1. REPRINT

PROOF, TRUTH, AND CONFUSION

Saunders Mac Lane
Max Mason Distinguished Service Professor of Mathematics
The University of Chicago

The 1982 Nora and Edward Ryerson Lecture
at The University of Chicago.

Saunders Mac Lane
Introduction

by

Hanna H. Gray
President of the University

Each year, one of the annual events at the University is the selection of the Nora and Edward Ryerson Lecturer. The selection is made by a committee of faculty which receives nominations from their faculty colleagues. Each year, this committee comes up with an absolutely superb selection for the Ryerson Lecturer, and this year is triumphant confirmation of that generalization.

The selection emanates from faculty nomination and discussion, and it is analogous to the process of the selection of faculty in this University, representing a selection based on the work and contribution, on the high esteem for the intellectual imagination and breadth of a colleague. The peer review process, in this as in faculty appointments, stresses scholarship and research, stresses the contribution of a member of the faculty to the progress of knowledge. In addition, of course, the faculty appointment process looks also to teaching and to institutional citizenship.

If I had the nerve to fill out an A-21 form for Professor Saunders Mac Lane, I would, I think, be creating a new mathematics because I would award him 100 percent for research, 100 percent for teaching, and 100 percent for contribution or citizenship. Even I can add that up to 300 percent.

Now, of course, in the evaluation of younger scholars for junior appointments similar judgments are made. They are based on the same three categories, and they are judgments about the promise of continuing creativity, continuing growth, continuing intellectual contribution. That judgment of the young Saunders Mac Lane was made a very long time ago in Montclair, New Jersey. Montclair, New Jersey is the home of the Yale Club of Montclair. I once had the enormous privilege of being invited to the Yale Club of Montclair where I was given something called the Yale Bowl, which had on it an inscription testifying that I had earned my “Y" in the “Big Game of Life."

In 1929, the young Saunders Mac Lane went to the Yale Club of Montclair, I was told—he was then finishing his senior year at Yale—and there was a young dean of the Yale Law school who was leaving in order to go to the University of Chicago as president. And the Yale Club of Montclair, which usually gave its awards to football players, decided on this occasion to give recognition to a young mathematician and to a young law school dean, and that was where Mr. Hutchins and Mr. Mac Lane met.

As Saunders Mac Lane graduated from Yale, Mr. Hutchins encouraged him personally to come to Chicago. And Saunders came. He had, however, neglected to take steps that are usually taken when one travels to enter another university, and the chairman of the Department of Mathematics, Mr. Bliss, had to say to him rather directly, “Young man, you’ve got to apply first."

He did apply, and fortunately he was accepted. Within a year, he had received his M.A. from Chicago and had come into contact with lots of extraordinary people, but two very extraordinary people in particular. One was the great mathematician E. H. Moore, and the other was a graduate student in economics named Dorothy Jones, who was in 1933 to become Mrs. Mac Lane.

Now, those of us who know Saunders think of him as a Hyde Parker, and indeed as a Hyde Parker forever. And, of course, he is a Hyde Parker, and a Hyde Parker forever, but he had a period in his life, I have to tell you, after he had taken his M.A., when he became something of an academic traveler. We really ought to have been able to trace those travels when we think about Saunders’ choice of costume.

Now, that’s not easy to figure out today because I think that necktie came out of a safe this morning.

But if you think about the flaming reds, for example, that Saunders affects, you are perhaps reminded of Cambridge, Massachusetts. If you think of the Scottish plaids which he affects, that's a harder one, because I would say that that has to do with the great tradition which took him to New England and made him for a time a resident of Connecticut. And then, of course, there is the Alpine hat which could only have come from Ithaca, New York. Saunders received his doctorate from the University of Göttingen in 1934. He had spent the years 1933-34 again at Yale as a Sterling Fellow. He then spent two years at Harvard. He then spent a year at Cornell. Then he came to the University of Chicago for a year. And then again he moved, called back to Harvard as an assistant professor, and there he rapidly went through the ranks. Fortunately, in 1947, he returned to the University of Chicago and, in 1963, became the Max Mason Distinguished Service Professor. Between 1952 and 1958, he succeeded Marshall Stone as chairman of the Department of Mathematics for two three-year terms, and he has served the University as he has his department with total dedication.

Saunders has extended his role beyond our University, serving primarily and prominently in a number of national scholarly organizations and institutions devoted to large questions of the relationship of learning to policy. He was president of the Mathematical Association of America and received its Distinguished Service Award in 1975 in recognition of his sustained and active concern for the advancement of undergraduate mathematical teaching and undergraduate mathematics. He was also president of the American Mathematical Society in 1973-74. He has been a member of the National Science Board and vice-president of the National Academy of Sciences.

His work in mathematics, of course, has been widely recognized. Alfred Putnam, who studied with Saunders at Harvard, had this to say of Saunders in a biographical sketch that he has published. He wrote, "Beginning as a graduate student with a brief exposure to group extensions, I've watched the development of Saunders Mac Lane's mathematics through homological algebra to category theory. Saunders Mac Lane belongs in a category by himself."

And so he does. So he does as a mathematican, as an academic citizen, as a spokesman for the fundamental values and principles of the University, and, of course, in sartorial wonder.

Now, it is to this category that we look for the Nora and Edward Ryerson Lecturers. When the Trustees established the lectureship in 1973, they sought a way to celebrate the relationship that the Ryersons and their family have had with our University—a relationship of shared values and a commitment to learning at the most advanced level.

Mr. Ryerson was elected to the Board in 1923 and became Chairman of the Board in 1953. Nora Butler Ryerson was a founding member, if not *the* founder, of the University's Women's Board. Both embraced a civic trust that left few institutions in our city untouched, and they passed to future generations of their family the sense of engagement and participation.

Saunders Mac Lane, through his staunch loyalty to our University, his broad interest in the community of scholars and their work, his distinguished scholarly career, represents these values for us in a special way, and, of course, he is entirely uncompromising also in his commitment to them. It is a pleasure to introduce this year's Ryerson Lecturer, Saunders Mac Lane.

Proof, Truth, and Confusion

Saunders Mac Lane

I. The Fit of Ideas

It is an honor for a mathematician to stand here. Let me first say how much I appreciate the initiative taken by the trustees on behalf of the Ryerson family in providing for this series of lectures, which afford opportunity for a few fortunate faculty members to present aspects of their scholarly work which might be of interest to the whole university community. In my own case, though the detailed development of mathematics tends to be highly technical, I find that there are some underlying notions from mathematics and its usage which can and will be of general interest. I will try to disentangle these and to relate them to the general interest.

This intent accounts for my title. Mathematicians are concerned to find truth, or, more modestly, to find a few new truths. In reality, the best that I and my colleagues in mathematics can do is to find proofs which perhaps establish some truths. We try to find the right proofs. However, some of these proofs and the techniques and numbers which embody them have turned out to be so popular that they are applied where they do not belong—with results which produce confusion. For this, I will try to cite examples and to draw conclusions.

This involves a thesis as to the nature of mathematics: I contend that this venerable subject is one which does reach for truth, but by way of proof, and does get proof, by way of the concatenation of the right ideas. The ideas which are involved in mathematics are those ideas which are formal or can be formalized. However, they are not purely formal; they arise from aspects of human activity or from problems arising in the advance of scientific knowledge. The ideas of mathematics may not always lead to truth; for this reason it is important that good ideas not be confused by needless compromise. In brief, the ideas which matter are the ideas that fit.

However, the fit may be problematical. A friend of mine with a vacation home in Vermont wanted to suitably decorate his barn, and so asked the local painter to put on the door "the biggest number which can be written on the broad side of a barn door." The painter complied, painting on the barn door a digit 9 followed by as many further such digits as could be squeezed onto the door (Figure 1.a). A competitor then claimed he could do better by painting smaller 9's and so a bigger number. A second competitor then rubbed out the first line and wrote instead: The square of the number 9, ... (Figure 1.b). Even that didn't last, because another young fellow proposed the paradoxical words, "One plus the biggest number that can be written on the broad side of this barn door" (Figure 1.c). At each moment, this produces a bigger number than anything before. We may conclude that there is no such biggest number. This may illustrate the point that it is not easy to get the ideas that fit—on barn doors or otherwise.

