLOGY

SOME ASPECTS OF NORMAL AND ABNORMAL CELL SURFACE GENETICS. Edward A. Boyse and Lloyd J. Old in *Annual Review of Genetics,* Vol. 3, pages 269–290; 1969.

THE G_{IX} SYSTEM: A CELL SURFACE ALLO-ANTIGEN ASSOCIATED WITH MURINE LEUKEMIA VIRUS; IMPLICATIONS REGARDING CHROMOSOMAL INTEGRATION OF THE VIRAL GENOME. Elisabeth Stockert, Lloyd J. Old and Edward A. Boyse in *The Journal of Experimental Medicine,* Vol. 133, No. 6, pages 1334–1355; June 1, 1971.

SERUM-MEDIATED LEUKEMIA CELL DESTRUCTION IN AKR MICE: ROLE OF COMPLEMENT IN THE PHENOMENON. Robert L. Kassel, Lloyd J. Old, Elizabeth A. Carswell, Nancy C. Fiore and William D. Hardy, Jr., in *The Journal of Experimental Medicine,* Vol. 138, No. 4, pages 925–938; October 1, 1973.

CELL SURFACES AND MALIGNANCY. Edited by Peter T. Mora. DHEW Publication No. (NIH) 75–796, U.S. Government Printing Office.

IMMUNOLOGICAL SURVEILLANCE AGAINST NEOPLASIA. George Klein in *The Harvey Lectures,* Series 69, pages 71–102; 1975.

AN ENDOTOXIN-INDUCED SERUM FACTOR THAT CAUSES NECROSIS OF TUMORS. E. A. Carswell, L. J. Old, R. L. Kassel, S. Green, N. Fiore and B. Williamson in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 72, No. 9, pages 3666–3670; September, 1975.

THE CASE OF THE MISSING SUNSPOTS

THE MAUNDER MINIMUM. John A. Eddy in *Science,* Vol. 192, No. 4245, pages 1189–1202; June 18, 1976.

EXPLORING THE HERBARIUM

GREEN MEDICINE: THE SEARCH FOR PLANTS THAT HEAL... Margaret B. Kreig. Rand McNally & Company, 1964.

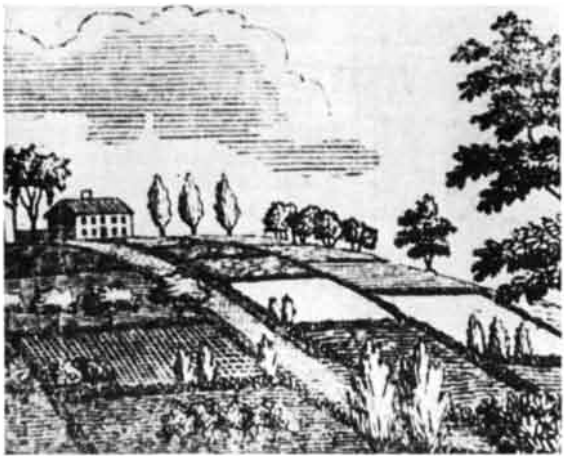
DICTIONARY OF ECONOMIC PLANTS. J. C. Th. Uphof. Verlag Von J. Cramer, 1968.

DRUGS AND FOODS FROM LITTLE-KNOWN PLANTS: NOTES IN HARVARD UNIVER-

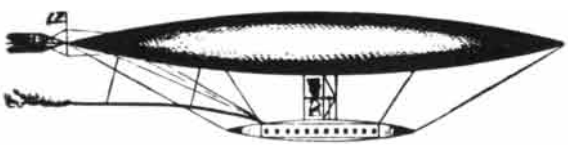
SCIENTIFIC AMERICAN
Filmstrips

Rufus Porter
A Scientific American

"We have been frequently solicited by enterprising mechanics to engage in the publication of a new scientific paper, for the advancement of several new, curious and useful arts which have recently been discovered and introduced . . ."



"It is not infrequently the case that the productions of self taught artists far surpass in excellence those of regular bred artists who have studied with the most popular masters . . ."



"It may be anticipated that within a few months these aerial machines may be seen soaring in various directions and at different elevations, some above the clouds, and others, like swallows, sailing leisurely just above the surface of the earth."

Rufus Porter—artist, inventor, journalist and founder, in 1845, of SCIENTIFIC AMERICAN—should be better known to his countrymen. His exuberant murals adorn 100 New England homes and are finding their way into museums as treasures of American art. His long list of inventions was crowned before 1850 by his demonstration of a self-propelled flying machine. His SCIENTIFIC AMERICAN spoke for the buoyant rationalism that committed this country to industrial revolution from early in the 19th Century.

The 20-minute filmstrip for DuKane projectors *Rufus Porter, A Scientific American* [$22, shipping included] brings this Yankee Leonardo back again to life. In his own colorful words and pictures it recounts the story of his untiring enterprise and places him in the context of the era in American history he so well personified.

