Apr. 23, 2022

Mathematics

Proofs

Conjectures and Proofs

For any conjecture in mathematics or computer science, there are the following possibilities:

(1) The truth of the conjecture is provable.
(2) The falsity of the conjecture is provable.
(3) The truth or falsity of the conjecture is not provable.

But now there arise new possibilities:

For (1):
The truth of the conjecture that the conjecture is provable, is provable.
The falsity of the conjecture that the conjecture is provable, is provable.
The truth or falsity of the conjecture that the conjecture is provable, is not provable.

and similarly for (2) and (3)

Why do we have any confidence at all, when we approach a conjecture, that there will not be "many" such recursions? Why do we believe that there are not an *infinite* number of such recursions, meaning that we can never prove anything?

In Lewis Carroll's dialogue, "What the Tortoise Said to Achilles", the Tortoise "proves" that we can never get from the premises,

A: "Things that are equal to the same are equal to each other."
B: "The two sides of this triangle are things that are equal to the same."

to the conclusion,

Therefore Z: "The two sides of this triangle are equal to each other."

because we must always insert a justification that the previous statements imply the conclusion. Thus we have:

(1): "Things that are equal to the same are equal to each other."
(2): "The two sides of this triangle are things that are equal to the same."
(3): (1) and (2) imply (Z).
(4): (1) and (2) and (3) imply (Z).
...
(n): (1) and (2) and (3) and (4) and ... and (n – 1) imply (Z).

...

Therefore (Z): "The two sides of this triangle are equal to each other."

— "What the Tortoise Said to Achilles," Wikipedia, Aug. 22, 2012.

The Tortoise thus mimics, in logic, one of the paradoxes of Zeno of Elea, which states that we can never move from one physical point to another, because first we have to move half the distance to the second point, but before that, we have to move half of half the distance, but before that we have to move half of half of half the distance...

Geometry "Proofs" by Measurement

What is the probability that we would get erroneous results if we did geometry proofs by measurement — "by test"? A theorem that asserted a certain property held for all figures of a certain type would be "proved-by-test" by the following procedure:

1. Find a general example of the figure — a figure chosen "at random" from a finite set of these figures. This finite set could be obtained either by considering all possible figures of the type in the problem that could be drawn with vertices on a specified grid or by simply drawing a figure so that it does not have any qualities favorable to what we are trying to prove, e.g., if a triangle were specified in the proof, then, say, a right triangle would not be drawn.
2. Measure the relevant dimensions.
3. If the property holds for some finite percentage of these measurements, we assert that the theorem has been "proved-by-test".

Thus, for example, to prove Archimedes' famous theorem that the volume of a sphere is equal to 2/3 the volume of its enclosing cylinder, we would simply select a sphere "at random", draw the enclosing cylinder, fill the cylinder with (mathematical) water, then immerse the sphere in the cylinder and see if the amount of water that spilled out was equal to 1/3 the volume of the cylinder.

What statements can we make about the probability of error in such "proofs"? How does this idea relate to the probabilistic proofs developed in the late eighties and early nineties?

Geometry Proofs Based on Movement of Figures

Years ago, in all high school courses in plane geometry and trigonometry, students were warned not to attempt proofs by continuous movement of lines or other figures. Students were not to make arguments of the form, "Lines x, y, z, and angles a, b, c, have the following relationships. Now if we slowly rotate line x clockwise around the point A, then ..." I seem to remember being told, or having read somewhere, that the Greeks had strictly forbidden such arguments.

The question is, why?

In 1997, it seemed that exactly such a proof was done in a program on PBS TV, the proof apparently arguing that, if this holds in this case, and we move that, then it holds in that case. What has changed in the teaching of geometry and trigonometry, at least on PBS, and why? The type of argument seemed similar to that used in the form of programming proving set forth in Dijkstra's *A Discipline of Programming*, in which a certain condition must be shown to remain true for each passage through the loop. So, this method of doing geometry proofs seems to be:

1. Define the relevant relations between relevant parts.

2. Pick a case you can easily prove is true.

3. Move things to arrive at the case or cases in question, showing that the relevant relations continue to hold between the relevant parts.

Morris Kline, in his *Mathematical Thought from Ancient to Modern Times* (a book I am inclined to call the best history of mathematics written in the 20th century) has the following to say on the subject:

"[The] second leading theme [of Poncelet (1788-1867)] is the principle of continuity. In his *Traité* [(1822)] he phrases it thus: 'If one figure is derived from another by a continuous change and the latter is as general as the former, then any property of the first figure can be asserted at once for the second figure.' The determination of when both figures are general is not explained. Poncelet's principle also asserts that if a figure should degenerate, as a hexagon does into a pentagon when one side is made to approach zero, any property of the original figure will carry over into an appropriately worded statement for the degenerate figure.

"The principle was really not new with Poncelet. In a broad philosophical sense it goes back to Leibniz [(1646-1716)], who stated in 1687 that when the differences between two cases can be made smaller than any datum in the given, the differences can be made smaller than any given quantity in the result. Monge [(1746-1818)] began the use of the principle of continuity to establish theorems. He wanted to prove a general theorem but used a special position of the figure to prove it and then maintained that the theorem was true generally, even when some elements in the figures become *imaginary.* Thus to prove a theorem about a line and a surface he would prove it when the line cuts the surface and then maintain that the result holds even when the line no longer cuts the surface and the points of intersection are imaginary. Neither Monge nor Carnot [(1753-1823)], who also used the principle, gave any justification for it...

"The other members of the Paris Academy of Sciences criticized the principle of continuity and regarded it as having only heuristic value... [But] the principle...was accepted during the nineteenth century as inutively clear and therefore having the status of an axiom. The geometers used it freely and never deemed that it required proof." — Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, N.Y., 1972, pp. 843-845.

Proofs in "Inductive Domains"

A related idea is that of "inductive domains". Frequently, in computer programming, we write a little program and then test it on a few inputs that we believe will show up any errors. If the program is correct for each such input, our confidence increases about its correctness for all values, even though we are prepared for our confidence to be shaken by later inputs.

Are there program *forms* such that we can, in fact, say, that if such a filled-in form works for "a few" values, then it will work for all? This question is discussed at length in my paper, "Occam's Razor and Program Proving by Test", which is accessible as a down-loadable .pdf file on the web site www.occampress.com.

It is certainly worthwhile investigating whether the idea can be applied in mathematics. By way of review: a frequently-used proof technique is that of induction. In this type of proof, we first prove that the statement we are trying to prove holds for the smallest, or first, element in the sequence of elements that we are trying to rove the statement true for. We then assume that the statement is true for all elements up to the kth element, and try to prove that this implies that it is true for the $(k + 1)$st. If we are successful, then the statement is true for all elements.

The question is, can we save ourselves most of this labor if we know that the set of elements we are dealing with has the property that if certain types of statement hold for one element, then

they hold for all elements? Then we do not need to do an inductive proof each time. Such sets of elements I will call an "inductive domain". The proving that a set of elements is an inductive domain is like doing all the inductive proofs once and for all. Thereafter, it is simply a matter of checking that the statement we want to prove is of the appropriate type.

A branch of mathematics where this idea, if it is valid, would seem to save a great deal of labor, is algebraic topology, in which many facts are proved about *n*-simplexes, which are the building blocks of a large class of geometric objects. (A 0-simplex is a point; a 1-simplex is a straight line; a 2-simplex is a triangle; a 3-simplex is a tetrahedron, etc.) We could pick an *n*-simplex — say, a 3-simplex — that made clear the statement we were trying to prove, and then, if the statement were the appropriate type, we would know that it held for all simplexes.

Proofs that Converge to a Proof

We are familiar with the phenomenon of an infinite sequence of numbers converging to a specific number, or, more generally, of an infinite sequence of points in a topological space, converging to a specific point. But why shouldn't there be an analogous phenomenon for proofs? An example is given in step 2 of the proposed proof of Conjecture 0.2 in "Approach by Induction on Inequalities", "'Arithmetical' Version of the Approach by Induction on Inequalities" in my paper "Is There a 'Simple' Proof of Fermat's Last Theorem?", available on the web site www.occampress.com.

Skeptical Thoughts on Computerized Proof-Checkers

Every once in a while I receive an email from a person who says that, since I am not an academic mathematician (my degree is in computer science, and for most of my career I have done research in the computer industry), the only way I will be able to get my papers published is by having at least the important proofs checked by a computerized proof-checker. No editor, the person says, will reject a paper if the important proofs have been checked and deemed correct by such a computer program.

The assertion is nonsense (see email below from an expert in such software). In my experience, it arises from a refusal, on the part of the email sender, to spend time or effort on a paper by a writer of little or no prestige. But on the other hand, the sender doesn't want to think of himself as one who merely dismisses papers by outsiders, and so, to maintain, and in fact increase, his opinion of himself, he says that he *would* be willing to believe my paper is worthwhile, *if* that were confirmed by an infallible authority, namely, a computerized proof-checker. "I cannot be enthusiastic about your paper because my standards are too high."

If the proofs in my papers were long and if they involved advanced mathematics, perhaps the sender would have a point. But since each of the proofs in my papers is less than four pages long, the sender's standards are laughable. Especially since all the mathematics in my papers is straightforward, even though the ideas are unorthodox and new.

Proponents of computerized proof-checking seem to believe that the programmers of the proof-checking software, and the persons who help the mathematician translate his original proof into language that can be accepted by the proof-checker, are somehow less likely to make mistakes than the mathematician who wrote the proof. And yet the complexity of the translation process alone should dispel this idea.

It is always possible that (a) the proof-checker will overlook an actual error, or (b) that the actual proof that the proof-checker checks, differs, however slightly, from the one the mathematician had in mind! In other words, that the wrong proof is in fact correct!

There are two extremes concerning inputs in computerized proof-checking: one is to put all or most of the translation burden on the proof-checker and supporting experts. In this case, the software must be continually monitored and updated to keep up with the varieties of word-processor in which mathematicians can write their proofs. This seems a formidable, never-ending, task.

The other extreme is to require that all proofs be submitted in one and the same format, for example, the equivalent of what is known in the computer industry as top-down (structured) format.

I once wrote the following email to a mathematician who had an intimate acquaintance with computerized proof-checking:

"The top-down (structured) format is straightforward, and doesn't vary between proofs."

"You said, "No, that's not close to being true. There are many structures…"

[I then said] "Perhaps I should spell out in detail what I mean by "top-down (structured) format". I mean the format that is the equivalent of the top-down (structured) format in computer programming.

"In the case of proofs, Level 1 consists of a few steps, say, less than seven, such that if each of the steps is correct, then they prove what the mathematician states they prove.

"Level 2 consists of a proof, following a similar rule, of the steps in Level 1 .

"Etc., until a Level is reached such that all the steps at the Level are known to be true.

"There are, of course, many formats for presenting proofs, but there is only one format such as I have described. In my experience, it is remarkably effective at revealing errors in logic (without a computerized proof-checker being involved).

"Of course, different proofs of one and the same theorem can certainly be presented in this one format."

He replied with an enclosure of a student exercise that he claimed showed that his proof-checker was capable of checking proofs using the format I have described. But the exercise only was one Level deep, so I was skeptical of his claim.

I then asked him, "What percentage of mathematics papers that were accepted for publication in 2017, would you say were accepted because at least some of their proofs had been checked and deemed correct by computerized proof-checkers?"

He replied,

"I don't have actual statistics, but I would say almost none. There is a movement by a few mathematicians to get to the point where we can verify all proofs submitted to journals in this way, but the technology to do that efficiently without excessive effort on the part of the author or journal does not yet exist.

"... our software is not designed or intended to check proofs submitted for journals. It is intended instead for educational purposes like teaching a beginner what constitutes a valid mathematical proof...."

How Can Contradictions Even Be *Expressed*?

In mathematics, a contradiction expresses a state of affairs which cannot exist. And yet the *expression* of that non-existent state of affairs, certainly does exist! (If *p* is a proposition, then "*p* **and not**-*p*" expresses a contradiction.) The expression is part of the (or, rather, of *a*) language in

which the mathematical subject is presented. It is certainly possible to write ungrammatical strings over the alphabet of the language. ("**and and and** *p* **not**" might be such an ungrammatical string.) In other words, contradictory states of affairs do not exist, but grammatical strings expressing them exist, along with ungrammatical strings expressing nothing at all in a given language.

It seems odd that we can talk so precisely about what does not exist. Somehow, we ought not to be able to do this: the pencil should simply stop moving, or fall off the table, when we attempt to write something contradictory; we should not be able to press down the keys on the word-processor. It should be *impossible* to say something contradictory. Yet it isn't.

Suppose we had a meta-language in which to talk about formal languages. Then we would probably have a way of saying, "the string *s* is ungrammatical in the language L", or, possibly, even, "the string *s* does not exist in the language L". But *s* certainly exists *somewhere*, probably even as a grammatical string in another language — or in many other languages (certainly in the language consisting of all finite strings over a given alphabet).

Is it possible to define a non-trivial language in which all contradictions can only be expressed by ungrammatical strings? If not, why not?

"3.032 It is as impossible to represent in language anything that 'contradicts logic' as it is in geometry to represent by its co-ordinates a figure that contradicts the laws of space, or to give the co-ordinates of a point that does not exist." — Wittgenstein, Ludwig, *Tractatus Logico-Philosophicus*, Routledge & Kegan Paul, London, 1961, p. 19

A related question is: In the language expressing a mathematical subject, there must be sequences of strings to cover all proofs by contradiction. What can we say about all such sequences of strings? "Where" are they in the language? Do they have a particular grammatical characteristic that separates them from all other strings? Upon being told truthfully that grammar G generates the language for a mathematical system, but not being told what any of the symbols stand for, is it possible to isolate all those strings that belong to proofs by contradiction?

Finally, "where" is the realm occupied by the entities in indirect proofs (or, rather, the realm occupied by entities having the *relationships* assumed and deduced in such proofs)? At least at the beginning of such proofs, we have no difficulty conceiving the assumed negation of what we want to prove. "Assume that *x* is a *y*. Then..." We are easily able to follow our argument, to have a more or less precise idea of what we are talking about, even to draw pictures. And yet, all this is about something that, we ultimately find out, doesn't exist!

A Few Words About Material Implication

During the course of an extended discussion concerning the validity of a proof I was proposing, it became clear that I and the other person were unsure about certain aspects of material implication, i.e., of the statement form, **if** *p* **then** *q*, which is also expressed as:

q is a necessary condition for *p*,
p is a sufficient condition for *q*,
p **implies** *q*,
q **if** *p*,
p **only if** *q*,
q provided *p*,

q whenever *p*,
q when ***p***.

That these expressions make sense can be seen if we let *P* denote a non-empty proper subset of *Q*, and then let *p* denote "*x* is an element of *P*" and *q* denote "*x* is an element of *Q*". Thus, e.g., clearly *x* is an element of *P* **only if** *x* is an element of *Q*.

The following is my attempt at resolving the uncertainty that the other person and I confronted.

The standard truth table definition of material implication is as follows. Here, "T" denotes a true proposition and "F" denotes a false proposition.

(I) **if** T **then** T is true;
(II) **if** T **then** F is false;
(III) **if** F **then** F is true;
(IV) **if** F **then** T is true.

In practice, line (I) is by far the most commonly used line in the truth table. I will call (I) "factual" implication, or "knowledge-building" implication, because typically, if someone has proved a lemma or theorem of the form, **if** *p* **then** *q*, then that lemma or theorem is used, in subsequent proofs, to establish the truth of *q* once the truth of *p* is established. That is, the truth of *p*, and the truth of the lemma or theorem, implies the truth of *q*.

For example, consider Fermat's Little Theorem:

if *r* is prime **and** $(a, r) = 1$ **then** $a^{(r-1)} \equiv 1 \bmod r$.

A typical use of this Theorem might be the following:

"Since by what has been established *r* is prime and $(a, r) = 1$, we have, by Fermat's Little Theorem, that $a^{(r-1)} \equiv 1 \bmod r$. But then, multiplying through the congruence by a^2, we have $a^{(r+1)} \equiv a^2 \bmod r$, and thus..."

There is another common use of material implication, or, I should say, of the statement *form*, **if** *p* **then** *q*, and that is in the restatement of definitions. This use I will call the "definitional" use of (the form of) material implication. Here, the rules are more limited. Consider, for example, the statement,

If an integer *n* is odd, then it does not contain the factor 2.

This statement is true by definition of "odd". (In order to emphasize that this statement is not material implication, I do not bold-face "if" and "then".) The statement,

If an integer *n* is odd, then it contains the factor 2,

would be regarded by readers in the mathematical community as being unequivocably false, even though the truth table for implication allows the statement to be true if in fact the integer n is even. It is unequivocably false because it contradicts the definition of "odd".

Lines (III) and (IV) are used very rarely in normal mathematical practice. (I will welcome examples that contradict this statement.) The reason is that we cannot use these lines for "factual" or "knowledge-building" purposes: if the antecedent is false, then we can only conclude that the consequent is true or false.

These two lines are also counterintuitive to many people, at least to many students. Stoll, in his *Sets, Logic and Axiomatic Theories*, attempts to make the lines at least plausible. He argues that we surely want the statement

if (p **and** q) **then** p

to be true regardless of the truth values of p, q. But then if p is false (and q is either true or false) we get Line III, and if p is true (and q is false), we get Line IV. An aid to understanding is to let P, Q denote sets with a non-empty intersection, and let p denote "x is an element of P" and q denote "x is an element of Q". Then, clearly:

if ((x is **not** an element of P) **and** (x is an element of Q)) **then** (x is **not** an element of P), is true (Line III), and

if ((x is an element of P) **and** (x is **not** an element of Q)) **then** (x is an element of P), is likewise true (Line IV).

The Length of Theorem Statements

Why don't we feel compelled to attach to every theorem, a statement, "This theorem can be expressed in n symbols in the language we are using," where n is the exact number of symbols used in the statement of the theorem?

Is mathematics ultimately nothing more or less than the study of significant truths that can be expressed in relatively "few" symbols? Is it possible that, far more important than the content of a mathematical lemma or theorem, is the fact that that content can be expressed in the number of symbols it is expressed in?

Why do we believe that all important theorem and lemma statements are of "manageable" length? Why do we not believe that there are important theorem and lemma statements that are longer than can be contained in any paper, any book, or computer memory?

Formal Languages and Gödel's First Incompleteness Theorem

A formal language is a set of finite strings of symbols. The strings are generated according to the rules of a formal grammar. The grammar specifies which strings of symbols can be replaced by which other strings of symbols.

Each mathematical subject can be represented by a formal language. All the lemmas and theorems in the subject are represented by strings in the language.

Gödel's First Incompleteness Theorem (1930) states that in each subject large enough to contain arithmetic, there are propositions that cannot be proved. In particular, in the subject there must be the proposition, "This proposition cannot be proved."

It would seem that we have a string in a formal language *L* that represents the statement, "This string does not exist in the language *L*." We ask how that could be possible.

Each proof of a lemma or theorem in a mathematical subject is a sequence of strings in the formal language representing the subject, the last string in the sequence being a representation of the lemma or theorem statement. What would force us to realize the fundamental limitation this would be — what would force us to realize the truth of Gödel's First Incompleteness Theorem?

Against Self-Reference

After reading popularizations of Gödel's First Incompleteness Theorem (see previous section) and related formal logic, I find I have grown contemptuous of the subject of self-reference, e.g., "This sentence is false." In the hands of authors (like Rudy Rucker, in *Infinity and the Mind*) who have too high an opinion of themselves, the subject has come to seem primarily a way that authors can appear to be profound thinkers, namely, by confronting what is claimed to be a profound mystery in formal logic.

Even Turing, in his proof of the unsolvability of the Halting Problem[1], uses self-reference. But for me, the existence or non-existence of a Turing machine to solve the Problem is only of minor interest. What is far more interesting is knowing for what classes of Turing machine the Haling Problem *is* solvable, and if the answer is "None", then why that is so.

Similarly, for me what is important in formal logic is not what can be said about "This sentence is false" and similar sentences, it is the problem of finding the class of formal sentences whose truth or falsity can be determined by formal procedures, e.g., computer programs.

Number of Proofs vs. Number of Truths

Let *S* denote the set of all *n*th degree polynomials with complex-number coefficients, where $n \geq 1$. The number of polynomials in *S* is uncountable because the number of complex-numbers is uncountable.

"The roots of the polynomial equation $p(x) = 0$ are $r_1, r_2, ..., r_n$", provided the roots are correct, can legitimately be called a "truth". The proof is the written-out solution of $p(x)$.

Since each proof is a finite string of characters, the number of proofs is countably infinite.

But since there are uncountably many polynomials in *S*, there are uncountably many truths. Therefore, there are truths that cannot be proved.[2]

Berry's Paradox

Berry's Paradox can be described by the following example: "'*The least integer not nameable in fewer than nineteen syllables*' is a phrase which must denote the specific number, 111777. But

1. The Halting Problem asks if there is a Turing machine (computer program) that, given another program and an input to that program as input, will always determine whether or not .that program will halt on that input. The answer is No.

2. A fact that was proved in a much different way in 1930 by Gödel in 1930 in his First Incompleteness Theorem.

the italicized expression ... is itself an unambiguous means of denoting the smallest integer expressible in nineteen syllables in the English language. Yet, the italicized statement has only eighteen syllables! Thus we have a contradiction, for the least integer expressible in nineteen syllables can be expressed in eighteen syllables." — Newman, James R., *The World of Mathematics*, Vol. 3, Simon and Schuster, N.Y., p. 1951.

Exercise: Discuss cases where Berry's Paradox *doesn't* occur, e.g., "the smallest integer that can be expressed in *n* words", where *n* is *less than or equal to* the number of words in the phrase.

Exercise: Try to arrive at an estimate of the frequency of Berry's Paradox over the natural numbers.

Possible Problem-Solving Techniques

"If you don't believe a problem has a simple solution, you probably won't find one."

We begin with what I regard as the fundamental question concerning problem-solving techniques, namely:

Why Are There Difficult Problems?

I have never come across a discussion of this question though I am sure that some, perhaps many, exist. One answer might be that the formal grammar in which the subject in which the problem exists can be represented, simply has an extremely long sequence of grammatical strings of symbols before the string that constitutes a solution, is reached.

Another answer might be that the idea underlying a solution, is very difficult to discover. But how exactly does that relate to the previous answer? Does the idea enable us to find a much shorter sequence of grammatical strings that terminate in the string that constitutes a solution?

At the very least, it would be worthwhile if a collection of answers to the question were made widely available.

Use of the "Fixed-Set" Where One Exists

The *Fixed-Set* is the set of cases that remains the same whether or not a counterexample to a conjecture exists. Thus, for example, Fermat's Last Theorem (FLT), which asks for a proof that there do not exist positive integers x, y, z, n such that $x^n + y^n = z^n$ for $n > 2$, had been proved for all n up to 4,000,000 by the early 1990s. (The Theorem was finally proved, by Andrew Wiles, for all n soon after.)

So no $x^n + y^n$, where $n =$, say, 3, was equal to a z^3 prior to Wiles' proof, whether or not a counterexample to the Theorem existed. The Theorem applied to all $(x, y, z, 3)$ before, during, and after the proof of the Theorem, and so all $(x, y, z, 3)$ were in the Fixed-Set for FLT.

The Fixed-Set problem-solving technique involves considering members of the Fixed-Set that are "near" an assumed counterexample. Thus elements of the set $\{x^k + y^k - z^k \mid 3 \le k < n\}$ are "near' the assumed counterexample. Or, in a four-dimensional grid in which the point with coordinates (x, y, z, n) "contains" the value of $x^n + y^n - z^n$. The technique is used in "Is There a "Simple" Proof of Fermat's Last Theorem?" (Part 1) on occampress.com, in the sections:

"Approach Using 'Neighbor' of Assumed Counterexample,
"Vertical Approaches Based on Pythagorean Theorem",
"Approaches Based on Inner Products".

See also the section below, ""Movement" from a Known Solution to an Unknown Solution" on page 46. The technique is also used in the proofs of the $3x + 1$ Conjecture in "A Solution to the $3x + 1$ Problem" on occampress.com.

Assignment of Coordinates to Every Possible Solution

To show that a desired solution does not exist, assign coordinates to every possible solution, then show that there are no coordinates for the desired solution, therefore there is no solution.

Show that the Construction of a Solution Is a Process That Never Ends

To show that a desired solution does not exist, show that the process of constructing the desired solution, never ends.

"Movement" from a Known Solution to an Unknown Solution

In order to describe the basic idea, let me repeat two paragraphs under "A Thought on Differential Equations" on page 105:

"Consider standard 2-dimensional Cartesian coordinates. But instead of regarding an ordered pair of integers, $<x, y>$, as defining a *point*, i.e., the intersection of a vertical grid line and a horizontal grid line, let the ordered pair define a *square.* Specifically, we define a new set of coordinates in which points have been "expanded" to squares, all squares being of the same size. Thus, $<x, y>$ now denotes the location of a square.

"We can fill this new grid of squares with the values of functions taking two integers as arguments, and returning integer values, e.g., the ordinary arithmetic functions addition, subtraction, multiplication, and division. Let us consider the case of multiplication. In the square $<x, y>$ we place the value of $x \bullet y$. Now, observing this grid of values, we see that the value in the square $<x + 1, y>$ is (obviously) simply the value in the square $<x, y>$, $+ y$. The thought may now occur to us that, once we have gone through the labor of computing the value in square $<x, y>$, it only takes "a little more" labor to find the value in square $<x + 1, y>$. And not much more labor to compute the value in square $<x + 2, y>$, or in square $<x, y + 1>$, or in square $<x, y + 2>$, etc. We don't have to compute each value from scratch."

We saw one version of this approach in the above-mentioned sub-section. Another version is given in "Geometry Proofs Based on Movement of Figures" on page 37. We will here mention several other versions.

Consider the set S of all sums, finite and infinite, of terms

$$cx_1^{a_1}x_2^{a_2}x_3^{a_3}\ldots$$

where c and all x_i and all a_i are integers, and where the number of x_i terms is infinite. Clearly, the set S contains all polynomials and all forms, e.g., binary quadratic forms $c_1xy + c_2x^2 + c_3y^2 = c_1x_1x_2 + c_2x_1^2 + c_3x_2^2$, since we can always let an infinity of the x_i terms have value 1.

Each term has an integer value, and each term has a location in some infinite dimensional grid like the one we described above for multiplication. We can now "move about" in this grid, and observe how the values of the terms change, and possibly discover rules governing the changes. Would this be an aid in solving problems? We discuss this as an approach to a possible simple proof of Fermat's Last Theorem (FLT) in our paper, "Is There a 'Simple' Proof of Fermat's Last

Theorem?" on occampress.com. For example, we can ask if, starting from the assumption of a counterexample, i.e., x, y, z, n such that $x^n + y^n - z^n = 0$, we move to a known case, say, $3^3 + 4^3 - 5^3 = -34$, we find that our value is not -34. If that should be true, then our assumption has led to a contradiction, and FLT is true.

Suppose it were possible to "plot" all the lemmas and theorems in a given subject, i.e., assign each a "location" such that it was always legitimate to move from one theorem to any of its immediately neighboring locations. Then a proof would entail beginning at one location and moving to another containing the desired lemma or theorem. In principle, of course, one could start anywhere and eventually arrive at the location of the desired lemma or theorem, but some of the paths might be considerably longer than others. Or is this technique essentially the same as that of beginning with a known lemma or theorem and finding a path to the desired lemma or theorem in the graph structure representing every proof of every possible lemma and theorem in a subject?

A related idea is that of the topology of *properties*. Mendeleev's familiar table of the elements is probably the simplest example. Here, certain properties of atoms are so arranged that it was easy for scientists to see what atoms were missing from the table, and then to initiate searches for them. Is it possible to set up equivalent tables in mathematical subjects so that, for example in topology, seeing what spaces possess a certain property, might suggest which other spaces might also possess the property?

In physics, suppose we are to compute a certain electromagnetic property of some object. We would begin with an object for which it was easy to compute the property, then slowly deform the object into the desired one, keeping track of the corresponding change in the property. Needless to say, the deforming process will probably have to be "continuous". Or maybe the original object and the final object will have to be homeomorphic. It is tempting to think about this idea in connection with much of potential theory, where various types of surfaces, with various boundaries, are studied in relation to various types of forces. Could we develop techniques to begin with *this* surface and boundary, for which we have a solution, and then gradually convert surface and boundary into *that* surface and boundary, for which a solution is sought?

Consider a computer program that (1) displays a geometric object and along with it one or more properties, e.g., volume, surface area; (2) allows us to instruct the program to gradually change the object while the program simultaneously displays the new values for the properties as the change takes place. There are already programs to graphically display, e.g., an ellipse with specified axes and there certainly must be programs that can rotate a given ellipse around the major or minor axis and then display the volume and surface area of the resulting object.

But since we are talking about relative "nearness" of things, we might want to bring in topology. Let us review the two extremes among topological spaces: one extreme is the indiscrete space, in which every point is as close as possible to every other point; the other is the discrete space, in which every point is as far as possible from every other point. In between are the useful topological spaces. Consider Chaitin's algorithmic information theory, in which a binary string is random if its shortest description is roughly the length of the string itself. (Thus the string "10" repeated a million times is not random, because its description, which has just been given, is much shorter than the 2,000,000-digit string it describes.) Can we say that a set of random strings belongs to a discrete topological space, in that each string is equally far from each other string? In

other words, because it takes the same amount of "work" to get from one string to another, unlike in our multiplication space described above?

Use of Continuity to Show That a Function Can Have a Certain Type of Value

We illustrate the technique with an example.

Let $f(k)$ denote the function $x^k + y^k - z^k$, where x, y, z are constituents of an assumed minimal counterexample $x^p + y^p - z^p = 0$ to Fermat's Last Theorem, and k is real and ≥ 1. Clearly, f is continuous and has a derivative $f'(k)$ with respect to k for all k.