II. Truth and Proof

I return to the "truth" of my title. When I was young I believed in RMH–which sometimes stands for Robert Maynard Hutchins, who to my great profit first encouraged me to come to Chicago—and which sometimes stands for the slogan, "Reach Much Higher." At any rate, when young I thought that mathematics could reach very much higher so as to achieve absolute truth. At that time, *Principia Mathematica* by Whitehead and Russell seemed to model this reach; it claimed to provide all of mathematics firmly founded on the truths of logic. The logic in *Principia* was elaborate, symbolic, and hard to follow. As a result, it took me some years to discover that *Principia Mathematica* was not a *Practica Mathematica*—much of mathematics, in particular most of geometry, simply wasn't there in *Principia*. For that matter, what *was* there didn't come exclusively from logic. Logic could provide a framework and a symbolism for mathematics, but it could not provide guidelines for a direction in which to develop.

This limitation was a shocking discovery. Logic, even the best symbolic logic, did not provide all of absolute truth. What did it provide instead? It provided proof—the rigorous proof of one formal statement from another prior statement; that is, the deduction of theorems from axioms. For such a deduction, one needed logic to provide the rules of inference. In addition, one needed the subject matter handled in the deductions: the ideas used in the formulation of the axioms of geometry and number theory, as well as the suggestions from outside mathematics as to what theorems might usefully be proved from these axioms.

Deductive logic is important not because it can produce absolute truth but because it can settle controversy. It has settled many. One notable example arose in topology, a branch of mathematics which studies qualitative properties of geometric objects such as spheres. From this perspective, a smooth sphere and

a crinkly sphere would have the *same* qualitative properties—and we would consider not just the ordinary spheres—two-dimensional, since the surface has two dimensions—but also the spheres of dimensions 3, 4, and higher (Figure 2). For these spheres, topologists wished to calculate a certain number which measures the connectivity—a measure "two dimensions up" from the dimension of the sphere. The Soviet topologist L. Pontrjagin in 1938 stated that this desired measure was one. Others thought instead that the measure was two. In a related connection, the American reviewer of another paper by Pontrjagin wrote, "Both theorems (of Pontrjagin) contradict a previous statement of the reviewer. It is not easy to see who is wrong here." Fortunately, it was possible to see. With careful analysis of the proof, Pontrjagin did see who was wrong—and in 1950 published a statement correcting his 1938 error: that the measure of connectivity two dimensions up is *not* one, but two.

A few years ago, the *New York Times* carried an item about a similar fundamental disagreement between a Japanese topologist and one of our own recent graduate students, Raphael Zahler. Analysis of the deductions showed that Zahler was right. There lies the real role of logic: it provides a formal canon designed to disentangle such controversies.

Truth may be difficult to capture, but proof can be described with complete accuracy. Each mathematical statement can be written as a word or sentence in a fixed alphabet—using one letter for each primitive mathematical notion and one letter for each logical connective. A proof of a theorem is a sequence of such statements. The initial statement must be one of the axioms. Each subsequent statement not an axiom must be a consequence of prior statements in the sequence. Here "consequence" means "consequence according to one of the specified rules of inference"—rules specified in advance. A typical such rule is that of *modus ponens*: Given statments "S" and "S implies T," one may infer the statement "T."

This description gives a firm standard of proof. Actual proofs may cut a few corners or leave out some obvious steps, to be filled in if and when needed. Actual proofs may even be wrong. However, the formal description of a proof is complete and definitive. It provides a formal standard of rigor, not necessarily for absolute truth, but for absolute proof.

There is a surprising consequence: no one formal system suffices to establish all of mathematics. Precisely because there is such a rigorous description of a "proof" in a "formal system," Kurt Gödel was able to show that, in each such system with calculable rules of inference, one could formulate in the system a sentence which was not decidable in the system—that is, a sentence G which can neither be proved nor disproved according to the specified rules of inference. More exactly, this is the case for any system which contains the numbers and the rules of arithmetic, and in which the rules of inference can be explicitly listed or numbered in the fashion called "recursive."

In such a system, all statements are formal and are constructed from a fixed alphabet. Hence we can number *all* the possible proofs. Moreover, we can formulate within the system a sentence which reads, "n is the number of the proof of the statement with the number k." On this basis, and adapting ideas illustrated by the paradox of the barn door, one then constructs another sentence $G = G(p)$ (with number p) which reads, "There is no number which is the proof of the sentence number p." This means in particular that this very sentence G cannot be proved in the system. This is because G itself states that "there is no proof in the system for me"—hence G is true (Figure 3). Hence, unless the system is inconsistent, it can contain no refutation of G. Thus in such a formal system we can write one statement (and hence many) which, though true, is simply undecidable, yes or no, within the system.

This result is startling. It may seem catastrophic—but it turns out to be not quite so disastrous. It shows that there is an intrinsic limitation on what can be proved within *any one formal* system; thus proof within one such system cannot give all of truth. Very well then, as we shall see, there can be more than one formal system and hence more than one way in which to reach by proof for the truth.

III. Ideas and Theorems

Some observers have claimed that mathematics is just formalism. They are wrong. A mathematical proof in a given formal system must be *about* something, but it is not about the outside world. I say it is about ideas. Thus the formal system of Euclidean geometry is about certain "pictorial" ideas: point, line, triangle, and congruence; in their turn, these ideas arose as means of formulating our spatial experiences of shape, size, and extent and our attempts to analyze motion and symmetry.

Each branch of formal mathematics has a comparable origin in some human activities or in some branch of scientific knowledge. In each such case, the formal mathematical system can be understood as the

realization of a few central ideas.

Mathematics is built upon a considerable variety of such ideas—in the calculus, ideas about rate of change, summation, and limit; in geometry, ideas of proximity, smoothness, and curvature. To further illustrate what I mean here by "idea," I choose a small sample: The related ideas of "connect," "compose," and "compare."

To "Connect" means to join. There are different ways in which mathematicians have defined what it means for a piece of space to be connected. One definition says that a piece of space is connected if it does not fall apart into two (or more) suitably disjoint pieces. Another definition says that a piece of space is *path-connected* if any two points in the piece can be joined *within* the piece by a path—that is, by a continuous curve lying wholly in the piece. These two formal explications of the idea of "connected" are not identical; a piece of space which is path-connected is always connected in the first sense, but not necessarily vice-versa. This simple case of divergence illustrates the observation that the same underlying pre-formal idea can have different formalizations.

"Compose" is the next idea. To compose two numbers x and y by addition is to take their sum $x + y$; to compose them by multiplication is to take their product xy. To compose one motion L with a second motion M is to follow L by M to get the "composite" motion which we write as $L \circ M$. Thus to rotate a wheel first by 25° and then by 45° will yield after composition a rotation by 70°. To compose a path L connecting a point p to a point q with a path M connecting q to a third point s is to form the longer path $L \circ M$ which follows first L and then M, as in the top of figure 4. In all such cases of composition, the result of a composition $L \circ M \circ N$ of three things in succession depends on the factors composed and the sequence or order in which they were taken—but *not* on the position of the parenthesis. Thus arises one of the formal laws of composition, the associative law:

$$L \circ (M \circ N) = (L \circ M) \circ N.$$

However, $L \circ M$ may very well differ from $M \circ L$! The order matters.

The third sample idea is "Compare." One may compare one triangle with another as to size, so as to study congruent triangles. One may compare one triangle with another as to shape, and so study more generally similar triangles. Another comparison is that by deformation: Two paths in a piece of space may be compared by trying to deform the first path in a continuous way into the second—as in Figure 4, the composite path $L \circ M$ from p to r can be deformed smoothly and continuously into the path K also joining p to r.

Ideas such as these will function effectively in mathematics only after they have been formalized, because then explicit theorems about the ideas can be proved. The idea of composition is formalized by the concept of a group, which applies to those compositions in which each thing L being composed has an "inverse" thing or operation L^{-1} so that $L \circ L^{-1} = 1$. One readily sets down axioms for a group of "things" with such composition. The axioms are quite simple, but the concept has proven to be extraordinarily fruitful. There are very many examples of groups: Groups of rotations, groups of symmetry, crystallographic groups, groups permuting the roots of equations, the gauge groups of physics, and many others. There is a sense (analyzed by Eilenberg-Mac Lane in a series of papers) in which any group can be built up by successive extensions from certain basic pieces, called the "simple" groups. Specifically, a group is said to be *simple* when it cannot be collapsed into a smaller group except in a trivial way. A long-standing conjecture suggested that the number of elements in a finite simple group was necessarily either an even number or a prime number. About twenty years ago, here at Chicago, Thompson and Feit succeeded in proving this to be true (and I could take pleasure in the fact that Thompson, one of my students, had achieved such a penetrating result). The Thompson-Feit method turned out to be so suggestive and powerful that others have now been able to go on to explicitly determine *all* the finite simple groups. For example, the biggest sporadic one has $2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 41 \cdot 47 \cdot 59 \cdot 71$ elements (that number is approximately 8 followed by 53 zeros). This simple group is called the "Monster" (Figure 5). Another one of our former students has been able to make a high dimensional geometric picture which shows that this monster really exists. He needed a space of dimension 196,884.

