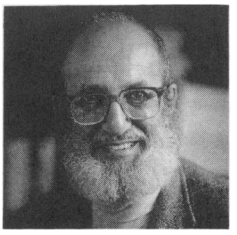
Has Progress in Mathematics Slowed Down?*

PAUL R. HALMOS, *Santa Clara University*



PAUL HALMOS has three degrees from the University of Illinois; soon after getting the last one he became, for a couple of years, assistant to John von Neumann. Since then he has taught at many universities (including Chicago, Michigan, and Indiana) and has visited many others (including Miami, Montevideo, Hawaii, Edinburgh, and Western Australia); he has been on the faculty of Santa Clara University since 1985. His mathematical interests include ergodic theory, algebraic logic, and operators on Hilbert space.

Prologue. Do we know anything that Dedekind didn't know? We should. Dedekind died in February, 1916. Six weeks before that, late Friday afternoon on New Year's Eve 1915, the MAA was born in Columbus, Ohio. In connection with the celebration of its diamond anniversary, in August 1990, in Columbus, Ohio, it became my mission to report on whether and how mathematics has changed during the 75 years of the MAA's existence, and what follows is an attempt at such a report.

I am not trying to teach any mathematics in this report, nor even any history of mathematics—all I am trying to do is share an interesting look at the growth of mathematics in the last 75 years. Everybody could find out everything I found out by spending a few months looking at all extant volumes of *Mathematical Reviews* and a few dozen other journals, but everybody hasn't done it, and I have, and I am ready to tell what I have found.

The question that I set out to answer might be phrased this way: if you had a time machine to take you back to Dedekind, what could you teach him about progress in mathematics since his day? In an attempt to organize the possible answers, I propose to put them into three classes: concepts, explosions, and developments.

The first class consists of the new concepts that Dedekind could not have predicted or expected, the new words that all mathematicians alive today have heard over and over again and that many of us wish we knew more about. (Sample: catastrophe theory.) By an explosion I mean a piece of mathematical progress that is genuine mathematics, so recognized by the whole profession, but that is at the same time the answer to old problems of such great fame that it is hot news not only for the *Transactions*, but also for the *Times* for a day, for *Time* for a week, and for student mathematics clubs for many months. (Sample: the four-color theorem.) The third proposed class consists of the deep and in some cases even breathtaking developments (but not explosions) of the kind that may not make the *Times*, but could possibly get Fields medals for their discoverers. (Sample: the independence of the continuum hypothesis. Incidentally: not all the developers mentioned below got Fields medals, and not all Fields medalists of the period are mentioned below, but qualitatively the two classes are essentially the same.) Many (but not all) of the

*A greatly condensed version of this paper was presented in August 1990 at the Columbus meeting of the Mathematical Association of America.

subjects have been and some of the others no doubt soon will be treated in expository articles in the MONTHLY.

I am listing a total of 22 subjects: 9 concepts, 2 explosions, and 11 developments. If someone else had picked 22, the chances are that the overlap would have been considerable (10? 18?) and so would the difference. There is nothing wrong with that. Some of the subjects that could (should?) have been included but were not are: exotic spheres, the Hauptvermutung, NP completeness, pseudodifferential operators, and the simplex method. One thing is for sure: no matter what subjects were treated, not all would be of equal importance. Some of the topics that I discuss are yes-or-no theorems (such as the solution of Hilbert's fifth problem), some are tools (such as the fast Fourier transform), some are attitudes (such as nonstandard analysis), and some are gigantic theories (such as the Atiyah-Singer index equation). Each topic that is listed is here because either it is important or it is amusing or, at the very least, it has been well publicized, and, in any case, it deserves attention. All the topics do have at least one feature in common: none of them belongs to what is sometimes called applied mathematics and (in particular?) none of them belongs to computer science. I had two reasons for that decision: one is that I don't know the subject and the other is that it will surely be represented in other reports.

The articles vary in length from one paragraph to about a dozen. The length was determined, in each case, by how complicated the subject is, by what I was able to learn of it, and, of course, to some extent, by personal preference. It is unusual in research mathematics to find something that is both interesting and elementary; when I did find something like that, I took advantage of the find and digressed somewhat from the main point.

If one of the pieces of mathematics that I am reporting on is one that you know about, you won't learn anything from what I say, but if not, you are in danger of learning something. You will probably not learn the precise statement of a theorem (mathematics is, after all, not a collection of theorems but a collection of ideas), and you will certainly not learn a proof. The purpose of the articles is not so much to explain things as to put them into a context. That can mean many different things; one of them is that an article about a generalization of something should at least refer to the easiest nontrivial manifestation of that something. Another thing it means is that the articles are written in prose; they contain very few of the customary abbreviating symbols of mathematical exposition. In a recently much publicized phrase, the articles are intended to contribute to the mathematical version of "cultural literacy"; they don't say what a subject really is, but they say enough about it to make casually dropped tea-time references more comprehensible than they were before.

A bibliography for a report such as this would almost be a bibliography for all mathematics. As a compromise between that and nothing, I offer exactly one reference at the end of each section. The understanding of a reader who consults one of these references would, I hope, be increased by the explanation there. In addition, such a reader would, I know, find further references to the articles and books that treat the subject of interest.

The unique reference for each section is either to

The Encyclopedic Dictionary of Mathematics
(abbreviated EDM)

(prepared by the Mathematical Society of Japan, M.I.T. Press, 1980) or to the journal

The Mathematical Intelligencer

(abbreviated MI).

EDM references appear in the form

EDM 146/B

(referring to the article labelled 146/B), and MI references appear in the form

MI 8/1/40

(referring to volume 8, number 1, page 40).

That's it; now let's get down to work, and see what these concepts, explosions, and developments really are.

Concept 1: Moore-Smith limits. Modern general topology, and in particular its manifestations in the so-called weak topologies of certain function spaces, taught us that limits of sequences are no longer a strong enough tool for analysis. In classical analysis we learn that a set is closed if and only if it contains the limits of all convergent sequences in it, but there are many important and useful topological spaces for which that statement is not true. If, however, sequences are replaced by "generalized sequences" (the streamlined word is "nets"), and, correspondingly, ordinary limits of sequences are replaced by Moore-Smith limits of nets, the classical proofs work again, often with no changes except terminological ones, and they yield results just as useful as the classical ones. (Example: a set is closed if and only if it contains the Moore-Smith limits of all convergent nets in it.)

The Moore of the seminal 1922 Moore-Smith paper is the great E. H. Moore (one of the teachers of R. L. Moore) and the Smith is the otherwise largely forgotten H. L. Smith. The word "net" was first used, so far as I know, by J. L. Kelley in 1950.

Reference: EDM 89/H.

Concept 2: Distributions. The set $\mathbb{S}$ of infinitely smooth functions $\mathbb{R}^n$ with compact support is in an obvious way a vector space. Let us say that a sequence $\{\varphi_m\}$ of such functions converges to 0 if the supports of all the φ_m are covered by some fixed compact set, and not only does the sequence converge uniformly to 0, but, in fact, so do all its mixed partial derivative sequences.

If now f is a complex-valued integrable function on $\mathbb{R}^n$, then the equation

$$T_f(\varphi) = \int \varphi(x) f(x)\, dx$$

defines a continuous linear functional on $\mathbb{S}$, and the same is true if f is only locally integrable. If μ is a complex finite measure in $\mathbb{R}^n$, then the equation

$$T_\mu(\varphi) = \int \varphi(x)\, d\mu(x)$$

defines a continuous linear functional on $\mathbb{S}$, and the same is true if μ is required to be finite on compact sets only. If, for instance, $\mu(E) = 1$ when the origin belongs to E and $\mu(E) = 0$ otherwise (in other words, μ is a point mass at 0), then $T_\mu(\varphi) = \varphi(0)$ for every φ—which may remind you of the curious behavior of something that used to be called the Dirac delta function (but which, of course, never was a function).

All continuous linear functionals on $\mathbb{S}$ are called *distributions*; the inspiration of Fields medalist Laurent Schwartz (and the other giants on whose shoulders he stood) was to realize that the concept of distribution generalizes the concept of function and that in many classical situations, in which no function exists that satisfies certain prescribed conditions, a distribution satisfying them may exist and is for all practical purposes just as good. A typical kind of "prescribed condition" is a partial differential equation to be solved—a distribution solution often gives as much applicable information as a classical honest one.

Reference: EDM 130/B.

Concept 3: The Monte Carlo method. Mark a table with a system of parallel lines spaced, say, two inches apart from one another, and consider a needle (a mathematically idealized needle, in other words a line segment) of length one inch. Drop the needle on the table at random and ask: what is the probability that the needle will not land between a pair of the parallel lines but will cross one of them? This is the famous Buffon needle problem, and the method of working it out is not especially difficult; it is an exercise in many elementary probability texts. The answer turns out to be $1/\pi$.

Is that a bit of a surprise? How does π get into the answer to *this* question? That's not the main issue right now, but as long as it's been raised a comment on it might be in order. The answer to the probability question depends, of course, on how the needle gets on the table—or, to say the same thing more precisely, it depends on the probability distribution that is tacitly or otherwise assumed to be the one that the fall of the needle is subject to. Different possible distributions have been studied, and, to nobody's surprise, yielded different answers. One possibility is to let the probability distribution of the location of the center of the needle be uniform on a line segment perpendicular to the parallel lines and of length three inches, say, and, independently, to let the probability distribution of the angle at which the needle is inclined to that perpendicular segment be uniform between 0° and 180°. Once angles are explicitly mentioned that way, it should be no surprise that π enters the answer.