By an elementary fact of the calculus, the derivative $f'(k) = x^k(ln\ x) + y^k(ln\ y) - z^k(ln\ z)$. We would like to know if $f'(k)$ has a non-zero integral value for some k in the segment $p - 1 \leq k \leq p$. We know, from other results, that $f'(p - 1) = 0$, that f' is negative over the segment indicated, and that there exists a k in the segment such that the derivative has a negative *integer* value of less than –1,000. We ask, Is it possible for $f'(k) = x^k(ln\ x) + y^k(ln\ y) - z^k(ln\ z)$ to have a negative integer value anywhere over the segment $p - 1 \leq k \leq p$.

At first sight, the presence of the natural logarithms, which are almost always irrational, may incline us to be skeptical that the derivative can have an integer value. But the following arument shows that our skepticism is not justified. For, since $f'(k)$ is continuous over the segment in question, and is always negative, and has an initial value of 0, and a value of less than –1,000 in the segment, it follows that $f'(k)$ must, over the segment, take on the integer values –1, –2, –3, ..., –1000, ... Of course, we do not know for which values of k these integer values occur, but that was not the question we were attempting to answer.

Use of Continuity to Solve the "Chair-Moving" Problem

Assume there is a square floor measuring, say, 20 feet by 20 feet. The floor is uneven but smooth, meaning there are no sharp points or sharp edges. Assume that no point of the floor rises higher than, say 2 incches above a horizontal "mid-plane", or 2 inches below the horizontal mid-plane.

Assume a square chair measuring, say, 2 feet by 2 feet, is situated on the floor. The four chair legs are all the same length, each descends vertically below a corner of the seat. The chair corners are labelled A, B, C, and D in clockwise order.

Assume that the feet of chair below the corners A, B, and D are touching the floor. (Three feet of the chair can always be made to simultaneously touch the floor.) The foot under the corner C is sticking up in the air.

Prove that if the chair is moved continuously around on the floor, always with feet A, B, and D touching the floor, there must be a time when all four feet are simultaneously touching the floor.

Proposed solution:

It is certainly possible for us to place the chair somewhere on the floor so that feet B, C, and D are touching the floor, and A is sticking up in the air. (Foot A is diagonally opposite foot C.) Then, since the floor is smooth, the movement of the chair on the floor is continuous, and thus we can arrive at this second position of the chair by pushing it around, which means that it was necessary that at some point, all four feet were simultaneously touching the floor.

Note: I recall seeing this problem presented in an issue of *Scientific American*, but I don't know which one.

The Nature of Mathematics

What Are We Doing When We Do Mathematics?
\What Are We Doing When We Do Mathematical Research?

What are we doing when we try to find a proof? Syntactically, the answer is simple: we are trying to find a path through the tree of all possible strings in the formal grammar representing the subject. But what are we doing semantically? Why are some proofs hard, others easy? Where do proofs exist in the world of semantics? In some cases, a picture makes proofs of certain elementary facts unnecessary. Is there such a picture for each mathematical fact, if only we could find it? Why are proofs necessary at all?

We know that not all problems are solvable by algorithm. Yet when humans do problem solving, they are not merely reaching at random into a set of proofs and hoping to pull out one that works. Why does human insight, creativity, so often work when no exhaustive mechanical procedure does?

We are not creating something out of nothing. "We are finding relationships," the reader might reply. But they already exist! (The statue exists in the stone.) What are we creating? What are we...changing? I ask this last question because I think most mathematicians, most thinkers *about* mathematics, fail to realize the amount that mathematics disturbs the universe. I don't mean through its application to physics, engineering, and other subjects, I mean through the amount of physical space — occupied by paper and by computer memories — that is required to store all existing mathematical knowledge, and that is required to teach and carry out mathematical research. But then we must also take into consideration all the energy and physical changes in the earth that are required to produce all the machinery that goes into the printing of math books and into the manufacture of computer memories.

As of now, I feel that the best answer to the question, "What are we doing when we do mathematics?", is an old one — and one that is not satisfying to me — namely, that we are exploring new realms, just as astronomers and terrestrial explorers do. When we create a new subject, we create a realm that lies before us just as an unexplored mountain range does, or a previously-unknown region of the cosmos. We then ask questions about that realm, and try to answer them.

But wait: Astronomers and explorers investigate what is already *there.* In what sense are all the facts about, say, the positive integers, already *there* in the set {1, 2, 3, ... }. It seems rather that, in mathematics, we investigate some of the things we can *construct* out of the simple elements, e.g., the positive integers, that we begin with.

Grammars are a necessary consequence of the use of formal logic, but like all syntactic matters, they are of secondary importance. Furthermore, the following question must be addressed regarding grammars: suppose in, say, 1850, computers existed having the speed and memory capacity of the most advanced ones of 2009. Suppose, further, that mathematicians of the time had created a grammar that they agreed was capable of generating all lemmas and theorems of all known mathematics at that time, and suppose this grammar were implemented as a computer program. Now the question arises: is it reasonable to believe that all mathematics of, say, the next 100 years would eventually have been generated by that computer program if it were able to run for a sufficiently long time? In particular, would Cantor's discoveries, e.g., in set theory and concerning countable vs. uncountable infinities, have eventually appeared among the lemmas and theorems produced by the grammar? I have a hard time believing that the answer is

yes. (Of course, these discoveries would *not* have appeared if the grammar did not allow completed infinities, that is, infinite sets.) Would Lebesgue integration have appeared? Or the concept of open sets in topology? Or topology itself, for that matter? Or Gödel's incompleteness theorems? Or Greg Chaitin's definition of randomness? (A finite string of binary digits is random if its minimum description is roughly the length of the string itself.)

Tentative assertion: just as all sequences of characters that would result from a formal grammar of the English language, are not meaningful statements in any of the sciences, so all sequences of characters that would result from a formal grammar of a mathematical subject, are not mathematically significant in that subject (or any other subject).

Another question that arises is the following: since we certainly are not interested in *every* string that the grammar generates, and since the grammar would almost certainly be what is known in formal language theory as a "Type 0" grammar, meaning a grammar such that short strings can be generated at any time — i.e., after any number of previous long strings had been generated, so that it would not be possible to generate all strings of length 1, then all strings of length 2, then all strings of length 3, etc. — what string length would mathematicians of the time have decided was an upper bound on the length of strings to be saved for human inspection, and how would they have arrived at their decision?

Still another question is: how is a lemma that is used many times in a subject, represented in the set of strings producible by a grammar?

But to return to the rather mundane idea that what we are doing when we do mathematics is carrying out an exploration: let us assume we have made certain definitions in the domain of three-dimensional real space ($\mathbf{R}^3$), which is to be the domain of our subject. We now want to explore that subject. Ultimately, what are we doing? One answer is: from the set of all subsets of the domain, we are exploring those that can be described in an acceptably small number of words and symbols. We are not exploring those subsets that take, say, millions of years even to name. So our explorations amount to a proof that "We can say this about that."

Why do we believe that all important theorem and lemma statements are of "manageable" length? (One fact that each theorem and lemma states is, "This theorem (lemma) can be expressed in the number of symbols in which it is expressed.") Why do we not believe that there are important theorem and lemma statements that are longer than can be contained in any paper, and book, and computer memory?

But we still have not explained the relationship of proofs to the subsets we explore. Set-theoretically, what is a proof? Where do proofs "live"? It seems inadequate to reply that they live in the countable infinity of finite sequences of symbols over some alphabet. What has that got to do with sets of subsets? If proofs "establish relationships", what is the domain of relationships? How exactly does it relate to domains of sets of subsets?

The reader will perhaps better understand one motivation for our questions by considering the following passage from Russell's *Principles of Mathematics*[1]:

"...there can be no greatest cardinal. Yet one would have supposed that the class containing everything would have the greatest possible number of terms. Since, however, the number of classes of things exceeds the number of things, clearly classes of things are not things..."

1. W. W. Norton & Company, Inc., N.Y., p. xiii.

So the set of all subsets of $\mathbf{R}^3$ cannot somehow "live" in $\mathbf{R}^3$. But the set of all statements and their proofs in our subject is at most only a countable infinity of finite strings of symbols, and these *can* "live" in $\mathbf{R}^3$. So: there is no room in the domain of our subject for everything we might like to talk about, but there is room for everything we can say.

A further thought regarding mathematics as the discovery of relationships will be found below under "The "Theory of Everything" in Physics vs. the "Theory of Everything" in Mathematics" on page 72.

Mathematics Consists Mainly of Establishing That "This" Is a "That".

At least that is the thought that occurs to me more often the more I continue to study the subject.

What Percentage of Theorems Are Simply Statements of a Fact About Many Things?

Fermat's Last Theorem (FLT) can be viewed as a statement of the fact that $x^n + y^n - z^n$, where x, y, z, n are positive integers, and $n \geq 3$, is never equal to zero. [1]

The Fundamental Theorem of Algebra states that any polynomial equation $a_n x_n + a_{n-1}x^{n-1} + ..., + a_0 = 0$, where $a_n \neq 0$, and where all coefficients are complex numbers, has at least one complex root.

One of Cantor's Theorems states that the cardinality of the reals is greater than the cardinality of the rationals

The quadratic formula states that there is one formula for the roots of any quadratic equation, and shows what the formula is.

Many theorems are of the form "There does not exist a ..."[2], "There exists a ..." , "The ... has the property ..."

Conjectures are likewise often simply statements of a fact about many things:

The Riemann Hypothesis states that all non-trivial roots of a certain equation have real part = 1/2.

Goldbach's Conjecture states that each positive integer greater than 2 is the sum of two primes.

The 3x + 1 Conjecture states that the $3x + 1$ function terminates with 1 on all positive integers.

---

1. If we didn't know that, then it could happen that in attempting to solve a problem, we needed to know if $x^n + y^n - z^n = 0$. And so we would have to somehow figure out if this was true. Perhaps a huge amount of labor and computer time would be required But since we know that FLT is true, we know immediately that $x^n + y^n - z^n$ never equals zero. So FLT can be thought of as stating a fact about an infinity of things.

2. Gödel's Second Incompleteness Theorem states that there does not exist a proof by which a consistent axiomatic system which includes Peano arithmetic can prove its own consistency.

A strategy that has enabled me to discover possible solutions to three very difficult problems[1], is in accordance with the idea set forth in this section. The strategy is: find a structure containing all possibilities (the "many things"), and that shows important relationships between them.

When God Created Mathematics, Did He Create the Proofs First?

Or did he create the important statements first, and leave the proofs to us humans? Perhaps all he created was the idea of a formal grammar, and the fact that a formal grammar underlies all axiom systems in mathematics, and then washed his hands of the rest, saying to us, "If you people decide that certain short strings produced by these grammars are important (say, the short string that expresses Fermat's Last Theorem, or the short string that expresses the Riemann Conjecture), then *you* figure out the sequence of strings that led to these short strings. I have more important things to do..."

What Is Needed To Make All of Mathematics Efficiently Accessible?

Only a professional mathematician can continue to believe that the huge quantity of present-day mathematics, with tens of thousands of new theorems being published each year worldwide in all disciplines, requires no change in how mathematics is accessed — namely, via a tree listing the major subjects, and then via textbooks that are organized the same way that Euclid's was in 300 b.c.

A radically different and, I think, much more time-saving approach to mathematical subjects, is presented in William Curtis's *How to Improve Your Math Grades* on occampress.com. The following is a brief outline of his main idea.

We can think of each mathematical subject as a set of "entities", and by an entity he means anything that is represented by any mathematical term. Associated with each entity name is a "template" which includes: definition of entity, representations of entity, common operations on the entity including ways of determining if two entities are equivalent, ways of creating more of the entity, and of breaking down the entity, and (in some cases) arithmetic functions involving the entity. In addition there is a tree representing the types of the entity. And a list of theorems and lemmas that contain the entity in their statements. Finally, there is a list of other entities to which the given entity is closely related — in the ideal case, there is a reference to each and every other entity in which the given entity appears in theorems and lemmas.

The set of entities includes commonly-occurring algebraic expressions, even though some fields of the template may be empty. But at the minimum, associated with each expression is a list of other entities (theorems and lemmas) in which the expression occurs.

The entities are listed in alphabetical order in paper presentations. There is, in addition, a complete index of all symbols that are part of the subject.

Such a presentation of a subject Curtis calls an "Environment".

In my experience, an Environment saves a great deal of time, one reason being that it is now no longer necessary to know everything in order to do something. The Environment makes it much easier for the reader to go as deeply into knowledge of an entity as he or she desires, which is exactly what is needed when there is far too much mathematics to learn in the old way. An Environment makes possible what has sometimes been called "just-in-time learning".

---

1. See the papers, "A Solution to the $3x + 1$ Problem", "Is There a 'Simple' Proof of Fermat's Last Theorem?", and "A Possible Proof of Goldbach's Conjecture", on occampress.com.

But even if (when) a complete Environment exists for each subject, the problem of efficient access to all mathematical subjects will hardly be solved. And so we must step back and consider how much mathematics there will be 10, 20, 30, ..., 100 years from now, and ask what might be the least inefficient way of accessing items in all this knowledge.

First, we can assume that the goal of all access is to solve a problem. Second, we must proceed throughout on the assumption that each user has a specific minimum knowledge of mathematics — say the equivalent of an undergraduate degree's worth. One thought is to have a thoroughly cross-referenced index of *concepts*, with a hierarchy of sub-concepts in each case. This is fine if the user knows the concept he wants further information about, but not if he or she doesn't. For example, suppose the user wants information on what amounts to fiber bundles, but doesn't know that term. So we need to think of possible ways to express esoteric concepts in some minimum language. Thus, e.g., if the user had never heard of topology, but wanted information regarding the abstract concept of "nearness", then he or she could find "topology" via a cross-reference from "nearness".

A short-cut to finding a concept whose name we may not know is simply to have (1) a web site where users can give descriptions of the concepts they want more information about, and (2) in each specialty, have at least one expert who routinely scans the descriptions on the web site and if one of them seems to be related to his specialty, he contacts the person seeking the information.

Is Mathematics As Difficult for Mathematics To Do As It Is for People To Do?

The question was motivated by a passage in a book on Buckminster Fuller:

"'I'd learned at school that in order to make a sphere, which is what a bubble is, you employ *pi*, and I'd also learned that *pi* is an irrational number." So 'when', he recalls asking, 'does nature have to fudge it and pretend it comes out even and then make some kind of compromise bubble?'" — Kenner, Hugh, *Bucky: A Guided Tour of Buckminster Fuller*, William Morrow & Company, Inc., N.Y., 1973, p. 132.

Think of some of the back-breakingly difficult problems that have existed in mathematics. And yet many of these, perhaps most, are easy to state: the three great problems left by the ancient Greeks — the squaring of the circle, doubling of the cube, and trisecting of an angle using only ruler and compass in each case — ; Fermat's Last Theorem; the Riemann Conjecture, the Poincaré Conjecture, and others. Does the subject of mathematics (or God) have to work as hard to *create* these truths as mathematicians have to work to prove them? Or does mathematics simply run out all paths in each possible logical grammar, and then we discover that, in some cases, at the end of some very long paths, there is a short string? (This possibility is known to exist in some kinds of formal grammar.)

Is Mathematics Really "Language-Like"?

Is mathematics primarily language-like? If so, then we can view, "Can you find a proof of this theorem?" as meaning "Can you 'say' this theorem in the language of this subject?" The growing feeling, as our study of a subject progresses, that we can solve more and more problems, do more and more proofs, is then equivalent to the feeling, when learning a natural language, that we can say more and more things.

Language vs. Geometry in Technical Subjects

Imagine that we had a language for speaking about the locations of objects in a rectangular frame. For example, we could say things like, "Well, the blue triangle is below and a little to the right of the red circle." Or, we could give precise coordinates of things. Speaking about any new arrangement of plane figures would then be easy, because our language would already be in place: "Just look at the picture and talk." Now consider the problem of a person who understood the syntax and semantics of the language, but was never given the picture. His labors to "understand", i.e., to deduce the picture, would be much greater than those of a person with the picture before him who was translating into words and/or numbers the arrangement of the pieces. So, in Artificial Intelligence, why not first develop a language for talking about pictures, then map the state of affairs into a picture, then talk about it?

Why is a picture worth a thousand words? Consider tables that represent facts and ask how they actually replace prose. What does prose really "do"? What do symbolic logic strings do relative to the underlying semantic structure?

Viewing Any Mathematical Subject As a Version of Another Mathematical Subject

To what extent is it possible to view a given mathematical subject as a different subject? To begin with, we can, and usually do, view a mathematical subject in terms of sets. We can view any mathematical subject as a formal language for which a model exists. So far so good. But can we, e.g., view elementary number theory as a special case of the theory of vector spaces, in which arguments of functions are represented as vectors, with the value of the function being represented by another vector? Category theory is one way of viewing one subject as another. Is it the best way, and if not, why not?

(This is a question about how self-similar mathematics is. Homomorphisms and isomomorphisms, the beginning student of modern algebra is too seldom reminded, are a way of saying, "This is really like that.")

What does it mean to say, e.g., "Group theory is not applicable here."

"Why Not a Course in All Things That Are Equal to the Number 2?"

Students of technical subjects are familiar with the basic courses in mathematics: analytic geometry, calculus, linear algebra, number theory, ... And yet in each of these courses, certain terms, certain quantities and relations recur again and again. So why not a course in each one of these? A course in all the things that equal 2, all things that equal 3, all things that equal 4, up to some maximum based on frequency of occurrence in undergraduate courses? Why not a course in all things called "linear"? In all things that are less-than-or-equal-to other things? ...

Why not a proof technique in which, if we are trying to prove the conjecture $x = y$, we simply go to a table of all things that are y and see if any of them are equal to x. If the answer is yes, then we have a proof. If the answer is no, then we have a disproof.

Does a KWIC Index of a Mathematical Subject, Tell Us Anything Important About the Subject?

A KWIC (Key Word In Context) is one in which each entry is alphabetized by each significant term in the entry. Thus, e.g.,

Theorem: *The eigenvalues of a matrix A are the roots of its characteristic polynomial,*

is indexed

under "C" as
Theorem: *The eigenvalues of a matrix A are the roots of its* ***characteristic polynomial****.*

and under "E" as
Theorem: *The* ***eigenvalues*** *of a matrix A are the roots of its characteristic polynomial.*

and under "M", as
Theorem: *The eigenvalues of a* ***matrix A*** *are the roots of its characteristic polynomial.*

and under "P" as
Theorem: *The eigenvalues of a matrix A are the roots of its characteristic* ***polynomial****.*

and under "R" as
Theorem: *The eigenvalues of a matrix A are the* ***roots*** *of its characteristic polynomial,*

Several questions suggest themselves"

(1) If we ask, "What subject is the Theorem in?" what should our answer be?

(2) If we make a KWIC index for each mathematical subject, what, if anything, will the indexes reveal about the relationship of each subject with other subjects?

(3) Is anything gained by regarding the boldfaced terms as "coordinates" of the Theorem in the body of mathematical knowledge?

(4) Is anything gained by regarding a mathematical subject, not as a linear structure, but as a set of statements occurring simultaneously in several different "subjects", where, here, we regard, e.g., "polynomial", "root", "eigenvalue", etc. as separate "subjects". See previous sub-section.

"What = Where"

"Roger Bacon died unaware that future historians would give him the title 'Doctor Mirabilis', the Wonderful Teacher, for whom every book had a place that was also its definition, and every possible aspect of human knowledge belonged to a scholarly category that aptly circumscribed it." — Manguel, Alberto, *A History of Reading*, Viking, N.Y., 1996, p. 197.

Suppose that the value of something equals its address, its location, in some system of coordinates. In these cases, we can say that "what = where", or that "semantics = syntax". The periodic table of the elements is probably the best example of such a scheme. Here, the chemical properties of an element are determined by the location of the element in the table.

Consider a binary tree representing binary numerals: "0" denotes descent down a left branch of a node, "1" denotes descent down a right branch, so that each node represents a unique numeral. Then the description of how to get to a specific node — i.e., the sequence of branches to take — constitutes the *numeral* that is found at the node. However, this does not quite give us the *semantics*, since the same *number* will have an infinite set of representations (numerals), e.g., the number 5 is represented by the numerals 101, 0101, 00101, 000101, ...

However we *do* get the unique number if we declare that the paths define binary proper *fractions*.

Similarly for the real number line, where each fractional value of a number is a unique location (point) on the line between the integral part of the number, and the next integer.

Suppose we were able to freeze in time all the mathematics in the world at present — all the math papers in all the math journals, all the textbooks, all the correct mss. in circulation and on the Internet. Suppose that then for each term — each number, each expression — we listed all the places where that term appears. What would be a good order in which to list the numbers and expressions? Lexicographical? But given any listing, what, if anything, could we learn from merely knowing all the places that each number or expression or term appears? Well, a definition of, say, the number 2 would be a listing of all the places where 2 appears. ("The meaning is the use." — Wittgenstein)

No matter what someone, genius or not, is doing in math, he or she is always somewhere in the set of all possible deductions.

Is there a meaningful, useful way that we can classify *steps* in proofs — all proofs in all subjects? Or is the best we can hope for a categorization based on the type of formal logic statement?

Is there any possibility that we can devise a topology for lemmas and theorems — and, for that matter, for problems in general — such that those which are "close" to each other have similar solutions (proofs)?

Suppose you were given the task of classifying all the exercises in a calculus textbook. You might reply that the textbook itself provides such a classification, in that all exercises pertaining to the subject of each chapter are typically given at the end of the chapter. But how would you further classify the problems at the end of each chapter?

Every statement — including every lemma, every theorem — has a location. "Where does this statement 'go'?" is an important question.

Why shouldn't there be "specialists" for each important term, and number, in mathematics? You would go to such a person and, e.g., say that you were having trouble with a proof that involved the term such-and-such. The specialist would then give you a summary of all the places in the mathematical literature where that term appears. You might ask for more detail about this or that place.

"During the past fifty years, more mathematics has been created than in all previous ages put together. There are more than 1,500 mathematical research journals, publishing some 25,000 articles every year (in over a hundred languages). " — Stewart, Ian, *The Problems of Mathematics*, Oxford University Press, New York, 1992, p. 19.

As the body of mathematical knowledge grows at this rate, it becomes more and more desirable to be able rapidly to understand the most important results in a given paper. Imagine that mathematics were a subject whose sole aim was to discover new real numbers. At the top of each paper would be the following statements:

"The previously-known largest decimal number for the integer 5 was 5.712.
This paper proves that 5.7128 exists."

Certainly that would enable us quickly to understand what the paper had accomplished. Is it possible to come up with a code that would vastly increase the speed of our understanding what a real paper in modern mathematics has accomplished?

Mathematical Subjects Whose Structure Is "Flat"

One characteristic of the calculus is the large amount that the successful student needs to know in order to perform well on exams. Integration affords just one example. Proofs in the typical advanced text often require knowledge drawn from every nook and cranny of the subject. (The practice of textbook authors of not always giving specific justifications for the statements in proofs does not help, of course.)

So we may be inclined to describe the structure of subjects like the calculus as "flat", meaning that a great deal of knowledge is often required to solve even elementary problems. The structure of a subject like elementary congruence theory, on the other hand, does not seem to have this property.

Is there a specialty that investigates the structure of mathematical subjects? Such a specialty might begin by counting the number of facts a student needs to know in order to get an A on a final exam. Then the specialty might develop formal grammars for each subject, and compare these.

Mathematical Subjects That Are Not Modular

I will call a subject *modular* if it can be regarded as consisting of modules such that, informally, a change in one does not cause a change in another. I will call a subjed *non-modular* if a change "here" does cause a change "there", and in fact in many places in the subject. Mathematics students run into the non-modular properties of a subject when what looks like a solution isn't one because it would produce, say, a falsehood in another part of the subject. "We thought this might work, but then we realized that it would imply that such-and-such was the case over there, and it's not. "

Is is possible to make these concepts rigorous, so that subjects might be classified by the degree to which they are not modular?

Patterns in the Symbols

In mathematics and the hard sciences, unlike the humanities, what counts is semantics, not syntax. *What* is said is important, not *how* it is said. Even though a good notation can make clear what a bad notation conceals, the truth of a mathematical statement does not depend on the particular symbols used to express it, only on their meanings, their definitions.

Nevertheless, we must also acknowledge that a page of mathematics has an esthetic of its own, regardless if you know what the symbols mean. A person who knew nothing of the meanings of the symbols could, by laborious examination, come up with a grammar to describe the allowed sequences of symbols. In the standard language for elementary algebra, for example, the sequence of symbols "$= + - x$" standing separately, is not grammatical, but the sequence "$z = x - y$" is.

But now consider the reverse situation, i.e., the writing down of sequences of symbols just because they look nice, even though they may in fact express nothing of interest, much less of use in the subject. Is there perhaps a "higher esthetic" in which beautiful sequences, beautiful-looking pages, always, or usually, express important truths? Does great mathematics look great, and ugly (i.e., unimportant) mathematics look ugly, to one who is in touch with the right esthetic? Are there people who understand the higher syntactic esthetic, and can apply it?

I once saw a woman on *60 Minutes* who suffered from a mental disability, but who nevertheless was able to write part of a string quartet while talking to reporter Leslie Stahl. The woman made clear that she was not hearing the music as she wrote it. The music was as new to her when it was later played as it was to any other listener. Perhaps she had somehow internalized the syntax of written music, and was just creating more, based on this internal grammar.

One suspects that something similar might have been going on in the case of the mathematian Ramanujan, who often saw mathematical expressions and equations in dreams.

And we must ask if this higher syntactic sense might not underlie the extraordinary ability at determining the primality of large numbers possessed by two otherwise severely mentally disabled twins, as reported by Oliver Sacks. Could it conceivably be that they saw something in, say, the appearance of the *numerals*, that enabled them to "deduce" that a given large number was prime?

Perhaps the most amazing example of writing whose appeal is purely syntactic, purely visual, is the *Codex Serafinianus* (Abbeville Press, New York, 1983), which is an encyclopedia about a non-existent country, full of non-existent creatures and machines and other objects, hand-written in a non-existent language which, as far as I know, has never been deciphered, if in fact, that term is even appropriate. It seems possible that one could study the text and develop a grammar for the language, and then write in it, without having the slightest idea of what, if anything, the writing meant.

Ramanujan and "Esthetic Grammars"

"Ramanujan generated formulas which he felt to be true on the basis of intuition and the checking of some special cases. He generally did not provide a rigorous proof of his results. Generally he was not strong in establishing such rigorous proofs." Watkins, Thayer, "Srinivasa Ramanujan, A Mathematician Brilliant Beyond Comparison", www.sjsu.edu/faculty/watkins/ramanujan.htm, Apr. 3, 2012.

We must recognize that there are different types of mathematical ability, and that these types range from the conceptual to the deductive to the computational to the pattern-recognizing-and-creating type exemplified by Ramanujan.

I cannot imagine Ramanujan, for all his genius, ever coming up with proofs like Cantor's that showed, e.g., that the cardinality of the rationals is the same as that of the integers, but that the cardinality of the reals is not the same, or that "most" reals are transcendental. I cannot imagine Ramanujan developing set theory.

Although he was mathematically precocious, his obsession with mathematics began when, at about the age of 17, he "came into contact with the book, *A Synopsis of Elementary Results in Pure and Applied Mathematics* by G. S. Carr. This rather eccentric book is essentially a huge collection of formulas and theorems compiled for students preparing for the celebrated Mathematical Tripos examination at Cambridge."[1]

Calculus students can perhaps get an idea of the beginning of Ramanujan's obsession by recalling their cramming for exams that involved lots of formulas, as in, say, integration. They learned a few rules, e.g., integration by parts, and then they became good at applying them to many different problems, and doing so rapidly, aided by lots of memorization. They may recall how, once they were into the effort, they began to see solutions quickly. They became good at solving that type of problem. Perhaps they began to feel that they had entered into a higher level — a kind of frenzy — as a result of the intensity of their efforts and their growing skill. Perhaps a similar frenzy is what young chess players enter into when they become obsessed with the game.

But even students who are destined for careers in mathematics do other things in the subject: they work on proofs and try to master difficult concepts. Ramanujan apparently stayed with the formulas, becoming ever more skilled at developing new ones from the ones he had developed. Perhaps this limitation is why Morris Kline, in his *Mathematical Thought from Ancient to Modern Times*, one of the best histories of mathematics ever written, does not mention him.

If I were going to spend a large amount of time studying Ramanujan, I would not spend it on trying to prove his results, but rather on investigating two types of grammar: the first I will call an "esthetic grammar" . This grammar would *not* be the formal logical grammar underlying the subjects he worked in. Rather, it would be a very illogical, even bizarre, set of possibilities, perhaps along the lines of "...if this kind of expression is in the numerator, and it contains a 5, and there are parentheses in the expression *u* in the numerator, then it is possible that the denominator contains this other kind of expression, along with the power *n*, multiplied by a factor *w* containing the expression ..." The grammar may be based on the frequency of occurrence of various strings in other formulas.

Computer Generation of (Some of) All Possible Strings in a Subject

A mathematical subject is a Type 0 formal language. The grammar for such a language is capable of producing a shorter string of symbols from a longer one. Therefore if we tell the grammar to generate all possible strings less than or equal to a specified length, we have no guarantee that all the strings will be generated in a finite time. However, we can at least tell the grammar to show us all the strings it generates that are within the length bound we specify, and let the grammar keep running until we reach the end of our computing resources.

A string can be an expression in the mathematical subject, or it can be a statement. In the latter case, the sequence of strings that led to the statement constitutes a proof of the statement.

Such a grammar could be developed for the subjects that Ramanujan worked on, and therefore it might be possible to generate some of the strings (statements) that Ramanujan discovered. By studying the sequence of strings that lead to each, we might discover something about Ramanujan's thought processes. Of course, we might discover some important statements that he did not discover. (*Note*: the type of grammar I am discussing here is almost certainly not an esthetic grammar, as described in the previous section.)

A Note on Infinite Sequences

---

1. *The Princeton Companion to Mathematics*, Gowers, Timothy, ed., Princeton University Press, Princeton, N.J., 2008, p. 808.

We must never forget that it is *we* who read infinities into the symbols we write. All we have before us is a short string of symbols: "1, 2, 3, ... " This string only *represents* an infinite set. Similarly, we need to remind ourselves every once in a while that even in the most breathtaking perspective painting or drawing, the distant mountains are in fact as close to us as the canvas on which they appear.