Groups also serve to measure the connectivity of spaces. In particular, there are certain homology groups which count the presence of higher dimensional holes in space. To start with, a piece χ of space is said to be *simply connected* if any closed path in the space χ can be deformed into a point. For example, the

surface of a sphere is simply connected, so its first homology group is zero; however, it has a non-zero second homology group—meaning the "hole" represented by the inside of the sphere. These properties characterize the two-dimensional sphere. Long ago, the French mathematician Poincaré said that the same should hold for a three-dimenional sphere (Figure 6). This famous conjecture has not yet been settled—but some years ago, Smale showed that the characterization was true for a sphere of dimension 5 or higher. Just during the last year, the Californian Michael Friedman, in a long proof, showed that it is also true for a sphere of dimension 4. Except for a solution which was announced on April 1, nobody yet knows the answer for a three-dimensional sphere. Proof advances, but slowly.

As already indicated, Whitehead and Russell, by *Principia Mathematica*, had suggested that all mathematical truth could be subsumed in one monster formal system. Their system, corrupted as it was with "types," was too complicated— but others proposed a system based on the idea of a *set*. A set is just a collection of things—nothing more. Mathematics does involve sets, such as the set of all prime numbers or the set of all rational numbers between 1 and 2. Mathematical objects can be defined in terms of sets. For example, a circle is the set of all points in the plane at a fixed distance from the center, while a line can be described as the set of all its points. Numbers can be defined as sets—the number two is the set of all pairs; an irrational number is the set of all smaller rational numbers. In this way numbers, spatial figures, and everything else mathematical can be defined in terms of sets (Figure 7). All that matters about a set S is the list of those things x which are members of S. When this is so, we write $x \,\epsilon\, S$, and call this the "membership relation."

There are axioms (due to Zermelo and Fraenkel) which adequately formalize the properties of this membership relation. These axioms claim to provide a formal foundation–I call this the grand set-theoretic doctrine—for all of mathematics.

By 1940 or so this grand set-theoretic foundation had become so prominent in advanced mathematics that it was courageously taught to freshmen right here in the Hutchins college. This teaching practice spread nationally to become the keystone of the "New Math." As a result, twenty years later sets came to be taught in the kindergarten. There is even that story about the fond parents inquiring as to little Johnny's progress. Yes said the teacher, he is doing well in math except that he can't manage to write the symbol ϵ when x is a member of the set S.

Johnny was not the only one in trouble. The grand doctrine of the new math: "Everything is a set" came at the cost of making artificial and clumsy definitions. Moreover, putting everything in one formal system of axioms for set theory ran squarely into the difficulties presented by Gödel's undecidable propositions.

Fortunately, just about the time when sets reached down to the kindergarten, an alternative approach to a system of "all" (better "most") of mathematics turned up. This used again the idea of composition for functions $f : S \longrightarrow T$ sending the elements of a set S to some of those of another set T. Another function $g : T \longrightarrow U$ can then be composed with f to give a new function $g \circ f$ (Figure 8). It sends an element of S first by f into T and then by g into U. The prevalence of many such compositions led Eilenberg and Mac Lane in 1945 to define the formal axioms for such composition. With no apologies to Aristotle, they called such a system a "category"—because many types of mathematical objects did form such categories, and these properties were useful in the organization of mathematics. Note especially that the intuitive idea of "composition" has several different formalizations: category and group.

Then in 1970 Lawvere and Tierney made a surprising discovery: that in treating a function $f : S \longrightarrow T$ one could forget all about the elements in S and T, and write enough axioms on composition alone to do almost everything otherwise done with sets and elements. This formal system is called an "elementary topos"—to suggest some of its connections to geometry and "Top"-ology. Their success in discovering this wholly new view of mathematics emphasizes my fundamental observation: That the ideas of mathematics are various and can be encapsulated in different formal systems.

Waiting to be developed, there must be still other formal systems for the foundation and organization of mathematics.

V. Confusion via Surveys

The crux of any search for the right alternative to set theory is the search for the right concatenation of ideas—in the same way in which leading ideas in mathematics have been combined in the past to solve problems (the Poincaré conjecture on spheres) and to give new insights. Thus it was with our example, where the related ideas of connection, composition, and comparison came together in group theory, in the application of groups to geometry, and in category theory. But sometimes the wrong ideas are brought together, or the right ideas are used in the wrong way. Today the use of numbers and of quantitative methods is so pervasive that many arrays of numbers and of other mathematical techniques are deployed in ways which do not fit.

This I will illustrate by some examples. Recently, in connection with my membership on the National Science Board, I came across the work of one prominent social scientist who was promoting (perhaps with reason) the use of computer-aided instruction in courses for college students. However, the vehicle he chose for such instruction was the formal manipulation of the elementary consequences of the Zermelo-Fraenkel axioms for set theory—and the result was an emphasis on superficial formalism with no attention to ideas or meaning. It was, in short, computer-aided pedantry.

"Opinion Surveys" provide another example of the confusion of ideas. For some social and behavioral research, the necessary data can be obtained only by survey methods, and responsible scientists have developed careful techniques to help formulate the survey questions used to probe for facts. Unfortunately these techniques are often used carelessly—both because of commercial abuse, statistical malpractice, or poor formulation of survey questions. First the malpractice:

On many surveys the percentage of response is uncomfortably low, with the result that the data acquired are incomplete. This situation has led the statisticains into very elaborate studies of means for approximately completing such "incomplete data." One recent and extensive such publication (by the National Research Council) seemed to me technically correct but very elaborate—perhaps overdone, and in any event, open to the misuse of too much massaging of data that are fatally incomplete.

In opinion surveys touching directly on the academic profession some of the worst excesses are those exhibited by the so-called "Survey of the American Professoriate." Successive versions of this survey are replete with tendentious and misleading questions, often such likely to "create" opinion rather than to measure actual existing opinions. Despite heroic attempts by others to suggest improvements, the authors of this particular survey have continued in their mistaken practices in new such surveys—as has been set forth with righteous indignation by Serge Lang in his publication *The File: A Case Study in Correction*.

That otherwise useful publication *Science Indicators* from the National Science Board makes excessive use of opinion surveys. The most recent report of the series (*Science Indicators 1980*) coupled results from a new, more carefully constructed opinion survey with a simple continuation of poorly formulated questions taken from previous and less careful surveys.

The main new opinion survey commissioned for this NSB report used an elaborate design—but this design still involved some basic misconceptions about science and some questions about science so formulated as to distort the opinions which were to be surveyed. For example, its Question 71 first observes that "Science and technology can be directed toward solving problems in many different areas"—while I would claim that science cannot be "directed" in the fashion intended by government bureaucrats. The question then lists fourteen areas and asks "Which three areas on the list would you *most* like to receive science and technology funding from your tax money?" Of the fourteen areas, some had little to do with science or technology and much to do with the political and economic structure of society (for example, controlling pollution, reducing crime, and conserving energy). Only one of the fourteen dealt with basic knowledge. With an unbalanced list of questions like this, the report goes on to claim that the answers "suggest that the public interest tends to focus on the practical and immediate rather than on results that are remote from daily life." This may be so, but it cannot be demonstrated by answers to a survey questionnaire which itself is so constructed as to focus on the "practical and immediate."

To get comparisons of opinions across time, new surveys try to continue questions which have been used before—and so often use older questions of a clearly misleading character. In *Science Indicators*, a typical such previous question is the hopelessly general one, "Do you feel that science and technology have changed life for the better or the worse?" The current version of this question does still more to lead the respondent to a negative answer. It reads, "Is future scientific research more likely to cause problems than to find solutions

to our problems?" It is no wonder that this latter slanted question, in the 1979 survey, had only 60% answers favorable to science, while the earlier one had 75% favorable in 1974 and 71% in 1976.