If the answer is accepted, a curious consequence follows. Perform the experiment, repeatedly, many times, and count the ratio of the number of successes to the number of trials—the law of large numbers tells us that that ratio will be very nearly equal to $1/\pi$. Conclusion: the value of π can be determined from a physical (probability) experiment, with no calculation (except the formation of a reciprocal at the end). This was known and appreciated a couple of hundred years ago—it was a forerunner of the modern technique known as the Monte Carlo method, introduced around 1945 by Ulam and von Neumann.

The Monte Carlo method has had frequent applications when the obvious calculations that a question demands seem formidable—replace the question by a probability question that has the same answer and find the answer to the probability question either by actual experimentation, or, more likely, by computer simulation of random experimentation. An easy example is the evaluation of a definite integral $\int_0^1 f(x)\,dx$ where f is a bounded function (for definiteness, $0 \leqq f(x) \leqq 1$ whenever $0 \leqq x \leqq 1$). Proceed as follows: choose (or let your computer choose for you) a large set of pairs (x, y) of random numbers, with $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$, and then evaluate the ratio of the number of them for which $y \leqq f(x)$ to the total

number. That ratio is approximately equal to the average $\int_0^1 f(x)\,dx$—and, lo and behold, the integral has been (approximately) evaluated.

This example is a little too naive for the theory, but the genuine real-life applications are in spirit pretty much the same. To calculate something too messy to calculate, whether or not the something is already a probability, replace the question by a probability question (or a different probability question) that is more easily susceptible of experimental (or, rather, computer imitated experimental) evaluation.

Why the name Monte Carlo? Presumably because that's the name associated with one of the most famous gambling places of the world. What the name suggests is that if you can't solve a problem, try gambling with it—that is, replace it by a gambling problem that has the same solution and that you have a better chance of being able to solve.

Reference: EDM 378/B.

Concept 4: Categories. If you understand vector spaces and linear transformations, you are 90% of the way toward understanding categories. Mathematicians have been viscerally aware for a long time that vector spaces and linear transformations behave "just like" groups and homomorphisms, and they in turn behave just like topological spaces and continuous mappings—the parallelisms are great and all pervasive. If you know what subgroups are, you know what subspaces are, and if you understand how to form quotient groups, you are in a good position to discover (to rediscover) how to define quotient spaces.

Category theory is a formalization of that visceral understanding; it was born (to Eilenberg and Mac Lane) in 1945. A category is a class of "things" of two kinds, objects and morphisms, satisfying three easy axioms. Brutally summarized, those axioms say that morphisms can be composed, provided only that their domains and ranges match properly, that for each object X there is an identity morphism from it to itself, and that if $\mathrm{Hom}(X, Y)$ denotes the set of all morphisms from X to Y, then the only way $\mathrm{Hom}(X, Y)$ and $\mathrm{Hom}(X', Y')$ can have any elements in common is to have $X = X'$ and $Y = Y'$.

Vector spaces, and groups, and topological spaces fit under this scheme, and so do abstract sets, and rings, and differentiable manifolds, and modules, etc., etc. More special concepts than just morphisms can be defined within each category (such as monomorphisms, subobjects, direct products, and duality); they are all pleasant and handy and none of them holds any surprises.

An important part of the theory of categories is the theory of functors. A functor is a way of associating with the objects and morphisms of one category the objects and morphisms of another. Typical example: each vector space has a dual, and each linear transformation between vector spaces has an adjoint. Trivial but helpful example (called the forgetful functor): each topological space determines a set and each continuous mapping determines a mapping—just forget the topology and the continuity, and thus pass from the category of topological spaces to the category of sets.

Category theory started as (and continues to be) a convenient language in which to describe many phenomena, and for many of us that's all it is. For the fascinated specialist it is a subject of research that can discuss quotient categories, adjoint

functors, categories of categories, and other such towers of elaborate complications of great fascination—long may it wave.

Reference: EDM 53/A.

Concept 5: *K*-theory. If $\mathbb{U}$ and $\mathbb{V}$ are finite-dimensional vector spaces with the field of real numbers acting as scalars, then so is the direct sum $\mathbb{U} \oplus \mathbb{V}$ of $\mathbb{U}$ and $\mathbb{V}$. At a casual glance the formation of direct sums looks like a perfectly respectable addition process, but a second glance shows that there is something a little wrong: the process isn't associative. The direct sum $(\mathbb{U} \oplus \mathbb{V}) \oplus \mathbb{W}$ resembles the direct sum $\mathbb{U} \oplus (\mathbb{V} \oplus \mathbb{W})$ in many ways, but strictly speaking it's not the same; the elements of the first are ordered pairs whose *first* coordinates are ordered pairs, whereas the elements of the second are ordered pairs whose *second* coordinates are ordered pairs. The temptation to consider the correspondence that assigns to every element $((u, v), w)$ in the first direct sum the element $(u, (v, w))$ in the second is well nigh irresistible. That correspondence is an isomorphism between the direct sum spaces involved, and the natural way out of the non-associative obstacle is to identify two vector spaces if they are isomorphic. If that is done, and that is the normal thing to do in this situation, another small obstacle arises, but it's easy to make it too go away: it has to be checked that the addition process can be unambiguously defined for isomorphism classes of vector spaces as well as for just plain vector spaces. That is, what has to be shown is that if $\mathbb{U}$ and $\mathbb{U}'$ are isomorphic vector spaces and if $\mathbb{V}$ and $\mathbb{V}'$ are isomorphic vector spaces, then the direct sum $\mathbb{U} \oplus \mathbb{V}$ is isomorphic to the direct sum $\mathbb{U}' \oplus \mathbb{V}'$. That is true and almost obvious, and once it is on record, then the direct sum of two isomorphism classes of vector spaces can be defined by choosing a representative element of each class, forming the direct sum of those representative elements, and then forming the isomorphism class of that direct sum.

Very well—a sort of addition can be defined for vector spaces (yes, yes—they should be called isomorphism classes of vector spaces, but everyday mathematical idiom continues to call them vector spaces anyway, and mentally remembers that equality no longer means honest equality but isomorphism)—what can be said about the class of all vector spaces under this addition? Natural question: do they form a group? The first part of the answer is easy—there is an identity element, that is, there is a neutral vector space with the property that adding it to any other produces no changes—namely, the unique 0-dimensional vector space. The next part of the answer is dishearteningly negative: no, vector spaces under direct sum do not form a group, because inverses, negatives, do not exist.

The setback is not serious—it is no more serious than that encountered by a child who has learned to add positive integers, has discovered 0, and is discouraged by his inability to go backward—subtraction doesn't always make sense. To subtract u, say, from $u + v$ is something that can be done, but subtraction of an arbitrary positive integer from another may or may not be doable. Similarly, someone might be willing to say that subtracting $\mathbb{U}$ from $\mathbb{U} \oplus \mathbb{V}$ is permissible, and yields the answer $\mathbb{V}$, but what about the general case, and, in particular, what about subtracting $\mathbb{U} \oplus \mathbb{V}$ from $\mathbb{U}$?

The way out, for vector spaces, is the same as the way out was for integers—the integers that do not exist are created by fiat, or by an explicit set-theoretic construction: negative numbers, and negative direct sums (!) are adjoined to the system under the study, and the result is a perfectly good group.

What group is it? The answer is easy to state and easy to understand—it is, except for notation, the same as the additive group of all integers. The point is that if isomorphic finite-dimensional vector spaces are identified, then there is nothing to distinguish between two vector spaces except their dimension—which is a nonnegative integer—and, consequently, the process of adjoining "negatives" to vector spaces (or, rather, to their isomorphism classes) is exactly the same as the process of adjoining negatives to the nonnegative integers.

The procedure indicated above can be carried out, and turns out to be profitable to carry out, even if the field of real numbers is replaced by an arbitrary ring. Rings occur in "the real world" more often than fields do, but they constitute a more difficult subject. Algebraic K-theory is a partial attempt to face the difficulties. What makes the use of fields easier is that a finite-dimensional vector space has a finite basis. No, that's not a redundancy. To see that it is not, rephrase it this way: a finitely generated module over a field has a finite basis. (Recall that the definition of a module is just like that of a vector space except that an arbitrary ring is allowed to play the role of the coefficient field.) The point is that the corresponding statement for modules over rings instead of fields is not always true. It *is* true for "good" modules, but a precise definition of "good" is a technicality that can safely be omitted from an overview such as this one. (According to a popular definition the good modules are the ones that are finitely generated and projective.)

All right then, if vector spaces are replaced by modules, the group theoretic construction goes pretty much the same way. Addition is defined for "good" modules by forming direct sums, and then the definition has to be modified, as before, so as to apply to equivalence classes of modules instead of modules; the result is a lovely associative addition, with a zero element, but no inverses. (Caution: the useful notion of equivalence here is a slightly weakened version of isomorphism, but the principal features of the theory remain the same anyway.) Let inverses, negatives, be adjoined, by fiat, or by the usual construction of forming ordered pairs ("differences") and equivalence classes thereof, and the result is a group. The equivalence relation that defines the group assigns to each module over a ring R an element of the group—that element plays the role of the generalized dimension of the module. The group thus associated with the prescribed ring R is denoted by $K_0(R)$ and is called the Grothendieck group of R—and that association is the first step of what is called K-theory.

K-theory does more than that first step: in general it associates an abelian group $K_n(R)$ with every ring R and every nonnegative integer n, in ways that become more and more mysterious as n increases. For $n = 0$, the result is a generalization of dimension; for $n = 1$, it is a generalization of determinant; after that—don't ask. But, in any event, it can't do any harm to take a quick look at $n = 1$.