Sounds, Smells, Feelings in Mathematics

We take it for granted that at least some mathematical entities have visual representations. By the end of our high school years, assuming we have had good math courses, we are familiar with the idea of graphing functions as curves, surfaces. But we have other senses, in particular, the sense of hearing, and the sense of smell.

Would we gain anything by graphing functions as variations in the pitch of a musical tone, so that, e.g., the higher the pitch, the larger the value of the function? (We would need some sort of audible clock to indicate how rapidly the values were changing as we moved along the *x*-axis.) In the mid-nineties, I saw a PBS program that described the use of sound to represent deviation from statistical limits.

"Perhaps more surprisingly, the CMC [the Computer Music Center at Columbia University] is beginning to explore ways in which sound technology can be used to teach quantitative subjects like mathematics. While rapid increases in computer power allow even undergraduates to work with mathematical models in four, five, or higher dimensions, it is impossible to represent these complex geometries visually — a significant hurdle in trying to help students really understand the math involved. It does seem, however, that these geometries can be explained and understood through sonic modeling, which represents data as audio with specific sonic attributes. Certain features of large datasets can be heard more easily than they can be seen, and relatively new areas of mathematics such as fractal geometries and nonlinear dynamical systems are well represented through sound." — Levitt, Jesse, "The Sounds of Science", *Columbia*, Columbia University, N.Y., Winter, 2000, p. 36.

We sometimes say, "I smell a good idea here." Suppose it were possible, in a systematic way, to associate a smell with a math subject, and even with particular concepts and theorems, so that a sufficiently sensitive mathematican would be able to use these smells as a basis for intuitions. "It just seems to me that we might find a proof for this conjecture in subject *x*, although I can't tell you exactly how."

Those readers who find these reflections too fanciful to have anything to do with real mathematics, should take a look at John Conway's book, *The Sensual (quadratic) Form*[1], e.g., "The Second Lecture: Can You Hear the Shape of a Lattice?", "The Third Lecture:..and Can You Feel Its Form?", and "The Fourth Lecture: The Primary Fragrances."

Perhaps we can go even farther in connecting physiological responses to mathematical ones. Einstein once remarked that he sometimes experienced intuitions as "muscle tensions".

Is Mathematics "Continuous"?

If we conceive of mathematics as a mapping from strings representing problem statements, to strings representing solutions, then what sub-set of this mapping is "continuous" in the sense that two problem statements which are "almost the same" have solutions which are "almost the

1. Published by The Mathematical Association of America, 1997.

same"? It may be necessary to stipulate that "almost the same" means "having almost the same number of bits in the minimum description".

Take any infinite series that represents the solution to a problem, say, to a differential equation. There will be a rule by which corresponding exponents in each term increase with each successive term: the exponents may increase by 1 or 2 or some other number. Similarly, there will be a rule by which corresponding factors develop, e.g., in the first term there may be a factor $(n + 1)$, in the second, two factors, $(n + 1)(n + 2)$, in the third, three, $(n + 1)(n + 2)(n + 3)$, etc. We may be able to write down a rule that describes how exponents, factors, etc., are related in each term. Or perhaps the closed-form notation will do the job for us.

Now suppose we have such a rule, and suppose we make a "minor" change in the rule: e.g., if exponents increase by 2 with each successive term, we have them increase by 3, or 4. We ask not merely: How will the resulting value of the series change, but whether the series will be the solution of another equation and if so, which one? How "far away" is the new equation to which the modified series is a solution? (Of course, it may be infinitely far away, meaning, that there is no equation to which the modified series is a solution.)

A way to categorize series, though not a practical one, is to categorize them in increasing order of the length of the shortest description of each, where "shortest" means in number of symbols.

Consider the well-known Laplacian equation. In three dimensions, it is:

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 0$$

where U is a continuous function having continuous derivatives of the second order. A great deal is known about this equation. But I have never come across a discussion of the equations:

$$\frac{\partial^3 U}{\partial x^3} + \frac{\partial^3 U}{\partial y^3} + \frac{\partial^3 U}{\partial z^3} = 0$$

or

$$\frac{\partial^4 U}{\partial x^4} + \frac{\partial^4 U}{\partial y^4} + \frac{\partial^4 U}{\partial z^4} = 0$$

or

...

Why haven't mathematicians considered it as important to investigate these equations as they have the Laplacian? Are the solutions to these in some way "similar" to each other, and to the

Laplacian, just as their written forms are similar (they differ only in one positive integer)? Knowing the solution to one equation, would it be easy, or at least possible, to go from that solution to the solution to the equation preceding or following it in the above order?

When we organize technical information alphabetically by title, it is rarely the case that successive items correspond to meanings that are "similar". And if we were able to organize information semantically, so that things having similar, or close meanings, were close, then certainly the titles would not be alphabetical.[1] Can we every have it both ways? If not, why not? If so, how?

This notion of continuity is present in the theory of evolution, in that biologists assume that a slight difference in form corresponds to a slight difference in the time when an animal lived.

Is Mathematics Created or Discovered? An Answer

This question goes back to Plato at least. The following seems to me a compelling argument that mathematics is discovered. If intelligent life exists elsewhere in the universe, and if this intelligent life knows and uses mathematics, then, although the symbols in which the mathematics is written will almost certainly be different from ours, the truths the symbols represent will be the same as ours in areas where our knowledge and theirs coincide. But if this intelligent life has political organizations — say, the equivalent of our national states — and if any of these have tax codes, then it is almost certain that the tax codes will differ from ours.

How Much Mathematics Can There Be?

In the film *Annie Hall* there is the following scene. Alvy Singer is the Woody Allen character.

Alvy as young boy sits on a sofa with his mother in an old-fashioned, cluttered doctor's office. The doctor stands near the sofa, holding a cigarette and listening.
MOTHER (To the doctor) He's been depressed. All of a sudden, he can't do anything.
DOCTOR (Nodding) Why are you depressed, Alvy?
MOTHER (Nudging Alvy) Tell Dr. Flicker. (Young Alvy sits, his head down. His mother answers for him) It's something he read.
DOCTOR (Puffing on his cigarette and nodding) Something he read, huh?
ALVY (His head still down) The universe is expanding.
DOCTOR The universe is expanding?
ALVY (Looking up at the doctor) Well, the universe is everything, and if it's expanding, someday it will break apart and that would be the end of everything!
Disgusted, his mother looks at him.
MOTHER (shouting) What is that your business? (she turns back to the doctor) He stopped doing his homework.
ALVY What's the point?

— www.script-o-rama.com

1. But consider an "entity first, types second" way of writing terms. Thus, e.g., "finite simple group" would be written as "group, simple, finite", and "second order linear differential equation" would be written as "equation, differential, linear, order, second". With this rule for writing terms, terms that were semantically closely related, would be close to each other in the alphabetical list as well.

The following may seem as absurd a concern as Alvy's, but nevertheless I feel the question has to be raised, namely, "How much mathematics can there be in our physical universe?" "How much mathematics is there *room for*?"

Suppose we decided to expand π as far as the universe allows. Assuming one atom for each digit, then at most about 10^{80} digits would be possible, with no room for anything else except coincidentally.

In limited-memory domains, compromises have to be made: if you want to say more about subject *x*, if you want more accuracy *here*, then you have to settle for less room to talk about that *over there*.

How much more mathematics is there now than there was, say, in 500 a.d., as measured by, say, the total number of words and terms, including terms in equations, in all the mathematical works then and now? Suppose we were running out of paper and computer memory: what would we want to throw out first? Second? Third? Etc. One reply is, of course, that we could afford to throw out everything except all the axioms that enable us to derive all the mathematics then and all the axioms that enable us to derive all the mathematics now.

What exactly do we gain from a lemma or a theorem in terms of number of symbols used? Here is the set of all positive integers, represented, say, as $N = \{1, 2, 3, ...\}$ Suppose we think up the sequence of patterns which is $\{N \bmod 1, N \bmod 2, N \bmod 3, N \bmod 4, ...\}$ (meaning, *N* organized into the finite set of residue classes for each modulus). It has taken us additional symbols to define this sequence of patterns, but we now "know more" than we did before, namely, that such an organization is possible. At some point (quite soon, in fact!), the number of symbols needed to represent our knowledge of something — in this case, of the positive integers — will start to exceed the number of symbols we need to represent the something itself. What are we gaining? What are we giving up?

Is it possible that the ultimate limit to mathematics is the number of atoms in the universe? In other words, the number of atoms that can be used to represent mathematical knowledge?

What is "all of mathematics"? It is a countable list of axioms, definitions, lemmas, theorems, and proofs. But the number of subsets of just the positive integers is uncountable. If a subset represents a property, then we cannot even *list* all the properties of just the *positive integers*.

Are all the subsets of a set all of its properties? But all the subsets of a set themselves are a set. What about all the sets that are subsets of other sets? Etc.

Does a Proof Decrease the Complexity of Mathematics?

By "complexity" here I mean the number of bits needed to represent mathematics. Something that has a pattern, for example, "10" repeated a billion times, requires fewer bits to represent it (I just gave a representation) than something that doesn't have a pattern, for example, a typical sequence of two billion 1s and 0s where each 1 corresponds to a head being tossed by a fair coin, and a 0 corresponds to a tail being tossed.

Consider Fermat's Last Theorem (FLT), which, after 300 years of effort, was finally proved in the early 1990s. The Theorem can be viewed as a statement about a function, namely, the function $U(x, y, z, n) = x^n + y^n - z^n$, where *x*, *y*, *z*, *n* are positive integers, and $n > 2$. The Theorem asserts that *U* never has the value 0. Before the Theorem was proved, it would be necessary, for sufficiently large *x*, *y*, *z*, *n*, to attempt to find the value of the function using the computer and/or extensive reasoning, with no guarantee of finding the answer.

(Question: what, if anything, is known about the frequency of values of the function other than 0?)

Was the function *U* more complex before the proof of FLT than after?

What is the complexity of all of mathematics today, that is, what is the minimum number of bits required to represent all of mathematics today? Will all of mathematics eventually prove to have a finite complexity, or an infinite complexity?

How Do We Know When a Mathematical Subject Is "Finished"?

It is sometimes said that no major new developments can be expected in the theory of complex numbers, that the subject is mature — "finished". How is this determined? What are the criteria? Shouldn't mathematicians in each specialty be working toward fulfilling these criteria for their specialty? Are they in fact doing so? If not, why not?

Instead of trying to determine if a subject is finished, we might instead try to establish a measure of the degree to which it is finished. One way might be the following: using the initial definitions, establish a formal grammar of all well-formed expressions in the subject. For each string-length 1, 2, 3, ..., there exists only a finite number of well-formed expressions of that length. If we have proved or disproved all the well-formed expressions up to a given length *n* that are assertions, then we know that, if there are any new discoveries to be made in the subject, their statements must be of length greater than n. Unfortunately, the most general formal grammars will include rules that produce strings that are shorter than the ones to which the particular rule of the grammar is applied, and thus in principle we will not be able to tell when we have all the strings of a given length.

The "Big Picture" of Mathematics: Some Surveys

A few of us believe that, no matter how specialized our mathematical efforts become, it is important to maintain, throughout our careers, as broad a grasp of the whole of mathematics as possible. One reason is that there is absolutely no guarantee that the problems in our specialty will have answers within that specialty, and so we need to know "where the concepts are", or at least, where the leading concepts are. Therefore, as a service to others who believe as I do, I would like to offer a list of surveys of mathematics that seem to me to be useful. As of July, 2009, apparently none of the software available to academic booksellers was capable of doing searches under this category, e.g., "survey, mathematics", or "survey of mathematics", so I will welcome suggestions from readers as to books that might be added to the list.

Items in the list are roughly in order of *decreasing* mathematical background required. One general comment is that the indexes, like those of almost every mathematics book I have ever seen, are usually frustratingly incomplete. I am sure the reason is that the ancient, and now obsolete, linear paradigm continues to rule the presentation of mathematics. One must begin on page 1, remember the contents, then proceed to page 2, remember the contents, then proceed to page 3, ... William Curtis's important and pioneering book, *How to Improve Your Math Grades*[1], makes a convincing case that this paradigm is not viable in a world in which some 1,500 mathematics journals publish more than 25,000 papers a year containing more than 200,000 new theorems.

List of Surveys

Here is the list:

1. Available on the web site www.occampress.com

Mac Lane, Saunders, *Mathematics: Form and Function*, Springer-Verlag, N.Y., 1986.
The book is apparently out of print. I was told by the manager of an academic bookstore that if I could find a copy for under $135, I should grab it. I was able to buy one for $90 from abebooks.com.

Gårding, Lars, *Encounter With Mathematics*, Springer-Verlag, N.Y., 1977.

Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, N.Y., 1972.
This is the best history of mathematics I know of. It also can serve as an excellent survey of mathematics up to the time of its publication.

Aleksandroff, A. D.; Kolmogorov, A. N.; Lavrentev', M. A.; *Mathematics: Its Content, Methods, and Meaning*, tr. Gould, S. H. and Bartha, T., 3 Vol.s, The M. I. T. Press, Cambridge, MA 1963.
An outstanding work, despite the occasional baffling proof. My only complaint — and it is a minor one — is that it could lead the reader who doesn't know better into believing that just about all progress in modern mathematics was made by Russians.

***The Princeton Companion to Mathematics*, ed. Gowers, Tim, Princeton University Press, Princeton, N.J., 2008.**
A useful work that should be in the hands of anyone planning to devote at least part of his or her life to mathematics. Its faults include its very incomplete index, the lack of an index of symbols, and quite a few sections that could have been much clearer.

Courant, Richard, and Robbins, Herbert, *What Is Mathematics?*, Oxford University Press, N.Y., 1969.

Stewart, Ian, *Concepts of Modern Mathematics*, Penguin Books, Great Britain, 1981.

Newman, James R., *The World of Mathematics*, 4 Vols., Simon and Schuster, N.Y., 1956.

Davis, Philip J., and Hersh, Reuben, *The Mathemaical Experience*, Houghton Mifflin Company, Boston, MA, 1981.

Critique of Three of the Surveys
I will offer a critique of three books in the above list that I think have major shortcomings. The rest are acceptable. Courant's, Kline's and Newman's can fairly be called classics.

Mac Lane, Saunders, *Mathematics: Form and Function*
According to the Preface, the author wrote this book "as a background for the Philosophy of Mathematics". On the one hand the book is admirable for its range and depth given the number of pages (456). Also for its emphasis on the connection between mathematics and the real world.

On the other hand, the presentations of topics are by no means as clear as they could be within the same space, which is surprising, since the author is known for the clarity of his style, e.g., in *A Survey of Modern Algebra*[1]. Statements that are easily understood are followed by statements that require long proofs, with no warning to the reader. It is abundantly clear that the author made no attempt to test the clarity of his text on randomly-selected members of his intended audience.

His treatment of the Lagrange equations (pp. 267-274), which are representations of Newton's second law, should convince any skeptics as to the clumsiness, inefficiency, obscurity and hence unnecessary difficulty of the discursive presentation of lemmas and theorems. In this type of presentation, a long sequence of equations, with commentary, is given, at the *end* of which the reader is finally made aware of the lemma or theorem that all this effort was aimed at deriving. There are no sub-titles, no indication of definitions and important statements, e.g., via boldface or italic type, or by placing definitions and statements on separate lines. The following are some headings that would make this section easier to understand. I certainly do not say this is the best set of headings.

Derivation of the Lagrange equations
- Rectangular coordinates expressed in terms of polar coordinates
- Radial force
- Torque
- Kinetic energy
- Derivation of the Lagrange Equations from Newton's Second Law

The Lagrange equations hold in any coordinate system
The Lagrange equations and generalized force for *N* particles
The Lagrange equations apply to motion under holonomic constraints
The Lagrange equations apply to motion under non-holonomic constraints
The Lagrange equations apply to the motion of a simple pendulum

I urge the reader to consider how much clearer and more rapidly understandable the proofs would have been had they been presented in the format of structured proof as described in the chapter "Proofs" in Curtis's *How to Improve Your Math Grades* on the web site occampress.com. One of the characteristics of this format is that the goal (the lemma or theorem) is always given *first*.

The section has other shortcomings. At least one of the equations for the definition of the angular component of acceleration is wrong, as the reader can determine by simply drawing the related diagram. Fig. 1, a diagram of the relationship between radial and angular acceleration, is worse than useless for what it omits. I encourage the reader to see the correct diagram in any good calculus text, e.g., on p. 608 of Kline's *Calculus: An Intuitive and Physical Approach*[2].

Another example of what happens when a book like this is not tested on randomly-selected members of its intended audience, is the description of an ordered field in which the Archimedean law fails ("Comment", p. 102). The first part of this sub-section is perfectly clear, but then it is followed by a sequence of statements whose justification is utterly baffling. We cannot help concluding that during and immediately after the book was written, it was shown to a few trusted col-

1. Birkhoff, Garrett, and MacLane, Saunders, *A Survey of Modern Algebra*, A K Peters, Natick, MA, 1997.
2. Kline, Morris, *Calculus: An Intuitive and Physical Approach*, John Wiley & Sons, N.Y., 1977.

leagues who were specialists in the various sections. Each specialist knew backwards and forwards the material in the section that he or she was asked to review. And everything was perfectly clear! He or she was simply unaware of the missing justifications for statements, because statements and justifications were so familiar. The result is a book that is guaranteed to be frustrating for most of its readers. (Schorer's Law: Almost any presentation of a mathematical subject is clear and easy to understand for someone who already knows the subject.)

I challenge any reader who feels I am being unduly harsh in these criticisms, and who has not previously been exposed to adjoint mappings, to read the first few paragraphs under "Adjoints", pp. 200-201. I say it without a moment's hesitation: this exposition is a disgrace.

At the end of some chapters, the author includes tree-like diagrams that apparently are intended to show relationships between some of the concepts dealt with in the chapter. But these diagrams are utterly baffling. They seem almost to result from the author's free-associating on each term. What are we to gather from a line connecting "Mechanics" to" Quantity" and another connecting "Quantity" to "Dependence"? The lines are certainly not indicative of a subset relationship.

The number of typographical errors is a disgrace. In a third of the pages I have already found more than 30. On p. 196, there are references to statements (2.1) and (2.2), but there are no such statements in the chapter. The publisher, Springer-Verlag, apparently wanted to save money by not having a competent proof-reader make a final pass through the book before publication.

Some of the drawings are bizarre — e.g., the one representing the subset condition on p. 27.

The index, as is virtually always the case with mathematics books, is woefully incomplete. It has all the hallmarks of a mere afterthought. There is an incomplete index of symbols.

Gårding, Lars, *Encounter With Mathematics*

As with the Mac Lane book, statements that are easily understood are followed by statements that require long proofs, with usually no warning to the reader. Some of the drawings are utterly baffling.

Some of the proofs are so badly written as to justify being called inept. Consider the proof of Fermat's Little Theorem (p. 13), and keep in mind that the author is not writing a textbook, but rather a popularization for students "studying [mathematics] in the in the first year after high school." (Preface)

He proceeds as follows:

> The implication
>
> p a prime implies p divides $a^p - a$ is true for every integer a. (6)
>
> We prove the theorem the way Euler did it in 1736. [The author has previously proved (5) p divides $(a + b)^p - a^p - b^p$.]
> Combining a special case of (5), namely,
>
> p divides $(b + 1)^p - b^p - 1$
>
> with the hypothesis that (6) is true when $a = b$ gives the result that p divides the sum

$(b + 1)^p - b^p - 1^p + b^p - b = (b + 1)p - (b + 1)$.

Hence, since our implication (6) holds for $a = 1$, induction shows that it holds when a is any natural number.

Although I can certainly believe that the above is based on the *idea* underlying Euler's proof, I cannot believe that Euler's presentation of the idea was that bad. Here is a much better presentation.

Fermat's Little Theorem states:

If p is a prime and a and b are integers
Then p divides $a^p - a$.

Proof:

We use proof by induction.

Basis Step

By (5)[1] we know that p divides $(a + b)^p - a^p - b^p$.
So let $a = b = 1$. Then we have

p divides $(1 + 1)^p - 1^p - 1^p = 2^p - 2$.

Induction Step

Assume the theorem is true for all k such that $1 \leq k \leq a$. That is, assume that

p divides $2^p - 2$, $3^p - 3$, ..., $a^p - a$.

By (5) we know that

p divides $(a + 1)^p - a^p - 1^p$.

Therefore p divides $(a + 1)^p - a^p - 1^p$ and p divides $a^p - a$, that is,

p divides $(a + 1)^p - a^p - 1^p + a^p - a$.

Cancelling a^ps, we get

p divides $(a + 1)^p - (a + 1)$.

And we have our proof.

---

1. The author has previously proved: (5) p divides $(a + b)^p - a^p - b^p$.

Obviously, our improved version can be compressed, e.g., by not writing each equation on a separate line.

The index is worse than even the usual mathematics index. For example, on p. 149, a matrix is described as "singular". Keeping in mind that the book was written for students who are "studying [mathematics] in the first year after high school," it is reasonable to expect that the meaning of the term should be readily looked up. How long should that take? As long as it takes to find the word in the index and turn to the referenced page — a few seconds. Unfortunately, "singular" does not appear in the index, neither as an entry, nor as a sub-entry under "matrix", nor anywhere else. Clearly, the author expects each reader of his book to start on p. 1, read it, remember the contents, then go to p. 2, read it, remember the contents, etc. . In passing, I should mention that I have so far been unable to find the word "singular" anywhere in the book prior to p. 149, so the author clearly assumes that the first-year mathematics student has already had a course in linear algebra. It is also clear from other parts of the book that the author assumes that this student has already mastered advanced calculus.

Throughout the book, we find the disgraceful practice of authors of popularizations and textbooks, namely, the not giving justifications for *every* statement that is not part of the subjects that the author assumes that all readers will have knowledge of. (Keep in mind that Gårding is assuming his readers have only a knowledge of high school mathematics.) Students, trained never to question the competence of anyone who writes mathematics, bow their heads in shame at their inability to understand the statements lacking justifications, when in fact they should start an international protest (never revealing their names, of course) against this practice. A prime example is the section, "Implicitly defined functions", pp. 150-151. If you don't already know this material, keep a record of the total time it takes you to come to what you regard as an understanding.

A further indication of how out of touch with reality the author is, is his statement, "The classic *What Is Mathematics?* by R Courant and H. Robbins (Oxford 1947) is perhaps the best effort in this direction written for the general public." (p. 8) The general public knows next to nothing about mathematics, and regards it with fear and loathing. It is inconceivable that a reader with no technical training could read and enjoy Courant and Robbins' book. Certainly any kind of appreciation of the book requires at least a year or two of college mathematics.

I wonder what a superb writer of mathematics like Morris Kline[1] would have produced if he had been asked to write a Survey.

The Princeton Companion to Mathematics

As we said above, the book's faults include its very incomplete index, the lack of an index of symbols, and quite a few sections that could have been much clearer.

Some of these sections — for example, the one on partial differential equations — are little more than a list, in prose, of some of the more important types of the entity covered. Any author

1. As I have said elsewhere in this chapter, Kline is the author of what I regard as the best history of mathematics written in the 20th century, namely, *Mathematical Thought from Ancient to Modern Times* (Oxford University Press, N.Y., 1972), and one of the two best elementary calculus textbooks (*Calculus: An Intuitive and Physical Approach*, Dover Publications, Inc., Mineola, N.Y., 1976), the other textbook being Sherman Stein's *Calculus and Analytic Geometry*, 4th ed. (McGraw-Hill Book Company, N.Y., 1973).

who believes that this is the best way to present this type of material, should have no difficulty believing that prose is the best way to present the log tables. In the case of partial differential equations, the equations should first and foremost be presented in tree graphs and/or in tabular form with indents to show sub-sets. Prose commentary comes second. The fact that so many sections are presented as these prose lists is proof of the pathetically naive belief among technical authors that prose makes difficult subjects easier to understand for the non-specialist. The prose presentation is also clear evidence that the author never bothered to test his article on randomly-selected members of his intended audience. What did he hope his readers would be able to do after reading his article? Create the tree graphs and tables that he should have provided in the first place? What else? Did he even think to ask himself the question before he began writing?

The treatment of homology groups (pp. 389-92) is even worse, almost certainly because the author never bothered to ask himself, "What are my goals in this section? What do I want the reader to come away with? What tests will I give to randomly-selected readers to determine if I have succeeded?" At present, the reader comes away knowing that there are different homology group values, and feeling that he should understand certain drawings, e.g., of loops around toruses, although he does not. He is told at the start of the article that homology is a way of measuring how many holes there are in a topological space, but he is told essentially nothing about how this is done. I challenge the author to find even one student in a first-semester course in algebraic topology who can explain how the claims the author makes concerning the pinching of a circle or of a sphere follow from the definition of homology the student learned in class.

If we make the legitimate assumption that of far greater value than a collection of homology group values, is an idea of how these values are arrived at, then our first step is to decide what we will expect as minimum knowledge in any reader. (And yes, certainly, we should say what this knowledge is before the section begins.) It seems that the minimum knowledge should be the equivalent of, say, two undergraduate years in mathematics, including at least one semester of group theory.

Next, we need to state that homology is another means of answering the basic problem of topology, namely, determining if two topological spaces are homeomorphic or not. To do this, we first "triangulate" each space, that is, carve it up into k dimensional triangles, where a 0-dimenstional triangle is a vertex, a 1-dimensional triangle is an edge, a 2-dimenstional triangle is our familar planar triangle, a 3-dimensional triangle is a tetrahedron, etc.

There is a homology group for each k, its symbol being $H_k(X)$, where X is the topological space. $H_k(X)$ is a quotient group, just as the set Z/pZ of integers mod a prime p is a quotient group. Specifically, $H_k(X) = Z_k(X)/B_k(X)$. $Z_k(X)$ is the set of chains of k-dimensional triangles such that the chains are cycles, and $B_k(X)$ is the set of chains of k-dimensional triangles such that the chains are not only cycles, but cycles that are boundaries.

At this point, it is time for a carefully-selected example. Probably $H_1(X)$ will be the simplest and clearest. This requires a drawing. And then we need to show, first, how to determine $Z_1(X)$, then how to determine $B_1(X)$, then how to determine $Z_1(X)/B_1(X) = H_1(X)$. At some point, we should make crystal clear to the reader how the number of holes in X is determined from this process.

The presentation should avoid the kind of hand-waving that is so typical of textbook examples, in which geometric intuition is brought in whenever necessary in order to compensate for the author's laziness in specifying what happens in each step and why.

After the example, we can mention that cohomology is another numbering of homology groups, what its advantages are, and the fact that it is a dual to homology. Finally, we can say that there are homologies that use generalization to triangles that have curved sides, and then to cells.

I strongly doubt that a presentation such as I have described here would take more space than the one in the book. It would almost certainly be of greater value to the reader.

How Surveys Are Written Now

It seems clear from the above examples of Surveys that the way in-depth Surveys are written now is somewhat as follows. A mathematician — preferably one with a reputation as an author of popular textbooks — decides that it would be a service to those with any interest in mathematics at all, if he were to write a book that gives a survey of the major fields of mathematics at present. A commendable idea.

He has a vague notion of who his intended audience will be: mathematics students in their first year of college, adults with a technical degree. He sets to work in effect giving a prose summary of the general knowledge he possesses, believing (with incredible naivete) that prose somehow makes technical material less intimidating to the non-expert. Just the opposite is the case, since prose forces the reader to figure out relationships that could be easily and clearly presented via tables and other fixed formats. (There is a reason why the phone book is not written in prose.)

What the author writes is, in fact, a prose summary of major theorems, with perhaps some background on what led to the theorems, and a few biographical paragraphs on some of the mathematicans. In particular, in many cases, it is a prose description of one or more graphical *trees* that represent the hierarchy of types of entities in the subject, e.g., types of groups, types of rings, types of algebraic varieties, types of operator algebras, etc. — without the trees being given. .

He shows his book to colleagues, who, having been immersed in the subject matter all their professional lives, pronounce it clear and easy to understand. (If you already know a subject, then almost any presentation of it is clear.) His publisher likes the large audience that the mathematician claims he has written the book for, and publishes the book without even bothering to do careful proof-reading of the final version, because the publisher feels that even if a few errors slip through, the vast majority of the audience won't notice, and the cost-saving will increase his profits.

Readers in his intended audience are confused and humiliated because some things are readily understandable, others are not. They tell themselves, "I knew I wasn't intended to study this subject."

How Surveys Should Be Written

The project must begin with a clear definition of the audience, one that is based on minimum skills and knowledge that members of the audience are expected to have. Not "first year of college" but rather, e.g., "Knowledge of how to solve equations in one variable, knowledge of set theory and mathematical logic at the level presented in, e.g., Bittinger's *Logic and Proof*, basic knowledge of limits as presented in ...", etc.

Next come the crucial questions: What are the goals of the book? and What tests can indicate if the book is achieving those goals?

Then as each chapter is completed, the chapter must be tested on randomly selected members of the intended audience.

As the book proceeds, a complete, thoroughly cross-referenced index, including symbols, must be generated.

Standard notation — the most commonly-used notation in existing, popular textbooks — must be used throughout.

With these and related practices, the author has a fighting chance of producing something beyond a mere exercise in vanity.

The "Theory of Everything" in Physics vs. the "Theory of Everything" in Mathematics

In physics, there is our huge universe (perhaps only one of many!), and physicists are searching for the full description of it, which they believe can be contained in a shelf of volumes (relative to some audience). So they are seeking what is an infinitesimally small description of something huge — "infinitesimally small" in the sense that the number of atoms required to contain the description is infinitesimally small compared to the number of atoms in the universe.

But in mathematics, the situation is reversed. Here everything is small to begin with — merely the positive integers, 1, 2, 3, ..., which can be represented by only a few symbols — and then mathematicians build volume upon volume of truths about these integers, and the structures that can be obtained from them.

The reader may reply that there is certainly something huge about a countable infinity, e.g., of the positive integers. To which we might reply, but this something huge can be represented by only a few symbols. (We can write a short computer program that, in principle, can generate all the positive integers.)