Surveys also may pose questions which the respondents are in no position to answer. For instance, one question in this survey probed the respondents' expectations of scientific and technological achievements: "During the next 25 years or so, would you say it is very likely, possible but not too likely, or not likely at all that researchers will discover a way to predict when and where earthquakes will occur?" How can the general public have a useful or informed opinion on this highly technical and speculative question? The question brought answers of 57% "very likely," 34% "possible," and 7% "not likely." After giving these figures, the text obscures the careful tripartite posture of the question as stated by lumping the first two categories together in the following summary: "About 9 out of 10 consider it possible or very likely ..."

The other five questions asking for similar 25-year predictions (for example, a cure for the common forms of cancer) are not much better.

In sum, the public opinion surveys currently used in *Science Indicators* are poorly constructed and carelessly reported. By emphasizing remote and speculative uses of science, the thrust of the questions misrepresents the very nature of scientific method. (There are worse misrepresentations, for example, in a report for GAO (General Accounting Office), mistitled *Science Indicators: Improvements Needed in Design, Construction, and Interpretation*).

To summarize: Opinion surveys may attempt to reduce to numbers both nebulous opinions and other qualities not easily so reducible. It would be wiser if their use were restricted to those things which are properly numerical.

My own chief experience with other unhappy attempts to use mathematical ideas where they do not fit comes from studying many of the reports of the National Research Council (in brief, the NRC). I recently served for eight years as chairman of the Report Review Committee for this Council. This Council operates under the auspices of the National Academy of Sciences, which by its charter from the government is required to provide, on request, advice on questions of science or art. There are many such requests. Each year, to this end, the NRC publishes several hundred reports, aimed to apply scientific knowledge to various questions of public policy. Some of these policy questions are hard or even impossible of solution, so it may not be surprising that the desire to get a solution and to make it precise may lead to the use of quantitative methods which do not fit. This lack of fit can be better understood at the hand of some examples.

VI. Cost-Benefit and Regression

Before making a difficult decision, it may be helpful to list off the advantages and the disadvantages of each possible course of action, trying to weigh the one against the other. Since a purely qualitative weighing of plus against minus may not be objective (or at any rate can't be done on a computer), there has grown up a quantitative cost-benefit analysis, in which both the costs and the benefits of the action are reduced to a common unit—to dollars or to some other such "numeraire." The comparison of different actions and thus perhaps a decision between them can then be made in terms of a number, such as the ratio of cost to benefit.

In simple cases or for isolated actions this may work well; I am told that it did so function in some of its initial uses in decisions about plans for water resources. However, the types of decisions considered in NRC reports were usually not so straight-forward. I studied many such reports which did attempt to use cost-benefit analysis. In every such case which came to my attention in eight years, these attempts at quantitative cost-benefit analysis were failures.

In most cases, these failures could have been anticipated. Sometimes the intended cost-benefit analysis was not an actual numerical analysis but just a pious hope. For instance, one study tried to describe ways to keep clean air somewhere "way out west." In this case, there weren't enough dependable data to arrive at any numbers for either the costs or the benefits of that clean air. Hence the report initially included a long chapter describing how these costs and benefits *might* be calculated—although it really seemed more likely that there never would be data good enough to get dependable numbers for such a calculation.

There are also cost-benefit calculations which must factor in the value of the human lives which might be saved by making (or not making) this or that decision. In such cases, the value ascribed to one human life can vary by a factor of 10, ranging from one hundred thousand to one million dollars. Much of the variation depends on whether one gets the value of that life in terms of discounted future earnings or by something called implicit self-valuation of future satisfaction. However, I strongly suspect that whatever the method,

there isn't *any* one number which can adequately represent the value of human life for such cost-benefit purposes. Our lives and our leisures are too various and their value (to us or to others) is not monetary. The consequence is that decisions which deal substantially with actions looking to the potential saving of lives cannot be based in any satisfactory way on cost-benefit analysis.

Another aspect of cost-benefit methods came to my attention just yesterday, in the course of a thesis defense. Cost-benefit methods attend only to gross measures, in a strictly utilitarian way, and give no real weight to the distribution of benefits (or of costs) between individuals.

Another striking example of the problems attending the use of cost-benefit analysis in policy studies is provided by a 1974 NRC study, "Air Quality and Automobile Emission Control," prepared for the Committee on Public Works, U.S. Senate. That committee was considering the imposition of various levels of emission controls on automobiles; it requested advice on the merits of such controls, and in particular wanted a comparison of the costs and the benefits of such control.

Some benefits of the control of automobile emissions are to be found in cleaner air and some in better health (less exposure to irritating smog). A number of studies of such health effects had been done; the NRC committee examined them all and considered all but one of them inadequate. The one adequate study was for students in a nursing school in the Los Angeles area. Each student carefully recorded daily discomforts and illnesses; these records were then correlated with the observed level of smog in Los Angeles. The results of this one study were then extrapolated by the NRC committee to the whole of the United States in order to estimate the health benefits of decreasing smog! It was never clear to me why Los Angeles is typical or how such a wide extrapolation can be dependable. Just as in the case of saving lives, the benefits of good health can hardly be reduced to numbers.

Some other difficulties with this particular study concern the use of regression, a mathematical topic with a considerable history. Mathematics deals repeatedly with the way in which one quantity y may depend upon one or more other quantities x. When such a y is an explicitly given function of x, the differential calculus has made extraordinarily effective use of the concept of a derivative $dy/dx = y'$; in the first instance, the use of the derivative amounts to approximating y by a linear function, such as $y = ax + c$, choosing a to be a value of the derivative y'. The number a then is units of y per unit of x and measures the number of units change in y due (at x) to a one-unit change in x. For certain purposes these linear approximations work very well, but in other cases, the calculus goes on to use higher stages of approximation — quadratic, cubic, and even an infinite series of successive powers of x.

But a variable quantity y involved in a policy question is likely to depend not just on one x, but on a whole string of other quantities x, z, and so on. Moreover, the fashion of this dependence can be quite complex. One approximation is to again try to express y as a constant a times x plus a constant b times z and so on—in brief to express y as a linear function

$$y = ax + bz + \cdots$$

with coefficients a, b, ... which are not yet known. Given enough data, the famous method of "least squares" will provide the "best" values of the constants a, b, ... to make the formula fit the given data. In particular, the coefficient a estimates the number of units change in y per unit change in x—holding the other quantities constant (if one can).

This process is called a multiple "regression" of y on x, z, ... This curious choice of a word has an explanation. It was first used by Galton in his studies of inheritance. He noted that tall fathers had sons not quite so tall—thus height had "regressed on the mean."

This technique of regression has been amply developed by statisticians and others; it is now popular in some cost-benefit analyses. For example, with the control of auto emission, how does one determine the benefit of the resulting clean air?

Clean air cannot be purchased on the market, so the benefits of cleaner air might be measured by "shadow" prices found from property values, on the grounds that homes in a region where the air is clear should command higher prices than comparable homes where the air is thick. In the NRC study, the prices of houses in various subregions of greater Boston were noted and then expressed as a (linear) function of some thirteen different measured variables thought to influence these prices: Clean air, proximity to schools, good transportation, proximity to the Charles River, and so on. The constants in this linear expression of house prices were then determined by regression. In this equation, the coefficient a for the variable representing

“clean air" (of units of dollars per measure of smog-free air) was then held to give the “shadow price" for clean air. The resulting shadow price from this and one other such regression was then extrapolated to the whole U.S.A. to give a measure of the benefit of cleaner air to be provided by the proposed auto emission control.

In Boston: House $ = a (Smog) + b (Charles River) + $\cdots$ one dozen more

This is surely a brash attempt to get a number, cost what it may. In my considered judgment, the result is nonsense. It is not clear that buyers of houses monitor the clean air before they sign the mortgage. A nebulous (or even an airy) quantity said to depend on thirteen other variables is not likely to be well grasped by any linear function of those variables. Some variables may have quadratic effects, and there could be cross effects between different variables. That list of thirteen variables may have duplicates or may very well miss some variables which should be there. Moreover, the coefficients in that function are likely to be still more uncertain than the known costs the equation estimates. These coefficients are not shadow prices; they are shadowy numbers, not worthy of serious regard. They employ a mathematics which does not fit.

The difficulties which have been noted in interpreting the coefficients in some regressions are by no means new. For example, you can find them discussed with vigor and clarity in a text by Mosteller and Tukey, *Data Analysis and Regression*, kept here on permanent reserve in the Eckhart Library. I trust that such reserve has not kept it from the eyes of economists or other users of regression. What with canned formulas from other sources and fast computers, any big set of data can be analyzed by regression—but that doesn't guarantee that the results will fit!