The objects at the center of the stage this time are not modules, but automorphisms of modules. If the coefficient ring is a field, a good assumption to begin with, then automorphisms can be viewed as nonsingular matrices. When should two such objects be regarded as equivalent? One classically inspired possibility is to call two matrices equivalent just when they can be obtained from one another by a finite sequence of elementary transformations. Elementary transformation? In this context that means: add an arbitrary scalar multiple of any row (or column) to any other row (or column). Alternative definition: call a matrix elementary if it is either equal to the identity matrix or differs from it at exactly one non-diagonal

entry, and then define an elementary transformation as multiplication, fore or aft, by an elementary matrix. Since an elementary matrix obviously has determinant 1, it follows that every matrix in the group generated by the set of all elementary matrices has determinant 1, and that two equivalent matrices must have the same determinant. The two theorems in the preceding sentence are the weaker halves of two better theorems, which say that a matrix has determinant 1 *if and only if* it is a product of elementary matrices, and that two nonsingular matrices (over a field) are equivalent *if and only if* they have the same determinant. Summary: the quotient group of the group of all invertible matrices by the (normal) subgroup generated by the elementary matrices is equal (well, isomorphic) to the multiplicative group of nonzero elements of the coefficient field.

The point of this summary is that it is a nontrivial statement about determinants of matrices over fields in which the word "determinant" does not occur. Determinants of matrices over rings are hard to come by, but elementary transformations and elementary matrices make perfectly clear sense. The concept they lead to must cope with two kinds of matrix operations—the algebraic ones of sum and product and the geometric one of direct sum. The best way to do that is to apply similar considerations to infinite matrices each of which is in fact a direct sum of an infinite identity matrix and a finite invertible matrix (with entries in a prescribed ring R). The resulting quotient group (modulo the subgroup generated by the set of all elementary matrices) is denoted by $K_1(R)$ and is called the Whitehead group of R. It turns out to be abelian (a nontrivial theorem), and its elements are the generalizations of determinants (of matrices over fields) in the same sense in which the elements of $K_0(R)$ are the generalizations of dimensions (of vector spaces over fields).

The uses of K-theory in algebra and topology are many, and, in particular, the topological cousin of the algebraic theory discussed above is essentially involved in the proof of the Atiyah-Singer generalization of the Riemann-Roch theorem.

Reference: EDM 236/I.

Concept 6: The fast Fourier transform. The Fourier transform of a function f on the real line is usually defined to be the function g given by

$$g(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(y)\, e^{-ixy}\, dy,$$

assuming that the integral exists in some decent sense. The Fourier transform is a powerful tool in modern pure analysis and it is also a powerful tool in classical hard analysis and many of its concrete applications. In the applications it is, of course, desirable to minimize the computations that enter, and it is not surprising that various approximations are frequently introduced, such as replacing the infinite interval by a finite one and replacing the integral by a sum.

The discrete Fourier transform comes therefore to play a role—it is the transformation that assigns to an n-tuple $(x_0, \ldots, x_{n-1})$ of complex numbers the n-tuple $(y_0, \ldots, y_{n-1})$ given by

$$y_q = \frac{1}{n} \sum_{p=0}^{n-1} x_p\, e^{-2\pi i p q/n}, \qquad q = 0, 1, \ldots, n-1.$$

The so-called fast Fourier transform is a minor observation—an inspired minor

observation—whose discovery is variously attributed to J. W. Cooley and J W. Tukey (1965) and to C. F. Gauss (1805). The observation is that the sum, the discrete Fourier transform, can be evaluated by expressing it as a sum of clever subsums and making clever use of the algebraic properties of the roots of unity (the numbers $e^{-2\pi ipq/n}$) that enter. The popular name of those algebraic properties is "trigonometric identities". The idea that makes cleverness possible is that the Vandermonde matrix formed with the powers of a primitive nth root of unity can, for some n's, be factored in such a way that the factor matrices have many entries equal to 0, and that, as a result, the subsum evaluations need less arithmetic.

The evaluation of each y_q by the definition of the discrete Fourier transform involves the evaluation of $n - 1$ products (x_p times $e^{-2\pi ipq/n}$ for $p = 1, \ldots, n - 1$) followed by $n - 1$ additions (the partial sums). Since there are n values of y_q, the total work involves $n(n - 1)$ additions and the same number of multiplications.

The simplest version of the fast Fourier transform applies to the case in which n is represented as a product of two factors, $n = hk$. In that case the fast Fourier transform can involve as few as $n(h + k - 1)$ additions and $n(h + k)$ multiplications. If, for instance, $n = 100 = 10 \times 10$, then the direct method takes 9900 additions and 9900 multiplications, whereas the fast method takes 1900 additions and 2000 multiplications—a huge difference, which gets huger as n grows and the process is iterated.

It's a beautiful and useful part of what used to be called numerical analysis (before it came to be called computer science). Is it more than a trick? Does it teach us anything about pure analysis, about the group of roots of unity—does it have valuable analogues in other groups?

Reference: MI 7/3/49.

Concept 7: Nonstandard analysis. Leibniz used infinities and infinitesimals, but he admitted feeling slightly queasy about them. His successors banished such things from the mathematical heaven and worked with ε's and δ's instead. The modern theory of nonstandard analysis dredged the forbidden concepts up from the underworld and is trying to reinstate them at the right side of Cauchy's throne.

A possible attitude toward infinitesimals is to regard them as "ideal" elements, similar in that respect to the points at infinity in the projective plane. Once assumed, or defined, or constructed, the main tool in working with them is the so-called transfer principle, which says (roughly) that anything sayable in the appropriate formal language and true about the nonstandard universe remains true about the standard universe.

The basic standard universe consists (roughly) of "individuals" (Urelemente), and sets of them, and sets of sets, and sets of sets of sets, and so on ad infinitum. The nonstandard universe has many extra elements; they are (roughly) functions with values in the standard universe, or, slightly more precisely, classes of functions that have been identified according to a suitable equivalence relation. There is a sense in which the new elements are reminiscent of sequences. A constant sequence, whose constant value belongs to the standard universe, plays the role of that value, but there are sequences that are infinitely small (converge to 0) and others that are infinitely large (diverge to ∞), and in the democracy of the nonstandard universe they are all treated equally.

Consideration of such universes leads to ordered fields that are like the field of real numbers, and play the role of that field in the theory, but that are non-Archimedean—that is, they do not share with the real numbers the property that a small thing added to itself often enough becomes a large thing. For those who are comfortable with the use of formal languages, and believe that infinitesimals and infinities are intuitively clear and pleasant concepts, nonstandard analysis seems to be an efficient tool—such people can use that language to think in and, they say, they are led thereby to make otherwise elusive discoveries. Their greatest single victory so far (the only one that every proselyting article quotes) is the first proof (by Allen Bernstein and Abraham Robinson) of an invariant subspace theorem about certain special operators in Hilbert space—and, with considerable justice, they refuse to be fazed by the fact that soon after the first proof (nonstandard) other (completely standard) proofs came along.

The future of the subject is not yet clear to most non-believers—will it or won't it become an established part of mathematics—will it or won't it displace ε's and δ's?

Reference: EDM 274/E.

Concept 8: Catastrophes. Consider the graph of the parabola given by the equation $y^2 = x$ and consider the mapping $(x, y) \mapsto x$ (projection) from that graph to (the positive part of) the x-axis. With exactly one exception every point of the graph has a neighborhood that is homeomorphic to its image, which is an open interval on the line. The exception is, of course, the origin; in no neighborhood of the origin is the projection mapping one-to-one. This phenomenon is described by saying that the origin is a singularity of the (smooth) mapping from a 1-dimensional manifold (the graph) to another (the axis).

Consider the surface of a sphere, project it perpendicularly to a plane, and note that the local behavior of the projection varies from point to point. Each point of the northern hemisphere has a neighborhood in which the projection acts as a homeomorphism, and the same is true of the southern hemisphere, but at points of the equator the projection has a kind of singularity called a "fold." It looks like a fold, locally, doesn't it?: a small neighborhood of a point of the equator is, in effect, folded over on itself by the projection.

Consider next a suitable cubic surface (visualize it as a plane folded over a part of itself and then folded back again at a slightly different angle), project it perpendicularly to a plane, and note that this time the projection can misbehave in different ways. In the case of the sphere, any point of the plane that had an inverse image at all had either one or two inverse images; for the cubic surface that number can be 1, 2, or 3. The points for which it is 3 correspond to the regular points of the projection mapping, the points for which it is 2 correspond to a fold singularity, and the unique point for which it is 1 comes from a singularity called a "cusp". The reason for the terminology this time is that the "fold" points on the surface project onto a semicubical parabola in the plane whose cusp, in the usual sense of the word, comes from the cusp singularity.

The mathematics of catastrophe theory studies the problem of classifying the singularities of smooth mappings (such as the projections above) between smooth manifolds (such as the parabola, the sphere, and the cubic surface above). A typical result (for the 2-dimensional case) is that every smooth mapping of a surface to a plane has, after possibly an appropriate small perturbation, nothing

but cusps and folds for its singularities. In higher dimensions there are similar (and surprisingly short) complete lists of "elementary catastrophes". The theory goes back to the work of Whitney in 1955; since then it has been extensively cultivated and applied by several mathematicians. The list of names includes Thom in France and Arnold in Russia; both of them have made important contributions to both the theory and its applications. Legitimate applications exist, for instance, to wave propagation problems, elastic stability, and geometric optics.

Where does the term "catastrophe" come from? The answer seems to be that it was suggested (by Thom?) as an appropriate way to describe startlingly discontinuous changes in the world around us. Here is a simple example: suspend a weight from a horizontal bar, and ask for the maximum supportable load. Intuition and experience suggest that as the load is increased nothing much happens, except possibly a gentle bend, till, suddenly, the bar breaks and the weight falls—a catastrophe. A more complicated example that is mentioned by almost everybody who writes on the subject is the degree of aggressiveness of a dog that is approached by a possible enemy. The dog feels both fear and rage—which is greater? What will the dog's response be: flight or fight? Once again intuition and experience suggest that the dog will cower at first and back away, till suddenly, because the distance to the enemy has decreased too much and the threat has increased beyond the supportable point, the dog will snarl and attack—a catastrophe.