Is there — should there be — can there be — a "Theory of Everything" in mathematics? If so, wouldn't it be an encoding of something? Of what? Or are we ultimately trying to fill in "bins" — the 2 bin (containing everything that equals 2), the 3 bin (containing everything that equals 3),..., the true bin (containing all true statements), the false bin (containing all false statements), ..., the equals bin (containing all equalities), the non-equals bin (containing all inequalities) ... so that when we are done, we will be able to look up anything?

Gödel's First Incompleteness Theorem

Gödel's first Incompleteness Theorem (1930) states: "There are (first-order) statements about the natural numbers that can be neither proved nor disproved from Peano's axioms."[1]

Gödel's proof involves first creating a means of encoding all first-order statements, then showing that one such statement says, in effect, "This statement cannot be proved." If the statement could in fact be proved, that would be a contradiction, and so the only way to avoid the contradiction is by accepting the statement as true.

Popularizations of mathematics sometimes give readers the impression that Gödel's long, ingenious proof is necessary to prove that there exist mathematical truths that cannot be proved. However, that is not true. It is easy to prove that there are mathematical truths that cannot be proved. Here is one such proof:

1. The set of all subsets of the real numbers is an uncountable set.

---

1. *The Princeton Companion to Mathematics*, ed. Gowers, Tim, Princeton University Press, Princeton, N.J., 2008, p. 701.

2. Therefore, since there is only a countably infinite number of proofs, there are "many" truths concerning set membership that cannot be proved. (In fact there are "many" truths that can't even be written down, because we can't even write down "most" irrational numbers!)

A counterargument to this proof is given below in this sub-section, along with my reply.

I have never seen any mention, in histories of mathematics covering the early part of the 20th century, that this, or a similar, proof, was used to prove that there are mathematical truths which are not provable. Hilbert, prior to Gödel's incompleteness theorems, believed that all mathematical truths were provable. Why did the above simple countability proof not occur to him, especially since a similar countability argument was used by Cantor in the latter part of the 19th century to prove that "most" real numbers are transcendental, not algebraic? (An algebraic number is one that is a solution of a polynomial equation with integer coefficients.) Is it really possible that, after Cantor's proof, mathematicians did not immediately see that a similar countability argument could be used to prove that there are unprovable mathematical truths? An answer to these questions is the following:

"There is a common misconception that Gödel's theorem tells us that there are 'unprovable mathematical propositions', and that this implies that there are regions of the 'Platonic world' of mathematical truths...that are in principle inaccessible to us. This is very far from the conclusion that we should be drawing from Gödel's theorem. What Gödel actually tells us is that whatever rules of proof we have laid down beforehand, if we already accept that those rules are trustworthy (i.e. that they do not allow us to derive falsehoods) and are not too limited, then we are provided with a new means of access to certain mathematical truths that those particular rules are not powerful enough to derive." — Penrose, Roger, *The Road to Reality*, Alfred A. Knopf, N.Y., 2005, p. 377.

The above-mentioned counterargument to the above simple proof came from a reader who said:

"This proof leaves open the possibility that:

"(1) If for any particular real number r we might have in mind;

"(2) And any set S of real numbers we might also have in mind;

"that

"(3) Our proof methods could be strong enough to prove or disprove r's membership in S in the logic $LA(r,S)$, that is, the logic of arithmetic *augmented by* names for the number and set we're interested in."

My reply is simply that there is only a countable infinity of such logics LA. But there is an uncountable number of truths.

Another very simple proof that there are mathematical truths that cannot be proved is the following.

1. Let a, b, c be positive integers such that we, possibly with the help of a present-day computer, can determine if $ab - c = 0$ is true or not.

2. Then if our computations reveal that $ab - c = 0$, we have a proof that $ab = c$. If $ab - c \neq 0$, we have a proof that $ab \neq$ c.

3. But now assume that each of a', b', c' is irrational. Then a proof that $a'b' = c'$ is not possible by computational means, because, since irrational numbers have only infinitely-long representations which in general exhibit no patterns over their entire length, the computation would never end. However, if in fact $a'b' \neq c'$, then in principle this fact would eventually be known through computation (successive comparison of corresponding digits of $a'b'$and c').

We can further argue that no proof that $a'b' = c'$ is possible even by non-computational arguments, simply because a', b', c' can never be completely described by finite expressions.

However, we can definitely do computational, and other, proofs for finite approximations to a', b', and c'.

In passing, we note that, if a' is an irrational number, where $0 \leq a' < 1$, and if the decimal digit j appears n consecutive times in the decimal representation of a', then in principle we can determine that fact by testing consecutive digits in the decimal representation. But if j does *not* appear n consecutive times in the decimal representation of a', we can never determine that by testing. (Can we say that we cannot prove that j does not appear n consecutive times in the decimal representation of a'?)

The Real Numbers

A Note About the Following Sub-Sections On the Irrationals

After the following sub-sections on the irrationals were written, a mathematician informed me that most of those mathematicians who deny the existence of the irrationals do so for two reasons: (1) they deny the legitimacy of infinite processes, e.g., limits, and (2) they deny the legitimacy of completed infinites, e.g., the set of all rationals, the set of all reals.

When asked what the deniers believe that the traditional proof that $\sqrt{2}$ is irrational shows, he said they believe only that the proof shows that $\sqrt{2}$ is not rational. When asked what the deniers believe that Lambert proved in 1761 when he proved that π is irrational, he said they believe that a number must be known to exist before it can be proved to have any property, and since π is defined by a limit argument, for the deniers it has no claim to existence.

We *Can* Write Down All the Reals (In a Sense)

In Greg Chaitin's *Meta Math! The Quest for Omega*[1] the reader will find an interesting discussion of some of the consequences of the fact that "most" real numbers cannot even be written down. "Why should I believe in a real number if I can't calculate it, if I can't prove what its bits are, and if I can't even refer to it?" (p. 97). (Chaitin is the discoverer of algorithmic information theory.) His question suggests that perhaps Gauss was right in his refusal to countenance the completed infinite, e.g., the *set* of rational numbers, the *set* of real numbers, etc. If we limit ourselves to decimal numbers that have finite but arbitrarily long decimal representations, then Chaitin could, in principle, always calculate with these numbers, and could always prove what their bits are, and could always refer to them.

Dialogue 1:

1. Vintage Books, N.Y., 2005.

A: "You can't write down all the real decimal numbers!"
B: "Yes, I can: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9."
A: "But those aren't all the real decimal numbers!"
B: "Then tell me a real number that does not begin with one of those digits."
A: "That's not the point. You haven't written down sufficiently *many* digits."
[B now writes down the hundred possible pairs of decimal digits.]
B: "How's that? Is there a real number that doesn't begin with a pair in that set of a hundred?"
A: "Well, of course not, but..."

When we learn that the rationals constitute a countable infinity of numbers, but that the reals constitute an uncountable infinity, we may conclude that it is impossible even to *lay hold* of most of these reals. We may imagine them as infinite strings written in unknown symbols and floating in a vast, dark void — as though the void were populated by uncountably many unrelated entities: shoes, ships, sealing wax, cabbages, kings, Swiss Army knives, moon rocks, kitchen towels, butterfly wings, ... But a key difference between such a conglomeration and the reals is that with the former there is no obvious way to tell how much any two differ from each other, whereas in the latter case there is.

There are far too many reals to place in a single list, or finite set of lists, or a countable infinity of lists, since, as we know, all these lists have only a countable infinity of numbers. But the above dialogue suggests that this view is wrong. We can *lay hold* of the beginnings of *all* the reals, no matter how long we choose those beginnings to be. We can refer to them, determine their representation in bits, and calculate with them (see "A Remarkable Fact About Adding and Subtracting Irrationals" on page 94).

The argument that, in principle, we can write down any rational number, but, in principle, we can never write down an irrational number needs to be looked at more closely. Surely at any given time, there are rationals — in fact, integers! — that cannot be written down simply because to write any one of them down would require more computer memory than exists in all the computers in the world at any given time. For example, consider the set S of all integers each of which, r, expressed in binary digits, is random, meaning, according to algorithmic information theory, that the shortest program needed to write down r is essentially as long as r is itself. Now consider the subset S_1 of S such that the length of each element r of S_1 is, say, greater than a billion times the number of atoms in the universe. It seems to me legitimate to say that, although each such r is an integer, it cannot be written down.

But the same applies to our successive approximations to the reals. So, exaggerating only slightly, as far as basic arithmetic is concerned, as long as there is a physical limit to the size of integers we can write down, then there is no difference between rationals and irrationals.

An Argument Against the Claim that Irrationals Do Not Exist

Dialogue 2

A: "Tell me something that doesn't exist."
B: "Well, a real number that is the square root of – 1."
A: "Can you give me the first few digits of that number?"
B: "Of course not. It doesn't exist!"

A: "Tell me something else that doesn't exist."
B: "Well, the integer that lies between 0 and 1."
A: "Can you give me the first few digits of that integer?"
B: "Of course not. It doesn't exist!"

A: "Tell me something else that doesn't exist."
B: "Well, any non-computable irrational number[1]."
A: "Can you give me the first few digits of any non-computable irrational number?"
B: "Sure! Let *T* be an infinitely deep, ten-ary tree. Each branch is labeled 0, 1, 2, ..., 9, which is also the label of the node at the end of each branch. Let the root of the tree represent the decimal point.
"For any positive integer *n*, I can give you a finite set of all sequences of *n* decimal digits. This set will contain all the sequences that are the first *n* digits of any irrational number."
A: "Do you mean any arbitrarily large *n*?"
B: "Sure!."
A: "Well, I am impressed by your ability to approximate something that doesn't exist with such accuracy — in fact with unlimited accuracy!"

Question: suppose *s* is an infinitely long string of decimal digits. How many real numbers does it (or can it) represent? *Answer*: a countably infinite number, because there is one real number for each possible position of the decimal point in the string.

A Proof That Irrationals Exist

Note: This proof would not be accepted by finitists, since it contains infinitely-long paths.

Before I begin, I must put the question to those who don't believe that irrationals exist: "What do you think π and *e*, the base of the natural logarithms, are if not irrational, as they were proved to be long ago?"

Proof:

1. Let *T* be an infinite ten-ary tree. The root node is the decimal point. Descending from each node are 10 branches labeled 0, 1, 2, 3, …, 9. Each branch ends in a node. Level *n*, where $n \geq 0$, in the tree is the set of all paths from the root that are of length *n*. Each such path represents a decimal number of *n* digits.

2. Now let *S* denote the set of all infinite paths in the tree. We ask if the set is countable or uncountable. We now apply Cantor's reasoning. Assume countable. Then it is possible to make a list of all the infinite paths. Let *L* be any such list. But we can go down through that list changing the *i*th digit in the *i*th path to a different digit. But then the resulting path cannot be in the list. So the number of infinite paths is uncountably infinite. But the number of rationals is countably infinite. So numbers other than rationals exist. These numbers are represented by infinite paths of decimal digits, each path having the property that no finite sequence of digits in it repeats indefinitely. . We can call them *irrationals*. □

An Irrational Need Not Be a Single Infinitely Long Decimal Number!

1. A computable irrational is, for example, π, since closed-form expressions for it exist.

Perhaps the deniers of the existence of irrationals are simply denying the existence of (decimal) numbers consisting of an infinite number of digits. Suppose that an irrational is simply an infinite sequence of increasingly-long prefixes having the property that no finite sequence of digits is repeated infinitely often in this infinite sequence of prefixes.

But if this is what the deniers have in mind, then they must also deny the existence of rationals, since each proper fraction, represented as a decimal, has an infinite number of digits. For example 1/7 = 0.142857..., where "..." consists of 142857 repeated infinitely often. The deniers can, of course, argue that the decimal representation of each rational consists of an infinite sequence of increasingly-long prefixes...

Perhaps the deniers will argue that each rational has a finite representation a/b, in addition to its infinite decimal represent. But so do at least some irrationals. For example, $\sqrt{2}$ is a finite representation of an irrational. If someone were to ask us, "What does that representation denote?", we would reply, "A number that, when multiplied by itself, yields 2. Furthermore, that number is the length of the hypotenuse of a right triangle whose two sides are length 1."

A Program That Generates All the Irrationals

Chaitin points out somewhere that a short program can be written that, in principle, will generate all the non-negative integers, 0, 1, 2, 3, ... Similarly, a short program can be written that, in principle, will generate all the reals, including the irrationals, between 0 and 1. It can work as follows. Assume the root of the ten-ary tree described in "*Dialogue* 2" is unlabeled. The ten branches from the root are labeled 0, 1, 2, ..., 9. Then:

1. The program first generates, and stores, all paths of length 1.
2. It then generates, and stores, all paths of length 2, doing this by extending each path of length 1 by ten branches, each labeled 0, 1, 2,...,9.
3. It then generates, and stores, all paths of length 3, etc.

In the limit, the program generates paths representing all the reals from 0 to 1.

Chaitin remarks, regarding his program that generates the positive integers, that it is possible to find, for arbitrarily large m, an m-bit integer that requires a program at least as long as m bits to generate it and it alone. There is an infinity of cases where m will have more bits than the number of bits in the program that generates, in principle, all the integers! It seems strange that the whole (the set of all non-negative integers) should be "less than" (require a smaller program to generate it) than just one of its parts.

Similarly, a short program can, in principle, generate all m-digit strings of decimal digits, where m is small enough to allow all the strings to be stored in computer memory. We may regard each such string as the m-digit prefix of a real number, regardless whether the number is rational or irrational.

We could legislate an upper bound on the length of all decimal numbers used in mathematics or anywhere else, based on our current computing capacity. "Each decimal number must be less than a million digits long!" And then we could always raise this bound as our computing capacity increased.

There is a world of difference between *finite, finite but arbitrarily long,* and *infinite*! Select a positive integer. The number of its prime factors always was and always will be a fixed, finite

number. But every finite string of decimal digits can be extended to any length our computing facilities allow.

The Reals are Consecutive (In a Sense)

Mathematics students learn early in their studies what a "sequence" is. Thus, e.g., the positive integers form a sequence. A fundamental characteristic of a sequence is that there is exactly one term — a "next" term — following each term except in the case of the last term of a finite sequence. Students learn that the rationals and the irrationals, on the other hand, do not form a sequence because, e.g., given a rational a/b, there is no next rational.

But there is a way in which the rationals and the irrationals can be made elements of a sequence, and it derives from the material in the previous sub-section. Consider the infinite ten-ary tree in which the nodes at level 1 are labelled 0, 1, 2, 3, ..., 9. Each of these nodes has ten branches, the ends of which are nodes labelled 0, 1, 2, 3, ..., 9. These nodes are at level 2. Etc. A number in the interval [0, 1] is given by a finite sequence of nodes traversed downward through the tree. Thus, e.g., 107752 corresponds to a path, downward through the tree, that corresponds to the number 0.107752.

Then for *any* positive integer n, we can let S_n be the set of all downward paths of length n in the tree. This set corresponds to the set of n-digit numbers in [0, 1]. These numbers can be ordered *sequentially* in accordance with their magnitude. Thus, for any n, we can speak of the next n-digit real number.

A Naive Thought About the Binomial Theorem

First-year college math students are taught the Binomial Theorem, which for each expression $(a + b)^n$, where a, b, n are integers, with n positive, gives a polynomial whose coefficients are the elements of the nth row in Pascal's triangle. The proof is by straightforward induction. The students may also be taught that if n is rational, i.e., $n = r/s$, then there is an infinite series representing $(a + b)^n$.

As we contemplate both cases of the Theorem, the thought might suddenly occur to us, "Why do we make such a big deal out of the sum of two integers raised to a power? The sum is an integer, and every pocket calculator contains a button that gives y^x for any real y, x. It is true that, for non-integer x, the result is only an approximation, but if we need greater accuracy, we can just buy a better calculator, or use a full-fledged computer.

We may also observe that each x^y, where y is an integer, is really a finite set of binomials raised to the yth power, the set being $(1 + (x – 1))^n$, $(2 + (x – 2))^n$, $(3 + (x – 3))^n$, ... , $(x + (x – x))^n$. It would seem that, although the elements of Pascal's triangle are the same in all these cases, the values of the terms $m^i(x – m)^j$, where $i + j = n$, differ. Is that possible? If so, then how can all of the polynomials have the same value, as they must?

Is there a Trinomial Theorem, and a Quadrinomial Theorem, and a Quntinomial Theorem, and ...

A Minor Problem Concerning the Nature of Number

We know, if we are students of modern logic, that a number n (where here n is a positive integer) is the set of all sets that have a one-one match with a set containing n elements. Suppose $n = 3$. How can we be sure that a collection of things contains three elements? Suppose the things are so close together that even under our highest power of magnification, they look like one thing?

Suppose they are so far apart that we don't even know they constitute three things? It would seem that we cannot legitimately ask for a list of all sets containing *n* elements.

The Concept of "Nice"

"Often, when we are studying a subject, we come across theorems which, in essence, say that, in such-and-such circumstances, things go as we would like them to go; in other words, things exhibit "nice" behavior. For example, if a function is commutative, e.g., if $f(a, b) = f(b, a)$ for all a, b, that might be considered a "nice" property of the function; if the product of a finite set of topological spaces retains a property which each of the spaces has, that might be considered a "nice" property of the product; if a function called a "sum" behaves in a way analogous to the way the arithmetic sum behaves — e.g., the "sum" of a, b is always "at least as large" as the larger of a and b — then that might be considered a "nice" property of the sum.

"Then we may ask of a given subject, how much of the entire subject is "nice", and how much isn't? If we could make this concept more precise, then, it seems, our job as students — as *users* — of the subject, would be made much easier, in that we would know a great deal more about the subject at the start, with much less effort. " — Curtis, William, *How to Improve Your Math Grades*

Let me give a more extended example. Let us assume that a graduate student needs to quickly get an idea of the main concepts in the subject of algebraic topology because someone has told him that this subject might be useful in helping him to solve an important problem in his PhD thesis. He finds another graduate student who knows something about the subject, and asks him to give him a briefing. The other says he is very busy, but he'll be glad to talk to him during a walk across campus to his next class. The second graduate student then begins as follows:

"OK, so a fundamental problem in topology is telling if two spaces are homeomorphic — you know: can one of them be deformed into the other without gluing or tearing. To use the standard example, a coffee cup and a donut are homeomorphic, because you can deform a coffee cup into a donut, and vice versa. (Think of them as being made of modeling clay.)

"Well, there are various techniques for solving this fundamental problem of telling if two spaces are homeomorphic. And one of them was discovered, or rather glimpsed, at the end of the 19th century by several mathematicans, including Betti and Poincaré. And the basic idea is this: suppose you've got these two topological spaces. And they're really complicated, and furthermore they're not merely complicated *three*-dimensional objects, but *n*-dimensional objects, where *n* is much larger than three. You want to find out if the spaces are homeomorphic. So what these early mathematicians found is this:

"You can use lines, triangles, tetrahedrons, and their higher dimensional analogues as building blocks of each space. These can be thought of as being "minimal" building blocks because in each case they are made from the minimum number of points you can have to get an object of the specified dimension. In other words, speaking informally, two points is the minimum number of points you can have and still make a line; three points is the minimum number of points you can have and still make a two-dimensional object (a triangle), and four points is the minimum number of points you can have and still make a three-dimensional object (a tetrahedron). Etc.

"Then, for each dimension in each of the spaces, you can 'cover' (I am using the term informally here) the space in that dimension using the corresponding building blocks. So, in the second dimension you can cover the space with triangles, in the third dimension you can cover it with tetrahedrons, etc. OK?

"Now you can assign an orientation — clockwise or counterclockwise — to all the building blocks in each of the dimensions. In the line case, an orientation establishes in which direction you are to move down the line; in the triangle case, an orientation establishes if you go around the sides clockwise or counterclockwise. In the tetrahedron case, an orientation establishes if you go around *all* the triangles in the one direction or the other.

"Now once you've got an orientation assigned, you've got all you need to start talking in terms of groups.

"And from these groups you can make other groups, so that you wind up with a special group for each dimension (it's called a homology group).

"Now here's the payoff: these homology groups are topologically invariant. What does that mean? It means that if the two spaces are *homeo*morphic, then the homology groups at each dimension are *iso*morphic. Or, to turn that around (using the contrapositive) what you can say is that if the groups are *not* isomorphic at any one or more dimensions, then the spaces can*not* be homeomorphic. Neat, huh?

"Now, things turn out to be pretty much nice from here on. I mean you can generalize up from triangles and tetrahedrons and their higher dimensional analogues to cells and even more abstract things. And, let's see, you can make other groups out of the homology groups — e.g., you can make what are called cohomology groups. And, there are nice mappings (homomorphisms, naturally, since we're in an algebraic world here) connecting all this stuff. And some of these mappings yield or *induce* other mappings in a nice way. And if you take the product of several spaces, these products behave nicely in many cases. Etc.

"So that's pretty much it. Which is not to say that all the proofs are easy, of course. But that's the Big Picture."

If, at the conclusion of this informal explanation, the second graduate student were to hand to the first student a Venn-diagram-like map showing the various subsets of complexes and groups encountered in algebraic topology — singular complexes, simplicial complexes, chain complexes, ..., chain groups, homology groups, cohomology groups,... — the first graduate student would have obtained, in a matter of minutes, a view of the subject which would have taken him hours if not days or weeks to obtain from a textbook alone. Perhaps the concept of nice can be made more precise. I don't know. But even in the form presented by the above examples, it would seem to be enormously useful in shortening the amount of time required to obtain a certain kind of basic understanding of a subject.

A Tally of the Uses of Mathematical Subjects

An ongoing tally of all the actual instances of use of each mathematical subject, and important equations, formulas, lemmas, and theorems in each subject, would be of interest, I think, not only to students but to a few professionals as well. Suppose one could get a histogram of calculations made, throughout the world, using Maxwell's equations over the past month? Of applications of Dirichlet's integral? Of Rolle's Theorem? Of the Hardy-Weinberg equation?

The Essence of Mathematics...

Sometimes I think that ultimately, mathematics boils down to nothing more than a collection of complicated ways of saying, "This is a that."

Algebra

π Is Transcendental But...

It was proved at the end of the 19th century that π is a transcendental number, i.e., a number that is not the solution to any algebraic equation. But how can this be, given that π is the solution of the algebraic equation, $x - \pi = 0$? (Answer: by definition, the coefficients of an algebraic equation are rational numbers only.)

What, if anything, of interest can be said about the set of polynomials *all* of whose coefficients are transcendental numbers? Or the set *some* of whose exponents are transcendental numbers?

Cataloging All the Polynomials

The general form of an algebraic equation (in one variable) is:

$$c_0x^n + c_1x^{n-1} + c_2x^{n-2} + \ldots + c_{n-1}x + c_n = 0$$

where the c_i are complex numbers. Suppose we were librarians working in a library that stored nothing but polynomials, i.e., nothing but the left-hand sides of such equations. How might we go about cataloging these polynomials? One answer might be: "Lexicographically", meaning, in this case: put the one-term polynomials (monomials) first, then the two-term polynomials (binomials) next, then the three-term polynomials (trinomials) next, etc. For each type of polynomial, we might then order the polynomials by coefficients, treating each sequence of coefficients in a given polynomial of degree n as an $(n + 1)$-digit number in an infinite number system. (Each coefficient in the polynomial is one digit of the number.)

Although this cataloging scheme seems easy to understand, it leaves us with the problem that each kind of polynomial has an infinite, in fact, an uncountably infinite number of instances.

The thought may now occur to us that we can dispense with the division of the set of polynomials into monomials, binomials, trinomials, etc., and instead consider them all as n-nomials for some sufficiently large n to suit our purposes. Monomials are then simply n-nomials having 0 as coefficients for all terms except the first; binomials are simply n-nomials having 0 as coefficients for all terms except the first and second, etc.

Let us generalize. For each $k \geq 0$, and for each $n \geq 0$, there exists an infinite set of expressions,

$$c_1x_1^{u_{1_1}}x_2^{u_{2_1}}\ldots x_k^{u_{k_1}} + c_2x_1^{u_{1_2}}x_2^{u_{2_2}}\ldots x_k^{u_{k_2}} + \ldots + c_nx_1^{u_{1_n}}x_2^{u_{2_n}}\ldots x_k^{u_{k_n}}$$

where c_i, $1 \leq i \leq n$, is a complex number, and no two terms are symbolically identical. There is no requirement that the sum of the exponents in each term of a given expression equal a fixed number, e.g., n. If each x_i can be any complex number, then each expression can be regarded as defining a point in a complex-number space, where the coordinates of the point are the values of c_i, the x_i, and the u_{i_j}. Clearly, any set of forms, and any set of traditional, single-variable polynomials, is a subset of the set of expressions we have defined.

What good is all this? you ask. Well, our cataloging scheme — our system of "points" — enables us to assign a value to any given point (namely, the value of the expression defined by the coordinates of the point), and then, thereafter, if we want to find the value of another expression, we can find it by "getting to it" via appropriate changes in the coordinates. "If you know the value *here*, and you want to find the value *there*, why, then, you need to do only the *additional* calculation required, and not start from scratch." (See also "A Thought on Differential Equations" on

page 105.)

As far as coefficients are concerned, we can summarize our idea by saying, "A coefficient is simply an index of a point."

Toward Possible New Proofs of the Fundamental Theorem of Algebra

The Fundamental Theorem of Algebra (FTA) states that a polynomial with coefficients that are complex numbers has *all* its roots in the complex field[1].

Toward Possible New Proof 1

1. Assume that the FTA is false. Then there exists an n-th degree polynomial $P(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_0$ such that all coefficients are complex, but such that at least one root is not complex.

2. By Vieta's formulas, $x_1 + x_2 + ... + x_n = -a_{n-1}/a_n$, where the x_i are all the roots. But by assumption, one or more of the x_i is not complex, and so, by moving all the complex numbers to the right-hand side of the equation we have a sum of non-complex numbers (possibly only one non-complex number) on the left-hand side equaling a sum of complex numbers on the right-hand side. But this is impossible, since the complex numbers are closed under addition. This impossibility implies that our assumption was false and that FTA is true.

.

Toward Possible New Proof 2

1. For each sequence of the n coefficients (all of them complex) in each nth-degree polynomial $P(x)$ (the sequence running from left to right), where $n \geq 1$, there exists an nth-degree polynomial having the coefficients as roots.

Proof: if $r_1, r_2, ..., r_n$ are the coefficients, then the polynomial is $(x - r_1)(x - r_2)...(x - r_n)$. □

2. Let S denote the set of all pairs $<P, Q>$ where P is a polynomial with complex coefficients, and Q is the polynomial whose roots are the coefficients of P.

3. Assume to the contrary that there exists a polynomial Q at least one of whose roots is not a complex number. But then there must exist a polynomial P, in the pair $<P, Q>$, at least one of whose coefficients is not complex. But this is contrary to our definition of each P as a polynomial all of whose coefficients are complex. Therefore we have a contradiction that implies FTA is true.

Determinants

Determinants are an example of a simple representation (numbers in rows and columns forming a rectangle) but complicated, or at least tedious, rules for evaluation. Can each such case be converted into simple rules for evaluation at the cost of a complicated representation?

---

1. Herstein, I. N., *Topics in Algebra*, John Wiley & Sons, N.Y., p. 337.

On Investigating "How Abelian" (Commutative) a Finite Group Is

Soon after students begin their study of group theory, they are introduced to the type of group known as "abelian". This is a group in which all elements commute with each other, that is, for all a, b in the group, it is the case that $a \bullet b = b \bullet a$.

It is natural for students with the gift of idle curiosity to wonder if there is a way of determining, for any *non*-abelian group — or at least for any finite non-abelian group — just how "abelian" the group is, meaning, just how many pairs of elements do in fact commute with each other. The students know that there are always *some* elements of each group that commute with each other, namely, by definition, each element and its inverse, since for all elements a of a group, we have $a \bullet a^{-1} = a^{-1} \bullet a = e$, the identity element.

Students also learn, early in their study of groups, that sets have been defined that are directly related to the commutative property: the *centralizer* of an element a is defined to be the set of all elements of the group that commute with a. Similarly, one can speak of the centralizer of a set of elements. The *center* of a group is the set of elements that commute with all elements of the group. (Clearly, if the center of a group is the group itself, then the group is abelian.)

But the student may feel that these definitions are too limited. The student may want to think in terms of all the subsets of the group and ask of each, Are all the elements in this set commutative?

How can we organize an answer to the student's question?

Well, we can begin with the set of all subsets of the group. We know from elementary set theory that if the group has n elements, then the number of subsets is 2^n. We can arrange these into:

all subsets consisting of just one element;
all subsets consisting of just two elements;
all subsets consisting of just three elements; ...;
all subsets consisting of n elements, where n is the number of elements (i.e., the order) of the group.

We can now associate with each subset, the term $S(u, r, s, t)$, where u is a list of the elements of the subset, r denotes the number of elements in the subset, $s = 1$ if all the elements commute, 0 if not, and $t = 1$ if the subset is a subgroup, 0 if not (we know that if e is not an element of the subset, then the subset cannot be a subgroup).

It might be of interest to use the computer to generate all S terms for as many finite groups of increasing size, beginning with 1, as computer resources allow. Perhaps from these S's, we may learn something new about the nature of groups.

Toward a Poor Man's Proof That There Are No General Formulas in Radicals for the Roots of Algebraic Equations Beyond Degree 4

It is sometimes a good exercise to ask of some major mathematical problem that was solved in the past, "How would I have thought about solving this problem?" The question becomes more interesting if you still do not fully understand the solution that eventually was arrived at. The point of the exercise is not to discover a previously unthought-of solution to the problem, nor is the point to test to see if you are at least intelligent enough to re-discover what the mathematicians of that earlier time discovered. The point is simply to investigate your own thinking processes.

Take, for example, the theorem that was proved by Abel in 1830 that, for each n greater than 4, there does not exist a general solution in radicals to algebraic equations of degree n. We imagine that we are living in the early 1800s, and the mathematical world, or at least part of it, is talking about a proof of whether such general formula is possible for degree 5 and higher. Hoping to make a name for ourselves, we might begin reasoning as follows.