I have not studied the extensive academic literature on cost-benefit analysis, but these and other flagrant examples of the misuse of these analyses in NCR reports leave me disquieted. Current political dogma may create pressure for more cost-benefit analysis. In Congress, the House is now considering a “Regulatory Reform Bill" which requires that independent and executive agencies of the government make a cost-benefit analysis before issuing any new regulation (except for those health and safety regulations required by law). It is high time that academicians and politicians give more serious thought to the limitations of such methods of analysis.

The future is inscrutable. However, people are curious, so fashion usually provides some method for its scrutiny. These methods may range from consultation with the Oracle at Delphi to opinion polls to the examination of the entrails of a sacrificial animal. Now, thanks to the existence of fast computers, some economists can scrutinize the future without entrailing such sacrifice. The short-term predictions by econometric models can be sold at high prices, though I am told that some of these models deliver more dependable short-term predictions when the original modeler is at hand to suitably massage the output figures.

At the NRC, my chief contact with projection was on a very much longer time scale—econometric projections of the energy future of the United States going forward for fifty years or more. This was done in connection with a massive NRC study called CONAES (for the Committee on Nuclear and Alternative Energy Systems). For this study, there was not just one econometric projection of energy needs, but a half dozen such models, with a variety of time horizons. Now projections for a span of forty or fifty years cannot possibly take account of unexpected events such as wars, oil cartels, depressions, or even the discovery of new oil fields. Since the present differs drastically from the past, there is little or no hope of checking a fifty-year projection against fifty years of actual past development. Consequently, this particular NRC study did not check theory against fact, but just theory against theory—by asking just how much agreement there was between the half-dozen models. It hardly seemed reasonable to me to conclude that agreement—even a perfect agreement — in the results of several fictive models can be of any predictive value. In the case of the CONAES report, there was even a proposal to use the thirty-five-year projection of those models to assess the future economic value of the breeder reactor. Such assessment breeds total futility. All told, despite the use of fast computers and multiple models, the ambiguities of the models being computed still leave the future dark and inscrutable.

Projections over time into an unknown future are not the only examples of policy-promoted projection of the unknown. Many other types of extrapolation can be stimulated—for example, extrapolation designed to estimate risks. Since it is claimed society has become more risk-averse, there is great demand to make studies of future risks, as in the reports of the NRC Committee on the Biological Effects of Ionizing Radiation (BEIR for short). The third report of this committee, a report commonly known as “BEIR III," dealt with

extrapolation, another kind of projection. Data available from Hiroshima and Nagasaki give the numbers of cancers caused by high dosages of radiation. For present puropses one wants rather the effect of low doses, on which there are little or no data. To estimate this effect, one may assume that the effect E is proportional to dosage D—so that $E = kD$ for some constant k. Alternatively, on may assume that the effect is quadratic so that E depends both on D and on D^2. Then the curve giving E as a function of D is parabolic (Figure 10). The constants involved—such as the proportionality factor k—are then chosen to get the best fit of the line or the parabola to the high dosage data. The resulting formula is then used to calculate the effect at low dosage. Quite naturally, the linear formula and the quadratic one give substantially different results by this extrapolation; this is the cause of considerable controversy. Is the linear formula right? Does the choice of formula depend on the type of cancer considered? There is no secure and scientific answer to these pressing policy questions. In particular, the mathematical methods themselves cannot possibly produce an answer. Mathematical models such as these may be internally consistent, but that doesn't imply that they must fit the facts. Here, as in the case of regression, the assumption that the variables of interest are connected by a linear equation is gratuitous and misleading.

That BEIR III report deals with just one of many different kinds of risks that plague mankind. There are many others that might be estimated, by extrapolation or otherwise. From all these cases there has arisen some hope that there might be effective general principles underlying such cases–and so constituting a general subject of "risk analysis." The hope to get at such generality may resemble the process of generalization so successful in mathematics, where properties of numbers have been widely extended to form the subject of number theory and properties of specific groups have led to general group theory. However, I am doubtful that there can yet be a generalized such "risk analysis" — and this I judge from another current NRC report.

This report arose as follows: The various public concerns about risks were reflected in Congress, so a committee of the Congress instructed the National Science Foundation (NSF) to establish a program supporting research on risk analysis. The NSF, in its turn, did not know how to go about choosing projects in such a speculative field—so it asked the National Research Council for advice on how to do this. The NRC, again in turn, set up a committee of experts on risk analysis. This committee, in its turn, prepared a descriptive report on risk analysis "in general." The report also commented on specific cases of risk analysis. For example, there were extensive comments on the BEIR III report—but these comments did not illuminate the BEIR III problem of extrapolation and made no other specific suggestions. The report had little of positive value to help the NSF decide which projects in risk analysis to fund. Such a general study of risk analysis is clearly interdisciplinary, but I must conclude that it is not yet disciplined.

These and other examples of unsatisfactory reports may serve to illustrate the confusion resulting from questionable uses of quantitative methods or of mathematical models. But why are there so many cases of such confusion? Perhaps the troubled history of that report on risk analysis is typical. A practical problem appears; many people are concerned, and so is the Congress or the Administration. Since the problem is intractable, but does involve some science, it is passed on to the scientists, perhaps to those at the NRC. Some of these problems can be—and are—adequately treated. For others there is not yet any adequate technique—and so those techniques which happen to be available (opinion surveys, cost-benefit analysis, regression, projection, extrapolation, decision analysis, and others) get applied to contexts where they do not fit. Confusion arises when the wrong idea is used, whether for political reasons or otherwise.

There are also political reasons for such confusion. Our representatives, meeting in that exclusively political city of Washington, represent a variety of sharply different interests and constituencies. To get something done, a compromise must be struck. This happens in many ways. One which I have seen, to my sorrow, is the adjustment of the onerous and bureaucratic regulations of the OMB (Office of Management and Budget) about cost principles for universities. Their Circular A-21 now requires faculty members to report the percentage distribution of their various university activities, with results to add up to 100%, on a "Personnel Activity Report Form" (PAR!). Such numbers are meaningless; they are fictions fostered by accountants. Use of such numbers makes for extra paperwork—but it also tends to relocate some control of scientific research from universities to the government bureaucrats. For A-21, there was recently a vast attempt at improvement, combining all parties: the government bureaucrats, their accountants, university financial officers, and a few faculty. What resuted? A compromise, and not a very brilliant one.

Thus government policy, when it requires scientific advice on matters that are intrinsically uncertain, is likely to fall into the government mold: compromise. And that, I believe, is a source of confusion.

VIII. Fuzzy Sets and Fuzzy Thoughts

The misuse of numbers and equations to project the future or to extrapolate risks is by no means limited to the National Research Council. Within the academic community itself there can be similar fads and fancies. Recently I have been reminded of one curious such case: The doctrine of "fuzzy" sets.

How can a set be fuzzy? Recall that a set S is completely determined by knowing what things x belong to S (thus $x \epsilon S$) and what things do not so belong. But sometimes, it is said, one may not know whether or not $x \epsilon S$. So for a fuzzy set F one knows only the likelihood (call it $\lambda(x)$) that the thing x is in the fuzzy set F. This measure of likelihood may range from 0 (x is certainly not in F) all the way to 1 (x is certainly in F). Now I might have said that $\lambda(x)$ is the probability that x is in F, to make this definition a part of the well-established mathematical theory of probability. The proponents do not so formulate it, because their intention is different and much more ambitious: Replace sets everywhere by fuzzy sets!

By the grand set-theoretic doctrine, every mathematical concept can be defined in terms of sets, hence this replacement is very extensive. It even turns out that many mathematical concepts can be fuzzed up in several ways, say, by varying the fuzzy meaning to be attached to the standard set-theoretic operations (intersection, union, etc.) of the usual Boolean algebra of sets. And so this replacement doctrine has already produced a considerable literature: on fuzzy logic, fuzzy graphs, fuzzy pattern recognition, fuzzy systems theory, and the like. Much of this work carries large claims for applications of this fuzzy theory. In those cases which I have studied, none of the applications seem to be real; they do not answer any standing problems or provide any new techniques for specific practical situations. For example, one recent book is entitled *Applications of Fuzzy Sets to Systems Analysis*. The actual content of the book is a sequence of formal fuzzy restatements of standard mathematical formulations of materials on programming, automata, algorithms, and (even!) categories, but there is no example of specific use of such fuzzy restatement. One reviewer (in *Mathematical Reviews*) noted a "minimal use or lack of instructive examples—the title of the book purports applications." Another more recent book on fuzzy decision theory states as one of its six conclusions, "It is a great pity that there exist only very few practical applications of fuzzy decision theories, and even practical examples to illustrate the theories are scarce." This leads me to suspect that the initially ingenious idea of a fuzzy set has been overdeveloped in a confusing outpouring of words coupled with spurious claims to importance.