How does it happen that a deep mathematical theory, with genuine and valuable applications, becomes as controversial as catastrophe theory has become? There is nothing wrong with the theorems that Whitney, Thom, Arnold, and others have proved; what then can a controversy be about? The answer is that some of the proponents of the theory (notably Thom himself, and Zeeman) are describing and claiming applications that seem to others (notably Arnold) far-fetched and unsound. The applications claimed are to economics, embryology, linguistics, psychology, and other subjects, and include political elections, mental disorders, prison riots, heart beats, stock market crashes, and the outbreak of war.

Here are four of the many comments that Arnold makes.

"The mathematical articles of the founder of catastrophe theory, René Thom, were reprinted as a pocket book—something that had not happened in mathematics since the introduction of cybernetics from which catastrophe theory derived many of its advertising techniques."

"A particular aspect of Thom's work on catastrophe theory is his original style: he established a fashion in not giving even sketchy formulations of results, let alone proofs."

"When the mapping we are concerned with is known in detail we have a more or less direct application of the mathematical singularity theory to various natural phenomena... In the majority of works on catastrophe theory, however, more controversial situations are considered where not only are the details of the mapping not known, but its very existence is problematical."

"The deficiencies of... catastrophe theory are too obvious to discuss in detail. We remark only that articles on catastrophe theory are sharply distinguished by the catastrophic lowering of the level of the requirement of rigour and also of the requirement of the novelty of the published results. Although one can understand the catastrophe theorists' reaction against the traditional flow in mathematics of rigorous but dull epigonic works, nevertheless their lack of respect to their

predecessors (to whom belong the majority of concrete results) can hardly be justified."

Reference: EDM 410/K.

Concept 9: Chaos. To a mathematician the word "point" is virtually synonymous with "element of a set". Let me give a moderately complicated set whose "points" are quite a bit more complicated than the points that most students usually meet. Begin with the closed unit interval $I = [0, 1]$, and let X be the set of all finite disjoint unions of nondegenerate closed subintervals of I. ("Nondegenerate" means that single points are not to be considered as intervals.) A "point" in the following paragraph is to be an arbitrary element of X.

Let T_C be the mapping of the set X into itself that acts on each point P by removing the open middle third of each of the closed intervals whose union P is. Thus, for instance, since I itself is a point of X, the image $T_C(I)$ makes sense and is equal to the set (point) $[0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. What happens when the mapping T_C of X into itself is iterated? Start, for instance, with I, and form successively $T_C(I), T_C^2(I), T_C^3(I), \ldots$. That decreasing sequence of finite unions of closed intervals is familiar to anyone who has ever taken a course in the set theory of real numbers: the intersection of that sequence of sets is the classical Cantor set. (The suffix C is intended to be a reminder of Cantor.)

The transformation so described is an artificial but interesting and suggestive example of a dynamical system. According to the most general definition, a dynamical system is a mapping of a set into itself. That definition, however, is too general to be of much use. The concept becomes useful when the underlying set and the transformation that acts on it have some structure of mathematical interest (algebraic, analytic, or geometric), and in all extant studies of dynamical systems such extra structures are indeed present.

The concept of "chaos" depends on that of dynamical system; roughly speaking, the theory of chaos is the study of the behavior (misbehavior?) of dynamical systems at infinity. Take any transformation T of any set X into itself, form the successive iterates $T, T^2, T^3, \ldots$, and then ask an intelligent question that involves them. Example: let X be the real line, define Tx to be $\cos x$ for every x in X, and ask: what happens to the sequence $x, Tx, T^2x, T^3x, \ldots$ for various values of x? The answer is easy to derive, but if you have never done it, I would suggest that first you take a hand-held calculator that has a "cos" button, start with any original input x that you like, and keep pushing "cos" over and over again. (The result looks a little more interesting if you use radians instead of degrees.)

Another example is the dynamical system T_C above. For still another example, much deeper, let X be the closed unit interval with its end points identified (so that this time a "point" is a real number between 0 and 1 inclusive, except that 0 and 1 count as the same point), and let T_2 be the act of doubling modulo 1. Thus, for example,

$$T_2\left(\frac{1}{3}\right) = \frac{2}{3},$$

$$T_2\left(\frac{3}{4}\right) = \frac{3}{2} \bmod 1 = \frac{3}{2} - 1 = \frac{1}{2},$$

$$T_2\left(\frac{1}{2}\right) = 0,$$

and

$$T_2(\sqrt{2} - 1) = T_2(1.4142\ldots - 1) = 2\sqrt{2} - 2 \,(\mathrm{mod}\, 1)$$
$$= 2.8284\ldots - 2 = 0.8284\ldots .$$

What happens to the sequence $x, T_2x, T_2^2x, T_2^3x, \ldots$ for various values of x? Can it converge to some limit? Can its closure contain an open interval? This example, relatively innocent as it may look, is in fact quite deep; it is a close relative of one that is topologically important (the so-called Smale horseshoe) and one that is analytically important (the so-called unilateral shift).

An example of a dynamical system that is more reminiscent of classical analysis than the ones above is obtained as follows: choose a couple of constants a and b, and define $T = T_{ab}$ in $\mathbb{R}^2$ by

$$T_{ab}(x, y) = (y + 1 - ax^2, bx).$$

Transformations of this kind were introduced and studied by Hénon. The formula defining them is not enough to scare even a student of elementary analytic geometry, but the study of their asymptotic properties is delicate and interesting. Those properties depend, of course, on the parameters a and b. If, for instance, $a = 1.3$ and $b = 0.3$, then there exists a set of seven points in the plane that constitute a "periodic attractor". That means that the iterated images of a point in the plane will either tend to infinity or else tend to get closer and closer to one of the seven points, and then to a next one, and then a next, and then, at the eighth step closer still to the first, and then to the second, etc. If, on the other hand, $a = 1.4$ and $b = 0.3$, then the iterated images of some points tend to infinity, whereas for other points those images cluster at a complicated set of curves that constitute what is called a "strange attractor" (the "Hénon attractor"). A good definition of that concept is yet to be offered. The main property of a strange attractor seems to be that it is an infinite set with "sensitive dependence" on the initial conditions (whatever that means). Strange attractors occur in, for instance, the study of turbulence.

Why is the word "chaos" used? The reason seems to be similar to the reason that caused catastrophes to be called catastrophes—it seems to be a subjective (not really a mathematical) reaction to an unexpected appearance of discontinuity. A possible source of confusion is that the startling discontinuity can occur at two different parts of the theory. Frequently a dynamical system depends on some parameters (the way the Hénon T_{ab} depends on a and b), and, of course, the sequence of iterates $x, Tx, T^2x, T^3x, \ldots$ of a dynamical system T applied to any particular x always depends on the initial point x. The startling change of the Hénon family (from periodic attractor to strange attractor) is regarded as chaos—unpredictability—and the very existence of the Hénon strange attractor, not obviously visible in the definition of the dynamical system, is regarded as chaos —unpredictability. I would like to register a protest vote against the attitude that the terminology implies. The results of nontrivial mathematics are often startling, and when infinity is involved they are even more likely to be so. It's not easy to tell by looking at a transformation what its infinite iterates will do—but just because different inputs sometimes produce discontinuously different outputs doesn't justify describing them as chaotic.

A part of the theory of chaos has acquired a kind of popularity (notoriety?), and no discussion of the subject can omit at least mentioning that part. I refer to the things called "fractals".

The artificial dynamical system T_C with which this exposition began led to the Cantor set, which occurs in chaos theory frequently and non-artificially. A curious and important property of the Cantor set is that its part in the interval $[0, \frac{1}{3}]$ is similar to the entire set—just blow it up by the dilatation $x \mapsto 3x$ and the small set becomes identical with the large one. Similarly, the part of the Cantor set in the interval $[\frac{2}{9}, \frac{7}{27}]$ is similar to the entire set; in fact, every neighborhood of every point of the Cantor set includes a subset that is similar to the entire Cantor set. The iteration of transformations tends to produce sets that exhibit such "self similarity", that is, sets that are invariant under change of scale; the Hénon attractor itself is another example.

The phenomenon is not new to the mathematical world. It has long been known that the Cantor set is the prototype of a large family of sets obtained by various architectural changes. Example: omit middle fifths, say, instead of middle thirds. Another example: replace the discarded open middles by, say, equilateral triangles. Still another: perform middle omissions not on an interval but on rectangles or disks or other interesting sets in the plane. All such sets exhibit local properties that are pleasing and surprising when first encountered but that become, upon acquaintance, so nonsurprising as to be almost trite—they are like friends whose company is as desirable as ever but whose shape and whose gestures remain completely familiar even when they grow old or wear disguises.

The Cantor set is a fractal, and so are various variations on that theme. (No definition of the general concept is needed here, but I mention in passing that a standard one is this: a fractal is a set whose dimension in the usual sense of topology, which is always less than or equal to its Hausdorff-Besicovitch dimension, is in fact strictly less. The topological dimension of the Cantor set is 0, and its Hausdorff-Besicovitch dimension is $\frac{\log 2}{\log 3}$.) The reason for the popularity of fractals is that they can be pretty. Certain planar versions of the Cantor set, with or without "colorization", look good on your living room wall, and disguising them by changing their sizes and shapes can produce enough pretty pictures to fill the whole wall or even an entire room in a museum of modern art.