Initial Thoughts

Clearly there are some general solutions to *subsets* of all equations of degree n, where n is greater than 4. For example, 1 is a solution of every equation $x^n - x^{n-1} + ... + x^2 - x = 0$ for n even. And for each prime $p \geq 5$, each of the p pth roots of unity is a solution of the equation $x^p - 1 = 0$.

We must remember that all we have to do is show that, for each $n \geq 5$, there is *just one* case for which a general formula in radicals could not apply. So that suggests a proof by contradiction in which we assume a general formula, and then show that it will not work in one case for each $n \geq 5$.

Why Does the Quadratic Formula Always Work?

In high school we learned that the two roots of any quadratic equation, $ax^2 + bx + c = 0$ are given by:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

We also learn the derivation of this formula by the process called "completing the square".

But even though the derivation is correct, we might be curious as to how that formula always gives exactly 0 when it is plugged into the quadratic equation — no matter what the coefficients a, b, c may be: rational, irrational, or complex numbers. So let us plug the formula into the equation and see how things work out. We have on the left-hand side of the equation (considering just the positive square root for now):

$$a\left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)^2 + b\left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right) + c$$

or

$$a\left(\frac{-b}{2a} + \frac{\sqrt{b^2 - 4ac}}{2a}\right)^2 + b\left(\frac{-b}{2a} + \frac{\sqrt{b^2 - 4ac}}{2a}\right) + c$$

We want to see if this equals zero, no matter what the coefficients are, and no matter what the square root terms are: rational, irrational, or complex numbers. Expanding the squared term, we get:

(1)

$$a\frac{b^2}{4a^2} - 2ab\frac{\sqrt{b^2-4ac}}{4a^2} + a\frac{b^2}{4a^2} + a\left(-\frac{4ac}{4a^2}\right) + \frac{-b^2}{2a} + \frac{b\sqrt{b^2-4ac}}{2a} + c$$

We see that the first and third terms minus the fifth term of (1) gives us 0. That is:

$$\frac{b^2}{4a} + \frac{b^2}{4a} - \frac{b^2}{2a} = 0$$

Next, we see that the fourth term and the last term of (1) likewise give us 0. That is

$$a\left(-\frac{4ac}{4a^2}\right) + c = 0$$

And finally, we see that the second and the sixth terms of (1) give us 0. That is:

$$-2ab\frac{\sqrt{b^2-4ac}}{4a^2} + \frac{b\sqrt{b^2-4ac}}{2a} = 0$$

And thus we have shown that expression (1) = 0, which was our goal.

We should not hesitate to regard this result as rather remarkable — it applies regardless what kinds of numbers — rational, irrational, complex — happen to be involved.

By rights, we should carry out the same process that we just did on quadratic equations, on degree 3 and degree 4 equations. The calculations would be more complex, as the reader can see by looking at the general formula for degree 3 equations in "Appendix B — General Formulas in Radicals for Solution of Equations of Degree 1 Through 3" on page 134. Until we complete those calculations, let us consider an assumed degree 5 general formula.

A Look At An Assumed Degree 5 Formula in Action

Let us assume that a general formula to solve degree 5 equations could be written in the binomial form $u + v$ for each root, that is, that each of the five roots could be expressed as $u_1 + v_1$, $u_2 + v_2$, ..., $u_5 + v_5$, where $u_i = \sqrt[5]{w_i}$, and $w_i \neq r_i^5$. This is a reasonable assumption considering the general formulas for the degree 2 and degree 3 cases (see "Appendix B — General Formulas in Radicals for Solution of Equations of Degree 1 Through 3" on page 134). So the degree 5 equation could be written:

(2)

$$(u_i + v_i)^5 + a_1(u_i + v_i)^4 + a_2(u_i + v_i)^3 + a_3(u_i + v_i)^2 + a_4(u_i + v_i) + a_5 = 0$$

The following table represents the expansion of some of the terms of (2) according to the binomial theorem: (Pascal's triangle):

Table 1:

	u_i^5 terms	u_i^4 terms	u_i^3 terms	u_i^2 terms	u_i terms
$(u_i + v_i)^5$	u_i^5	$5u_i^4v_i$	$10u_i^3v_i^2$	$10u_i^2v_i^3$	$5u_iv_i^4$
$a_1(u_i + v_i)^4$		$a_1(u_i^4)$	$a_1(4u_i^3v_i)$	$a_1(6u_i^2v_i^2)$	$a_1(4u_iv_i^3)$
$a_2(u_i + v_i)^3$			$a_2(u_i^3)$	$a_2(3u_i^2v_i)$	$a_2(3u_iv_i^2)$
$a_3(u_i + v_i)^2$				$a_3(u_i^2)$	$a_3(2u_iv_i)$
$a_4(u_i + v_i)$					$a_4(u_i)$

Now we ask: if u^k and $u^{k'}$ are incommensurable (see next sub-section) when $k \neq k'$, then is it possible that the sum of the terms in each u^k column in the above table can equal zero? If the answer is no *for just one set of coefficients* a_1, a_2, a_3, *and* a_4, then we have a proof that a general solution in radicals is impossible for degree 5.

The Problem of Incommensurability

One thing that seems to be in our favor is the fact that for some j, k, where $j \neq k$, u_i^j and u_i^k in the above table are incommensurable. Thus if the terms in one of these powers are going to equal 0, it will have to be a result of the arithmetic on these terms alone. In the case of the quadratic formula, we had only the first power of the square root to deal with, and these terms canceled nicely.

We now show that, e.g., the u_i^4 and the u_i^3 terms in the above table are incommensurable. Assume the contrary. Then there exist integers a, b, c, d such that

$$a/b(w_i^{1/5})^4 = c/d(w_i^{1/5})^3.$$

We assume that all common factors in a, b, c, d have been canceled, that a/b is in lowest terms, and similarly for c/d, and that all 5th powers of primes in w_i have been factored out and canceled on both sides of the equation.

Then raising both sides to the 5th power we get

$$a^5d^5w_i^4 = b^5c^5w_i^3$$

or

$$M^5w_i = N^5$$

But since all 5th powers have been removed from w_i, this implies that there are prime powers in the left-hand side that are not 5th powers, which is not possible, since all prime powers in the right-hand side are 5th powers. So we conclude that $w_i^{4/5}$ and $w_i^{3/5}$ are incommensurable. Informally, they cannot together yield 0, regardless of the values of a_1, a_2, or a_3 or v_i. If our argument

can be made rigorous, then we have our proof that there is no general formula in radicals to solve degree 5 equations.

A Conversion of Polynomials That Is Related To Our Degree 5 Table

Suppose we have a polynomial $f(x) = -18 + 21x - 26x^2 + 22x^3 - 8x^4 + x^5$. One of the roots of $f(x)$ is 2. We can re-write $f(x)$ in terms of $(x - 2)$, and get $f(x) = 5(x - 2) - 6(x - 2)^2 - 2(x - 2)^3 + 2(x - 2)^4 + (x - 2)^5$. Observe that $f(2) = 0$ in each version of $f(x)$. To check that the two versions of $f(x)$ are in fact equal, we can simply expand each power of $(x - 2)$ using the binomial theorem, then gather all x terms, all x^2 terms, ..., and all x^5 terms and observe that the sum of their coefficients for each power of x equals the coefficient for that power of x in the first version of the polynomial[1].

But this is precisely the technique we propose in our degree 5 table, except that there we attempt to show that there are coefficients that do not allow some incommensurable powers of our assumed radical formula for a root, to equal zero.

Vieta's Formulas and Solutions to Polynomial Equations

Vieta's formulas state that for each complex coefficient a_i of a polynomial $x^n + a_1x^{n-1} + a_2x^{n-2} + ... + a^n$, $(-1)^i a_i$ is the sum of all products taken i at a time of all roots of the polynomial.

It is natural to ask if this fact can be used to solve the polynomial equation. I have never come across a consideration of this question in any algebra textbook. So I offer the following thoughts.

- There are n equations in n unknowns, which seems encouraging.

- But we must be clear on what we are hoping to accomplish. Given a specific polynomial with specific numerical coefficients, we are hoping to find numerical values for each of the roots x_i. This may not be possible. But nevertheless we can make the following observations:

- The number of terms in the ith equation is $\binom{n}{i}$.

- It is easy to see that:

for each i, where $1 \le i \le n$, there are n equations $x_i = ...$;
for each i, j, where $1 \le i, j \le n$, $i \ne j$, there are $n - 1$ equations $x_ix_j = ...$;
for each i, j, k, where $1 \le i, j, k \le n$, $i \ne j \ne k$, there are $n - 2$ equations $x_ix_jx_k = ...$;
...
for 1, 2, 3, ..., n, there is one equation $x_1x_2...x_n = (-1)a_n$.

- The nth equation is $x_1x_2... x_n = (-1)a_n$. If we know that all the roots are rational, then this implies that there exists at most a finite number of sets of possible roots of the equation. For each set, we can then try the roots on the equations listed in the previous dotted item, until we get an inequality, in which case we know that the set is not the set of roots of the equation. If we don't get an inequality, then we know we have the set of roots. Does this strategy apply if we assume that at least one root is a complex number?

---

1. Gouvêa, Fernando Q., "Local and Global in Number Theory" in *The Princeton Companion to Mathematics*, Princeton University Press, Princeton, N.J., 2008, p. 241.

• Each equation can be represented as several inner products, all having the same value. Are there known inner product facts that could help us here?

• We must not forget that if, for each root x_i, we multiply the ith sum by x_i^{n-i}, where $1 \le i \le n - 1$, and then add x_i^n and a_n, we get zero!

Topology

How Many Topologies?

How many possible topologies are there on a set of n elements, $n \ge 1$?

Topologies of Topologies

Is there anything useful to be gained by considering topologies of topologies, and if so, what? Very roughly, a topology defines what is "close" to what in a set of objects, e.g., numbers. A topology of topologies would tell which topologies are "close" to which topologies, i.e., which definitions of "closeness" are close to which others.

Radical Questions About Algebraic Topology

No can deny the fundamental importance of the ideas underlying algebraic topology, specifically , the ideas underlying homotopy groups, homology groups, and cohomology groups. However, the calculation of these groups is tedious, and most of the proofs of important lemmas and theorems require an extraordinary amount of memorization of facts in the subject, or, at least, a complete index that includes an index of every term and symbol that is peculiar to the subject. There are many hundreds of such terms and symbols. Two of the leading textbooks in the subject have no such index, and, in addition, frequently omit steps in proofs, and justifications for the statements that are included. See Appendix C, "Two Notably Bad Textbooks", in chapter 2, "Mathematics in the University", in William Curtis's *How to Improve Your Math Grades*, occampress.com. The result is that, for virtually all students, the subject can only be learned in the classroom, where the professor can fill in all the material that is missing from the textbooks.

Two questions that I am sure have never occurred to the authors of the above two books, or to the overwhelming majority of mathematicians in any subject, are

(1) "What does the number of terms and symbols peculiar to a subject, say about the nature of the subject — at least about the structure of the subject?", and

(2) "What, if anything, can be done to significantly reduce the number of terms and symbols peculiar to a subject?"

Another question that must be asked is based on the following facts. In homology theory, p-dimensional chains are sums of functions whose domain is the set of p-dimensional simplexes (simplices). In cohomology theory, p-dimensional cochains are sums of functions whose domain

is the set of p-dimensional chains. Does anything of value result if we continue this process of successively making the functions at level n, where $n \geq 1$, become the elements of the domain of functions at level $n + 1$?

I will welcome hearing from knowledgeable readers.

Knot Theory

Is it possible to devise a computer program that could slowly "deform" a knot according to one or more rules? (More precisely, we are asking for an algorithm that would always be able to determine if two knots are equivalent.)

Is it possible to devise a computer program that would act as a kind of inertial navigation device with a built-in recorder so that the device could travel, like a rocket, inside the tube (curve) defining the knot, keeping track of changes in direction and of crossings of other parts of the string? Then if the records of travel for two knots were "the same", the knots would be equivalent.

Why don't we routinely "rectify" knots — i.e., convert them into a sequence of lines parallel to the x, y, and z axes? Thus, let:

N (north) denote movement in the positive y direction;
S (south) denote movement in the negative y direction;
E (east) denote movement in the positive x direction;
W (west) denote movement in the negative x direction;
U denote movement in the positive z direction;
D (down) denote movement in the negative z direction.

Then a knot could be described by a finite sequence of the letters N, S, etc. If each letter denotes a movement of length 1 unit in the given direction, then there will be sequences of letters that allow for a line to pass under or over or to the left or right of another line. For example, for a line a traveling south to cross over a line b traveling west, the following letters represent the crossing by a: D,S,U,S...

Obviously, each knot would be represented by at least two sequences of letters, one for the "right-hand side" of the knot, the other for the "left-hand side".

We can stipulate that two knots always begin at the same level y, and make other restrictions to standardize the representation of knots.

What property of the sequences would imply the knots were in fact different?

A question in elementary knot theory is: are the trefoil knot and its mirror image "the same" (i.e, isotopic)? The answer is no, but the proof is not trivial. And yet the fact is that if we turn the mirror image over, we get the original trefoil knot! (Or, we could tape the mirror image knot to a window, and then go around to the other side of the window, and see the original knot.)

This tactic is not allowed because the proof that two knots are isotopic must rely on continuous deformations of one or both knots. But nevertheless, it seems to me worth investigating whether a generalization of "going around behind the knot" might prove to be a useful proof technique in other subjects.

A Question About the Jordan Curve Theorem

"In topology, a Jordan curve is a non-self-intersecting continuous loop in the plane...The Jordan curve theorem asserts that every Jordan curve divides the plane into an 'interior' region bounded by the curve and an 'exterior' region containing all of the nearby and far away exterior points, so that any continuous path connecting a point of one region to a point of the other intersects with that loop somewhere." — "Jordan curve theorem", Wikipedia, 8/26/16.

The question arises, Is it possible to tell quickly if a point not lying on the curve, is in the interior region or not? My tentative answer is Yes, it is possible. Simply draw a straight line from the point to a point in the exterior region. If the line crosses an even number of points on the curve, then the original point lies in the exterior region. If it crosses an odd number of points, then it lies in the interior region.

A Shortcut Through *n*-Space?

One way of turning something that is, say, right-handed in n space, into a left-handed version of the same thing in n space, is by rotating it in $n + 1$ space. But we can turn a right-handed glove in 3-space into a left-handed glove by simply turning it inside out! The same can be done with a glove in 2-space (as long as the opening for the two-dimensional hand is in fact open in 2-space). Is there something important about this ability to "save having to go through $n + 1$ space"? Is it a shortcut that we can exploit?

A Shortcut Through the Irrationals

The hypotenuse of a right triangle whose sides equal 1 is a shortcut through the irrationals. For, no matter how "narrow" our attempted approximation to the hypotenuse using rationals — i.e., no matter how small the steps we make out of increments parallel to one leg of the triangle and increments parallel to the other leg, in order to approximate the hypotenuse — the sum of all these increments is always 2. We always have to travel a distance of 2 in order to get from the extremity of one leg to the extremity of the other leg via these increments. But if we are allowed to use irrationals, then we can reduce the length of our travel to $\sqrt{2}$. *Exercise*: Give other examples in mathematics where the introduction of a new type of number allows one to shorten distances or to otherwise reduce a certain quantity

The Chicken Salad Sandwich Scene in the Film "*Five Easy Pieces*"

Some readers may recall the chicken salad sandwich scene in the 1970 film, *Five Easy Pieces*:

[Scene: roadside restaurant. Bobby, his girlfriend, hippie girl and her female friend at a table.]

Bobby [Jack Nicholson] to waitress: I'd like a plain omelette, no potatoes, tomatoes instead, a cup of coffee, and wheat toast.

Waitress: No substitutions.

Bobby: What do you mean, you don't have any tomatoes?

Waitress: Only what's on the menu. You can have a No. 2, a plain omelette. It comes with cottage fries and rolls.

Bobby: Now, I know what it comes with but it's not what I want.

Waitress: Well, I'll come back when you've made up your mind.

Bobby: Wait a minute! I have made up my mind. I'd like a plain omelette, no potatoes on the plate. A cup of coffee and a side order of wheat toast.

Waitress: I'm sorry, we don't have any side orders of toast. English muffins or a coffee roll.

Bobby: What do you mean you don't make side orders of toast? You make sandwiches, don't you?

Waitress: Would you like to talk to the manager?

Hippie girl: Hey, mac!

Bobby: [To girl] Shut up. [To waitress] You've got bread and a toaster of some kind?

Waitress: I don't make the rules.

Bobby: OK, I'll make it as easy for you as I can. I'd like an omelette, plain, and a *chicken* salad sandwich on wheat toast, no mayonnaise, no butter, no lettuce, and a cup of coffee.

Waitress [reading from her pad]: A No. 2, chicken sal san, hold the butter, the lettuce and the mayonnaise, and a cup of coffee. Anything else?

Bobby: Yeah, now all you have to do is hold the chicken, bring me the toast, give me a check for the chicken salad sandwich, and you haven't broken any rules.

Waitress: You want me to hold the chicken, hunh.

Bobby: I want you to hold it between your *knees*.

Waitress: You see that sign? Yes, you'll all have to leave. I'm not taking any more smartness or sarcasm.

Bobby: You see this sign? [Makes obscene gesture in his lap, then sweeps dishes and silverware off the table and gets up]

[Later, in car]

Hippie girl: *Fantastic* that you could figure that all out and lie [sic] that down on her so you could

come up with a way to get your toast. *Fantastic!*

Bobby: Yeah, well I didn't get it, did I?

Hippie girl: No, but it was very clever. I would have just punched her out.

The scene sets forth (however crudely) an important mathematical idea. One way of thinking of the idea is as a certain kind of continuous transformation. Assume you have a function $f(x_1, x_2, ..., x_n)$, defined everywhere in some domain. Let $f(a_1, a_2, ..., a_n) = a$ (this represents the situation described by the menu and rules). You want $f(b_1, b_2, ..., b_n) = b$ (this represents what Nicholson wants). But assume the only way you can achieve this is by continuously changing the value of each argument. In other words you simply can't replace the a_i values with the b_i values. You have to "get there". The idea is vaguely reminiscent of defining a circle as the limit, as n approaches infinity, of n-gons. Or is this simply the idea of homotopic transformation in topology, in which we continuously transform one function into another, e.g., a function defining any closed curve into a function defining an arbitrarily small circle about a point located in the region enclosed by the original curve? Topologies can be defined on finite sets, so something like the chicken sandwich transformation is legitimate.

On the other hand, one can look at the problem as one in graph theory, in which the arrows in a graph define allowed sequences of nodes, so that the problem is simply to find a path from one node, A, in the graph, to another node, B.

Calculus and Analysis

And Why Exactly Did Newton Need to Discover the Calculus on Order to Formulate His Theory of Gravity?

I have never seen a calculus textbook that explained why Newton needed to discover the calculus in order to formulate his theory of gravity. Sometimes we read that his first proofs were "geometric", and did not explicitly involve the calculus because of his fear that if they were presented in terms of that new subject, they would be discounted.

The only readily accessible history of the discovery of the calculus that I know of is Boyer's *The History of the Calculus and its Conceptual Development*[1]. Boyer devotes a few pages to Newton's mathematics in the book in which the latter set forth his theory (*Philosophiae Naturalis Principia Mathematica*) but not a word that I can find relating these calculations to the orbits of planets or the force of gravity.

The time is long overdue for a clear, understandable, historically-accurate, student-tested presentation of this fundamentally important part of the subject of the calculus. The only reason this hasn't been done long ago, it seems to me, is the shameful contempt for the history of mathematics on the part of academic mathematicians.

It is a worthwhile exercise to try to prove, from first principles, that if the force exerted by the sun on the planet is $F = (Gm_1m_2)/r^2$, there G is the gravitational constant, m_1 is the mass of the planet, m_2 is the mass of the sun, and r is the distance from the sun to the planet at each moment of time, then the planet is in an elliptical orbit around the sun as Kepler described,

1. Dover Publications, Inc., N.Y., 1949

Our first thought is probably that, when the planetary orbits were established, there must have been pieces of matter that simply plunged into the sun. So what were the conditions that led a piece of matter to start on an elliptical orbit?

Next, we can imagine the movement of the planet as being approximated by a sequence of finite straight-line movements, and the gravitational force as being approximated by a sequence of pulls on the planet, so that the planet makes a finite straight-line movement, then a gravitational pull is exerted on it, causing it to change direction of motion. Then the planet makes a finite straight-line movement in the new direction, a gravitational pull is exerted on it, causing it to change direction again.

Determining Volumes, Areas, and Curve Lengths Without Using the Calculus

In calculus courses we learn to:

determine the volume of an object by slicing it into ever thinner slices;
determine an area, e.g., the area under a curve, by dividing the area into ever narrower rectangles, and
determine the length of a curve by approximating the curve by ever shorter straight lines.

However, we should remember that there are other ways of determining these properties. We can

determine the volume of a container of any shape, e.g. of a vase, by filling the container with water, and then pouring the water into a graduated cylinder; to determine the volume of any solid object, we can fill a container with water, then submerge the object, then measure, using a graduated cylinder, the volume of water spilled;
determine an area by using a planimeter; and
determine the length of a curve by laying a string over the portion of the curve whose length is to be determined, then cutting the string at start of the portion, and cutting it at the end of the portion, and measuring the length of the string.

(Prior to the discovery of the calculus in the latter part of the 17th century, Descartes, for one, didn't believe it was possible to determine the length of a curve segment. The length of a curve segment was not "rectifiable".)

An Amusing Fact About the Area and Circumference of the Circle

Consider the function $A = \pi r^2$, i.e., the area of circles as a function of their radii. This is a parabola.

Now consider the derivative,

$$\frac{dA}{dr} = 2\pi r$$

The derivative of the area function is the circumference function!

Can Rationals Be Approximated by Irrationals?

Does it make any sense to speak of approximating rationals with irrationals? Consider the successive digits of the irrational.

Rationals Are the Limits of Infinite Series

Let c/d be any positive rational. Then there exists an infinite geometric series whose limit is c/d.

Proof: Since

$$\frac{x}{1-x} = x + x^2 + x^3 + \ldots$$

$$\frac{\frac{c}{b}}{\frac{b}{b} - \frac{c}{b}} = \left(\frac{c}{b}\right) + \left(\frac{c}{b}\right)^2 + \left(\frac{c}{b}\right)^3 + \ldots = \frac{c}{d}$$

where $b - c = d$. □

A Remarkable Fact About Adding and Subtracting Irrationals

An irrational number is represented by a decimal number such that there is no infinitely-repeating sequence of digits in the digits to the right of the decimal point. In other words, in general one cannot state what the nth decimal digit of an irrational number is, where n is arbitrarily large.

It is therefore plausible to say, for example, "One thing we know about arithmetic on the irrationals is that, except in trivial cases, no calculation can ever be completed." However, this is wrong. Consider any expression $(ab)/c$, where a, b, c are positive integers, and such that $ab = c$. Thus $(ab)/c = 1$. Take the natural logarithm of both sides of this equation. We get $ln\ a + ln\ b - ln\ c = ln\ 1$. Now it is well known that the natural logarithm of a positive integer is irrational. It is also well known that ln 1 = 0. And so we have the sum of two irrationals, minus a third irrational, equaling the integer 0.

This fact holds for any fraction whose numerator and denominator are products of positive integers, provided only that the numerator equals the denominator. In short, it holds for a countable infinity of cases.

How can irrational numbers have this remarkable property? Does it hold for any positive integer besides 0?

In passing, we remark that the fact that the natural logarithm of a positive integer is irrational, enables us to quickly answer the question, Are there irrationals a, b such that a^b is rational? The answer is yes, because, since the natural logarithm of any integer (a rational number), say, 5, is irrational, this means that $e^r = 5$, where e is the base of the natural logarithms, and r is the natural logarithm of 5. Both e and r are irrational.

Against a Claim of N. J. Wildberger Regarding Multiplication of Irrationals

In a series of lectures on YouTube (they were viewable in July, 2015), N. J. Wildberger claims that one argument against the existence of irrationals is that multiplying two irrationals is a com-

plex and uncertain process, because we have to multiply from left-to-right (since each irrational has infinitely many digits to the right) whereas in multiplying rationals, we can proceed in the usual way by multiplying from right-to-left. (Thus, e.g., to multiply a/b by c/d, we can multiply the integers a and c, then multiply the integers b and d, giving us the rational number ab/cd.)

But Wildberger's argument is not valid. For, to multiply two irrationals, I can proceed as follows:

1. Compute the product of the most significant digit of each irrational.
2. Replace the product obtained in step 1 with the product of the *two* most significant digits of each irrational.
3. Replace the product obtained in step 2 with the product of the *three* most significant digits of each irrational.

etc.

This is, of course, an infinitely-long process, as is Wildberger's multiplication procedure. But each step is simple, and there are no uncertainties. Furthermore, the limit of the succession of products approaches the product of the two irrationals.

The Missing Irrationals

We frequently read that every real number can be approximated by rationals. Some of the infinite series that approximate irrationals like π and e are well known. The question now arises whether a representation of such a series must be capable of being written down (i.e., must be capable of being represented by a finite string of symbols). If the answer is Yes, then it seems we face a problem of missing irrationals, since there is only a countable infinity of finite strings, but an uncountable number of irrationals.

Of course, the reader may question the requirement that a series have a finite representation, and argue that a series is merely an integer-valued function on the non-negative integers, and that there is an uncountable number of such functions. Is this in fact the answer to the problem of the missing irrationals?

The Irrational in the Real World

A meter is defined as the length of a certain metal rod in Paris. Suppose someone uses this rod, plus all the care humanly possible at this time, to create a right triangle each of whose legs is of length 1 meter, and whose hypotenuse is therefore $\sqrt{2}$ meters. Suppose someone else is looking for a unit to measure by. Communications get scrambled, and this second person is sent, not the 1 meter length, but the $\sqrt{2}$ length. He uses it to create a right triangle each of whose legs is $\sqrt{2}$. Of course, since he thought he started with a unit, he thinks that the hypotenuse is irrational, namely, $\sqrt{2}$. We know it is rational, namely 2.

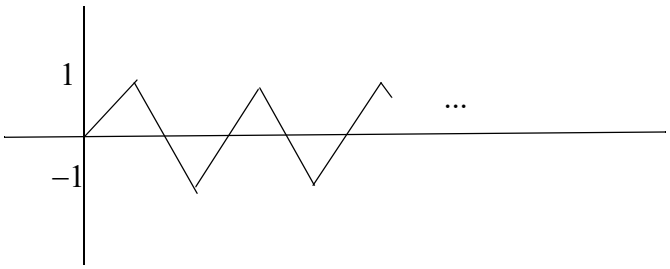
How do we know that any length we take to be a unit, no matter how accurately measured, is rational or irrational? Doesn't this uncertainty render suspect many calculations in physics?

On Δx, Δy, dx, dy and the Nature of the Infinitesimal

See the paper, "A New Insight Into an Old Calculus Mystery: Δx, Δy, dx, dy and the Nature of the Infinitesimal" on the web site occampress.com.

On Functions Continuous Everywhere and Differentiable Nowhere

In the mid-19th century, Weierstrass and others exhibited functions that were continuous everywhere and differentiable nowhere — a deeply troubling discovery for mathematicians of the time. But the definitions of these functions are difficult to understand, at least for me. The question is, why didn't a much simpler function suffice? For example, the limit f of an infinite sequence of functions f_n like the following:



Here, the period of each sawtooth waveform is, say, halved with each increment in n, the amplitude remaining the same.

A plausibility argument that f is nowhere differentiable is the following:

(1) for each $n > 1$ the function f_n has no derivative at any apex point, because the left-hand limit of the derivative (slope) as x approaches the x-value of the apex point does not equal the right-hand limit of the derivative (slope) as x approaches the x-value of the apex point.

(2) let $f_n(x)$ be a point on any straight-line portion of f_n. Clearly the derivative (slope) of $f_n(x)$ exists.

(3) let n increase. Then steps (1) and (2) apply, except that the absolute value of the derivative (slope) in step (2) is now greater than it was for f_n. Thus in the limit, the absolute value of the derivative (slope) is infinite. We conclude that f is differentiable nowhere.

It might be argued that, in the limit, every point in the function f is a vertex point, but this is not correct. Although the sequence {1/2, 1/3, 1/4, 1/5, ... } approaches zero, it never contains zero.

A Thought Regarding the Infinitesimal

A few calculus students (I was one), although they may be reluctant to admit it, were greatly bothered by the question, What is the nature of the infinitesimal? How can a number be arbitrarily small but not zero? In my first semesters of calculus, I walked the campus trying to imagine a long line of fractions extending indefinitely far ahead of me; I kept trying to get to the one that was arbitrarily small but not zero. Sometimes, for a moment, I thought I had succeeded, but then I realized that there is a smaller fraction beyond that one!

Naturally, I assumed that the nature of the infinitesimal was crystal clear to all the other students and, of course, to the professor. Only years later did I discover that the question had perplexed some of the best mathematicians in England and Europe for more than 150 years, until Cauchy found a formal definition for "limit" in the 1830s.

But I continued to try to imagine a number that was arbitrarily small but not zero. Then I heard that, in the 19th century, some mathematicians had proposed that the infinitesimal could be imagined as a function that asymptotically approached zero. The values of the function, taken as a whole, most certainly could be said to be "arbitrarily small but not zero". Of course, one had to be able to accept the idea of a "number" being an infinite *set* of numbers. But an ideal, in algebra, was such a number (e.g., all multiples of a given prime integer, say, 5, constitute an ideal). And,

of course, any fraction can be considered equivalent to a set of numbers, e.g., the fraction 1/2 can be considered equivalent to the infinite set of numbers 1/2, 2/4, 3/6, 4/8, ...

But still, in my idle moments, I couldn't help wondering if there might not be a single number that is the infinitesimal. I eventually heard about Abraham Robinson's discovery of non-standard analysis in the early 1960s, but I didn't understand it. It then occurred to me:

Why not do with the perplexing quantity, the infinitesimal, what mathematicians had done in order to deal with the perplexing quantity, $\sqrt{-1}$? Instead of spending more years trying to imagine what the square root of –1 could possibly be (what number can we possibly multiply by itself to get –1? how would that multiplication proceed?, etc.), mathematicians simply defined a new kind of number (called a *complex number*), whose general form is $a + bi$, where a and b are real numbers, and $i = \sqrt{-1}$, and built a consistent arithmetic (and calculus) upon it.