There are other examples—cybernetics, catastrophe theory— where an originally ingenious new idea has been expanded uncritically to lead to meaningless confusion.

IX. Compromise Is Confusing

But enough of such troubling examples of confusion. Let me summarize where we have come. As with any branch of learning, the real substance of mathematics resides in the ideas. The ideas of mathematics are those which can be formalized and which have been developed to fit issues arising in science or in human activity. Truth in mathematics is approached by way of proof in formalized systems. However, because of the paradoxical kinds of self- reference exhibited by the barn door and Kurt Gödel, there can be no single formal system which subsumes all mathematical proof. To boot, the older dogmas that "everything is logic" or "everything is a set" now have competition–"everything is a function." However, such questions of foundation are but a very small part of mathematical activity, which continues to try to combine the right ideas to attack substantive problems. Of these I have touched on only a few examples: Finding all simple groups, putting groups together by extension, and characterizing spheres by their connectivity. In such cases, subtle ideas, fitted by hand to the problem, can lead to triumph.

Numerical and mathematical methods can be used for practical problems. However, because of political pressures, the desire for compromise, or the simple desire for more publication, formal ideas may be applied in practical cases where the ideas simply do not fit. Then confusion arises — whether from misleading formulation of questions in opinion surveys, from nebulous calculations of airy benefits, by regression, by extrapolation, or otherwise. As the case of fuzzy sets indicates, such confusion is not fundamentally a trouble caused by the organizations issuing reports, but is occasioned by academicians making careless use of good ideas where they do not fit.

As Francis Bacon once said, "Truth ariseth more readily from error than from confusion." There remains to us, then, the pursuit of truth, by way of proof, the concatenation of those ideas which fit, and the beauty which results when they do fit.

If only Longfellow were here to do justice to the situation:

Tell Me Not in Fuzzy Numbers

In the time of Ronald Reagan
Calculations reigned supreme
With a quantitative measure
Of each qualitative dream

With opinion polls, regressions
No nuances can be lost
As we calculate those numbers
For each benefit and cost

Though his budget will not balance
You must keep percents of time
If they won't sum to one hundred
He will disallow each dime.

References

"Air Quality and Automobile Emission Control." A report prepared by the Commission on Natural Resources, National Academy of Sciences for the Committee on Public Works, U. S. Senate. September 1974. Volume 4. *The Costs and Benefits of Automobile Emission Control.*

Eilenberg, S. and Mac Lane, S. "Group Extensions and Homology," *Annals of Math.* 43 (1941), 758-831.

Eilenberg, S. and Mac Lane, S. "General Theory of Natural Equivalences" (category theory), *Trans. Am. Math. Soc.* vol. 28 (1945), 231-294.

Energy Modeling for an Uncertain Future: Study of Nuclear and Alternative Energy Systems. A series: Supporting Paper #1, 225pp. National Academy Press, 1978.

Feit, Walter and Thompson, John G. "Solvability of Groups of Odd Order," *Pacific J. Math.* 13 (1968), 775-1029.

Kickert, Walter J. M. Fuzzy Theories in Decision-making: A Critical Review. Martinus Nijhoff, Social Sciences Division, Leyden, 1978. 182 pp. Reviewed in *Mathematical Reviews* vol. 81f (June 1981), #90006.

Ladd, Everett C. and Lipset, Seymour M. The 1977 Survey of the American Professoriate.

Lang, Serge. The File: A Case Study in Correction. 712 pp. Springer Verlag, New York, 1981.

Lawvere, R. W. "Toposes, Algebraic Geometry, and Logic," *Springer Lecture Notes in Math.* No. 274 (1972).

Mosteller, F. and Tukey, John W. Data Analysis and Regression. 586 pp. Addison-Wesley, Reading, Mass. (1977).

Negoita, C. V. and Ralescu, D. A. Applications of Fuzzy Sets to Systems Analysis, 1975. John Wiley and Sons, New York. 191pp. Reviewed in *Mathematical Reviews* vol. 58 (1979), #9442a.

Science Indicators: Improvements Needed in Design, Construction, and Interpretation. Report by the Comptroller General of the United States. PAD 79-35, September 25, 1979.

Whitehead, A. N. and Russell, Bertrand. *Principia Mathematica*, vol.1, 2nd edition. Cambridge University Press, 1925. 674 pp.

The Ryerson Lecture was given April 20, 1982 in the Glen A. Lloyd Auditorium of the Laird Bell Law Quadrangle.

The Nora and Edward Ryerson Lectures were established by the trustees of the University in December 1972. They are intended to give a member of the faculty the opportunity each year to lecture to an audience from the entire University on a significant aspect of his or her research and study. The president of the University appoints the lecturer on the recommendation of a faculty committee which solicits individual nominations from each member of the faculty during the winter quarter preceding the academic year for which the appointment is made.

The Ryerson Lecturers have been:

1973-74: John Hope Franklin, "The Historian and Public Policy"
1974-75: S. Chandrasekhar, "Shakespeare, Newton, and Beethoven: Patterns of Creativity"
1975-76: Philip B. Kurland, "The Private I: Some Reflections on Privacy and the Constitution"
1976-77: Robert E. Streeter, "WASPs and Other Endangered Species"
1977-78: Dr. Albert Dorfman, "Answers Without Questions and Questions Without Answers"
1978-79: Stephen Toulmin, "The Inwardness of Mental Life"
1979-80: Erica Reiner, "Thirty Pieces of Silver"
1980-81: James M. Gustafson, "Say Something Theological!"

ARTICLE

DIGITAL MAPS AND DATA BASES: AESTHETICS VERSUS ACCURACY *

Robert F. Austin, Ph.D.
President
Austin Communications Education Services
28 Booth Boulevard
Safety Harbor, FL 34695-5242
Past-president, AM-FM International

I. Introduction

Of the many courses lectured by Immanuel Kant at the University of Königsberg, legend has it that one of the most frequently offered was a course on natural philosophy (that is, physical geography). It was argued that individuals could acquire understanding through three distinct perspectives: the perspective of formal logic and mathematics, the perspective of time (history), and the perspective of space (geography). The last of these — the perspective of space — acknowledges the importance of distance, site characteristics, and relative location in describing the relationships among several objects or facilities.

Maps are the primary means of representing such relationships. Maps are analytical tools which depict spatial relationships and portray objects from the perspective of space. The power of maps rests on their synoptic representation of complex phenomena. To paraphrase the Confucian wisdom, a map is worth a thousand words.

> "Recently it has become common to convert spatial phenomena to digital form and store the data on tapes or discs. These data can then be manipulated by a computer to supply answers to questions that formerly required a drawn map ... This stored geographic information is referred to as a [data base]." [1]

The maps produced from such data bases are termed digital maps. Computerized data bases, which may be queried and used by several people simultaneously, and digital maps are of immense value to engineers, comptrollers, planners, and managers. The combination of a digital map and data base is worth a thousand "mega-words."

The advantages of digital maps over manually drafted maps are most apparent in situations of frequent growth or change. Among these advantages are the ease and speed of revision and the fact that special purpose maps can be produced in small volume at reasonable cost. Moreover, digital maps offer greater precision in representation and analysis. "As more governmental bodies [and other agencies] expend the necessary one-time capital investment, and begin to reap the vast rewards of computer-assisted record and map keeping, others are likely to follow quickly."[2]

II. Basic Issues

After an agency or firm has decided to convert its manual records to digital map and data base form, several issues must be addressed.

The first, and most important question the agency must answer is related to the source documents to be used by the mapping firm. In general terms the choice is between cartographic sources and mechanically drafted sources.

Cartography is defined as:

> "The art, science and technology of making maps, together with their study as scientific documents and works of art. In this context maps may be regarded as including all types of maps, plans, charts, and sections, three-dimensional models and globes representing the Earth or any celestial body at any scale."[3]

In particular, cartography is concerned with the accurate and consistent depiction on a flat surface of activities occurring on a sphere.