"Art" is perhaps the operative word. A great mathematician conceives new theorems, discovers their beautiful proofs, and shares the results with the rest of us by publishing them. A great cook conceives new dishes, discovers their ingenious recipes, and shares the results with the rest of us by preparing them. An infinite Andy Warhol iteration of the label on a can of soup may be popular and it may be art, but it has very little to do with the genius of a cook. The infinite iteration of Cantor-like omissions may be a popular subject of conversation and it may be art, but it has very little to do with the insight of a mathematician. Paintings of cans of soup are not cooking, and paintings of Cantor sets are not mathematics.

Reference: MI 2/3/126.

Explosion 1: The four-color theorem. Everybody knows what the four-color problem is, but I'd like to begin by asking some elementary questions about it anyway, some questions that you may not have bothered to ask yourself before. The point is that the problem concerns the Euclidean plane, and the topology of the plane can exhibit some unexpected and weird phenomena.

Did you know, for instance, that there exist five regions in the plane ("the lakes of Wada") all of which have the same frontier? With "two" instead of "five"

everybody does know it: just remember the right and left half planes, that is, in terms of Cartesian coordinates, the set of points with positive x coordinates and the set of points with negative x coordinates. The common frontier is the y axis, and there is nothing surprising about that. (By the way, "region" in this context has its meaning familiar from a first course in complex variable theory: an open connected set.) What about "three" regions with a common frontier—is there a simple example of that? No, there is not; anybody who can construct an example with three can construct one with five just as easily, or, for that matter, with any other finite number.

Doesn't that sound like a negative solution of the four-color problem? The five regions represent five countries on a planar map, and since each two of them touch (at their common frontier), the only way to color the map correctly is to use five different colors.

Yes, that is a negative solution of the four-color problem, but only if the problem is carelessly formulated. If five regions have a common frontier, then they and their frontier must twist and interlock in an extraordinarily nasty manner. The classical four-color problem assumes (usually tacitly) that the countries represented on the map are healthy, not pathological. A simple way to rule out pathology is to demand that the frontiers of the countries be polygons (a finite number of line segments); the full depth and difficulty of the problem is present even in that strongly restricted case. If that seems too severe, it is all right to allow the frontiers to be pleasant continuous curves, but, in any event, as the five-region example shows, some geometric restriction is necessary.

A different kind of instructive example to look at is a map consisting of a circle in which a certain number of diameters have been drawn (say, for instance, three). Those diameters divide the circle into six countries each the shape of a slice of pie, and all six "touch" at the center. Does that mean that six colors are needed to color the map? No, it does not. That's a situation that the usual statement of the four-color problem does foresee and does rule out: two countries are said to "touch" only if they have a reasonably long piece of their frontiers in common, more than just some isolated corners.

Very well then: what is the smallest number of colors that is sufficient to color all decent maps? Could it possibly be two? One way to answer that question is to ask whether it is possible to design a map that represents three countries each of which touches both the others. The answer to that is easy: just make a circular map and make the three countries like three slices of pie with each of them having an angle of, say, 120° at the center: in that case each of them is "between" the other two. Conclusion: two colors are not enough to color all maps.

All right: is there a map with four countries so that each of them touches all three of the others? Yes—take the pie map just constructed and punch a hole in it —replace the center by a substantial circular country, with, say, half the radius of the entire map. The outside countries still have the property that each of them is between the other two, and, at the same time, each of them touches the center country. Conclusion: three colors are not enough to color all maps.

Does the technique used twice by now work for the next stage: is there a map with five countries each of which touches all others? In an attempt to see the answer, change the four-country map just studied in two ways: put the circle that it consists of (call it circle C) inside a larger circle (call it D), and cut the small circle in its center (call it B) into two parts. The resulting map has six regions: one is the

ring shaped region between D and C, two others are the two halves of B, and the other three were present before the changes. Call the five regions inside C five countries. A glance at the picture shows that three of them touch each of the other four, but two of them fail to do so. Adjoin the outside ring-shaped region (as a colony?) to that half of B that fails to touch each of the four countries—and, lo!, the result is a map of five countries every one of which touches all four others. A proper coloring of this map needs five colors—four are not enough. The technique is not difficult to generalize; for each positive integer n there exists a map of n countries each of which touches all $n - 1$ others.

Is something wrong? A possible objection is that (in the case $n = 5$) one of the five countries in the picture consists of two pieces—the country it represents is not connected. Is that wrong? Does that happen in the real world? Sure it does; the state of Michigan, for one, consists of two pieces. Does that mean that four colors are not enough for all maps? Yes, it does—provided that maps are allowed to contain disconnected countries. Once again the usual statement of the four-color problem foresees this obstacle, and rules it out; the problem concerns connected countries only.

If that five-color example is not acceptable, what would have to be done to prove that four colors are not enough to color all maps? An obvious possibility that the preceding discussion suggests is to try again to design a map with five countries, but this time connected ones, so that each of them touches all four of the others. Augustus de Morgan, in 1852, used ingenious topological arguments to prove that there is no such map (at least they would be called topological in the twentieth century)—but he was not allowed to jump to the conclusion that he had proved the four-color theorem. There was, perhaps, some temptation to jump to that conclusion, but it wouldn't have been right.

The point is this: to say that n colors are not enough to color a map is not the same as saying that it contains a submap of $n + 1$ countries each of which touches all n of the others—not at all. To see that, even for the already known case of $n = 3$, begin by contemplating a circular map with five pie-slice countries (each making an angle of 72° at the center). That map can be colored with three colors: just go A, B, C, A, B around the circle. Now punch a hole in the middle, that is, replace the center by a substantial circular country. The result is a map in which no submap of four countries has the property that each of them touches all three of the others—but, nevertheless, the entire map cannot be colored with only three colors. This argument proves again that three colors are not always enough, but this time the proof does not use four mutually touching countries. That is: many touching countries imply the need for many colors, but not conversely—many colors might be necessary for more subtle combinatorial reasons, and that's one of the sources of difficulty with the four-color problem.

In 1976 Appel and Haken offered a proof of the four-color theorem. Their main achievement was to reduce the problem to a finite one: they showed that once the answer is known for a finite list of specific maps (a very large list, to be sure), then it is known for all. That reduction is the first step of their proof; the second step is a computation. The reduction is standard mathematical reasoning—in principle (though definitely not in detail) it was known long ago. That is: it was known what kind of reduction was necessary, and many of its steps had already been taken. The computation was then carried out electronically, and, when the systematic exhaus-

tion procedure (which needed over 1000 hours of computer time on the first run through) encountered no obstacles, the computer ran up a flag and blew a fanfare of trumpets. Victory: the answer is yes, four colors suffice.

After the excitement died down, several grumbles and more severe complaints began to be heard: the program was never checked, some people said that it was uncheckable, and there were even dark rumors of errors, plain old-fashioned mathematical errors, gaps in the reduction proof. Grumbles and rumors such as that may be interesting, but right now they are beside the point. Assuming that the reduction is correct, assuming that the program is correct, and assuming that the electronic components functioned completely correctly (and they always do, don't they?), I want to ask: what did we learn from the proof? What do we know now that we didn't know before?

I do not find that an easy question to answer. To be sure: I am not going to spend my time looking for a counterexample to the four-color assertion. The printout had at least that practical effect: it discouraged attempts to prove it wrong. Except for that, however, I feel that we, humanity, learned mighty little from the proof; I am almost tempted to say that as mathematicians we learned nothing at all. Oracles are not helpful mathematical tools.

The most celebrated and difficult outstanding problem of mathematics is the Riemann hypothesis. There are many important theorems, with correct proofs on record, that are of the form "if the Riemann hypothesis is true, then . . . ." It would be good to know a proof (or disproof?) of the Riemann hypothesis, but till one is discovered the theorems in which it appears as a hypothesis can be illuminating and useful. That value judgment, however, does not change my opinion of oracles. If an oracle told me that the Riemann hypothesis is true, I don't think that my soul would be any richer for that single syllable.

The development of mathematics shortens proofs, gives insight, and deepens understanding by the discovery of ever new concepts and by the resulting subsumption of old ones under a suitably general theory that took years, decades, and sometimes centuries of labor to construct. Technical terms (Banach space, Artinian ring, fiber bundle) may frighten and alienate non-specialists, but they are just short phrases that abbreviate and clarify the effort, the work, and the thought of our great predecessors.

We may still be far from finding a "good" proof of the four-color theorem. We need a simple insight into a new and complicated kind of geometry or intricate algebra, and the distance from there to a purely conceptual, existential proof of the four-color theorem is probably just as great. I hold as an article of faith, however, that we have seen and travelled greater distances than that; the distance from the Peano axioms to the Atiyah-Singer index theorem could, for instance, be one of them. I believe that the computer (and, for another example, the 10,000 pages of published proof solving the simple groups problem) missed the right concept and the right approach. Their time will come. A hundred years from now both theorems (maps and groups) will be exercises in first-year graduate courses, provable in a couple of pages by means of the appropriate concepts, which will be completely familiar by then. Down with oracles, I say—they are of no use in mathematics.

I cannot, however, stop with that dictum—in all honesty, I should report that Appel and Haken do not completely share my religion. Speaking of their work,

they say: "We now know that a proof can be found. But we do not yet (and may never) know whether there is any proof that is elegant, concise, and completely verifiable by a (human) mathematical mind."

One final word. By an explosion I mean a loud noise, an unexpected and exciting announcement, but not necessarily a good thing. Some explosions open new territories and promise great future developments; others close a subject and seem to lead nowhere. The Mordell conjecture, the next explosion below, is of the first kind; the four-color theorem of the second.

Reference: EDM 165.