Why not do the same kind of thing with the infinitesimal? — decide on the general form for the infinitesimal, perhaps something like $a + bi$, and what the basic rules are that govern calculations with the infinitesimal, and leave it at that?

It turns out that someone has done this. In Google, enter

Non-nonstandard Calculus, I | The Everything Seminar

(This was written Jan. 9, 2021.)

Global vs. Local Approximation

In mathematics, I often have at least two choices: I can learn more and more about one specific local thing, e.g., a certain value of a function, or I can learn more and more about one global thing, e.g., an entire function. For example, the frequencies produced by a Fourier transform approximate an entire time function. They tell you something about the entire function, even though they do not tell you the precise value of any given point. Certainly a worthwhile project would be to collect, in one book, an overview of these two types of approximation, giving references to the various subjects in which they occur.

Global vs. Local More Generally

Suppose we have a grid in which each cell is either black or white.

Well-Ordering the Reals

So far, no one has been able to produce a well-ordering on the reals, even though such well-orderings exist. Suppose all mathematicians agree to start with any likely candidate, and develop the set as they need to. Then is it correct that there is no problem with this approach as long as no contradictions develop?

"Francisco Antonio 'Chico' Doria, a Brazilian mathematician...suggested [that] when mathematicians encounter an apparently undecidable proposition, they can create two new branches of mathematics, one that assumes the proposition is true, and one that assumes it is false. 'Instead of a limit of knowledge,' Doria concluded, 'we may have a wealth of knowledge.'" — Horgan, John, *The End of Science*, Broadway Books, N.Y., 1996, p. 231.

Integration Forever

Is there anything to learn from the study of the repeated integration of functions? In other words, in the case of functions of a single variable, for each such integrable function, *f(x)*, we integrate the function, then integrate the result, then integrate the result, repeating this an arbitrary number of times. What functions cannot eventually be "reached" in this way?

"It is Impossible to Write Down the Fourier Inversion Theorem!"

Every student of analysis learns the two equations that constitute the Fourier Inversion Theorem:

$$F(t) = \int e^{-ixt} f(x)\,dx$$

$$f(x) = \frac{1}{2\pi}\int e^{ixt} F(t)\,dt$$

But it would seem that neither of these equations can ever be written down in its entirety, because if we substitute the value of $f(x)$ into the right-hand side of the first equation, we then have to substitute the value of $F(t)$ into the resulting expression, and then $f(x)$ into the resulting expression, etc. And similarly beginning with the right-hand side of the second equation. What is wrong with this argument?

Taylor's Formula: the Derivation Made Clear

Taylor's Formula, which is familiar to most first-year calculus students, is a way to approximate functions $f(x)$ that have derivatives of all orders, by a simpler function $g(x)$, at and in the vicinity of a real number a. Taylor's Formula states:

$$g(x) = f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \frac{f'''(a)(x-a)^3}{3!} + \ldots + \frac{f^{(n)}(a)(x-a)^n}{n!}$$

Taylor's Formula should not be confused with Taylor's *Theorem*, which specifies the error in the Taylor's Formula, depending on how far from a the x is. We will not be concerned with the Theorem here.

Many students feel that although, on the one hand, the standard textbook derivation of Taylor's Formula seems simple and straightforward, it is nevertheless difficult to keep straight in the mind — it seems slippery, it seems almost to have a vicious-circle quality about it. The following is an attempt to provide a derivation that, I hope, the student will find more straightforward. The format is in accordance with one of the recommendations in William Curtis' *How to Improve Your Math Grades*[1], namely, that things that are "the same" should look "the same", even down to the structure of the sentences used. In this case, I have used a programming format to make clear exactly how the steps in the derivation are the same, and yet how they do not involve circular reasoning.

1. accessible on the web site www.occampress.com

I can do no better than begin by quoting from Morris Kline's *Calculus: An Intuitive and Physical Approach*[1], which is the best first-year calculus book I have ever come across.

"If we are to approximate a given function $f(x)$ by another, $g(x)$ say, then the second function should certainly be a relatively simple one so that we can integrate this second function or calculate its values, because these processes are the ones that cause trouble in the case of the complicated function. Now, the simplest functions to work with are the polynomials, and therefore we shall look into the question of the approximation of functions by polynomials.

"Generally, one is interested in the values of a function over some domain of x-values. This domain might be the interval over which one is integrating the function or it might be the domain over which one wishes to calculate the values of the function. Hence the problem we face is that of approximating a function over some domain of x-values. If we approximate a function very closely at and near some one value of x in that domain, we have some reason to expect that the approximation will still be good at values in the entire domain, at least if the domain is not large. Let us therefore look into the simpler problem of approximating a function around one value of x and then see what follows.

"Suppose that we have a function $f(x)$ and consider approximating it around or in the neighborhood of $x = 0$. Since, as already noted, polynomials are desirable approximating functions, let us consider the approximation

(1) $$g(x) = c_0 + c_1x^1 + c_2x^2 + c_3x^3 + ... + c_nx^n \ .$$[2]

What follows is now my rewriting, in the "programmatic" format mentioned above, of Kline's derivation.

Find the values of c_0, c_1, c_2, c_3, ..., c_n.

To do this:

0. Find the value of c_0.

To do this:

0.0 Find $g\,(x)$. We already know this. It is
$g(x) = c_0 + c_1x^1 + c_2x^2 + c_3x^3 + ... + c_nx^n$.

0.1 Set $x = 0$. This gives:
$g(0) = c_0$.

0.2 But certainly we want our approximating function $g(x)$ to agree with our function $f(x)$ at $x = 0$, so we set:
$g(0) = c_0 = f(0)$.

1. Find the value of c_1.

To do this:

1.0 Find $g\,'(x)$. By basic rules for derivatives, working from $g\,(x)$ in step 0., this is
$g\,'(x) = c_1 + 2c_2x^1 + 3c_3x^2 + ... + n\,c_nx^{n-1}$.

---

1. Dover Publications, Inc., Mineola, N.Y., 1977
2. ibid., p. 634

1.1 Set $x = 0$. This gives:
$g'(0) = c_1$.

1.2 But certainly we want the first derivative $g'(x)$ of our approximating function $g(x)$ to agree with the first derivative $f'(x)$ of our function $f(x)$ at $x = 0$, so we set:
$g'(0) = c_1 = f'(0)$.

2. Find the value of c_2.

To do this:

2.0 Find $g''(x)$. By basic rules for derivatives, working from $g'(x)$ in step 1., this is
$g''(x) = 2c_2 + 2\bullet 3c_3x^1 + ... + (n-1)n\, c_nx^{n-2}$.

2.1 Set $x = 0$. This gives:
$g''(0) = 2c_2$.

2.1 But certainly we want the second derivative $g''(x)$ of our approximating function $g(x)$ to agree with the second derivative $f''(x)$ of our function $f(x)$ at $x = 0$, so we set:
$g''(0) = 2c_2 = f''(0)$, or $c_2 = f''(0) / 2$

3. Find the value of c_3.

To do this:

3.0 Find $g'''(x)$. By basic rules for derivatives, working from $g''(x)$ in step 2., this is
$g'''(x) = 2\bullet 3c_3 + ... + (n-2)(n-1)n\, c_nx^{n-3}$.

3.1 Set $x = 0$. This gives:
$g'''(0) = 2\bullet 3c_3$.

3.2 But certainly we want the third derivative $g'''(x)$ of our approximating function $g(x)$ to agree with the third derivative $f'''(x)$ of our function $f(x)$ at $x = 0$, so we set:
$g'''(0) = 2\bullet 3c_3 = f'''(0)$, or $c_3 = f'''(0) / (2\bullet 3)$

...

n. Find the value of c_n.

To do this:

n.0 Find $g^{(n)}(x)$. By basic rules for derivatives, working from $g^{(n-1)}(x)$ in step $n - 1$., this is
$g^{(n)}(x) = 1\bullet 2\bullet ... (n-2)(n-1)n\, c_n + ...$ [terms in x].

n.1 Set $x = 0$. This gives:

$$g^{(n)}(0) = 1\bullet 2\bullet \ldots (n-2)(n-1)n\, c_n \,.$$

n.2 But certainly we want the *n*th derivative $g^{(n)}(x)$ of our approximating function $g(x)$ to agree with the *n*th derivative $f^{(n)}(x)$ of our function $f(x)$ at $x = 0$, so we set:
$g^{(n)}(0) = 1\bullet 2\bullet \ldots (n-2)(n-1)n\, c_n = f^{(n)}(0)$, or
$c_n = f^{(n)}(0) / (1\bullet 2\bullet \ldots (n-2)(n-1)n)$

And hence, we get, substituting in the values for the coefficients c_n we have found, and using factorial notation:

$$g(x) = f(0) + \frac{f'(0)x}{1!} + \frac{f''(0)x^2}{2!} + \frac{f'''(0)x^3}{3!} + \ldots + \frac{f^{(n)}(0)x^n}{n!}$$

We have derived Taylor's Formula at and in the vicinity of $x = 0$ How about in the vicinity of a, whether a is equal to 0 or not? We quote from Kline:

"...instead of approximating $f(x)$ by $g(x)$ at and near $x = 0$, we could equally well make the approximation at and near some other value of x, say $x = a$. If we proceeded to do this by using the form (1) for $g(x)$ and the above method of determining coefficients c_0, c_1,..., we would not succeed. We can see that we would be blocked in the very first step because for $x = a$

$$g(a) = c_0 + c_1a^1 + c_2a^2 + c_3a^3 + \ldots + c_na^n \,,$$

and we would like to have this expression equal $f(a)$. However, this time we do not obtain the value of c_0 at once as we did in the preceding case of $x = 0$. There might still be some way of proceeding with the form (1) of $g(x)$, but a wiser procedure is to recognize that the proper form of $g(x)$, which generalizes the form of (1), is

$$(6) \qquad g(x) = c_0 + c_1(x - a) + c_2(x - a)^2 + c_3(x - a)^3 + \ldots + c_n(x - a)^n \,.$$

That is, we recognize that the form of (1) is really of the form

$$g(x) = c_0 + c_1(x - 0) + c_2(x - 0)^2 + c_3(x - 0)^3 + \ldots + c_n(x - 0)^n$$

and that (6) is the proper generalization."

"With the form (6) for $g(x)$ we can use the procedure above to determine the coefficients so that at $x = a$, $f(x) = g(x)$ and each of the successive derivatives of $f(x)$ up to the *n*th derivative agrees with the corresponding derivative of $g(x)$."[1]

So we proceed as before:

---

1. ibid., p. 636

Find the values of c_0, c_1, c_2, c_3, ..., c_n.

To do this:

0′. Find the value of c_0.

To do this:

0′.0 Find $g(x)$. We already know this. It is
$g(x) = c_0 + c_1(x - a) + c_2(x - a)^2 + c_3(x - a)^3 + ... + c_n(x - a)^n$.

0′.1 Set $x = a$. This gives:
$g(a) = c_0$.

0′.2 But certainly we want our approximating function $g(x)$ to agree with our function $f(x)$ at $x = a$, so we set:
$g(a) = c_0 = f(\text{a})$.

1′. Find the value of c_1.

To do this:

1′.0 Find $g'(x)$. By basic rules for derivatives, working from $g(x)$ in step 0′., this is
$g'(x) = c_1 + 2c_2(x - a)^1 + 3c_3(x - \text{a})^2 + ... + n\,c_n(x - a)^{n-1}$.

1′.1 Set $x = a$. This gives:
$g'(a) = c_1$.

1′.2 But certainly we want the first derivative $g'(x)$ of our approximating function $g(x)$ to agree with the first derivative $f'(x)$ of our function $f(x)$ at $x = a$, so we set:
$g'(a) = c_1 = f'(a)$.

2′. Find the value of c_2.

To do this:

2′.0 Find $g''(x)$. By basic rules for derivatives, working from $g'(x)$ in step 1′., this is
$g''(x) = 2c_2 + 2 \bullet 3c_3(x - a)^1 + ... + (n-1)n\,c_n(x - a)^{n-2}$.

2′.1 Set $x = a$. This gives:
$g'(a) = 2c_2$.

2.2 But certainly we want the second derivative $g''(x)$ of our approximating function $g(x)$ to agree with the second derivative $f''(x)$ of our function $f(x)$ at $x = a$, so we set:
$g''(a) = 2c_2 = f''(a)$, or $c_2 = f''(a) / 2$

3′. Find the value of c_3.

To do this:

3′.0 Find $g'''(x)$. By basic rules for derivatives, working from $g''(x)$ in step 2′., this is
$g'''(x) = 2 \bullet 3c_3 + ... + (n-2)(n-1)n\,c_n(\text{x} - a)^{n-3}$.

3′.1 Set $x = a$. This gives:
$g'''(a) = 2\bullet 3c_3$

3′.2 But certainly we want the third derivative $g'''(x)$ of our approximating function $g(x)$ to agree with the third derivative $f'''(x)$ of our function $f(x)$ at $x = a$, so we set:
$g'''(a) = 2\bullet 3c_3 = f'''(a)$, or $c_3 = f'''(a) / (2\bullet 3)$

...

n′. Find the value of c_n.
To do this:
n′.0 Find $g^{(n)}(x)$. By basic rules for derivatives, working from $g^{(n-1)}(x)$ in step $(n-1)'$., this is
$g^{(n)}(x) = 1\bullet 2\bullet \ldots (n-2)(n-1)n\, c_n + \ldots$ [terms involving $(x - a)$].

n′.1 Set $x = a$. This gives:
$g^{(n)}(a) = 1\bullet 2\bullet \ldots (n-2)(n-1)n\, c_n$.

n′.2 But certainly we want the nth derivative $g^{(n)}(x)$ of our approximating function $g(x)$ to agree with the nth derivative $f^{(n)}(x)$ of our function $f(x)$ at $x = a$, so we set:
$g^{(n)}(a) = 1\bullet 2\bullet \ldots (n-2)(n-1)n\, c_n = f^{(n)}(a)$, or
$c_n = f^{(n)}(a) / (1\bullet 2\bullet \ldots (n-2)(n-1)n)$

And hence, we get, substituting in the values for the coefficients c_n we have found, and using factorial notation:

$$g(x) = f(a) + \frac{f'(a)'(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \frac{f'''(a)(x-a)^3}{3!} + \ldots + \frac{f^{(n)}(a)(x-a)^n}{n!}$$

A Challenge for Calculus Students: Laplace's Clever Series

In his history of mathematics, *Mathematical Thought from Ancient to Modern Times*[1], Morris Kline states (p. 1098), "The asymptotic evaluation of integrals goes back at least to Laplace. In his *Théorie analytique des probabilités* (1812) Laplace obtained by integration by parts the expansion for the error function

(1)

1. Oxford University Press, N.Y., 1972 This is the best history of mathematics I have ever come across.

$$Erfc(T) = \int_T^\infty e^{-t^2} dt = \frac{e^{-T^2}}{2T}\left(1 - \frac{1}{2T^2} + \frac{1 \cdot 3}{(2T^2)^2} - \frac{1 \cdot 3 \cdot 5}{(2T^2)^3} + \ldots\right)$$

,,

The first question is, why did Laplace go to the trouble of obtaining this series rather than simply using the standard one that is obtained by integrating term-by-term the series expansion of the integrand, i.e.,

$$\int e^{-t^2} dt = \int\left(1 + \frac{(-t^2)^1}{1!} + \frac{(-t^2)^2}{2!} + \frac{(-t^2)^3}{3!} + \ldots\right) dt$$

yielding,

(2)

$$\int e^{-t^2} dt = \left(t + \frac{(-t^3)}{3 \cdot 1!} + \frac{(t^5)}{5 \cdot 2!} + \frac{(-t^7)}{7 \cdot 3!} + \ldots\right)$$

The answer almost certainly is: because the upper limit of the definite integral in (1) is ∞, hence it is not obvious that the series in (2) will yield a finite value. And yet we know[1] that it must, because

$$\int_0^\infty e^{-t^2} dt = \frac{\sqrt{\pi}}{2}$$

Thus we know that, if $T \geq 0$,

$$\int_T^\infty e^{-t^2} dt \leq \frac{\sqrt{\pi}}{2}$$

So it is understandable that Laplace would look for another series. But how did he find the one in (1)? Kline says he used integration by parts, a rule that asserts that, if u, v are functions of x, then

$$\int u \ dv = uv - \int v \ du$$

However, when we try this in the obvious way on the integral in (1), we find it doesn't work. That is, letting

1. I do not know if Laplace knew this at the time he derived his series.

$$e^{-t^2} = u, \quad dt = dv$$

we get
(3)

$$\int e^{-t^2} dt = e^{-t^2} t - \int t \; e^{-t^2} (-2t) dt$$

which does not give us the desired series because, e.g., letting

$$e^{-t^2} = u, \quad (-2t) dt = dv$$

we get, for the integral on the right in (5),

$$\int t \; e^{-t^2} (-2t) dt = e^{-t^2} (-t^2) - \int (-t^2) e^{-t^2} (-2t) dt$$

etc.

So what was Laplace's trick? The answer is very simple , once we see it. Since some students may not want to see it before they have had a chance to discover it on their own, it has been placed in "Appendix A — Derivation of Laplace's Series" on page 132.

Linear vs. Non-linear

Why not think of a non-linear (single-argument) continuous function as a linear function in which certain ranges of the *x* axis have been compressed or expanded, e.g., as though the (rubberized) paper was scrunched together too much, or stretched apart too much, in these ranges?

A Thought on Differential Equations

Almost everyone who begins the study of differential equations is struck by the wealth of material to be learned — the fact that each equation, or each class of equations, seems to require its own methods. Is there any way the subject can be simplified?

Let me begin with a very simple example. Consider standard 2-dimensional Cartesian coordinates. But instead of regarding an ordered pair of integers, $\langle x, y \rangle$, as defining a *point*, i.e., the intersection of a vertical grid line and a horizontal grid line, let the ordered pair define a *square.* Specifically, we define a new set of coordinates in which points have been "expanded" to squares, all squares being of the same size. Thus, $\langle x, y \rangle$ now denotes the location of a square.

We can fill this new grid of squares with the values of a function taking two integers as arguments, and returning integer values, e.g., the ordinary arithmetic functions addition, subtraction, multiplication, and division. Let us consider the case of multiplication. In the square $\langle x, y \rangle$ we place the value of $x \bullet y$. Now, observing this grid of values, we see that the value in the square

$<x + 1, y>$ is (obviously) simply the value in the square $<x, y>$, $+ y$. The thought may now occur to us that, once we have gone through the labor of computing the value in square $<x, y>$, it only takes "a little more" labor to find the value in square $<x + 1, y>$. And not much more labor to compute the value in square $<x + 2, y>$, or in square $<x, y + 1>$, or in square $<x, y + 2>$, etc. We don't have to compute each value from scratch.

Let us see if this idea has any validity in the realm of differential equations. We can define an ordered n-tuple that allows us to represent each differential equation. Now the question is, assuming we know how to solve the differential equation in n-tuple t, is there a transformation that will give us the solution to the equations represented by adjacent tuples and do so with "less work" than we would have had to expend in figuring out from scratch how to solve these equations? And if not, why not?

The Application of *m*-Dimensional Matrices to Number Theory

The basic idea in the above section can be applied to integer functions in number theory. An example will illustrate how this can be done.

"Consider a four-dimensional matrix M such that cell (u, v, w, k) is occupied by the value of $u^k + v^k - w^k$, where u, v, w, k are positive integers. The matrix makes it possible to speak of the values of neighboring cells, given the value and location of a cell — if we know u, v, w, k, then we can compute the value of $u^k + v^k - w^k$, and then we can compute the value of, for example, $(u - 1)^k + v^k - w^k$, which is the value of one of the cells next to that containing $u^k + v^k - w^k$. In fact, there are 10 cells next to each cell except where one of the arguments = 1, because each of the arguments (or "coordinates") can be increased by 1 or decreased by 1.

"The matrix provides a framework for mathematical induction on any coordinate. We assume that a cell contains 0, which would be the case if a counterexample existed, and then compute the value of each neighboring cell such that at least one of the coordinates is decreased by 1. We then repeat this process until we arrive at a cell the value of whose contents is known from other results. If the values differ, then we know that the assumption of a counterexample was false, and thus FLT is proved.

"As an example of our inductive process, let us consider the cell (x, y, z, p), whose value, by our assumption of a counterexample, is 0. Does the adjacent cell $((x - 1), y, z, p)$ contain a negative or a positive value? We see immediately that it contains a negative value, because $(x - 1)^p + y^p - z^p + (x^p - (x - 1)^p) = x^p + y^p - z^p = 0$, and $(x^p - (x - 1)^p)$ is positive. Informally, if we add a number a to a positive number b and get zero, then a must be negative.

"We conclude that the cell $(1, (x - 1), (y - 1), z, p)$ contains a more negative number than $(1, (x - 1), y, z, p)$." — Schorer, Peter, section "Four-Dimensional Matrix Approach" in "Is There a 'Simple' Proof of Fermat's Last Theorem?", www.occampress.com.

It appears that, in general, the values of any integer function of a finite number m of values can be represented by such an m-dimensional matrix. Mathematical induction, including Fermat's "method of infinite descent" is captured by this scheme. Furthermore, we can move in any of m directions in order to build a proof.

The Complex Plane vs. the Real Plane

Granted, this is only a student question, but I have never seen an answer to it: what exactly is different, in terms of results, theorems, lemmas, between the complex plane and the real plane?

Can we pair up the results for each, so that we can say, "This is the real plane version of this complex plane theorem", etc.?

Graphics as an Independent Variable

Project: Devise a program that gives the equations of a surface as the surface is changed. Such a program would do just the opposite of what graphing programs normally do, namely, show how the surface changes as the equations change.

Curved Space: Is There a Simpler Approach?

Certainly tensor calculus is one of the most difficult mathematical subjects. We recall that Einstein, when trying to learn it for use in his General Theory of Relativity, was driven to write to his colleague, Marcel Grossman, "Marcel: Help me!" Not the least reason for its difficulty at present is the poor quality of presentation of the subject. One need only consider the definition of a tensor. Different texts seem to have entirely different definitions. No text that I am aware of provides proofs of the equivalence of the definitions. No text that I am aware of gives a formal definition of a general tensor, that is, one of arbitrarily large dimension, and of arbitrarily large rank $(n + m)$. No text that I am aware of explains what motivated the transformations that are used (in some cases) to define tensors, or why there are covariant, contravariant and mixed tensors, or what motivated the contraction operation.

Tensor calculus can be described as the calculus of curved spaces. This made it the appropriate branch of mathematics for Einstein's General Theory of Relativity, in which gravity curves space-time.

Let us for the moment step back from the details of tensor calculus, including the fact that, if a tensor equation is true in one set of coordinates, it is true in all sets of coordinates, which allowed Einstein to say that if a law of physics (expressed in tensors) is true in one set of coordinates, it is true in all sets of coordinates.

Let us imagine a metal bar with the ends bent downward. The bar is our curved space. Let us further imagine that, before the bar was bent, throughout it there were straight lines running the length of it. These lines are bent in conformity with the bending of the bar.

We ask, now, if it is possible to give a mathematical description of each of these lines without referring to coordinates. Because, after all, the bent bar, and its lines, are the same no matter what the coordinates of the space we might place it in.

Or suppose we imagine a long tunnel in space. The tunnel curves in various directions. Suppose our task is to guide a plane along the center of the tunnel. Then to remain in the center, we will have to cause the plane to turn in various directions. The angle of rotation of the steering wheel at each moment, and the speed of the plane at each moment, together give us a description of the curve — without coordinates.

Can we apply the idea of an inertial navigation system here?

The astute reader will point out that we will need to specify, using coordinates, the beginning the curve (the point that we consider the beginning, for our purposes), and similarly, the beginning of each line that runs the length of our metal bar. We can take each of these beginnings as the origin of a set of coordinates. In other words, in the scheme we are describing, there is an origin of the set of coordinates for each curve in space, or each line in our metal bar!

Is this a beginning of an alternate approach to the calculus of curved spaces? What about a mesh throughout curved space, the mesh having edges of any finite length we choose (all the lengths being the same)? Suppose we can choose between the edges being straight lines but the angles between them at the nodes being allowed to vary? Or suppose the edges can be curved?

Two Questions Re Parallel Transport of Vectors

In an *n*-manifold, the notion of vectors being constant (parallel) throughout the manifold, or a part of it, is important, but difficult to formalize. Textbooks typically explain some of the difficulties by moving a given vector around on the great circles of a sphere in 3-space.

Would anything be accomplished if the vector were held in place, and the sphere were rotated about its center?

Would anything be accomplished if we began with the desired vector *v* with its lower end positioned ("fastened") at a desired point *p* on a curve **C**? Then suppose we positioned a copy *v′* of *v* at a small distance *ds* from *v* on *C*, parallel to *v*, and fastened the copy at that point p′. Then suppose we repeated this process with copies of *v* as far along C as we wished.

When we were finished, we would have a "picket-fence" of parallel vectors fastened to a succession of points on *C*.

The Spinning Top

I have so far seen only three expositions of elementary facts about the child's toy, the spinning top: one by Saunders Mac Lane in his book *Mathematics Form and Function* (Springer-Verlag, N.Y., 1986, pp. 295-301), one by Michael Fowler in his lecture "Euler's Angles" (Google), and one by two physicists (see"An Example of (Part of) the Job Done Right" on page 110).

Overcoming Incompetent Presentations

Of the three, Mac Lane's is by far the worst[1], but we can get at the reason for the poor quality by asking the Student's Forbidden Question, namely, "Why is this difficult?"

The first answer is, "Because necessary diagrams are either missing altogether, or are very hard to understand". The second answer is, "Because we often do not know what the purpose is for all the complicated mathematics."

Both of these difficulties could be greatly reduced by a presentation in the following form:

• A large diagram showing the top with its axis *z* of rotation at an angle Θ from the vertical axis *Z*, and

The other two axes *x*, *y* of the top, and

The other two axes *X*, *Y* perpendicular to *Z*.

The location of the center of mass in the top, and

A labeled line showing the distance of the center of mass from the point on which the top spins.

• An informal description of the behavior of the top, beginning with it being spun with its axis z vertical, then eventually, the axis making an angle Θ with the vertical and rotating about the ver-

1. If someone had told me, in my youth, that the portion of Mac Lane's presentation of the spinning top on pp. 298-299 of his book (ibid.) was "real mathematics", I would have taken up sociology or literary criticism instead.

tical axis (*precession*), then the axis starting to bob (*nutation*), and the bobs becoming larger in amplitude, until the top falls over on its side.

The description would conclude with a statement to the effect that we will be considering this entire behavior of the top.

• A list of the principal properties of the top, with brief definitions. This list would include:

Euler's angles (these three angles fix the position of the top at any moment; one of the angles is the above Θ);

The rate of change of each of Euler's angles ;

The angular speed of rotation about each of the three axes x, y, z of the top.

The moment of inertia for each of the three axes of the top (the moment of inertia is the equivalent in rotational motion of mass in rectilinear motion);

The angular momentum about each axis of rotation;

The kinetic energy of the top at any moment;

• A more formal description of the behavior of the top over time (its *trajectory*), beginning at time $t = 0$ with the axis of the top vertical and the top given a specified starting angular speed of rotation. The time from $t = 0$ is divided into arbitrarily small increments Δt; at any time $t = n\Delta t$, where $n \geq 0$, we can observe the value of each principal property by looking at a curve of its values plotted on two-dimensional Cartesian coordinates, with the horizontal axis showing $n\Delta t$, and the vertical axis showing the units in which the property is measured. Thus there will be several different sets of two-dimensional Cartesian coordinates.

If at some point we find it desirable to know a speed or velocity of angular rotation in terms of other speeds or velocities of angular rotation, then we shall carry out the derivation. But it must be supported by clear diagrams.

However we will absolutely not include text like the following without explaining why it is not nonsense:

"[Vector representation of velocities of angular rotation] is used because it is effective in combining two angular velocities.; by calculating the composite of two rotations one can prove that the effect of combining the two such angular velocities is represented by the sum of the two vectors."[1]

Suppose I have an electric drill. It is rotating at a certain speed. The drill is pointing in a certain direction. It is represented by a vector. Suppose I have a second electric drill. It is rotating at a different speed and is pointing in a different direction. It is represented by a different vector. What does "the sum of the two vectors" mean? That as soon as I think of this sum, a third electric drill will come into being, rotating at the speed and pointing in the direction that is dictated by the sum of the two vectors?

And so the presentation must make crystal clear what it means to say that a given velocity of angular rotation has *coordinates* that are other velocities of rotation.

Unquestionably, the presentation should include an explanation of the *physics* governing the top's behavior — the physics making clear why the top's behavior from the initial spinning to its finally falling over on its side, is as it is.

---

1. Mac Lane, ibid., p. 297.

I refuse to believe that such a presentation would not be more appreciated by students, and be of more interest to them, and of more ultimate use to them, than the complicated, seemingly pointless, mathematics in the Mac Lane and Fowler presentations.

An Example of (Part of) the Job Done Right

One of the reasons that rotational motion is a difficult subject is that there are so many new symbols and technical terms, which, in keeping with the mathematicians' refusal to question if prose is always the best way to present mathematics, are always presented in a many-page sprawl of prose.

But sometimes someone sees the light — in this case, two authors of a standard textbook, *Physics*[1]. On p. 278 of Part I, in the first chapter on rotational dynamics, is a table comparing some of the basic properties of rectilinear motion with some of the basic properties of rotational motion about a fixed axis. Thus, for example, we can see *at a glance* that mass, M, in rectilinear motion, is analogous to rotational inertia, I, in rotational motion, and that linear momentum, Mv, where v is velocity, is analogous to angular momentum, $I\omega$, where ω is angular velocity. But the table should appear at the *start* of the chapter, not 19 pages into the chapter.