It is not possible to duplicate, without distortion, the features on the surface of a sphere on any object other than a sphere. A surface of constant positive curvature may be represented on a surface of zero curvature only if distortion is introduced in the representation. As a simple illustration of this fact, consider

the problem of "flattening" an orange peel: it will tear. If the orange was made of rubber, it would be possible to flatten it without tearing, but not without distortion of another kind — a topological transformation.

The methods by which cartographers represent the surface of the earth on a flat piece of paper are known as map projections. For any particular purpose, the selection of a particular projection (transformation) is based on the properties of a sphere that the projection loses or retains. Every method of mapping large areas is affected, whether it is continuous mapping or facet mapping. No coherent, distortion-free transformation exists, nor, given the theorems of mathematics, can it ever exist. [4, 5] However, cartographers can identify projections that suit a client's particular purpose.

Quite often appropriate cartographic source documents already are available to a public utility and mapping firm team. Indeed, such sources may have served as the base for the construction of existing records. In other cases, it may be necessary for the mapping firm to perform an aerial photographic survey and to translate these photographs into cartographic documents — a process known as photogrammetry.

It is also possible to produce maps from non-cartographic sources such as tax assessor sheets. Certainly the most common non-cartographic sources are mechanically drafted cadastral maps and engineering drawings or plans. These documents have been defined by the International Association of Assessing Officers:

> map, cadastral — A map showing the boundaries of subdivisions of land, usually with the bearings and lengths thereof and the areas of individual tracts, for purposes of describing and recording ownership.
>
> map, engineering — A map showing information that is essential for planning an engineering project or development and for estimating its cost. An engineering map is usually a large-scale map of a comparatively small area or of a route. [6]

Although such drawings have some value for small area design, engineering, and planning purposes, there are a number of problems associated with their use as source documents for large area mapping. The most critical of these is related to accuracy; tax assessor sheets in the United States, for example, are designed to be used as indices only and are subordinate to actual legal descriptions. They are highly stylized and, despite their appearance and name, highly inaccurate in terms of geographic placement.

Small plans "look" correct primarily because they correspond to the limited range of vision of human beings at ground level. However, these documents also suffer from the transformation problem. Non-uniform, interpretive, subjective corrections by a draftsman make this problem appear to vanish on individual sheets. But such corrections preclude accurately merging sheets for a large area.

In the language of the philosophy of science, the distinction is one between an iconic model and a symbolic model. An iconic model (the mechanically drafted plan) is designed to look, in some metaphorical fashion, like the object of study. Often, the closer the similarity in appearance, the less valuable the model for analytical purposes. A symbolic model (the map) is designed to facilitate quantitative measurements of characteristics of interest to analysts, managers, and engineers.

A second issue that must be considered by public utilities is the use to which the digital maps will be put. This will determine the type of output the mapping firm will generate. This also will determine the accuracy levels needed. [7] In general terms, the types of output products correspond to the types of input products: maps and plans. [8] In our experience, public utility clients generally have expressed a preference for digital maps related to spatial relationship data bases because they facilitate the more accurate analysis of physical plant attributes and distributions over a large area in a geographic information system.

The primary purpose of most digital map and data base conversion work is to provide a means to manage corporate assets. Often the actual maps produced are used for index only, not for scaled or direct measurement. This is in part a function of the distortion inherent to any mechanical production or reproduction process, in part a function of the demonstrated superiority of a fully digital, displayable linked-attribute data base management system (see Section 5), and in part a function of the distortion inherent to all map projections (the transformation problem previously discussed.)

In some cases, an agency may wish to construct a geographic data base that will support a computerized plan generation and facilities management system. As an example, consider the case where facility data will be superimposed on a merged cadastral and land base. The data base must guarantee the geographic locations of features and their connectivity, relationships, and other characteristics. The final digital plan and data base may include information on street, road, and highway names, centerlines, and rights-of-ways; political,

legal, and natural boundaries; township, range, and section lines; river, stream, and creek centerlines and names; and legal lot and parcel lines and numbers, among other data. [9]

III. Map Production

Regardless of the type of source document or output product, several stages in cartographic production remain relatively constant. These are considered first in a general manner and then as they apply to digital map production per se.

First, we must define the purpose and accuracy standards of the map. For example, will the map be used for scaled measurement? Or will the map be used as an index? Second, we must identify the features and activities to be mapped. The nature of these features will influence the amount of detail appropriate for the base map and the finished map. The strength of a map may be diminished by displaying too much detail.

Third, we must prepare or obtain a land base or base map. In this regard, it is important to consider the variety of map projections and coordinate systems available for particular tasks. Using a widely accepted system such as the UTM grid or latitude and longitude coordinates has a number of advantages, including ease of data exchange and reduced production time and cost.

The next step is to collect and compile the data to be mapped. The basic rule is to compile data at the most detailed level of measurement possible and to aggregate the data only at later analytical stages. Finally, we must design and construct the map. This is a two step process that involves: (a) the design of symbols, patterns, legends, and other cartographic devices, and (b) the location and actual placement of the features and activities.

IV. Digital Maps

As in traditional cartography the first step in constructing a digital map is to establish accuracy levels and to determine which attributes should be displayed and which should simply be stored.

A displayable linked-attribute data base system (discussed below) allows for the construction of a fully digital geographic information system for data management, as well as for the construction of index and general route maps. For such maps, placing items of plant “on the right side of the street" generally is adequate: in the real world, a utility pole 60 feet tall is clearly visible at an intersection. The critical attributes of each item of plant appear as numbers or words on the map and also as manipulable information in the data base. On the other hand, if the map is to be used for scaled measurement, accurate placement of items of plant is paramount. It should be noted that this second approach implies considerable supplementary manual adjustment and therefore substantially higher production costs. The next question is the method of land base construction. Land bases, or base maps, may be constructed in a variety of ways. It may be possible to develop an accurate land base by digitizing existing source documents. The information may be captured by board digitizing vectors (line strings and endpoints) or by raster scanning. If the quality of the source documents is high, these methods are extremely cost effective.

If the quality of the available source documents is unknown or suspect, it is common to conduct an aerial photographic survey and to compile the photographs into a “model" of the region (the air photos are rectified to generate a plane view of the photographed region similar to a map projection). These models are then stereo digitized and used as highly accurate source documents for land base construction using the digitizing or scanning methods noted above. (+ 10 feet accuracy is standard, but + 1 foot accuracy is possible.) Although more expensive than working from existing source documents, such an approach guarantees extreme accuracy. Moreover, the end product often has resale value which will offset the initial cost incurred by the end user.

Data sources also may be combined to form a hybrid land base. For example, individual assessor sheets at a wide variety of scales can be overlaid on a stereo digitized base. The stereo digitized street centerline network can be modified to agree with cadastral maps so the assessor data will scale properly and satisfy aesthetic criteria.

Because such a hybrid is, by definition, unique, it is appropriate to discuss in detail at the outset of the project the problems likely to be encountered in production. Disadvantages of such an approach include the substantial cost to make the map “look" like the source documents, many difficult production problems (e.g., warp of cadastral data and fitting cadastral information to an accurate land base), and the volume of source documents required.

Nevertheless, some utilities wish to use a hybrid approach because it gives them a product that is

internally acceptable from an aesthetic viewpoint. We have encountered many situations where end users are uncomfortable with the computer-plotted version of a geographic data base because it does not "look" like the product they have used for many years. The issue of user acceptance is of critical importance to the ultimate success of a conversion project. The cost and problems associated with a hybrid approach may be justified in the long run because the end user is comfortable with the "look" of the product. However, care must be taken to avoid simply computerizing existing problems.

The process of generating a hybrid combines land base construction and data compilation. When a data base management system approach is used, the distinction between these two production processes is much clearer. After a land base is assembled, a decision is made to define some selection of "attributes" of facilities or items of plant as displayable. Displayable attributes are those attributes that actually will be plotted on a map. Other attributes may be stored in the data base, but not, as a matter of course, be displayed on maps. (Information of interest to comptrollers may be of little use to engineers. Conversely, attributes that facilitate engineering procedures may be unimportant to comptrollers.)

Once agreement on the attributes to be displayed is reached, the data are coded and laid out on the source document. Some of the information will be recorded interactively at a board digitizing station. Other information will be keypunched and bulk loaded at the construction phase. After construction, updating normally is performed interactively. One significant advantage of a data base management system approach is its ability to grow or change with technical advances.