Explosion 2: Mordell's conjecture. Every grade school child knows that $9 + 16 = 25$, or, in other words, that $3^2 + 4^2 = 5^2$. A microsecond of thought reveals that the equation continues to hold when 3, 4, and 5 are replaced by $3k$, $4k$, and $5k$, no matter what k is. If the point of view is to look for integer solutions of the equation $x^2 + y^2 = z^2$, then there is really no virtue in distinguishing between $3, 4, 5$ and $6, 8, 10$. An efficient way of saying the same thing is to divide the equation through by z^2, so that it becomes $\left(\frac{x}{z}\right)^2 + \left(\frac{y}{z}\right)^2 = 1$, and then replacing $\frac{x}{z}$ and $\frac{y}{z}$ by x and y, so that the equation becomes $x^2 + y^2 = 1$, with the understanding that the solutions sought are rational numbers, not necessarily integers. From that point of view all the positive multiples of the $3, 4, 5$ solution become just one rational solution.

The equation $x^2 + y^2 = 1$ has many rational solutions, infinitely many—or, in geometric language, the curve that the equation defines has an infinite number of rational points on it. Those points are related to the so-called Pythagorean triples, and they are quite easy to find. The situation is different for the Fermat equations of higher degree; in fact, the celebrated infamous Fermat problem is to show that if $n > 2$, then the curve defined by the equation $x^n + y^n = 1$ has no rational points on it at all (except the trivial ones for which either x or y is 0).

The Fermat problem belongs to a broader context suggested by Louis Mordell in 1922. Mordell's genius in formulating the conjecture was to make the Fermat problem simultaneously harder and easier. Harder: enlarge the set of curves considered (so as to include all algebraic curves of genus greater than 1 over the field of rational numbers—a description that includes the Fermat curves of degree greater than 3). Easier: relax the conclusion to be drawn (so as to allow the existence of rational points, but, to be sure, only a finite number of them) Mordell's conjecture is a conjecture no longer—it was proved in 1983 by Gerd Faltings. Consequence: Fermat's problem may still turn out to have unexpected solutions, but, in any event, for each exponent it has only finitely many!

Reference: MI 6/2/41

Development 1: Ergodic theory. If you shoot a billiard ball perpendicularly into a side of the table, it will bounce straight back to the opposite side and then keep on bouncing back and forth between those two sides. If instead you shoot it straight from the middle of one side to the middle of an adjacent side, then it will bounce out at the same angle as it came in, hit the middle of the side opposite the one you started with, bounce onto the middle of the fourth side, and keep on going cyclically around like that. You can probably think of many other such

patterns—a carefully specialized original aim will produce a periodic pattern. What, however, happens if you shoot the ball at an angle that stands in an irrational relation to the sides of the table, or, more clearly put, if you shoot the ball at an angle that produces a nonperiodic orbit? Boltzmann, one of the founders of ergodic theory in the preceding century, thought about such things and formulated, among other things, what has come to be called the ergodic hypothesis. According to a primitive version of that hypothesis, the path of the (irrationally aimed) billiard ball will eventually pass through every point on the billiard table.

A little cogitation about the topology of a rectangle in the plane versus the topology of a curve in that rectangle might convince you that that version of the ergodic hypothesis is unlikely at the very least—the applicable laws of mechanics are not likely to produce plane filling curves. A modified version of the ergodic hypothesis came along: it was the conjecture that the path of the billiard ball, while it may not go through every point, will in any event get as near as you like to every point—that, in other words, it will be everywhere dense in the billiard table.

Statistical mechanics occupied itself with meditations such as these till G. D. Birkhoff came along with his seminal paper in 1931. What Birkhoff proved was that if you take any decent subset of the table, such as the right half, or the interior of a circle whose center is the center of the table—whatever—the average time that the billiard ball will spend in that subset, no matter where it started, is proportional to the area of the subset. The technically difficult part of his achievement is that it makes sense to speak of the average time—in other words that the limit involved in the meaning of that phrase actually exists—but the striking part of the conclusion is that the average time is a constant, independent of where the ball started. That conclusion implies, among other things, that if you look at the interior of any circle drawn on the table, no matter where it is drawn and no matter how large or how small it is, the billiard ball is certain to enter that interior over and over again—and hence, in particular, that the path of the ball is indeed everywhere dense.

Birkhoff's work was published well over 50 years ago, and since then ergodic theory has become a major part of mathematics that has rich connections with both classical and modern analysis and that is definitely one of the major steps forward that was taken in the last 75 years. (Family comment: G. D. Birkhoff's son is Garrett Birkhoff, both at Harvard in their days.)

Reference: EDM 146/B.

Development 2: Transcendental numbers. Is the number $2^{\sqrt{2}}$ rational? If not, is it at least algebraic (that is, a solution of a polynomial equation with integer coefficients)? The seventh of Hilbert's famous problems was about that question and the general context to which it belongs. A good way to formulate the general question is to ask when numbers of the form α^{β} are transcendental. If $\alpha = 0$, the question degenerates, and the same is true if $\alpha = 1$; similarly if β is either 0 or 1, the question is not one anybody wants to ask. If, more generally, β is rational, then the question transforms into easy and generally known subquestions concerning algebraic numbers. On the other hand, if either α or β is transcendental, then the answer sought for appears to be too close to the data. In view of these comments, the "right" question to ask is this: if α and β are algebraic numbers, with α different from both 0 and 1, and β irrational, does it follow that α^{β} is transcen-

dental? The answer turns out to be yes; it was obtained, more or less simultaneously and independently, by A. O. Gelfond ($\neq$ I. M. Gelfand) and Theodor Schneider in 1934. The question about $2^{\sqrt{2}}$ is covered: since $2 \neq 0$, $2 \neq 1$, and $\sqrt{2}$ is irrational, it follows that $2^{\sqrt{2}}$ is transcendental.

The theory didn't stop there. Here, for instance, is an n-fold generalization, a sample of some of the deep results obtained by Alan Baker in a sequence of papers in the late 1960's (for which he got the Fields prize). If $\alpha_1, \ldots, \alpha_n$ are algebraic numbers different from both 0 and 1, and if $\beta_1, \ldots, \beta_n$ are algebraic numbers such that $1, \beta_1, \ldots, \beta_n$ are linearly independent over the field of rational numbers, then $\alpha_1^{\beta_1}\alpha_2^{\beta_2} \cdots \alpha_n^{\beta_n}$ is transcendental.

Reference: EDM 414/D.

Development 3: The continuum hypothesis. In 1900 there was an International Congress of Mathematicians in Paris, and that's where Hilbert presented his list of 23 problems. The first of those problems was the continuum hypothesis, originally formulated by Cantor. The simplest statement of Cantor's continuum hypothesis (there are other, more general, statements of great interest) is that every uncountable set of real numbers is in one-to-one correspondence with the set of all real numbers, or, in Cantor's notation, that there is no cardinal number between $\aleph_0$ and $2^{\aleph_0}$.

Is the continuum hypothesis true? The question has often been likened to a similar one about Euclid's parallel postulate, and the answer has shocked and annoyed a lot of people, just as the Bólyai-Lobachevsky resolution of the status of the parallel postulate shocked and annoyed many of our grandfathers. In both cases there is a more or less pleasant axiom system (in the present case it is the Zermelo-Fraenkel system of axioms for set theory) and a less pleasant, more complicated, non-obvious additional axiom. If the extra axiom is a consequence of the basic ones, it is true, and all is well; if its negation is a consequence of the basic ones, it is false, and, for better or for worse, the question is definitively answered. The answer, awaited for a long time, turns out to be a subtle and profound intellectual achievement. Gödel proved in 1940 that the continuum hypothesis is not false—it is consistent with the other set-theoretic axioms—and Paul Cohen proved in 1964 that it is not true—it is independent of the other axioms, or, in other words, its negation is also consistent.

Both Gödel and Cohen argue by the construction of a suitable model, but use very different techniques. Gödel starts with a universe of sets that satisfies the Zermelo-Fraenkel axioms and shows that there exists a subuniverse that also satisfies them, and in which, moreover, the continuum hypothesis is true. Cohen's argument is similar but harder. It is reminiscent of Felix Klein's construction of a Lobachevskian plane that endows a Euclidean disk with a new metric. Cohen, like Gödel, starts with a model of set theory and then enlarges it, adjoins new objects to it, in such a way as to "force" the continuum hypothesis to be false. ("Forcing" has become an important technical word in the subject.)

Where does that leave the continuum hypothesis? Many people believe, as Gödel did, that despite the independence of the continuum hypothesis it is in some legitimate sense either true or false—that humanity has not yet thought of the right way to describe the full truth about set theory, and when the full truth comes to be known, when the appropriate additional axioms are found and adjoined to

the present ones, then the continuum hypothesis will become either provable or disprovable. Both schools of thought exist, and you are free to join whichever one you find more attractive. Would it influence your decision if you knew what Gödel believed? He thought it would become disprovable.

Reference: EDM 35/D.

Development 4: Lie groups. Hilbert's fifth problem asked whether some relatively mild assumptions about topological groups were enough to imply a strong conclusion. A topological group is a set that is both a group and a topological space in which the two structures are compatible in the sense that the group operations (multiplication and inversion) are continuous.

A typical example is the set of all 2×2 matrices of the form

$$\begin{bmatrix} x & y \\ 0 & 1 \end{bmatrix}$$

with $x > 0$; the topological structure is that of the right half plane (all (x, y) with $x > 0$), and the multiplicative structure is the usual one associated with matrices. This example has an important special property: it is "locally Euclidean" in the sense that every point has a neighborhood that is homeomorphic to an open ball in 2-dimensional Euclidean space. (Equivalently: every point has a "local coordinate system".) An even more important special property of the example is that the group operations, regarded as functions on the appropriate Euclidean space, are not only continuous but even analytic. That is obvious at a glance; if the matrix

$$\begin{bmatrix} x & y \\ 0 & 1 \end{bmatrix}$$

is identified with the ordered pair (x, y), then

$$(u, v)(x, y) = (ux, uy + v),$$

and

$$(x, y)^{-1} = \left(\frac{1}{x}, -\frac{y}{x}\right).$$

If a group is locally Euclidean, that is, if it can be "coordinatized" at all, then there are many ways of coordinatizing it; if at least one of them is such that the group operations are analytic, the group is called a "Lie group". Hilbert's fifth problem was this: is every locally Euclidean group a Lie group?