And, of course, the index of the book should include each and every symbol and technical term, so that the reader can find out in a matter of seconds the definition of I, ω and each of the many other symbols and technical terms in the subject of rotary motion (and indeed of all subjects covered in both Parts of the book).

Note: I'm afraid that the above authors at times slip into standard textbook-author-incompetence, as, for instance, in their derivation of the speed of precession on p. 297, in which we are asked to believe that there exist isosceles triangles in which one of the base angles, hence both, can be 90 degrees.

There Were Self-Similar Structures Long Before Fractals!

Roughly speaking, a self-similar structure is one that "looks the same" no matter how small a piece of it we view. Fractal geometry, which was discovered by Benoit Mandelbrot in the latter half of the 20th century, deals extensively with such structures. But they were known already to the ancient Greeks.

Consider the Golden Rectangle, which is a rectangle whose sides are in the ratio $(1 + \sqrt{5})/2 : 1$. Each Rectangle consists of a square and a rectangle which is another, smaller, Golden Rectangle. And similarly inside that Golden Rectangle, etc.

Or consider the Greek representation, using dots, of successive squares of positive integers. Begin with one dot. Make an L-shape out of 3 equally-spaced dots, and place that L-shape at the right-hand side of the initial dot so that a square, two dots on a side, is formed. Now make an L-shape out of 5 equally-spaced dots, and place that L-shape at the right-hand side of the square consisting of four dots. We now get a square of three dots on a side. Etc.

By this sequence, we see that the sum of the first n odd positive integers equals n^2.

The L-shapes are called "gnomons". Other gnomons are known in plane geometry for the successive expansion of parallelograms..

We can consider the geometric squares inside the succession of Golden Rectangles to be gnomons.

---

1. Resnick, Robert, and Halliday, David, *Physics*, John Wiley & Sons, Inc., N.Y., 1966.

Number Theory

Geometrical Domains for Number Theory Functions

In my research on several difficult number theory problems, I have found that considerable progress can be made by mapping (in an informal sense) number theory functions onto appropriate geometric domains. (The term "domain" here does not mean the domain in the formal definition of any function.)

This idea is different from the mere graphing of the function, as the reader will see in the following examples. I will begin with one of the simplest.

The Multiplication Plane

Let x be the horizontal axis and y the vertical axis in the traditional Cartesian axes. We here consider only non-negative x and y. But now, instread of (x, y) denoting a point, we let it denote a square (all squares are the same size), and for the multiplication function $f(x, y)$, we let the square (x, y) contain the product $x\,y$. Thus, e.g., the square (124, 7965) contains 987,660. But now observe: if we want to know the product of 124 and 796*6*, we do not need to carry out the entire multiplication again. We need simply add 124 to 987,660, yielding 987,784. And similarly for the product 12*5* and 7965, we simply add 7965 to 987,660, yielding 995,625. In the case of 123 or 7964, we subtract.

The point is that if we know the contents of a square, we can easily determine the contents of an adjacent square.

The 4-Dimensional Space for Fermat's Last Theorem

The same idea can be carried out with number-theoretic functions of more than two variables. Consider, for example, the function $x^k + y^k - z^k$, where x, y, z, k are positive integers. Here, instead of a plane, we use a 4-dimensional integer space such that the contents of the 4-dimensional box (x, y, z, k) contains $x^k + y^k - z^k$. Once again it is easy to see that if we know the contents of such a box, we can easily determine the contents of an adjacent box, and this might be useful if we assume that a counterexample $x^k + y^k - z^k = 0$ to Fermat's Last Theorem exists. In particular, it might allow us to invoke an inductive argument that proceeds downward through decreasing values of x, y, z, or k. Further details will be found in the section, "Four-Dimensional Matrix Approach" in Part (1) of the paper, "Is There a 'Simple' Proof of Fermat's Last Theorem?" on occampress.com.

Geometrical Models of Congruence

It is possible to represent geometrically the integers mod m, where m is any positive integer greater than 1, in at least three ways, (1) the wheel-and-spokes representation; (2) the helical representation; and (3) the lines-and-circles representation. All can easily seen to be equivalent. Here are the details.

The Wheel-and-Spokes Representation of Congruence

Divide a circle into m equal segments, making the segments such that one begins at the lowest point of the circle; place a tick mark at the end of each segment; at each tick mark, draw a straight line (a "spoke") outward on a radius of the circle through the mark. Place tick marks at intervals

of a constant distance d on each spoke. Now begin at the base of the lowest spoke and mark it 0. Proceed counterclockwise around the circle, marking the base of each successive spoke, 2, 3, 4, ..., m - 1. (We say that these numbers are on level 0). Now proceed to the next level (level 1) of tick marks, and proceed counterclockwise, marking each tic mark m, $m + 1$, $m + 2$, ..., $2m - 1$. Then proceed to the next level (level 2), etc.

The Helical Representation of Congruence

Here we again begin with a circle, but we orient it horizontally, then use it to define an infinite cylinder in the vertical directions upward and downward from the circle. We now select a 0 point on the circle, and define a helix passing through that point and running upward and downward around the cylinder. We place tick marks at regular intervals on the helix so that they define m vertical lines (corresponding to the spokes in the wheel-and-spokes model). The positive (and here, negative) integers are marked at the tic marks in a manner analogous to that for the wheel-and-spokes model. Levels are numbered similarly.

The Lines-and-Circles Representation of Congruence

This is virtually identical to the helical representation, and is described in my paper, "Is There a 'Simple' Proof of Fermat's Last Theorem?", accessible as a downloadable PDF file on the web site www.occampress.com.

These representations make certain relations in elementary congruence theory much easier to understand. For example, $x \equiv y \bmod m$ simply means that x and y lie on the same spoke or the same vertical line. $x = qm + r$ simply means that x lies at the intersection of level q and spoke (or line) r.

These models also suggest approaches to the solutions of problems (see, e.g.,"Is There a 'Simple' Proof of Fermat's Last Theorem?" on occampress.com).

Furthermore, eliminating the rods for the moment, but keeping the integers with their associated tick marks, we can imagine the radius of the cylinder increasing continuously, and then ask questions about the change in location (their angle) of various integers as this process occurs. Clearly, if the radius increases without limit, then the integers will eventually line up as for mod 2, mod 3, mod 4, ..., mod m, ... for any m. Given that the definition of equality in modular arithmetic is, $a = b$ iff $a \equiv b \bmod m$ for all m, we might begin to wonder how equality is even possible when viewed in terms of the continually increasing diameter of the cylinder!

Another question that arises is whether we can generalize the notion of modulus, remainder, etc., to other operators. In the case of modular arithmetic, we have addition, subtraction, multiplication, and, in certain cases, division. Suppose we take as our fundamental operation, exponentiation relative to a base m, as opposed to addition relative to a modulus m. Then corresponding to 0, m, $2m$, $3m$, ... on the 0 spoke (or 0 line) would be m^0, m^1, m^2, m^3, ... on the 0 spoke (line). Corresponding to remainders mod m would be fractional powers except that here we would have to limit the number of decimal places for each such fractional power in order to ensure only a finite number of such fractions between each successive power of m.

Ackermann's function, well known to computer scientists, might lend further insight into the idea of higher level moduli.

The Prime Numbers Plane

Here, we map (informal sense) all the positive composite numbers and all multiples of each prime, onto to the positive plane. Details will be found in the section "Graphing the Primes" on page 114 .

The Tree of Tuple-sets for the 3*x* + 1 Problem

The 3*x* + 1 function can be mapped (informal sense) into an infinite tree of planes, each plane containing what I have called a "tuple-set". The tuple-set structure as of now seems to make possible a remarkably simple proof of the 3*x* + 1 Conjecture. See the paper, "A Solution to the 3x + 1 Problem" on occampress.com.

Mappings Between Moduli

Why is it that there are so few results, in elementary congruence theory, concerning mappings between moduli ? What can we know about mappings from smaller moduli to larger, and vice versa? How many such mappings are there from a given modulus? Which ones are of particular interest? (We are really talking here about mappings between arithmetic series, because each residue class mod *m* is in fact an arithmetic series.)

Divisibility

In determining if an integer *y* is evenly divisible by an integer *x*, is it ever advantageous, computationally, to convert *y* into a number in base *x* and then simply check if the least significant digit is 0? Is the answer yes only if it is necessary to determine divisibility of a large number of *y* by a given *x*?

A counterargument is probably that the computational labor in converting *y* into a number in base *x*, is greater than the computational labor of simply dividing *y* by *x* and seeing if there is a remainder.

But suppose there were a list of a very large number of positive integers in base 2, and list of a very large number of positive integers in base 3, and ...

Then to determine if a given positive integer *n* were prime, one would only need to look at the *n*th item in each list for all lists up to the smallest base greater than the square root of *n*...

The second-hand of a watch or clock is seen to land in between the seconds divisions at each tick. What condition would be necessary for the hand eventually to land at every point on or between divisions?

Counting Made Difficult

Instead of counting 1, 2, 3, ... why not count *interestingly*, e.g., (complicated integral that equals 1), (complicated something else that equals 2), (complicated something else that equals 3), etc. Then everyone's counting would be individualized, and, in particular, counting would be a challenge again. Well, no, it probably wouldn't, because once everyone knew that the activity of counting was going on, they wouldn't have to figure out each number. But they would if it became the practice to express every integer, whether in a counting sequence or not, via a complicated expression.

Calculating With Waves

What kinds of calculations can we perform using waves? For example, I can determine the least common multiple of any two positive integers by superimposing waves having frequencies

corresponding to the two integers, then (assuming the waves are in phase) looking for the points at which their peaks coincide. We can add or subtract any two numbers by representing the numbers by the amplitudes of in-phase waves and determining the amplitude of the resulting wave. Calculating with waves would allow the loss of many points of the waves without rendering the result ambiguous.

A Challenge for the Creative Few

Make a useful calculator out of a piece of cloth the size of, say, a men's handkerchief. The cloth can be printed with any desired figures, lines, colors.

Number Popularity

Let us say that powers of integers — squares, cubes, etc. — are popular because the corresponding functions — $y = x^2$, $y = x^3$ — are popular functions. The question is, are there some numbers which are "lonely" because they occur in hardly any functions? Begin with the lexical ordering of all Turing machines under some Turing machine formalism, tally the numbers which they produce (out to some reasonable maximum) for the lexical ordering of all inputs.

On Determining if a Number is Prime

Preliminary Remark

The literature on determining if a number is prime is extensive and profound. An unsophisticated reader could almost imagine that the primes occur randomly in the sequence of positive numbers. And yet the fact is that the primes result from an extremely orderly process: it is known as the Sieve of Eratosthenes and is as follows. The first prime is 2. Delete from the set of positive integers all multiples of 2 (except 2 itself). Now go to the next larger number that is not a multiple of 2. That is 3. Delete from the set of positive integers all multiples of 3 (except 3 itself). Now go to the next larger number that is not a multiple of 2 or 3. That is 5. Delete... etc. The numbers left by this process are all and only the primes.

I assume that the only reason that this process can not be automated via a computer program is that the number of known primes at present is so large that determining the next larger number that is not a multiple of all known primes is beyond the powers of modern computers (and computer memory).

But can the Sieve not be the basis of a means for determining if a number is prime?

Graphing the Primes

Consider a line graph of all the positive composite integers and all the positive primes created as follows. Let the vertical left-hand axis contain the positive integers. All of the following vertical axes are evenly spaced from each other and from the vertical left axis. The first axis to the right contains 2 and all the positive multiples of 2. Each number on this axis is on a horizontal line drawn extending from the same number on the vertical left axis. Each multiple of 2 is marked by a heavy dot.

The next axis to the right contains 3 and all the positive multiples of 3. Each number on this axis is on a horizontal line drawn extending from the same number on the vertical left axis. Each multiple of 3 is marked by a heavy dot.

The next axis to the right contains 4 and all the positive multiples of 4. Each number on this axis is on a horizontal line drawn extending from the same number on the vertical left axis. Each multiple of 4 is marked by a heavy dot.

The next axis to the right contains 5 and all the positive multiples of 5. Each number on this axis is on a horizontal line drawn extending from the same number on the vertical left axis. Each multiple of 5 is marked by a heavy dot.

Etc.

Then, to determine if any given positive integer u is a prime, simply move horizontally from u on the vertical left axis rightwards to the vertical axis with u at its base. If you do not intersect any heavy dots in the process, then u is prime. If you do, then u is composite.

For example, if u is 12, then in moving rightwards from 12, we intersect a heavy dot on the 2 vertical axis ($2 \times 6 = 12$), so we know that 12 is composite; then, moving rightwards, we intersect a heavy dot on the 3 vertical axis ($3 \times 4 = 12$); then, moving rightwards, we intersect a heavy dot on the 4 vertical axis ($4 \times 3 = 12$); then, moving rightwards, we intersect a heavy dot on the 6 vertical axis ($6 \times 2 = 12$). We intersect no further heavy dots until we reach 12 ($12 \times 1 = 12$). We stop.

In the case of $u = 13$, we intersect no heavy dots until 13, and conclude that 13 is prime.

This is an extremely simple way to determine the primality of positive integers! The only drawback is that it requires unlimited space (computer memory) for the graph. On the other hand it allows us to implement a rule of "no repeats of the same calculation!" That is, up to the limits of our computer memory, and the spare computer time between other tasks, we could have the computer calculate the heavy dots. Once they were in the graph, they would not have to be calculated again.

The question now is: What is the largest square matrix that can be stored in a modern computer, if each element of the matrix is either empty or contains a heavy dot? Such a matrix could be filled with heavy dots according to the rule described above, in times when the computer was idle. This could, and should by done, via a sequence of smaller matrices of successively increasing sizes, growing from the lower left-hand corner, e.g., the 1 x 1 maze, then the 2 x 2 maze, then the 3 x 3 maze, then ... A test to see if a number u is prime would then begin with a query as to size of the largest maze that had been filled. If the size was greater than or equal to u x u, then the above-described method for determining if u could be carried out.

In passing, we raise the following question: considering the extensive computations that are currently needed to determine primaility, and the simplicity of the above graphical method, is it possible that we can view exxtebsuve computations as a way of attempting to deal with too-limited space? Are computations what we must do to "get around" our having too little space to make them simple — to "spread things out"?

We also ask if it is possible that a graphical representation like this is what enabled the remarkable twins described by Oliver Sacks to quickly determine if large integers were prime or not.

Questions Regarding Determining If a Number is Prime

When we think of the question of whether a given number c is prime, we regard it as a question about its factors, hence a question about *division*. Is it possible that an efficient way to determine if c is prime might be to compare the result when c is *multiplied* by a prime and the result when c is *multiplied* by a composite? Is it possible that an efficient way to determine if c is a prime might be to observe properties of all a, b such that $a + b = c$?

But this seems to require much more computation than simply dividing c by the primes in succession, up to the square root of c.

A Technique for Determining a Probability That a Number Is a Prime

Let c be a positive integer. Divide c by the largest known prime p such that $p \leq \sqrt{c}$. (The largest factor of c must be $\leq \sqrt{c}$.)

If there is no remainder, then we know that c is composite. If there is a remainder, i.e., if $c = qp + r$, where $1 \leq r < q$, then we may think that all we know is that p is not a factor of c. But we know more.

Consider the segments, $s_0 = \{0, ..., p - 1\}$, $s_1 = \{p, ..., 2(p - 1)\}$, $s_2 = \{2p, ... \ 3(p - 1)\}$, etc. The integer c must be in one of these segments. In fact, c is in the segment s_q. Now consider all primes $p_1, p_2, p_3, ..., p_k$, where p_k is the largest prime $< p$. Then the segment s_q contains multiples of each of these primes. Let S_p denote the number of all these multiples that are in the segment s_q. We can therefore conclude that the probability that c is prime $\leq (p - S_p)/p$.

Unfortunately, the amount of calculation needed to implement the above procedure, is not less than that required to divide c by each prime $\leq \sqrt{c}$, and thus the procedure is impractical.

A Technique for Determining if a Number Is a Prime

Let c be a positive integer. If all primes p such that $p \leq \sqrt{c}$ are known, then the following procedure might be a computationally-feasible way to determine if c is a prime.

1. Compute the product P_1 of all successive primes 2, 3, 5, ..., p_k, where p_k is the largest prime such that P_1 can be computed in an acceptable time.

2. Compute the greatest common divisor (P_1, c) of P_1 and c. If $(P_1, c) \neq 1$, then stop: c is composite. Otherwise, go to step 3.

3. Compute the product P_2 of all successive primes $p_{k+1}, ..., p_{k+j}$, where p_{k+j} is the largest prime such that the product can be computed in an acceptable time.

Repeat steps 2 and 3 with appropriate modifications and with the successive primes increasing in the obvious way.

If the greatest common divisors $(P_i, c) = 1$ for all products P_i up to the one that includes the largest prime $\leq \sqrt{c}$, then c is prime. Otherwise c is composite.

Obviously, the P_i can be computed beforehand and stored. The only limitation on the technique, then, is the computation time of the greatest common divisors. The computations of these can be parcelled out among several computers — ideally, so that each computer is only burdened with one such computation.

A Technique for Determining the Factors of a Number

Assume that we have a very simple hardware circuit, which we will call a *divider*, that does nothing but repeatedly subtract a given prime divisor p from a given dividend n, pausing after each subtraction to determine if the result is greater than or equal to 0. If a value less than 0 is reached, then the divider prints, "n is not divisible by p". If 0 is reached, then the divider prints, "n is divisible by p".

Our circuit thus performs division without being concerned about a quotient. Thus it does not require any memory except that required to hold the dividend n (the result of each subtraction of p is stored back into the memory that held n) and the divisor p, and, of course, the very simple subtraction circuitry. We assume that this division can be very fast, giving the current state of hardware technology.

At the time of this writing (April, 2013), numbers consisting of more than 100 decimal digits are considered difficult to factor. Assume that the number of binary digits is roughly four times

the number of decimal digits. So we need at least 400 bits to hold the dividend n and, as a worst case, at least 400 bits to hold the divisor p. Assume another 200 bits to hold the subtraction circuitry. Thus each divisor requires at least 1,000 bits.

A large number of divisors can be wired in parallel on a single integrated circuit chip, and then a large number chips can be wired in parallel. We assume, initially, that current technology permits at least 1,000,000 such dividers — in other words, we assume that 1,000,000 trial divisions (each consisting of subtractions only) can be made simultaneously. We assume that each trial division can be done in no more than one second, hence we can perform 1,000,000 trial divisions in no more than one second.

Let k denote the number of dividers.

An outline of an algorithm for determining the factors of a positive integer n (and hence for determining if n is prime) is the following.

1. Load n as dividend into each of our k dividers. Load the jth prime as divisor into the jth divider, for $1 \leq j \leq k$.

2. Start all the dividers. Each divider then repeatedly subtracts its divisor p from n until it returns "n is not divisible by p" or "n is not divisible by p." We remark in passing that as soon as the latter message appears, if in fact such a message ever appears in our tests of n, we know that n is not prime.

As each divider produces its result, the next untested prime is loaded as divisor into the divider, and the process is repeated. It continues until all primes less than or equal to $\sqrt{n}$ have been tested. (Obviously the algorithm will have to deal with the possibility that the largest known prime is less than $\sqrt{n}$.)

According to the Prime Number Theorem, the number of primes less than a number x is given by $(x/(\log x))$, where "log" here is the natural log.

Assuming we test 1,000,000 primes a second, that means we can test 86.4 billion primes, or 86.4×10^9 primes in 24 hours. Assume n = is a 100-digit number, that is, that $n = 10^{100}$. Then $\sqrt{n} = 10^{50}$. For ease of calculation, use base 10 log instead of natural log. The number of primes less than 10^{50} is then about $10^{50}/50$, or more than 10^{48}. Sadly, our one day of prime testing will not bring us anywhere near the number of primes that need to be tested.

The reader will find a more sophisticated discussion of the problem of determining if a number is prime in *The Princeton Companion to Mathematics*, ed. Gowers, Tim, Princeton University Press, Princeton, N.J., 2008, pp. 348-362 .

.

How Odd That There Is Not An Odd Number Like 2!

If we want to make an even number out of an odd number, we need simply multiply it by 2. Yet if we want to make an odd number out of an even number, there is no integer we can multiply it by to make it odd. Instead, we have to divide it by the largest power of 2 that is one of its factors. Why isn't there an integer that serves the same purpose as 2 does in the first case?

"What Is the Next Number in the Following Sequence of Numbers...?"

Intelligence tests almost always have questions in which the test taker is asked to write down the next integer in a given sequence of integers.

I have always been bothered by these questions, because I think that, except in the most obvious cases, e.g., when the given sequence is something like 2, 4, 6, 8, 10, what the test designer requires as the correct answer is nothing more than what he has decided is the correct answer. Yet, in fact, there are at least two ways of deciding objectively what the correct answer is. However, each one requires far more time (and computing power) than the test taker will have available to him or her during the test. For this reason, I am sure that the test designer's decision as to the correct answer is never based on either of the two ways.

The first of the two ways is the following: (1) choose a Turing machine formalism that is capable of representing every computable sequence of integers; (2) of all the Turing machines that generate the given sequence of integers, find the shortest one — that is, the machine that is defined by the shortest string of symbols; (3) the next integer that is generated by this machine is the correct answer to the intelligence test question.

The second way uses a basic procedure in the mathematical subject called *Finite Differences*. The procedure begins with one writing down the difference between each successive pair of integers in the sequence in the test question, then the difference between each successive pair of those differences, etc., until one arrives at a sequence of identical integers. From the number of sequences of differences, one is able to write down a certain polynomial. One then goes through a process that yields, from the polynomial, a set of j simultaneous equations in j unknowns. The solutions then allow one to determine the next integer in the original sequence, and hence the answer to the test question.

I do not know of any proof that the answers arrived at by these two ways are always identical.

Is There A Number Too Large To Be Written Down?

In his book, *Littlewood's Miscellany*[1], J. E. Littlewood asks, "could there be a case in which, while pure existence [of a number] could be proved, no numerical X could be given *because any possible value of X was too large to be mentioned*?"

Suppose N is the smallest value of X, i.e., the smallest number that is too large to be mentioned. ("N", of course, is simply a name of the number.) Have I just mentioned a number that is too large to mention? Any irrational number, e.g., the base e of the natural logarithms, cannot be written down in its entirety, but we can write down approximations of it. Is there a difference between *mentioning a number* and *writing it down* and *writing down an approximation of it*?

Littlewood describes a recursive sequence, based on exponents, that yields ever larger numbers. We can stop the sequence at any time, and consider the number thus expressed (mentioned). It is certainly true that a sufficiently large such number cannot be written down as a normal integer in decimal notation at any given time because there is not sufficient paper in the world, or sufficient computer memory.

There is, of course, an infinity of integers so large that the shortest known representation of such an integer requires more binary digits than, say, the number of atoms in the universe. For example, we know, from algorithmic information theory, that there are finite binary sequences

1. ed. by Béla Bollobás, Cambridge University Press, Cambridge, 1990, p. 112.

whose shortest representation is essentially as long as the sequence itself. These sequences are called *random*. So let S be the set of all binary sequences of length, say, the number of atoms in the universe. Each sequence represents an integer in binary. A subset of S consists of sequences that are random. Let s be the sequence in the subset representing the smallest integer of all the integers represented by sequences in the subset. Then I have described a specific, very long sequence, but I can't even write down, say, the first five bits of it.

Here is a tentative proof that there is no number that is too large to be written down.

Assume, to the contrary, that there is such a number. Assume that N is the smallest such number. We now ask if $N - 1$ can be written down. It it can, then certainly $(N - 1) + 1 = N$ can also be written down, contrary to our assumption. So $N - 1$ can't be written down. But then $N - 1$, and not N, is the smallest number too large to be written down. This contradiction gives us our proof.

A question in passing is: suppose N is the largest number that can be written down. Then how does the person, or computer, acquire that number, so that it can be written down? Presumably, it must be written down and then copied. But then how did the person, or computer acquire that number so that it could be written down? Etc.

On Inner Products and Finite Sums of Powers of Consecutive Positive Integers

In mathematics, there is a set of finite sums of powers of consecutive positive integers, for example, $1^3 + 2^3 + 3^3 + ... + k^3$. Formulas are known for the sums.

In contemplating such a sum, we might see it as an inner product, namely, $\langle u, v \rangle$, where $u = (1, 2, 3, ..., k)$ and $v = (1^{n-1}, 2^{n-1}, 3^{n-1}, ..., k^{n-1})$. (In our example, $n = 3$.)

It is a well-known fact that

(1)

$$\cos \Theta = (\langle u, v \rangle)/(||u|| \, ||v||),$$

where $||u|| = \sqrt{\langle u, u \rangle}$, and similarly for $||v||$, and Θ is the angle between the vectors u and v.

We observe in passing that (1) implies the Cauchy-Schwartz Inequality, namely, $|\langle u, v \rangle| \leq ||u|| \, ||v||$.

Can we make anything of the fact that a finite sum of powers of consecutive positive integers, can be represented by an inner product?

We observe that our set of finite sums is a function $F(k, n)$. We can list the ordered pairs of this function in an infinite matrix, in which row headings are the values of k (namely, 1, 2, 3,...) and column headings are the values of n (again, 1, 2, 3, ...). The matrix cell (k, n) contains the value of $F(k, n)$. If we compute a sufficiently large number of these values, do we observe any patterns that enable us to find values of $F(k, n)$ much more rapidly than any of the formulas do?

Set Theory

A Question About Categorizing a Finite Set of Things

Our question is simply: What is the "best" way to categorize a finite set of things?

We might ask this question when, e.g., designing a company organization chart (to be distributed on paper) that includes departments and job titles. Programmers and users often disagree on the best categorization: "It's better if you organize them like this." Sometimes they put the same subset inside several other subsets, e.g., "programmers" inside "marketing", "programmers" inside "engineering", "programmers" inside "manufacturing".

What are they attempting to optimize? What are we saying when we say, "This categorization is better than that one?" Merely that the access time, or memory usage, is less? Or that the categorization is easier to comprehend for humans? Is anything known about the complexity, ala algorithmic information theory, of different categorizations? How are different categorizations even possible? Because each thing has more than one property, and a property defines a set? What is the property of each subset in the power set of a set? Is it simply the property of having only its elements?

Or suppose we want to make a tree showing the subsets of a class of mathematical entities, say of rings (these are sets of numbers or other entities that behave like the integers under addition, subtraction, and multiplication). We would like the top of our tree to have just two branches. So we might choose to have one branch for, say, commutative rings and the other for non-commutative rings. But we could also choose one of the top two branches to be for rings with the unique factorization property, and the other for rings without this property. Or we could choose to have the top two branches represent quadratic rings and non-quadratic rings. Etc. What is the "best" way?

Or suppose we want to categorize (sort) all the subsets of a finite subset? What is the "best" way?

A Tentative Answer to the Question

Our tentative answer to the question is the following. Let n be the number of objects in the set. Number the objects 1, 2, 3, ..., n. Suppose each object can have from 0 to k properties numbered 0, 1, 2, ..., k. Associate with each object a binary string of length k. If an object has the property k, then let the kth digit from the right in the string be 1; otherwise let it be 0.

We see immediately that there are many ways that we can categorize the objects in the set by their properties. We can begin by collecting into a subset, all objects having a certain property (any property), then, within that subset, we can collect all objects having another property, and then, within the resulting subset, we can collect all objects having still another property, etc.

The "best" categorization is then a matter of our personal wishes and needs. We might decide on a property that is most important for us, and collect all objects having that property, then continue down through subsets of that subset. Or we might want the categorization that has the most levels in the vertical direction (the "deepest" categorization). Or we might want a categorization that has the least levels in the vertical direction (the "flattest" categorization). (There may be more than one categorization in each of these cases.) Etc.

A Historical Question About Infinite Sets

It is easy to show that the number of points in a long line is exactly the same as the number in a short line: simply center the short line above the long one so that the two are parallel. Then from a point above both it is possible to draw a straight line through any point of the longer line that intersects a point of the shorter line, and vice versa. Or consider two concentric circles: the radius through any point on the smaller circle meets a point on the larger circle, and vice versa. The Greeks were perfectly capable of recognizing these facts. Why didn't they? If you reply that

they would have considered it nonsensical for a smaller-diameter circle to have as many points as a longer-diameter circle, I reply that they could have proven the fact by using a technique they were familiar with and accepted, namely, approximating a circle by a sequence of regular polygons of increasing number *n* of sides. For some *n*, inscribe a regular *n*-gon in the smaller circle, and then inscribe an *n*-gon in the larger concentric circle so that corresponding vertices of each *n*-gon lie on the same radius. Then regardless how large *n* is, it is clear that the points in the smaller circle match one-one with those in the larger. Unfortunately, this argument holds only for a countable infinity of points in each circle, whereas the number of points in each circle, as we now know, is uncountable.

"Proclus [412-485], the commentator on Euclid, noted that since a diameter of a circle divides it into halves and since there is an infinite number of diameters, there must be twice that number of halves. This seems to be a contradiction to many, Proclus says, but he resolves it by saying that one cannot speak of an actual infinity of diameters or parts of a circle. One can speak only of a larger and larger number of diameters or parts of a circle. In other words, Proclus accepted Aristotle's concept of a potential infinity but not an actual infinity. This avoids the problem of a double infinity equaling an infinity.

"Throughout the Middle Ages philosophers took one side or the other on the question of whether there can be an actual infinite collection of objects. It was noted that the points of two concentric circles could be put into one-to-one correspondence with each other by associating points on a common radius. Yet one circumference was longer than the other.

"Galileo struggled with infinite sets and rejected them because they were not amenable to reason." — Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, N.Y., 1972, p. 993.

Is the Set of All Sets "Like" a Klein Bottle?

I will assume that the reader is familiar with the Klein bottle. Its form suggests the following "demonstration" that the set of all sets contains itself.

Imagine the set of all sets inside a gourd-shaped container that has a hole in the end of the narrow neck. Now make a small hole in the side of the container, then extend the narrow neck and curve it around so that the hole in the neck can be inserted into the hole in the side. The set of all sets now contains itself.