> "Maps today are strongly functional in that they are designed, like a bridge or a house, for a purpose. Their primary purpose is to convey information or to 'get across' a geographical concept or relationship ... The mapmaker is essentially a faithful recorder of given facts, and geographical integrity cannot be compromised to any great degree. Nevertheless, the range of creativity through scale, generalization, and graphic manipulation available to the cartographer is comparatively great."[10]

Given its pragmatic character, it may be surprising to learn that the physical appearance of a map is a common point of disagreement. Most frequently such disagreement arises because different sets of aesthetic principles have been applied by the client and the mapping firm.

The general question of aesthetics is not at all simple; as the art historian Ivins has argued, the notion of simple geometric relationships is not invariant in aesthetic assessments. [11] Indeed, aesthetic issues are involved in both the creation and the appreciation or perception of a work of art. Within cartography, the term "aesthetics" is reserved for consideration of the placement of elements such as compass rose, legends, and scales; the balance of these elements vis-a-vis the map object; the selection of type faces from the range of standard or customary fonts; and similar elements of visual display. To equate "correct" with cartographic aesthetics is inappropriate, except in the limited sense that some features are required by cartographic convention (e.g., italic type fonts for bodies of water). Styles as to what is "correct" from a creative standpoint also vary from one academic discipline to another: the geographer prefers a fine-line drawing while the urban planner uses heavy lines to focus attention and the landscape architect employs pictorial symbols. Differences in styles of map creation help to condition the manner in which maps are appreciated by consumers; that is, aesthetics is more than simply giving the consumer what he is used to. For example, many public utilities have developed a sense of aesthetics conditioned by experience with manually drafted, subjective, and highly symbolic plans. Unless discussed at the beginning of a conversion project this point can become most difficult, because a mapping firm must assume that map accuracy takes precedence over map symbology, visual appearance, and the superficial aesthetics of the perception of appearance. Often in cases of this sort, the aesthetics of appreciation of the consumer give way to concerns of accuracy on the part of the mapping firm which in turn may rest on the aesthetics of cartographic creation.

V. Computerized Data Bases

> "Computer systems are increasingly used to aid in the management of information, and as a result, new kinds of data-oriented software and hardware are needed to enhance the ability of the computer to carry out this task. [Data base systems are] computer systems devoted to the management of relatively persistent data. The computer software employed in a data base system is called a data base management system (DBMS)." [12]

Of the several methods of classifying data bases used by software engineers, one most important dichotomy is that between network (hierarchical) and relational data bases. [13] Certainly the most useful

approach for many users is the relational data base, because this approach permits a larger variety of queries. Regardless of the approach, the method of manipulating the data base remains a critical issue.

As noted in the quotation of Blasgen [12], the software used with a data base is known generically as a data base management system (DBMS). Such a system generally will have provisions for data structure definition as well as for data base creation, maintenance, query, and verification. Blasgen observed that in 1981 an estimated 50 companies were marketing 54 different DBMS packages. [14]

In our experience, as noted earlier, most public utility clients prefer working with a "displayable linked-attribute" DBMS. This term describes a system in which selected attributes or characteristics of the company's physical plant are stored in the data base, where they may be manipulated by users and also displayed on the digital map. For example, the age, length, size, and identification number for a piece of cable may be stored and displayed. Because the length of cable in a given span is known from installation and stored in the data base, scaled measurement is unnecessary.

A second approach to building and maintaining a data base is termed the "hybrid" approach. Consider the following example.

Analytics technicians select National Geodetic Survey monuments and photo identifiable points which provide a network for accurate control of the area. On-site field survey crews accurately survey these points and target them for aerial photography. After the flight, analytics technicians assemble these points into an accurate control network and place them in a digital file. Using the network file, features defined in contract specifications are stereo digitized from the photography in a digital format. The major features are portrayed in the form of a detailed centerline network of interstate highways, public roads and private roads. The file then is divided into facets (corresponding to individual maps) and plotted at a scale of 1:100'.

Tax assessor sheets and the 1:100' stereo digitized centerline plots are joined at the next production phase. The 1:100' plot is overlaid on individual assessor sheets, intersections are held for control, and the stereo digitized road centerline network is adjusted to match the cadastral maps. The revised centerlines and additional features such as rights-of-ways, lot lines, boundaries, and text are board digitized.

The product is the best of the digital and mechanical worlds. It is accomplished in digital format so future modifications can be made easily, and it is aesthetically pleasing — it looks like an engineering drawing or cadastral map. The information is accurate, the data can be scaled, and bearings can be extracted.

The success of endeavors of this type depends on an excellent vendor and client relationship. For such a project, it is recommended that a test or pilot study in a pre-selected area be completed prior to final contract negotiations. It is the responsibility of the client to define his needs as accurately as possible and convey these requirements in understandable language to the vendor. Vague terminology and ambiguous specifications can compound production problems. The vendor, based on his background and experience, must make meaningful suggestions to the client as early in production as possible as alternative options may be considered.

VI. User Community

In summary, a comprehensive digital map and data system has many advantages. In order to fully realize these advantages, users must consider and resolve several questions related to actual needs and aesthetic conditioning. Users should understand that digital maps and data bases constructed using highly accurate aerial photographic source documents will not, as a rule, look like familiar graphic products. They must consider the distinctions between mechanically drafted products and cartographic products and decide at the outset of the project whether they are comfortable with the graphic specifications.

Close communication between the utility company and the mapping firm is critical. The more carefully specified the project is at the beginning, the fewer the changes that will be required. Changes in specifications made during production, no matter how trivial they appear, generally affect production schedules and costs adversely.

Thus, the creation of a digital map involves not only the mastery of current technology, in order to produce an "accurate" map, but it also involves an awareness of aesthetics, as well. Attention to aesthetics, as appreciation of the map by the consumer will ensure a satisfied client; to this end, considerable education of the client with attention to close communication is appropriate. At the other end, the mapping firm needs to consider the aesthetics of map creation. When the aesthetics of creation help to guide the choice of technology, an accurate and satisfying digital map is generally the end product.

References

1. Robinson, A., R. Sale and J. Morrison (1978), *Elements of Cartography* (4th edition), New York: John Wiley, p. 4.

2. Robinson, A., et al., p. 272.

3. International Cartographic Association (Commission II, E. Meynen, Chairman) (1973), *Multilingual Dictionary of Technical Terms in Cartography*, Wiesbaden: Franz Steiner Verlag GMBH.

4. P. W. McDonnell, Jr. (1979), *Introduction to Map Projections*, New York and Basel: Marcel Dekker, Inc.

5. H. S. M. Coxeter, (1974), *Projective Geometry*, (second edition), Toronto; University of Toronto Press.

6. American Congress on Surveying and Mapping and American Society of Civil Engineers, Joint Committee (1978), *Definitions of Surveying and Associates Terms* (rev.), p. 101.

7. "Symposium on the National Map Accuracy Standard," *Surveying and Mapping*, 1960, v.20, n.4, pp. 427-457, and M. M. Thompson and G. H. Rosenfield, "On Map Accuracy Specifications," *Surveying and Mapping*, 1971, v.31, n.1, pp. 57-64.

8. Cuff, D.J. and M.T. Mattson (1982), *Thematic Maps: Their Design and Production*, New York and London: Methuen.

9. For additional discussion of this approach, see Easton, C.H. (1975), "The Land Records Information System in Forsyth County, North Carolina," pp. 261-267 in *International Property Assessment Administration*, Chicago: International Association of Assessing Officers.

10. Robinson, A., et al., p.7. See also: Amheim, R. (1976), "The Perception of Maps," *The American Cartographer*, 1976, v.3, n.1, pp. 5-10.

11. Irvins, W. Jr. (1964), *Art and Geometry: A Study in Space Intuitions*, New York: Dover (reprint of 1949 Harvard University Press edition). 12. Blasgen, M. W., "Data Base Systems," *Science*, 1982, v.215, 12 February.

13. Codd, E., "Relational Data Base: A Practical Foundation for Productivity," *Communications* of the ACM, 1982, v.25, n.2.

14. Blasgen [12].

⋆ **Acknowledgments**

An earlier version of this essay appeared as a Chicago Aerial Survey Production Report. Comments on the previous version by Mr. Harold Flynn are gratefully acknowledged, as is the support provided by Geonex Corporation during its preparation. The author wishes to thank an anonymous referee for comments useful in preparing the current version.