The problem is analogous to one in complex function theory. It is relatively elementary that a twice-differentiable function is analytic; the hard theorem is that the conclusion holds under much weaker assumptions. Similarly, it has been known for a long time that if a topological group has sufficiently differentiable coordinates, then it has analytic ones; the Hilbert problem is to draw the same conclusion under much weaker assumptions.

Immediately after the discovery of Haar measure, von Neumann (1933) applied it to prove that the answer to Hilbert's question is yes for compact groups. A little later Pontrjagin (1939) solved the abelian case, and Chevalley (1941) solved the solvable case. The general case was solved in 1952 and 1953 by Gleason, and Yamabe, and, jointly, by Montgomery and Zippin; the answer to Hilbert's question is yes. Gleason offered a new characterization of Lie groups; Montgomery and

Zippin used geometric-topological tools (and Gleason's result) to reach the desired conclusion.

Reference: EDM 406/N.

Development 5: Simple groups. Every group has two obvious normal subgroups, namely the group itself and, at the other extreme, the subgroup that consists of the identity element only. A group is called simple if these are the only normal subgroups it has.

Simple groups are like prime numbers in two ways: they have no proper parts, and every finite group can be constructed out of them. Consider, indeed, an arbitrary finite group, and look at a maximal normal subgroup of it, that is, a normal subgroup that is not included in any other proper normal subgroup. If the group is simple, then that maximal normal subgroup is just the identity, but in any event, no matter what that subgroup is, its maximality implies that the quotient group obtained by dividing out by it is simple. The relation among group, normal subgroup, and quotient is sometimes described by saying that the original group is an extension of the quotient group by the subgroup. In this language, every finite group (except the trivial group, the identity) is an extension of a simple group by a group of strictly smaller order. That statement is the group-theoretic analogue of the number-theoretic one that says that every positive integer (except 1) is the product of a prime by a strictly smaller positive integer.

If the maximal normal subgroup is not trivial, then the procedure just used can be used again; the result is a maximal normal subgroup of the maximal normal subgroup such that the original subgroup is an extension of the second quotient by the second subgroup. The procedure can be repeated so long as it produces non-trivial subgroups; the end product is a decreasing chain (a composition series) of subgroups of the original group with the property that each quotient group obtained by dividing a term of the chain by its successor is simple. A great part of the problem of getting to know all finite groups reduces in this way to the determination of all finite simple groups.

The abelian ones among the finite simple groups are easy to determine—it's a classroom exercise to show that they are just the cyclic groups of prime order. That's the only easy part of this subject. What's hard is to find all non-abelian simple groups. Some examples of simple groups are not hard to come by; among permutation groups, for instance, the most famous ones are the alternating groups of degree 5 or more. For a long time the known simple groups did not exhibit any pattern, and even the simplest questions about them resisted attack. Burnside, for instance, conjectured (in 1911) that every finite non-abelian simple group has even order, and that conjecture stood as an open problem for more than 50 years.

In a spectacular display of group-theoretic power, Feit and Thompson settled Burnside's conjecture (in 1963)—it is true. The proof occupies an entire issue (over 250 pages) of the *Pacific Journal*. It is technical group theory and character theory. Some reductions in it have been made since it appeared, but no short or easy proof has been discovered. The result has many consequences, and the methods also have been used to attack many other problems in the theory of finite groups; a subject that was once pronounced dead by many has shown itself capable of a vigorous new life. Thus, for example, Burnside's dream is by now completely realized: by a gigantic collaborative effort of many mathematicians all over the planet all finite simple groups have been found and can be explicitly described.

Reference: EDM 160/D.

Development 6: The Atiyah-Singer index theorem. How can a linear transformation on a finite-dimensional vector space fail to be invertible? There are two obvious answers: it can fail to be injective (one-to-one) or else it can fail to be surjective (onto). The first kind of failure is the existence of a nontrivial kernel, and the second kind is the existence of a nontrivial cokernel. The kernel of a transformation T is, of course, the null-space, the inverse image of 0; the cokernel is, roughly, the complement of the range, or, precisely, the quotient of the entire space modulo the range. It is an elementary fact of linear algebra that the two kinds of failure always occur simultaneously, and that, in fact, the numerical measures of the two kinds of failure, that is, the dimensions of the kernel and the cokernel, are always equal.

For infinite-dimensional spaces the story is different. If, for instance, T is the unilateral shift on the infinite-dimensional sequence space l^2, the transformation defined by

$$T\{\xi_1, \xi_2, \xi_3, \dots\} = \{0, \xi_1, \xi_2, \xi_3, \dots\},$$

then ker T is the trivial subspace $\{0\}$, but ran T is the subspace of all vectors whose first coordinate is 0, and consequently dim ker $T = 0$ and dim coker $T = 1$.

The *index* of any linear transformation T (including the one just described) is defined by

$$\text{index } T = \dim \ker T - \dim \operatorname{coker} T,$$

whenever that makes sense (that is, whenever it is not equal to $\infty - \infty$). For linear transformations on finite-dimensional spaces the index is always 0, but on infinite-dimensional spaces the index (being a measure of the extent to which a transformation is irreparably non-invertible) can give interesting information.

The Atiyah-Singer index theorem has to do with the "analytic" index just defined. The first instance of it that most of us run into is the Cauchy integral formula thought of as the computation of a winding number. Another instance is the Riemann-Roch theorem. Compact Riemann surfaces such as the sphere and the torus arise in complex function theory, and the Riemann-Roch theorem (which goes back to a paper by Roch in 1865) is about the dimensions of certain vector spaces of meromorphic functions on such Riemann surfaces; its conclusion is a formula for that dimension. The general Atiyah-Singer theorem is a generalization of the Riemann-Roch theorem. It deals with smooth compact manifolds more general than Riemann surfaces. Elliptic differential operators acting on smooth functions defined on such manifolds have two numbers associated with them. One of them is the analytic index, as above, and the other is the topological index, which is something else. The topological index is related to K-theory, and, in particular, in more classical contexts, to the Euler characteristic. The achievement of the Atiyah-Singer theorem (1963) is that those two indices have the same value —that, in other words, a property of great analytic importance, with a purely analytic definition, is in fact almost completely determined by the topological properties of the underlying manifold.

The equation just referred to is only a small part of the joint work of Atiyah and Singer. The results of that work are the deepest and broadest, and, for me as a reporter, much the hardest of the contents of this report; they are not just a theorem but a theory, a context, a point of view that enters, influences, and is influenced by many parts of mathematics. Writing about the spectacular successes of research in differential geometry during the last fifty years, Osserman called the

Atiyah-Singer index theorem "a grand synthesis of analysis, topology, and geometry leading, in particular, to a new way of viewing the Gauss-Bonnet theorem: not as an isolated result, but as one instance of a larger scheme of things." That synthesis was probably a major factor in Michael Atiyah's being knighted in England and Isadore Singer's being awarded a presidential medal in this country.

Reference: EDM 236/H.

Development 7: Fourier series. It is a historical misfortune (which was responsible for almost 200 years of barking up the wrong tree) that Fourier series were discovered before convergence. Fourier series are a vital part of both classical and modern analysis; they are important for both abstract theory and concrete applications. They arise in topological groups and in operator theory; they have their origins in problems about vibrating strings and about heat conduction.

In their most classical manifestation, Fourier series have to do with numerical-valued functions on the line (it is usually best to let them be complex-valued) that are periodic of period 2π and integrable on the interval $[0, 2\pi]$. The Fourier series of such a function is the infinite linear combination of the exponential functions e^{inx}, $n = 0, \pm 1, \pm 2, \ldots$, with coefficients determined by integrating the given function against them. (Alternatively the Fourier series uses sines and cosines running in one direction only; the complex version, however, is algebraically simpler to manipulate.)

Trigonometric polynomials (in either real or complex form) are familiar objects and are computationally accessible; surely nothing but good could come out of representing more difficult functions as limits of such polynomials. It seemed natural, therefore, to hope that the "sum" of the Fourier series associated with a function f would be "equal" to f, and, in any event, to ask for which functions that would occur. The answer, it was hoped, was that good functions would have good series, and the history of the subject has been strongly influenced by that hope.

When limits began to be understood, "sum" and "equal" were interpreted in the sense of pointwise convergence; the more fruitful and usable concepts of weak convergence and convergence with respect to a norm came along only after the mathematical community was irretrievably committed to research in the pointwise direction.

How good does a good function have to be? Differentiability is good enough, but, it turns out, continuity is not; there are continuous functions whose Fourier series diverges at a point (du Bois Reymond, 1876), and, in fact, at many points. Must the Fourier series of a continuous function converge almost everywhere? That was an unsolved problem for many years.

Kolmogorov showed that if all that is assumed is that f is integrable on $[0, 2\pi]$, then it could happen that the Fourier series of f diverges almost everywhere (1923), or even everywhere (1926). The biggest question along these lines was asked by Luzin, and it remained unanswered for 50 years: if f is square integrable on $[0, 2\pi]$, does it follow that the Fourier series of f converges to f almost everywhere? Repeated failure to prove the affirmative led to the official state religion among the cognoscenti in the 1950s and 1960s: the answer must be no.

The answer is yes. The first proof is due to Carleson (1966). A remarkable feature of Carleson's achievement is that it uses no unknown techniques; it just uses the known ones better. It depends on an ingenious push-me-pull-you way of

selecting subintervals. It is as if Carleson had power enough to replace everyone else's ε by ε^2, and that did the trick.

Reference: EDM 167/H.