Carrying the Theory of Types to Extremes

Consider the concept of *the number of numbers*, e.g., three 3's. Would anything useful result if we imposed a strict hierarchy of types on everything we did that was concerned with the non-negative integers, so that, e.g., we would speak of a 3 of Type 0 (a 3 standing alone), a 3 of Type 1(a 3 that was counting, or multiplying another number), a 3 of Type 2 (a 3 that was counting or multiplying another number which in turn was counting or multiplying another number), etc.? Would some things become clearer? Suppose numbers were "color-coded" to represent their type in a given context.

The Borges Set

Exercise: make Venn diagrams to represent a few reasonable interpretations of the classification of animals described in the following passage:

"These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote pages it is written that animals are divided into (a) those that belong to the emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance." — Borges, Jorge Luis, "The Analytical Language of John Wilkins," 1941, quoted in Rucker, Rudy, *The Fourth Dimension: A Guided Tour of the Higher Universes*, Houghton Mifflin Company, Boston, 1984, p. 197.

Cryptography

A mathematician I know claims that every code, except, possibly for those used in one-time messages, can be broken.

But J. E. Littlewood, one of the great mathematicians of the first half of the 20th century, said:

"The legend that every cipher is breakable is of course absurd, though still widespread among people who should know better, ... [although] it is sufficiently obvious that a single message cannot be unscrambled."[1]

Littlewood's book was first published in 1953, before the development of the high-powered computers that are used today to break codes. Yet even so, I doubt if, even now, he would concede that every code can be broken. I certainly do not believe that.

The correct answer to the question, "Is this coded message decipherable?", is "It depends". Thus, e.g., sufficiently numerous messages sent using a simple substitution code, will eventually result in the breaking of the code by well-known statistical facts concerning the language in which the message is written. But a message requiring very large computer resources, including time, to be deciphered, might not be deciphered before the message has become obsolete. Which is as good as saying that the message was not decipherable.

It seems to me that the proper way to think about codes is to think in terms of the values of items in a table of parameters such as the following:

1. How much of the message the recipient already has (see "Item 1 Notes" below);
2. How often messages in a single code are sent, and/or how long the average message is;
3. The code used;
4. How much of the context of the message is known to the potential code-breaker (see "Item 4 Notes" below);
5. The decoding equipment available to the potential code-breaker (pencil and paper vs. high powered computer with sophisticated software);
6. The maximum amount of time available to break the code, after which the message is obsolete;
7. Concealing from the potential code-breaker, the starting point of the message. Suppose that the sender sends a sequence of characters starting at 12:05 a.m. each day, and continuing until 12 midnight that day. The recipient would have been given, in advance, the date and a positive integer *n* to indicate that, on that date, the message starts with the *n*th character sent that day.

---

1. Littlewood, J.E., *Littlewood's Miscellany*, Cambridge University Press, New York, 1990, p. 43.

8. Possibility of the potential code-breaker finding out the location of the recipient, and capturing him and torturing him to make him reveal the code;

9. The electromagnetic frequency over which the message is to be sent, and the speed at which the characters will be sent. This would seem to require a very large investment in equipment (how many possible frequencies are there, at any given time, over which a message could be sent?).

Item 1 Notes

Suppose that, in World War II, the Allies had a spy in France. The complete plans for the possible attack at Normandy, and the complete plans for the possible attack at Calais, had been delivered to him by hand, along with the fact that, at a certain specified time, at a certain specified wavelength, a 1 would be sent if the attack was to be at Normandy, and a 0 if it was to be at Calais.

Clearly, this one-bit message could not be deciphered — not even if the Germans knew that the bit referred to which plan was to be implemented!

So that is one extreme. The other would be if the recipient had no idea what the coded message was to tell him.

In war at least, the recipient usually has a very clear idea what the message will be about — a prospective attack by his side, a change in physical direction and speed for his vehicle(s). So why not number all the possibilities beginning at 1, where one possibility might be, say, "Attack [name of town] at 5 a.m.", another might be "Attack [name of town] at 6 a.m.", etc. Since there are only a finite number of towns and a finite number of attack hours, a number can be assigned to each possibility. Similarly for other commands, e.g., "Continue on present course until 9 a.m., then turn left 45 degrees", "Continue on present course until 10 a.m., then turn left 45 degrees", etc. The number of the possibility can then be buried in a pre-agreed-upon location in a code text whose other numbers and letters are randomly chosen.

Or suppose there are, say, five attack plans, each numbering many pages. The code recipient has a copy of each. Then to send the message, "Attack according to plan no. 3", only the number 3 need be sent (at a pre-agreed-upon location in a text otherwise filled with randomly-selected characters)[1].

Why wouldn't such codes be harder to break than traditional ones?

(We might call this the "pre-sent messages" approach to cryptography.)

Item 4 Notes

Thus, e.g., the breaking of the code that the Germans used with their Enigma machines (one of which had been stolen by the Allies in Poland early in the war), was certainly made easier by the fact that all the messages pertained to military activity in Europe and in the Atlantic (submarine warfare).

---

1. One reader whose extensive knowledge of advanced cryptographic techniques made it impossible for him to think about simple things, told me that the coding idea in this section was trivial and unimportant. I asked him how trivial and unimportant Hitler would have considered knowing which of two numbers, 1 or 0, referred to the plans for the D-Day invasion, where, say, 1 referred to the plans for the Normandy invasion, and 0 referred to the plans for the invasion via Calais (which was also under consideration at the time). Even even if he had a copy of each set of plans, and even if somehow he was able to locate the number (1 or 0) in the stream of characters he intercepted, he would still not know where the attack was to take place without knowing which number had been assigned to which set of plans.

Other Thoughts

If I were asked to break a code, I would think about the problem topologically and in terms of algorithmic information theory. Suppose you have a text to be encoded so that the enemy can't read it. Suppose that you decide to substitute letters, but what you will do is, letter by letter, select its substitution at random. So, in effect, if the first word were "This" you would randomly select a character from the alphabet and substitute it for "T". Then do the same for "h", "i", "s", and continue this process throughout the message, so that the substitution for a given letter will in general depend upon where it appears in the message. The trouble is that the only way you can tell your agent the code is by sending him the original text along with the replacement characters, and if you can send him the original text, why bother with the replacement characters? On the other hand, if the enemy somehow manages to get only the encoded text, they can *learn* nothing about the code, because it is new for every character. Putting it topologically, nothing is "near" to anything else. If I know something, I don't also know something else. So, in my coding school, we begin by looking at all possible topologies that are applicable to strings of characters.

In ordinary replacement ciphers, such as appear on the puzzle page of Sunday papers, e.g., the *San Francisco Chronicle*, what are the chances we are wrong once we have found a substitution that works for two words? Three? Four? Etc. This is a question about how "prickly" the English language is from a topological point of view. See my paper, "Occam's Razor and Program Proving by Test", on the web site www.occampress.com .

Consider the following coding system.

1. Assume there are 40 alphanumeric symbols. Now consider a tape with, say, 1,000 divisions marked on it. (Obviously, such a tape can be represented in a computer.)

2. Divide the tape up into 40 segments of arbitrary length.

3. Randomly associate each segment with an alphanumeric symbol.

4. A message is a sequence of three-digit integers in the range 000 to 999, where 000 denotes the first division after the start of the tape, and 999 denotes the end of the tape.

The alphanumeric character indicated by the three digits *xyz* is the character that was randomly assigned to the interval containing the division *xyz*.

A message consists of a sequence, without breaks, of three-digit sequences in the above range.

Observe that *the same text can be represented by several different messages*, since more than one three-digit sequence can represent the same character if there is more than one division in the interval associated with that character.

It seems to me that this would be a difficult code to break, since the enemy does not know (1) that we are using such a scheme in the first place; (2) how many divisions there are in our tape, (3) what divisions are associated with what intervals, and (4) what alphanumeric character is associated with each interval.

In principle, the code can be broken, via the following program:

For all numbers-of-divisions *d* up to the limits of the computer do
for all sequences of 40 intervals each of any length as long as the total length of all the intervals = *d* do
for all possible assignments of 40 alphanumeric characters to the intervals do
compute the resulting sequence of alphanumeric characters.
Compare all the sequences of alphanumeric characters, and choose the one most likely to be the original text.

Cluster Theory

In a math class for liberal arts students, the following was a homework problem:

Suppose a student got the following grades on different tests: 80, 90, 85, 90, 50, 78, 84. What is the best single number to assign as a grade for this student?

The professor asked his students to consider mean, median, mode and midrange for these grades, as possibly providing the answer. But insofar as that implies that one or more of these measures can be used in general to give the best grade, it is misleading. I can only hope that, in going over the homework problem, the professor pointed out that a good answer can be arrived at by computing the center-of-gravity of the grades. He could ask the class to imagine a metal rod, say, 100 inches in length. Each grade could represent a point on this rod. Thus, e.g., the grade of 50 is 50 inches from the left-hand end, the grade of 78 is 78 inches from the left-hand end, etc. At the point that each grade represents, a string can be tied to the rod and, say, 1 inch below it, a 1 pound weight attached. (There would be a two one-pound weights below 90.) The professor could then explain that there is a point on the rod which, if the rod were held aloft on the end of a finger placed exactly at that point, the rod would balance. The best grade – the most representative grade – is the distance from the left-hand end of the rod to that point.

He could then state that the formula for the center-of-gravity *cg* is:

$$cg = \frac{x_1 m_1 + x_2 m_2 + \ldots + x_n m_n}{m_1 + m_2 + \ldots + m_n}$$

where x_i in the present case is the distance from the left-hand end of the rod to a grade, and m_i is the number of pounds suspended at that point (all the m_i are 1 except for 2, below the grade of 90).

Filling in the values in the above formula we get *cg* = 79.57. The reader can verify that the sum of the clockwise torques to the right of the *cg*, exactly equals the sum of the counterclockwise torques to the left of the *cg*. Hence the rod, if held aloft on one's finger at that point, exactly balances, i.e., does not rotate about the point either clockwise or counterclockwise.

Reversing the Names of Courses

A major subject in mathematics is algebraic topology. Why is there no subject, topological algebra? What might such a subject consist of? We can ask equivalent questions about algebraic geometry and arithmetic geometry.

Troubled Thoughts About Symmetry

In elementary mathematics courses and popularizations, symmetry is taught using symmetric geometrical figures, for example, the square. The vertices of a square are labeled, then an operation is performed that demonstrates the symmetry of the square, for example, rotating it clockwise by 90 degrees about its center, The rotated square is shown to be superimposed on the original square, and the vertices of the rotated square are shown to be different than the vertices of the original square. This gives rise to a discussion of permutation of vertices, and an answer to the question, How many different symmetric operations of the square are there?

But a square is not a geometric figure with labeled vertices. Given a square without labeled vertices, we can perform various operations that result in the same square being present, and superimposed on the previous square. We can make a list of all the operations that have this effect, and we can then observe that arbitrarily long finite sequences of these operations will always result in the same square being present, and superimposed on the previous square.

I want to say: *and that is all we can legitimately do*! To talk of labeled vertices, and how the vertices move as a result of an operation, is to talk, not of a square, but of something we should call a l*abeled square*. Which is not the same thing.

Kakeya's Problem

Kakeya's Problem is to "find the region of least area in which a segment of unit length can turn continuously through 360° (minimize area swept over)."[1]

Littlewood says that it was long thought that the minimum area was $\pi/8$. This was achieved by a "triangle" of certain dimensions with the sides curved inward by the same amount. The unit segment *s* could then be slid along each side and pivoted appropriately when one end reached a vertex of the "triangle".

But Littlewood then says, "... A. S. Besicovitch (*Math. Zeit.* **27**(1928), 312-320) showed that the answer is zero area (unattained). : given an arbitrarily small area ε the area swept can be less than ε. As ε tends to 0, the movements of the segment become infinitely complicated and involve excursions towards infinity in all directions."[2]

As a personal challenge, I have deliberately not looked up the solution. However, Ian Stewart in his *The Problems of Mathematics*[3], pp. 173-174, says that, in the course of the development of Lebesgue measure, "it transpired that sufficiently messy and complicated sets may not posses well-defined areas and volumes at all."

In 1924, the Banach-Tarski Paradox was published. It states that it is possible to dissect a solid sphere into six pieces, which can be reassembled, by rigid motions, to form two solid spheres each the same size as the original and each having the volume of the original. "The trick is that the pieces are so complicated that *they don't have volumes*."

---

1. Littlewood, J. E., *Littlewood's miscellany*, Cambridge University Press, New York, 1990, p. 38.
2. ibid.
3. Oxford University Press, N.Y., 1992

The zero-area solution to Kakeya's Problem was published in 1928. My guess is that it is an application of the measure theory that yielded the Banach-Tarski Paradox.

The Game of FreeCell

The computer game of FreeCell (Windows 7.0 version): 52 cards are dealt randomly in 8 rows so that there are 4 columns containing 7 cards each, and 4 columns containing 6 cards each. On top are 4 free cells on the left, and 4 final cells. The goal, as in solitaire, is to wind up with the 4 final cells containing ace through king, with ace on bottom, for each suit. Same moves are allowed as in solitaire.

"Help" for the game says all games are winnable but it seems easy to come up with an initial layout that is not, e.g., all four aces at top of leftmost four columns; in front of them all four kings; all four fours at top of rightmost columns, then all four threes, then all four twos, then successive rows of same color as possible.

Does the game have a winning strategy?

The Windows 10.0 version allows the player to choose the level of difficulty, from Easy, Medium, Hard, and Expert. How are these levels determined?

Suppose we sorted all possible games. Each possible initial layout of cards would be the root of a tree. The first level [not the same as a level of difficulty] would be the result of all possible single moves. The second level would be the result of all possible single next moves. The third level would be ...

Clearly, we could use the computer to find all possible successful games.

A Question About Strategy and the Sorting of Games

A *Game*, e.g., chess, checkers, a card game, a board game, consists of various *games* each of which is defined by the starting positions of the pieces, or the initial distribution of the cards, etc.

The question is: if we know how to sort the games in order of increasing difficulty (more than one game may have the same level of difficulty, of course), do we have a strategy for winning each game?

Posed Problems

Although I have never come across these problems in the course of reading or study, I make no claim that any of them is original.

The Cross-Number Problem

In cross-number puzzles such as appear on the Sunday puzzle page of the *San Francisco Chronicle*, what is the minimum number of rows and/or columns and/or diagonals we have to check before we know our answer is correct? What is the minimum number of correct numbers we have to have in place in order to force a correct solution?

A Pendulum Problem

Suppose that a pendulum were suspended inside a given pendulum, how would this second pendulum behave as the first swung back and forth? Consider several cases, e.g., (1) if the initial position of the second were the same as the initial position of the first; (2) if the initial position of

the second were the opposite (i.e., far left if the first were far right, etc.); (3) various lengths of the second relative to the length of the first; (4) a third pendulum inside the second, etc.

> "*Build your own chaos machine*. One of the best non-computer projects I know for observing chaos involves building a double pendulum — a pendulum suspended from another pendulum. The motion of the double pendulum is quite complicated. The second arm of the pendulum sometimes seems to dance about under its own will, occasionally executing graceful pirouhettes, while at other times doing a wild tarantella. You can make the double pendulum from wood. At the pivot points, you might try to use ball bearings to ensure low friction...
>
> Place a lead weight at the bottom of the first pendulum so that the pendulum will swing for a longer time. (The weight stores potential energy when the pendulum is lifted.) The second pendulum arm can be about half the length of the first. You can place a bright red dot, or even a light, on one end of the second pendulum so that your eye can better track its motion. Note that your pendulum will never trace the same path twice, because you can never precisely reposition it at the same starting location, due to slight inaccuracies in knowing where the starting point is. These small initial differences in position are magnified through time until the pendulum's motion and position become unpredictable. Can you predict where the lower pendulum will be after two or three swings? Could the most powerful supercomputer in the world predict the position of the pendulum after 30 seconds, even if the computer were given the pendulum's precise equations of motion?" — Pickover, Clifford A., *A Passion for Mathematics*, John Wiley & Sons, Inc., Hoboken, N.J., p. 190.

The Parking Place Problem

What is the smallest parking place a car can be driven out of? Assume that all cars are the same width (though lengths can vary), that they are all rectangular when viewed from above and that they are all parked parallel to the curb. The turning radius of the car is given, i.e., for each position of the steering wheel, the forward and backward trajectories of the car can be determined. Tentative answer: a car can get out of any parking place whose length is greater than the diagonal of the car.

The Income Tax Problem

Assume you paid, say, $5,000 in income taxes last year. Late in the year you are amazed to hear about something actually constructive the federal government did, e.g., that it gave several million dollars to a company to develop solar power generators. You tell a friend how much tax you paid, and then you say, "It feels good to know my money is helping solve our energy problems." The friend shakes his head and says, "But you don't know that your $5,000 went to the solar project. As a matter of fact, I am quite sure it went to the Iraq war." Are you both wrong, and why? Or, if one of you is right, which one and why?

The Cassette Tape Problem

Write an equation to describe the speed of a tape casette wheel containing tape which is being wound onto another wheel of identical dimensions, assuming the second wheel runs at constant speed, s, and the tape is of some fixed thickness, t.

Describe the behavior of a tape casette as a group, in which the group elements are tape unwinding (tape rewinding), and turning over of the cassette (always beginning with the tape flat

on the table, with the longer edges parallel to the front of the table), the direction of rotation of the cassette always being the same. (We may imagine that when we are listening to the tape, i.e., running it in the forward direction, a person in a mirror universe is rewinding it, and vice versa.)

The Laundry Bag Problem

Suppose shirts cost $\$w$ each, and each shirt can be worn x days (not necessarily x consecutive days) before it needs to be washed. You must wear a shirt every day. Shirts wear out and must be discarded after u washings. The washing machines at the local laundromat can hold at most y shirts per load. Washing a load (or partial load) costs $\$z$. How many shirts should you own if you want to spend the least amount of money on shirts (buying and washing)?

The Leaf-Raking Problem

Your task is to rake the leaves on a large rectangular lawn and place them in a compost pile not on the lawn. How should you go about performing the task if you want to perform it with the least amount of work?

The problem definition can be made more precise if, instead of leaves, we consider n stones, all of uniform size and weight, distributed at random on the lawn, our task being to move all of them to a pile not on the lawn. Assume that the task is to be performed by only one person, and that the person can carry at most m stones at a time, where m may be less than, equal to, or greater than n. Assume further that the work involved in carrying a stone from point A to point B is c times the distance between A and B, regardless of the number of stones (less than or equal to m) being carried. (c is a constant > 1.) Thus, the work involved in carrying k stones from A to B is k times c times the distance from A to B. Assume that the work in simply walking from A to B empty-handed is equal to 1 times the distance between A and B.

The Hedge Clipping Problem

Normally we clip the entire hedge when we feel it needs it, and do no clipping in between. Why not instead clip just those branches that pass a certain length, or grow beyond a certain boundary we define in advance? Which is the more efficicient, i.e., which, in the long run, requires the least time, including the time required to get the shears out of the garage and put them back? Clearly the strategies are the same if all branches grow at the same rate, because then on a certain day they all pass the limit and must all be trimmed. But suppose all the branches do not grow at the same rate. To go to extremes, suppose all branches but one don't grow at all. Then our only choice is to cut that one branch when it passes the limit. What is the point — what is the percentage of faster-growing branches — at which it just becomes more efficient for us to clip them all?

The Tree Cutting Problem

Given a handsaw with saw teeth along only one side of the blade, and with a handle that is thicker than the saw blade, and ignoring energy requirements, fatigue, etc., can you always cut a tree in two, regardless of its diameter, assuming the cut must always be perpendicular to the axis of the trunk? The answer is yes, because you can always cut sufficiently thin slices so that you create a wide enough slot for the handle and your hand to move in. Assuming this technique is not allowed, for a given length of saw, what is the largest diameter trunk the saw can cut in two?

The Carrier-Based Aircraft Problem

Carrier-based aircraft are to fly an attack mission that is at the limit of their flying range. After launching their torpedoes, the aircraft will be a few hundred feet above the surface of the ocean. Should they immediately gain altitude so that if they run out of gas, they will be able to glide the remaining distance to the carrier, or should they stay at their initial altitude and thus save the gas they would otherwise have to burn in order to gain altitude? Assume that all relevant data are known, including glide ratio, gas consumption for each rate of climb including the zero rate. Initial thought: if the glide ratio is sufficiently large that it outweighs the cost in range of achieving high altitude, then they should gain altitude.

The Netflix Problem

In the queue of previously-chosen DVDs that each Netflix subscriber sees when he accesses his account in the Netflix website, there is a button at the left-hand end of each line containing the title of a DVD. It is labeled "Top". If it is clicked, that DVD is moved to the top of the queue, and thus will be the first DVD sent when the subscriber has at least one less DVD at home than the maximum he is allowed.

What is the most efficient algorithm, using only the "Top" button, for moving any DVD in the queue to any other position in the queue without disturbing the relative order of the others?

The Lunacy of Martin Gardner

Martin Gardner had a well-deserved reputation as an expert on mathematical games and puzzles — for many years he edited the "Mathematical Recreations" department in *Scientific American*. He was also known as a popularizer of various subjects in mathematics and science, and as a debunker of pseudo-scientific claims. His *The Annotated Alice*, which is an annotated edition of Lewis Carroll's classics, *Alice's Adventures in Wonderland* and *Through the Looking Glass*, is a masterpiece of scholarship.

However, readers of his essay, "The Irrelevance of Conan Doyle" in his *Science Good, Bad and Bogus*, are in for a shock, because in this essay he claims, in apparent complete seriousness, that "[Arthur Conan] Doyle had almost nothing to do with either Homes or Watson." The basis for his claim is that, since Doyle believed in spiritualism — the possibility of communicating with the dead — he couldn't possibly have created a character as logical and as dedicated to facts, as Sherlock Holmes. Gardner also denies that Cervantes wrote *Don Quixote*, claiming instead that the books were written by — Sancho Panza!

I have so far been unable to find, on the Internet, any explanation by Gardner for these bizarre beliefs. However, he is by means the only person with a strong technical mind who has also had bizarre beliefs. For example, there is a well-known mathematician who is convinced of the validity of pyramid-power, which is the belief that, e.g., sitting under an object that is shaped like a pyramid, will result in long life and other benefits.

"Never imagine yourself not to be otherwise..."

As a public service, I herewith attempt a parsing of the Duchess's "simplification" of her moral, "Be what you would seem to be", in *Alice in Wonderland*. The passage is as follows:

" 'Oh, I know!' exclaimed Alice... ' [mustard is] a vegetable. It doesn't look like one, but it is.'

" 'I quite agree with you,' said the Duchess; 'and the moral of that is — "Be what you would seem to be" — or, if you'd like it put more simply — "Never imagine yourself not to be otherwise than what it might appear to others that what you were or might have been was not otherwise than what you had been would have appeared to them to be otherwise."'" — Carroll, Lewis, *Alice's Adventures in Wonderland*, in the edition with introduction and notes by Gardner, Martin, *The Annotated Alice*, New American Library, N.Y., 1960, p.122.

It is possible that the "simpler" version of the Duchess's moral was derived from Carroll's recollection of a passage in Book II, XII, of Cicero's *De Officiis*: "Quamquam praeclare Socrates hanc viam ad gloriam proximam et quasi compendiaram dicebat esse, si quis id ageret, ut, qualis haberi vellet, talis esset." ["And yet, as Socrates used to express it so admirably, 'the nearest way to glory — a short cut, as it were — is to strive to be what you wish to be thought to be.'"] — Cicero, *De Officiis*, w. English tr. by Walter Miller, William Heinemann Ltd., London, 1956, pp. 210-211.

We can quite easily figure out a meaning for the first part of the Duchess's "simplification" if we add an assumed concluding independent clause: "Never imagine yourself not to be otherwise than what you want to appear to others to be otherwise than". Now "Never imagine yourself not to be otherwise than" is equivalent to "Always imagine yourself to be otherwise than..." And so the independent clause, in our modified form, is:

"Always imagine yourself to be otherwise than what you want to appear to others to be otherwise than."

Thus, e.g., if you want to appear to others to be a truth-teller, then you should always imagine yourself to be otherwise than a liar.

So it must be that "what it might appear to others that what you were or might have been was not otherwise than what you had been would have appeared to them to be otherwise" is equivalent to "what you want to appear to be otherwise than".

I will welcome hearing from readers.

Appendix A — Derivation of Laplace's Series

1. Laplace's Series is:
(1)

$$Erfc(T) = \int_T^\infty e^{-t^2} dt = \frac{e^{-T^2}}{2T}\left(1 - \frac{1}{2T^2} + \frac{1 \cdot 3}{(2T^2)^2} - \frac{1 \cdot 3 \cdot 5}{(2T^2)^3} + \ldots\right)$$

It will make our explanation easier if we multiply the right-hand term through by the term to the left of the parentheses. This gives us:
(2)

$$\int_T^\infty e^{-t^2} dt = \frac{e^{-T^2}}{2T} - \left(\frac{e^{-T^2}}{2T}\right)\left(\frac{1}{2T^2}\right) + \left(\frac{e^{-T^2}}{2T}\right)\left(\frac{1 \cdot 3}{2T^{2^2}}\right) - \left(\frac{e^{-T^2}}{2T}\right)\left(\frac{1 \cdot 3 \cdot 5}{2T^{2^3}}\right) + \ldots$$

2. We use integration by parts, which asserts that

$$\int u \ dv = uv - \int v \ du$$

We obtain the following equation. Explanation of the terms follows:
(3)

$$\int_T^\infty \left(\frac{1}{(-2t)}\right)(e^{-t^2}(-2t)dt) = \frac{e^{-t^2}}{(-2t)}\Bigg|_T^\infty - \int_T^\infty e^{-t^2}\frac{1}{2t^2}dt$$

where, on the left-hand side of the equation,

$$\left(\frac{1}{(-2t)}\right) = u$$

$$(e^{-t^2}(-2t)dt) = dv$$

We have taken v to be e^{-t^2}. The derivative of v is the above term, dv. To obtain it, we had to multiply by $(-2t)$, and to compensate, we must multiply by the inverse, $(-1/2t)$, which we called u, above.

The first term on the right-hand side is easily seen to be uv, which evaluates to the *first* term on the right-hand side of the series (2).

The leftmost term, e^{-t^2}, of the integrand on the right of (3) is v, as we have said. The remainder of the integrand is the derivative of $u = (-1/2t)$, as the reader can verify.

3. We now iterate our process. The right-most term in the previous step becomes the basis for the left-hand term in the following equation. Explanation of the terms follows:

(4)

$$\int_T^\infty \left(\frac{1}{(-2t)}\right)\left(-\frac{1}{2t^2}\right)(e^{-t^2}(-2t)dt) \;=\; \frac{e^{-t^2}}{(2t)(2t^2)}\bigg|_T^\infty - \int_T^\infty e^{-t^2}\frac{-3}{(2t^2)^2}dt$$

where, on the left-hand side of the equation,

$$\left(\frac{1}{(-2t)}\right)\left(\frac{-3}{(2t^2)^2}\right) \;=\; u$$

$$(e^{-t^2}(-2t)dt) \;=\; dv$$

As in the previous step, we have taken v to be e^{-t^2}. The derivative of v is the above term, dv, as in the previous step. To obtain it, we had to multiply by $(-2t)$, and to compensate, we must multiply by the inverse, which, with the term we began with (from the last integrand in the previous step) gives us the indicated term for u.

The first term on the right-hand side of equation (4) is easily seen to be uv, which evaluates to the *second* term on the right-hand side of the series (2).

The leftmost term, e^{-t^2}, of the integrand on the right is v, as we have said. The remainder of the integrand is the derivative of

$$\left(\frac{1}{(-2t)}\right)\left(-\frac{1}{2t^2}\right) \;=\; u$$

as the reader can verify.

Each succeeding term of the series in (2) is obtained by a similar iteration.

Laplace's trick is not mentioned in elementary calculus texts, as far as I have been able to determine. This is a shame, since it would seem to be a powerful tool for expanding the application of integration by parts.

Appendix B — General Formulas in Radicals for Solution of Equations of Degree 1 Through 3

Degree 1

The general equation is:

$$a_1 x + a_0 = 0$$

And the general formula to solve that, which we learn in jr. high school, is very simple:

$$x = -\frac{a_0}{a_1}$$

Degree 2

The general equation is:

$$a_2 x^2 + a_1 x + a_0 = 0$$

And the general formula to solve that, which we learn in high school, is only a little more complicated than that for degree 1:

$$x = \frac{(-a_1) \pm \sqrt{a_1^2 - 4a_2 a_0}}{2a_2}$$

Degree 3

The general equation can be written:

$$x^3 + a_1 x^2 + a_2 x + a_3 = 0$$

Viete's formula for the solution is

$$x_1 = \sqrt[3]{t} - \sqrt[3]{u}$$

Now the general formula gets more complicated. Cardan's formulas are as follows: Let:

$$p = a_2 - \left(\frac{a_1^2}{3}\right)$$

and

$$q = \frac{2a_1^3}{27} - \frac{a_1 a_2}{3} + a_3$$

and let

$$P = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{p^3}{27} + \frac{q^2}{4}}}$$

and

$$Q = \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{p^3}{27} + \frac{q^2}{4}}}$$

Then, with proper choice of the cube roots, the roots of the equation are:

$$x_1 = P + Q - (a_1/3)$$

$$x_2 = \omega P + \omega^2 Q - (a_1/3)$$

$$x_2 = \omega^2 P + \omega Q - (a_1/3)$$

where $\omega \neq 1$ is a cube root of 1. (I have here used the presentation of the formulas in Herstein, I. N., *Topics in Algebra*, John Wiley & Sons, N.Y., p. 251.)

$$x_2 = \omega \sqrt[3]{t} - \omega^2 \sqrt[3]{t}$$

$$x_3 = \omega^2 \sqrt[3]{t} - \omega \sqrt[3]{t}$$

where t, u are terms that are derived and ω is a complex root of $x^3 - 1 = 0$. (See Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, N.Y., 1972, p. 267.)

The formula for degree 4 is even more complicated, and will not be reproduced here.