Recommended for High School and Junior High School classes in American History, Social Studies and Art History. Available in filmstrip and tape cassette for DuKane projectors.

SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue, New York, N.Y. 10017, Dept. RP

Please send me _____ copies of *Rufus Porter, A Scientific American,* a filmstrip presentation for DuKane projectors, at $22 (shipping included). I understand I may return same and receive a full refund if I am not completely satisfied.

Enclosed is my check for $ _____.
Please include appropriate sales tax.

Name (please print) _____

Institution (if any) _____

Address _____

City and State _____ Zip _____

SITY HERBARIA. Siri von Reis Alt schul. Harvard University Press 1973.

EMPIRICAL AZTEC MEDICINE. Bernard Ortiz de Montellano in *Science,* Vol 188, No. 4185, pages 215–220; April 18, 1975.

RAT SOCIETIES

RATS. S. A. Barnett in *Scientific American,* Vol. 216, No. 1, pages 78–85 January, 1967.

AGGRESSION AND SOCIAL EXPERIENCE IN DOMESTICATED RATS. David Lucian and Richard Lore in *Journal of Comparative and Physiological Psychology* Vol. 88, No. 2, pages 917–923; February, 1975.

ROLE OF RESIDUAL OLFACTORY CUES IN THE DETERMINATION OF FEEDING SITE SELECTION AND EXPLORATION PATTERNS OF DOMESTIC RATS. Bennett G. Galef, Jr., and Linda Heibe in *Journal of Comparative and Physiological Psychology,* Vol. 90, No. 8 pages 727–739; August, 1976.

STEIN'S PARADOX IN STATISTICS

AN EMPIRICAL BAYES APPROACH TO STATISTICS. Herbert Robbins in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability—Vol. I: Contributions to the Theory of Statistics,* edited by Jerzy Neyman. University of California Press 1956.

STEIN'S ESTIMATION RULE AND ITS COMPETITORS—AN EMPIRICAL BAYES APPROACH. Bradley Efron and Carl Morris in *Journal of the American Statistical Association,* Vol. 68, No. 341 pages 117–130; March, 1973.

EMPIRICAL BAYES METHODS APPLIED TO ESTIMATING FIRE ALARM PROBABILITIES. Grace M. Carter and John E Rolph in *Journal of the American Statistical Association,* Vol. 69, No. 348 pages 880–885; December, 1974.

BIASED VERSUS UNBIASED ESTIMATION Bradley Efron in *Advances in Mathematics,* Vol. 16, No. 3, pages 259–277 June, 1975.

DATA ANALYSIS USING STEIN'S ESTIMATOR AND ITS GENERALIZATIONS Bradley Efron and Carl Morris in *Journal of the American Statistical Association,* Vol. 70, No. 350, pages 311–319; June, 1975.

MATHEMATICAL GAMES

THE METHOD OF MATHEMATICAL INDUCTION. I. S. Sominskii, translated from the Russian by Hallina Moss Blaisdell Publishing Co., 1961.

MATHEMATICAL INDUCTION. Albert A Blank in *Enrichment Mathematics for High School.* National Council of Teachers of Mathematics, 1963.

A CAR THIS LUXURIOUS DOESN'T HAVE TO COST $12,000.

INTRODUCING CHRYSLER LEBARON. $5,758. AS SHOWN.

Until now, you could often define a luxury car by its high price.

But today, Chrysler LeBaron has established a totally new class of automobiles.

LeBaron is a lighter, leaner Chrysler, yet luxurious in its comfort. A complete list of options, including genuine leather seating, rivals that of any luxury car.

LeBaron is powered by Chrysler's most sophisticated Lean Burn Engine, computer controlled for smooth performance. This engine is standard equipment on every LeBaron.

The list below will show you much more standard equipment. The total at the bottom of the list will show you how little the price of all this luxury really is.

Whether you choose the two-door or four-door, your Chrysler dealer will be happy to show how little it takes to buy or lease the Chrysler LeBaron of your choice.

CHRYSLER

A PRODUCT OF CHRYSLER CORPORATION

CHRYSLER LEBARON. THE BEGINNING OF A TOTALLY NEW CLASS OF AUTOMOBILES.

Four-Door LeBaron Medallion	
Base Sticker Price	$5,594
318 cu. in. V-8	Std.
Padded Vinyl Roof	Std.
Power Steering	Std.
Power Front Disc Brakes	Std.
Automatic Transmission	Std.
Wire Wheel Covers	$35
Whitewall Tires	$43
Bumper Guards, Front	$21
Light Package	$65
TOTAL	$5,758*

*Sticker price, including options as shown. Taxes and destination charges extra.

Standard 60/40 velour seating offered in LeBaron Medallion only.

Optional leather seat $208 extra. LeBaron Medallion only.