Development 8: Diophantine equations. Hilbert's tenth problem concerned the solvability of Diophantine equations. The problem was to design an algorithm, a computational procedure, for determining whether an arbitrarily prescribed polynomial equation with (positive) integer coefficients has (positive) integer solutions. (The modifier "positive" is technically convenient and does no harm.)

What does it mean to say there is an algorithm for deciding solvability? A reasonable way to answer the question is to offer a definition of computability for sets and functions, and then to define an algorithm in terms of computability. The concept of computability has received a lot of attention; it has several different but logically equivalent definitions all in accord with the intuitive notion that the word suggests.

Suppose now that $\{E_1, E_2, E_3, \dots\}$ is an effective enumeration of all the polynomial equations under consideration, and let S be the set of those indices k for which E_k has a solution. Hilbert's problem (is there an algorithm?) can be expressed by asking whether S is a computable set. The answer is no. The answer was a long time coming: it is the result of the cumulative efforts of J. Robinson (1952), M. Davis (1953), H. Putnam (1961), and Y. Matijasevič (1970).

The central concept in the proof is that of a Diophantine set, and the major step proves that every computable set is Diophantine. The techniques make ingenious use of elementary number theory (e.g., the Chinese remainder theorem, and a part of the theory of Fibonacci numbers). The proof exhibits some interesting Diophantine sets whose Diophantine character is not at all obvious (e.g., the powers of 2, the factorials, and the primes).

One way to prove that S (the index set of the solvable equations) is not computable is by contradiction. If S were computable, then it would follow (by a slight bit of additional argument) that each particular Diophantine set (i.e., the solution set of each particular equation) is computable, and hence (by the "major step" mentioned above) that the complement of every Diophantine set is Diophantine. The contradiction is derived by exhibiting a Diophantine set whose complement is not Diophantine.

The last step uses a version of the familiar Cantor diagonal argument. The idea is "effectively" to enumerate all Diophantine sets, as $\{D_1, D_2, D_3, \dots\}$ say, to prove that the set $D^* = \{n: n \in D_n\}$ is Diophantine (that takes some argument), and, finally, to prove that the complement of D^* is not Diophantine—and that's where Cantor comes in.

Reference: EDM 100.

Development 9: Banach bases. In calculus (beginning in the 17th century) we learn how to find maxima and minima of numerical functions defined on familiar domains such as intervals in the line and rectangles in the plane. In a later subject called the calculus of variations (beginning in the 18th century) we try to find maxima and minima of numerical functions whose domains consist of sets of functions. Most famous example: for each path joining two points in space there is a definite time that it takes for a particle to slide along the path from the first point to the second—what is the minimum of all those times? (The minimum time

is attained for the celebrated path called the brachistochrone—*brachisto* for shortest and *chrone* for time.) Problems such as that were among the ones that gave rise to functional analysis, the part of analysis in which the domains of the functions studied are sets of functions, subsets of function spaces, and in which algebraic and topological methods are freely used.

Another point of view on functional analysis is that it is an infinite-dimensional generalization of linear algebra. The first idea was to replace vectors and the additions involved in applying matrices to them by functions and the integrations involved in applying kernels to them. The first systematization of the subject (around the 1920's) was proposed by Banach and Wiener (independently)—the result was the theory of Banach spaces, and its many wildly generalized outgrowths. The study of Banach spaces is a typical instance of the axiomatic method in mathematics—abstract and general, with roots in the concrete and special. It is called too general by some and not general enough by others. In any event it is still a living part of mathematics, and major progress was made in it relatively recently.

One of the earliest questions about Banach spaces was the basis problem, raised by Banach himself in his book (1932). A sequence of elements in a Banach space is a Schauder basis for the space in case every vector has a unique expression as a convergent infinite linear combination of the terms of the sequence. The countability built into the definition of the word "sequence" implies that if a Banach space has a basis, then it is separable (that is, has a countable dense set). The basis problem, which was outstanding for 40 years, was the converse: does every separable Banach space have a basis? Each space that ever came up in analysis had one, and yet a proof that that had to be so remained elusive.

A classically important concept in the study of Banach spaces is that of a compact (completely continuous) operator, that is, a linear transformation between Banach spaces with the property that the closure of the image of the unit ball is compact. The easy compact operators are the ones of finite rank (that is, the ones for which the range is finite-dimensional); the next easiest ones are the (uniform) limits of operators of finite rank. If a Banach space is "good", then every compact operator mapping into it is such a limit (in technical language: the space has the approximation property), and, in particular, if a Banach space has a basis, then it has the approximation property.

The basis problem was solved by Per Enflo in 1973. The solution turned out to be negative: there exists a separable Banach space that does not have the approximation property. The technique is constructive; it is a combinatorial way of constructing and putting together infinitely many finite-dimensional Banach spaces.

Reference: EDM 39/A.

Development 10: Manifolds. A 2-manifold is a topological space that is locally like 2-dimensional Euclidean space (that is, every point has a neighborhood homeomorphic to an open disk in the plane) with the properties that the locally Euclidean patches are glued together continuously and that there are not too many of them (meaning that the entire space is separable). The definitions of 3-manifolds, 4-manifolds, 5-manifolds, etc. are exactly the same—just replace 2 by 3, 4, 5, etc., (and replace "plane" by 3-space, 4-space, 5-space, etc.).

A 2-manifold can be large (it can, for instance, be the entire plane), and even if it doesn't look large it may be intrinsically large in the sense that many paths in it converge to points that don't exist in it—as is the case with an open disk. The manifolds whose study is more promising are the compact ones.

Can all possible compact 2-manifolds be listed? Yes, and the simplest ones to list are the so-called orientable ones. One of them is simply connected (the sphere), another has genus 1 (the torus), another has genus 2 (a doughnut with two holes), etc.—and all possible orientable 2-manifolds are like one of these (in the sense of homeomorphism). The classification problem for compact 2-manifolds is classical and was solved long ago. (The 1-dimensional case can safely be left as homework.)

For higher dimensions things are tougher to see and tougher to do. Surprisingly, however, when the dimension is 5 or greater, much of the answer is known—simply connected 5-manifolds are understood and classified by homotopy theory. For 3-manifolds the problem is unsolved—that's what the celebrated Poincaré hypothesis is about. For 4-manifolds the story is different. Two great victories were won by Michael Freedman and Simon Donaldson in 1982 (and each was awarded a Fields medal at the Berkeley Congress in 1986).

It turns out that every oriented 4-manifold is associated with a symmetric integer matrix ("the intersection matrix") of determinant ± 1, and Freedman showed that all such matrices can indeed occur. The correspondence between matrices and manifolds is one-to-one half the time and two-to-one the rest of the time. The final result is a complete classification of all simply connected oriented 4-manifolds. If we think of a basketball (the surface of a sphere) as a 2-dimensional object (never mind that we usually see it embedded in 3-dimensional space), Freedman's work gives us as much insight about 4-dimensional basketballs as we have about the honest-to-goodness ones.

Donaldson, on the other hand, showed that if a simply connected oriented 4-manifold has a smooth differential structure and if its matrix is positive definite, then the matrix must be equivalent to the identity matrix. That's a very strong conclusion; it shows that the theories of topological and differential 4-manifolds are radically different.

Presenting Freedman's work, John Milnor said that the proofs of the results are extremely difficult. Presenting Donaldson's work, Michael Atiyah said that it opened up an entirely new area, and he went on to say that Donaldson's youth and mathematical power are "an indication that mathematics has not lost its unity, or its vitality".

Reference: MI 5/3/39.

Development 11: The Bieberbach conjecture. Some problems are interesting not because they are interesting but because they are elusive. The Fermat question is probably the best case in point. Nobody really wants to know the mere answer to that one—what mathematicians want to know is why they don't know the answer. The Bieberbach conjecture about schlicht functions was, in context, perfectly natural to the specialist—to most outsiders, however, it seemed like a curious technicality whose main claim to fame was that it remained unanswered.

What it concerned was the class of all those functions that are given by power series of the form

$$z + \sum_{n=2}^{\infty} a_n z^n$$

and that are schlicht (meaning analytic and injective or, in the more classical word, univalent) in the open unit disk. The class constitutes a "normal family" (a

compactness property), which implies that, for each n, the coefficient a_n remains bounded throughout the class. Bieberbach, studying the subject in 1916, proved that in fact $|a_2| \leqq 2$ for every function in the class. One particular function in the class, the Koebe extremal function defined by $a_n = n$, shows that the upper bound 2 is best possible (and is, in fact, attained). The Bieberbach conjecture was that, more generally, $|a_n| \leqq n$ for all n.

Löwner proved in 1923 that the conjecture is true for $n = 3$; in 1955 Garabedian and Schiffer proved it for $n = 4$; in 1968 Pedersen and Ozawa proved it for $n = 6$; in 1972 Garabedian, Pedersen, and Schiffer proved it for $n = 5$; and in 1973 Ozawa and Kubota proved it for $n = 8$. Progress was slow and not promising.

The breakthrough that transformed the conjecture into a theorem came in 1984 when Louis de Branges offered a proof of the general case, a proof of several hundred pages, based on his theory of square summable power series. His original proof contained minor errors, but they were correctable and soon corrected. The experts were still a little bothered, but their discomfort didn't last too long. What they didn't like was the reliance of the proof on the seemingly irrelevant methods of functional analysis, and, sure enough, ultimately the Leningrad seminar on geometric function theory produced a proof that Bieberbach would have liked. The proof is shorter and more perspicuous than its special predecessors for the cases $n = 5$ and $n = 6$. Square summable power series are gone, but the main structural insights of de Branges are still the ones that make things go. We don't, glory be to glory, lose them all.

Reference: MI 8/1/40.

Epilogue. The answer to the question in the title is clearly and decisively no.